Aalto University
School of Economics

Gibrat's law revisited - A study on Gibrat's law with models of industry dynamics

Economics

Master's thesis

Philip Relander

2011

Department of Economics
Aalto University
School of Economics

Abstract

This thesis is a literature review that studies Gibrat’s law and the firm dynamics with the help of conventional models of industry dynamics. Robert Gibrat formulated one of the first models of industry dynamics already in 1931 and in this model he used the assumption of law of proportional effect that is today understood as Gibrat’s law. A common interpretation of the Gibrat’s law presented in many articles is that a firm’s growth rate and its size are independent of each other. This thesis uses this interpretation of the law. Put differently, the law states that small firms grow at the same rate as large firms. The power of Gibrat’s law is the primary research question and a common theme that is carried throughout this thesis. In addition, this thesis touches briefly on what creates firm growth.

In general, the extensive literature has rejected the law, but various studies have found that the law is valid for certain subsamples or time periods. Therefore, this thesis argues that the question is not whether Gibrat’s law is valid, but rather when and with what restrictions it is valid. This thesis also tries to understand why Gibrat’s law should be accepted. One of the aims of this study was to identify testable hypothesis that would more accurate. Hence, they would further clarify the role of Gibrat’s law.

These questions are studied with help of the theory on firm and industry dynamics. This makes the approach more unique as existing studies focuses heavily on empirical testing. As primary material, this thesis uses the existing empirical literature on Gibrat’s law and four models of industry dynamics by Hopenhayn (1992), Jovanovic (1982), Cooley & Quadrini (2001) and Murto & Terviö (2010). Hopenhayn’s model is analyzed more thoroughly than the rest of the industry dynamic models.

The conclusion of this study is that in the majority of the cases small firms indeed grow faster than large firms. This is supported both by theoretical and empirical evidence. It can be case that sometimes the growth is observed as stochastic, but it would seem that the underlying process is indeed deterministic as there are profit-maximizing firms that act and make decisions. These findings could explain why sometimes studies reject the law and sometimes they accept it. Hopenhayn’s model is a stand-alone model, but it opens the possibility for other models were Gibrat’s law could be a special case. Thus, the conclusion is that Gibrat’s law can’t be a valid. The three other models that were presented more or less confirmed that the law can’t be valid.

Furthermore, it was possible to find new testable hypothesis. For example Gibrat’s law could be tested for an industry where there has been a large increase in cost of entry, preferably over a shorter period time. The expected result is that after the increase in cost of entry, there should be less deviation from Gibrat’s law. Finally, this thesis views that superior productivity creates firm growth.

Keywords: *Gibrat’s law, firm growth, growth-size relationship, industry dynamics, productivity shocks.*

Contents

List of pictures

1 Introduction

Does size have an impact on growth? Do smaller firms grow faster than larger firms? Overall, what stimulates the growth of firms? These are nontrivial questions that have been central in the study of industrial economics for a good reason for a long time. Growth has a comprehensive impact on different levels. The firm's shareholders would certainly be interested in knowing what induces growth. Also, the growth-size relationship is important from a policy perspective. Substantial amount of money and effort is spent on encouraging firms to grow. According to Toivanen et al. (2010) Finland invested 3.6 per cent of GDP – €5 billion - on R&D in 2001. In 2009, Tekes invested 343 million euro in R&D projects done by firms. 60 % of this funding went to small and medium-sized enterprises (SME)[1]. If growth were a random process, then it would be questionable that the subsidies are targeted on small and medium-sized firms.

The first formal model on industry dynamics and growth was presented already in 1931 by Robert Gibrat whom argued that firm growth is independent of firm size[2]. Put differently, Gibrat stated that small firms grow at the same rate as large firms. The power of Gibrat's law is the primary research question and a common theme that is carried throughout this thesis. In general, the extensive literature has rejected the law, but various studies have found that the law is valid for certain subsamples or time periods. Therefore, this thesis argues that the question is not whether Gibrat's law is valid, but rather when and with what restrictions it is valid. This thesis also tries to understand why Gibrat's law should be accepted. These questions are studied with help of the theory on firm and industry dynamics. This makes the approach more unique as existing studies focuses heavily on empirical testing. In other words, this thesis is a literature review.

Growth has been studied extensively and various models and frameworks have been presented. This thesis focuses on models were selection and firm-level heterogeneity are important factors. In addition, in these models the agent's primary action is profit maximization. Given the assumptions,

[1] Tekes (teknologian ja innovaatioiden kehittämiskeskus) is a Finnish government official that subsidies firms, organizations and universities in R&D-projects.

[2] When discussing firm size, evidently the first step would be to define what is meant by term. However, this is not maybe as easy as one might imagine. As Sutton (1997) states “"size" can be measured in a number of ways, and these arguments have been variously applied to measures of annual sales, of current employment, and of total assets. Though we might in principle expect systematic differences between the several measures, such differences have not been a focus of interest in the literature.” In general, this thesis will continue to treat size in a similar manner. One could assume that difference between different size measures move in the same direction and using a certain measure changes only the distance between the firms and not the order of the firms. Therefore, it is assumed that some kind of universal and unique measure exist that defines the firm size unambiguously. However, in certain cases it is assumed that some specific measure of size is used. In these cases, the choice of measure is indicated clearly.

the model's different results follow from this profit maximization. These holistic models have been defined as the industry dynamics literature[3]. One of the reasons why this family of models is chosen is that they are more conventional economics models. In addition, the models have a strong stochastic component. Gibrat has been justly criticized due to the model's lack of economics and it is interesting to see how the results change when economics is combined into a stochastic process. The models are also rich in implications and they provide an opportunity to touch on what actually creates growth. However, this is not the primary research question and the matter is discussed only briefly.

Hopenhayn's (1992) industry dynamics model is presented as the base case for the holistic and more traditional profit-maximizing models. In Hopenhayn's model there is a continuum of firms that produce a homogenous product. Firms can enter the industry once they have paid the cost of entry and they will exit once the firm's value drops below zero. The exogenous productivity shock, that follows a Markov process, is the model's central assumption. The key result is the existence of a stationary equilibrium where there is positive entry and exit. The stationary equilibrium is also a competitive one, meaning that there is no welfare loss. In Hopenhayn's model, firms grow because they are more productive than their competitors. Also, within the model smaller firms grow faster than larger firms. This is due to selection as small firms are more likely to exit. The reason for these exits is that small firms that don't grow fast don' have incentives to continue in the industry.

Naturally, Hopenhayn's (1992) model is not perfect and there are some unrealistic assumptions and factors that are omitted. Therefore, models that extend or are similar to Hopenhayn's model are presented. Jovanovic's (1982) model introduces the growth dependency on age and Cooley & Quadrini (2001) presents a model where size and age dependencies exist simultaneously. The model by Murto & Terviö (2010) is also presented. It is a hybrid of Hopenhayn's and Cooley & Quadrini's model. These models clarify the profit-maximizing firm's central role as the initial source of all results. Hence, the models further strengthen the view that Gibrat's law should be rejected. For example Jovanovic (1982) shows how the model can in certain special cases produce Gibrat-like results, but still growth is not random because it is a result of firms maximizing their profits. In other words, Gibrat's law is a special case of more elaborate models of industry dynamics as these models can explain Gibrat's results, but Gibrat's model can't replicate the results of the different industry dynamic models. Based on more elaborate models, new and more specific hypothesis on

[3] For comparison, the oligopoly strategy is another strand. Check from Sutton (1997)

the validity of Gibrat’s law can be formulated and further studied. Some of these hypotheses will be presented in the different subsections.

The thesis is organized in the following manner. Section 2 comprehensively presents Gibrat’s law including information on how the law has been studied empirically. Section 3 then presents the Hopenhayn’s model as a base case. Section 4 provides alternative explanations for why firm size could vary. Section 5 provides concluding remarks and suggestions for further research.

2 Gibrat’s law

The study of industrial organization has had a long history and within this field of research one of the fundamental questions has been market structure and the relationship between firm size and growth. Dynamics has had a central role when exploring the different relationships. This strand of industrial organization can be defined as the study of industry dynamics. According to Sutton (1997) it was Robert Gibrat that presented the first formal model of the dynamics of firm size and industry dynamic. The law of proportional effect and the subsequent model was presented in Gibrat’s book, Inégalités Économiques, which was published already in 1931.

This chapter will present Gibrat’s basic model and the implied results. It will also present some of the extensions and versions that have been proposed. The aim is to give a coherent and complete presentation on the mindset that is embedded into Gibrat’s model and its different variants. Further, this chapter explains how Gibrat’s law has been tested and subsequently what should be taken into account when testing the law. In general, Gibrat’s law is rejected, but there are also studies that would accepted the law at least for a subsample. In addition, there is some ambiguity in what is the correct way to test. Therefore, the outcome of this chapter is to conclude that the theory of firm dynamics should be revisited instead of continuing generically testing the law. Having another look at the theory should help clarify why the law could be valid.

2.1 The law of proportional effect and the stochastic firm size

Gibrat’s original work is based on the observations of skewed distributions in different areas of economics such as income or plant distributions. Gibrat proposed that the distribution of the firm’s size is lognormal and his aim was to give a theoretical and empirical justification for this proposal. He further argued that if the variable in question is transformed with an appropriate function, it could be shown that the transformation generated a normal distribution. According to Sutton (1997)

Gibrat followed the arguments of astronomer Kapteyn, who also had been interested in skewed distributions albeit in different environments. Nevertheless, according to Sutton (1997) Gibrat's main argument was that the skewed distribution consisted of a large number of small variables that were additive and independent of each other. The skewed distribution could then be converted into a normal distribution by transforming the initial variable x with an appropriate function into a variable z. Gibrat argued that the appropriate function for firm's size would be a logarithmic function. Gibrat tested his proposal with both income and plant size in the manufacturing sector and according to Sutton (1997) the goodness of fit was striking.

What today is understood as Gibrat's law is slightly different from what the original argument was. According to Loti et al. (2003) a common interpretation of the Gibrat's law presented in many articles is that a firm's growth rate and its size are independent of each other. It is good to note that this was only an assumption in Gibrat's model. For the sake of clarity, in this thesis Gibrat's law refers to the common definition that the growth rate is independent of firm size.

Mansfield (1962) presents a concrete example of Gibrat's law: a firm with sales of $100 million is as likely to double in size during a given period as a firm with sales of $100 thousand. Evidently, the firm with the $100 million sales will have a higher growth in absolute terms. Similarly, Sutton (1997) states that the "expected value of the increment to a firm's size in each period is proportional to the current size of the firm". It is important to distinguish between absolute and relative growth and therefore Gibrat's law states only that the relative growth is independent of the firm's size.

2.1.1 A formal presentation of Gibrat's law and argument

Gibrat's model is presented formally using Kalecki's (1945) presentation as an example. If one denotes number of workers at a certain date X_0 and assumes that number of workers undergoes a series of small random independent proportionate changes $m_1, m_2, \dots, m_t$, then at the end of the period the variable X_t is

1) $$X_t = X_0(1 + m_1)(1 + m_2) \cdots (1 + m_t).$$

The assumption that a variable undergoes a series of random independent proportionate changes is known as the law of proportional effect. According to Kalecki (1945) the law has been long known before Gibrat. As stated above, the distribution of equation 1 will be lognormal and if one wants to transform the distribution into a normal distribution then size has to be studied on a logarithmic scale. It is good to note that when Kalecki proves the transformation, the firm size is measured in a relative manner with respect to the average market size rather than looking at the firm's absolute size. The result should not be different whether size is measured in relative or absolute manner. It only has an impact on how the argument is constructed. Hence, if one would denote Y_0 as the deviation between the logarithm of X_0 and the mean of $\log X_0$ and if one would denote y_t, as deviation between $log\,(1 + m_t)$ and the mean of log (1+m$_t$), then the evolution of the firm's relative size on logarithmic scale could be presented as in equation 2.

2) $$Y_t = Y_0 + y_1 + y_2 + y_3 \cdots y_t$$

According to Kalecki (1945) it can be shown that "whatever the distributions of Y_t at the initial date, with the lapse of time the distribution of Y_t approaches normality more and more". The reason is that initial deviations' impact will diminish as time goes by and the distribution of $y_1 + y_2 + y_3 \cdots y_t$ will be approximately normal if the standard deviation of different components in the sequence have only a small impact on the sequences' standard deviation[4].

In addition to the above, there is couple of other assumptions that are needed in order to ensure that Gibrat's law holds. Rodríguez et al. (2003) point out that Gibrat assumes that there is no serial correlation. This means that the previous disturbance terms don't have an impact on the current disturbance term. If there were positive serial correlation, then past growth would simply generate higher growth in the future. Hamilton et al. (2002) note that "the variance, or volatility, of growth rates should be constant across all firm sizes for any given sector". If the variances would be different between large and small firms, then naturally it would lead to differences in the growth rates[5].

Bechetti & Trovatto (2002) state that Gibrat's law implies that, after controlling industry characteristics, the expected growth rates should not be affected by any other variable. Rodríguez et al. (2003) makes another interesting point that if the variance of logarithmic firm size increase with time, as Gibrat's law implies, and the number of firms stay constant then the industry concentration increases. Finally, Hamilton et al. (2002) point out that Gibrat didn't define the length of 'period'

[4] According to Kalecki (1945) Laplace-Liapounoff theorem guarantees this, which is a version of central limit theorem.

[5] The variances should be homoscedastic, not heteroscedastic.

for when the law of proportional effect was valid. The law could be valid for yearly growth, but as well as for growth measured during a decade. This could have implications on testing Gibrat's law.

Sutton (1997) states that Gibrat's aim was to convince his readers that the lognormal distribution and underlying purely stochastic process was a statistical regularity sufficiently sharp to provide a basis for serious mathematical modeling. This is what happened as Sutton (1997) continues that "during the 1950s and 60s, a substantial class of models appeared which combined "Gibrat's law" with a range of ancillary assumptions". Similarly Mansfield (1962) noted that Gibrat's law is a basic ingredient in many mathematical models designed to explain the shape of the size distribution of firm. Although the law was used in many models, the model was criticized quite early. For example when Mansfield (1962) tested Gibrat's law he rejected it four or seven times out of ten depending of different test specifications.

2.1.2 The variants, versions and mindset of Gibrat's law

It was already Kalecki (1945) that questioned Gibrat's model as it had some implications that seemed to be unrealistic. According to Kalecki (1945), the main problem is that the standard deviation (or variance) of the logarithmic variable increases with time and in many cases such an increase is not apparent. Since then, variants or extensions of the Gibrat's stochastic mindset have emerged in order to adjust the model into a more realistic direction. Kalecki himself presented one such variant where the variance of the relative logarithmic size was constant, but the size distribution remained lognormal[6]. This is done by assuming a linear negative dependence between the logarithms of relative growth and logarithms of relative size. There are also other extensions. For example according to Laitinen (1999) "the incorporation of the birth-and-death process in Gibrat's law leads to a Yule distribution when there is a constant rate of birth (Simon's model) or serial correlation between periodic growth rates (Ijiri-Simon's model)"[7]. In other words, the exact form of the distribution depends on assumptions made. In Gibrat's original proposal there were no "additional" assumptions and distribution was lognormal.

In addition to the different variants of Gibrat's law, it has been argued that the law is valid only for certain subsamples. Currently, three different versions have been identified. The standard version assumes that the law holds for all firms in a given industry. This includes also those firms that exit

[6] Kalecki (1945) also showed that the lognormal distribution could be transformed into a normal distribution given certain conditions.

[7] The distributions in the two models refer to the distribution of size, not the transformed distribution.

the industry within the study period[8]. The second version takes into account exit. Mansfield (1962) presented that the law should be only valid for firms that have survived. However, in sections 2.2 and sections 3.3.2 it will be discussed why the second version of Gibrat's law can never hold as small and slow growing firms are more likely to exit and this will result in a nonrandom sample. This sample selection bias will be discussed in more detail in the following section. Finally, the third version is that Gibrat's law applies only for larger firms. Formally this means that firm growth rate is independent of size only for those firms that have surpassed the level of minimum efficient scale (MES). The idea was put forward by Simon and Bonini (1958).

As Laitinen (1999) and Rodríguez et al. (2003) points out, it is essential to note that there is no optimal size for the firm in Gibrat's model and in the different versions, because there is no additional benefit from a specific size. This illustrates well the mindset that is embedded into the models. The firm's size is simply a stochastic process. The size evolves randomly over time and this leads to a skewed firm distribution. The distribution's exact form then simply depends on the assumptions. In this setting, managers can't generate additional growth by simply splitting a large company into smaller parts and public policies aimed at supporting growth by subsidizing small companies would be useless.

Rodríguez et al. (2003) notes that randomness could be generated from several factors that act in multiple fashion. They list executives' aversion to risk and industrial or political trends as possible factors favoring expansion in some cases and in others a reduction in size. Goddard et al. (2006) note that "Gibrat's law does not preclude the possibility that ex post, strong growth performance can be attributed to 'systematic' factors such as managerial talent, successful innovation, efficient organizational structure or favorable shifts in consumer demand". However, they further state that these factors can't be used to predict which firms grow, because "these factors are themselves distributed randomly across firms."

The logic is that the random variable completely captures the impact of all possible underlying factors. In other words, Gibrat has not categorized the different factors or their impact but instead assumed that these factors cannot be measured. Therefore, the best estimate for growth is a random variable that does not have any relationship with the firm size. In Gibrat's own empirical studies the rough estimate had a striking goodness to fit as Sutton (1997) pointed out, but in many studies, such

[8] Usually, in empirical studies the proportional growth rate of the firms that exit equals -1. At least according to Lotti et al. (2003), it is rather disputable whether this procedure is correct or not.

as Mansfield (1962) or Chesher (1979), the results were questioned. The empirical testing of Gibrat's law and the associated problems are the topic of following section.

2.2 How is it studied?

Gibrat himself originally tested his proposition by looking at the distributions of firm sizes and examining whether the distributions were lognormal or not. As pointed out, the results were impressive. After this, there has been an extensive amount of research on Gibrat's law during the past 50 – 60 years. The existing literature focuses mainly on testing the proposition with empirical methods. Most of the recent studies reject the law, although the results are not as straightforward as one would expect. There has been more diversity in the results as sensitivity analysis has been done. For example for certain time periods or for certain industries, studies fail to reject the law. In the following sections, the results are presented in more detail.

The past research can be categorized into two different classes. The first continues to test the law using Gibrat's original approach by evaluating a given distribution of firms and accept the law if it seems that the distribution is lognormal. Simon & Bonini (1958) and Reichstein & Morten (2006) are examples of the first type of research. The second strand studies whether Gibrat's law is valid or not by estimating whether the change is truly stochastic or not. This is usually done by looking at a panel of firms. The first strand was the dominating approach during the first 20 or so years. As more advanced econometrical methods have been developed, the second strand's popularity has increased considerably. Now, the majority of studies related to Gibrat test whether growth is stochastic or not. Examples of these studies will be presented below.

It is not only the development of more advanced methodology that has shifted the attention to the latter strand as there is also a principal reason in favor of it. Namely, a lognormal distribution does not guarantee that Gibrat's law would be valid. Weiss (1998) argues that the power of the first type of test is low since the relationship of growth rates to size is not explicitly investigated. Almus & Nerlinger (2000) go even further by stating that "testing whether the distribution of firm size is approximately log normal is not enough to verify Gibrat's law". Therefore, the following subsections will concentrate on presenting the results only from the second strand of literature as it is more relevant.

2.2.1 The foundation

The ideas put forward in Mansfield (1962) and Chesher's (1979) seminal paper has evolved into a foundation that majority of the studies use as a starting point. Equation 3 presents the methodology in its most basic form, the original logarithmic specification of Gibrat's Law. Size for company i at time t is presented by $S_{i,t}$. β_0 can be thought as an estimate of average growth rate for the whole industry or all firms. β_1 is the independent variable's coefficient that can be estimated for example with ordinary least squares (OLS) method[9].

3)
$$\log S_{i,t} = \beta_0 + \beta_1 \log S_{i,t-1} + \varepsilon_{i,t}$$

According to Lotti et al. (2003) "if both sides of equation 1 [in this thesis it is equation 3] are exponentiated, it becomes clear that if β_1 is equal to unity, then growth rate and initial size are independently distributed and Gibrat's Law is in operation. By contrast, if $\beta_1 < 1$ smaller firms grow at a systematically higher rate than do their larger counterparts, while the opposite is the case if $\beta_1 > 1$." In case large firms would grow faster than smaller firms, it would lead to explosive growth. This seems unrealistic at least for a longer period of time.

The primary null hypothesis is that β_1 is 1. In addition, there are usually two additional null hypotheses that are tested. These are the absence of serial correlation and homoscedasticity. According to Chesher (1979) failure of any of the three necessary conditions is sufficient to reject the law. The statistical significance and the validity of the null hypothesis are usually confirmed with a t-test or an F-test. Mansfield (1962) used Chi^2 for null hypothesis testing. This study will not go deeper into the merits of the different methods regarding hypothesis testing as it is not within the scope of the study.

Although the foundation seems to be simple, there are major challenges in testing that can have impact on the results and therefore they should be taken into account properly. The three major challenges are serial-correlation, heteroscedasticity and sample selection bias and they are presented in more detail in the following section.

[9] Equation 3 can be compared to Mansfield's (1962) equation 10.

2.2.2 Problems relating to empirical testing

It was already Chesher (1979) that showed how serial correlation can be a problem if there is no proper control. He argued that the serial correlation in the disturbance terms "may render least-squares estimators of β inconsistent, even though estimation proceeds using cross-sectional data"[10]. According to Dougherty (2002) "a consistent estimator is one that is bound to give an accurate estimate of the population characteristic if the sample is large enough, regardless of the actual observations in the sample". If there is positive serial-correlation meaning that past growth generates future growth, then the estimator overstates the estimated variable β_1[11]. The length of the time period over which growth is examined, ceteris paribus, has also an impact on the size of inconsistency. The overestimation decreases when the length of the period is increased. The main trouble is that the null hypothesis is accepted due to serial correlation even though it shouldn't be.

So it is not only that there shouldn't be serial-correlation, but if there is serial correlation and it is not taken into account properly, then the probability that the law is accepted increases. Chesher (1979) proposed equation 4 to control serial correlation where y_1 is $\beta_1 + \rho$, y_2 is $-\beta_1\rho$ and y_0 is the average growth rate of all firms. The null hypothesis is that y_1 is 1 and y_2 is 0[12].

4) $$\log S_{i,t} = \gamma_0 + \gamma_1 \log S_{i,t-1} + \gamma_2 \log S_{i,t-2} + \varepsilon_{i,t}$$

It is not clear how large an impact serial-correlation has on the results. Wagner (1992) found that the serial correlation was significant. Kumar (1985) noted that "there was some persistency in firm growth over time, but it was considerably weaker than was found for earlier periods." In contrast, Dunne and Hughes (1994) didn't find any evidence of serial correlation in their sample.

[10] According to Chesher (1979) the inconsistency can be derived "under the assumption that the process generating company sizes is observed a finite period of time after its commencement and that the parameters of the process remain constant as time passes." In addition, only first order serial correlated disturbance terms are taken into account.

[11] Serial correlation is formally defined in the equations below where ε refers to the disturbance term in equation 3. ρ is the variable that indicates of the serial correlation and u is the disturbance term. The ρ-values will range between -1 and 1. If ρ is larger than 0, then there is positive serial correlation and the estimate is overstated and vice versa.

$$\varepsilon_{k,i} = \rho\varepsilon_{k-1,i} + u_{i,k}$$

[12] Chesher (1979) elaborates that OLS regression to cross-sectional data on $S_{t,i}$, $S_{t-1,i}$ and $S_{t-2,i}$ may be expected to yield consistent estimators. Further, the estimates on β and ρ may be obtained from the equation below. However, sample information alone will not tell which of the estimates is which on the RHS, but one can assume that β should be close to unity.

$$(\check{\beta}, \tilde{\rho}) = \frac{1}{2}\{\tilde{\gamma}_1 + (\check{\gamma}_1^2 + 4\tilde{\gamma}_2)^{1/2}\}$$

Nevertheless, serial-correlation can be a problem, as showed above, so it should be controlled in order to ensure robust results.

Heteroscedasticity is another issue that should be taken into account. Heteroscedasticity means that the disturbance term's variances are not constant but rather a function of for example size or time. According to Evans (1987a) "previous studies have found that the variability of firm growth decreases with firm size, which suggests that u [the disturbance term], is not constant across firms". Dunne and Hughes (1994) report that "the source of this heteroscedasticity is often thought to lie in the greater stability of larger diversified firms compared to their smaller brethren", but according to them it is not the sole reason for heteroscedasticity. Evans (1987a) supports this view as he points out that failure of the heteroscedasticity test may indicate a number of possible problems. Nonlinearity is one of Evans (1987a) explanations.

Dunne & Hughes (1994) argues that there is a difference between the managerial talent between young and old firms as it would be logical to assume that young firms have less managerial talent and thus are more prone to error. Hence, heteroscedasticity could be induced by that large firms are typically old and this interdependence could create the differences in the variances.

As stated in the previous section, if Gibrat's law were too accepted, then there should be no heteroscedasticity. However, if there is no control for heteroscedasticity then there is a risk of a false positive meaning that the null hypothesis is accepted even though it shouldn't be. There are two reasons why heteroscedasticity causes false positives in standard OLS regression. The first is that the estimators of the standard errors of the regression coefficients will be wrong rendering the hypothesis testing invalid. In addition, heteroscedasticity causes inefficiency in the OLS estimators. Efficiency is a measure of reliability meaning that on average an efficient estimator gives more accurate results than an inefficient estimator.

According to Dougherty (2002) the magnitude of the problem depends on the nature of the heteroscedasticity and there are no general rules. He presents an example where the OLS estimated coefficient is doubled when there is no proper control for heteroscedasticity and in addition standard errors of the OLS estimator are underestimated[13]. The results on heteroscedasticity are varying.

[13] In the example the standard deviation is proportional to independent variable and the result has been estimated using OLS technique. Even though the disturbance term is argued to be decreasing in size by various studies such as Evans (1987a), the given example shows that heteroscedasticity can be a major problem if not properly taken into account.

Evans (1987a) didn’t find results on heteroscedasticity, but Hall (1987) on the other hand found evidence of heteroscedasticity. Wing & Yiu (1996) found that heteroscedasticity was significant.

As presented above, heteroscedasticity can be a problem, but fortunately there are ways to control it. The control proposed by White (1980) is used by majority of the studies. According to Dougherty (2002) ”White (1980) demonstrated that a consistent estimator of $\sigma^2_{b_2^{OLS}}$ [the variance of estimated coefficient when OLS] is obtained if the squared residual in observation i is used as an estimator of $\sigma^2_{u_i}$ [The variance of the disturbance term in i].” In other words, the control ensures that the t-statistic used for null hypothesis testing is accurate for large samples. Even though White’s test can make the estimator’s variance consistent, the OLS estimator(s) remains inefficient.

The final major problem is the sample selection bias. This issue was already mentioned in section 2.1.1, but it is covered here, in the following section and in section 3.3.2 in more detail as the issue is rather important. The main problem is how to treat firms that exit during the examination period. Mansfield (1962) was the first one to bring up this issue. He actually found out that there was a clear difference in results when only remaining firms were included into the dataset. Mansfield rejected Gibrat’s law in 70 % of the studied industries when he tested the standard version, but when he tested the law on only surviving firms the rejection percentage had decreased to 40 %[14]. Even though the amount of industries where the law is accepted increased, Mansfield rejects the law in general, because the OLS regression shows that the β_1 does equal unity in half of the cases. The intuition why the second version should always be rejected is presented in section 3.3.2.

Firms that exit wouldn’t be a problem if the number of firms that exit would be evenly distributed. However, if exit is not evenly distributed and for example small or young firms exit more frequently than large or old firms, then there will be problems when there are no proper controls. Lotti et al. (2003) sum up that “if survival is not independent of firm’s initial size – that is, if smaller firms are more likely to exit than their larger counterparts – the empirical test can be affected by a sample selection bias and estimates must take account of this possibility.“ They further point out that the sample selection bias naturally applies in particular to new and small firms, for which the hazard rate is generally high. The overrepresentation of small and fast-growing

[14] Mansfield (1962) tested the law first by classifying firms by their initial size and computing the frequency distribution of growth rate within each of these classes. He then uses a χ^2 test to determine whether the frequency distributions are the same in each class.

firms is not the main cause for sample selection bias. The real issue is that due to some other determinant the sample includes also some small and slow growing firms. These other determinants are captured by the disturbance term leading to correlation between the dependable variable and the disturbance term. In other words, it is the classical omitted variable bias.

In other words, the essential question is whether small firms are more likely to exit. The evidence is in favor that this would be case, but there are some results that contradict this. Evans (1987a,) found "the positive relationship between survival and size holds for 81 percent of the industries and the positive relationship between survival and age holds for 83 percent of the industries". Dunne & Hughes (1994) have similar findings as Evans, but they note that the relationship between firm size and death rate is not as straightforward as it seems. As a matter of fact, Dunne & Hughes (1994) find that the relationship is U-shaped rather than a linear one. Audretsch et al (1999) didn't find any evidence on small firms exiting with a higher probability. Calvo (2006) states that "there are big differences in size between those firms that survived and those that did not: the mean size of the surviving firms is four times that of those that closed".

The non-surviving small firms are not the only reason for sample selection bias. As Hall (1987) points out "some of the most rapidly growing and successful small firms may not be present at the beginning of the period, which will produce biases in the other direction". In other words, there are considerable incentives to control for sample selection bias and naturally there are many ways to do so. The approach presented by Hall (1987) and Evans (1987 a, b) has been favored by many studies. The three studies use a sample selection model which belongs to the Tobit model family. The model will be explained in the next section.

2.2.3 The Tobit models and multivariate models

In the "foundation" -approach presented above, the firms that didn't survive were given an arbitrary growth rate such as -1 or were left out of the sample depending on which version of Gibrat's law was studied. Exiting firms can be studied more elaborately. According to Verbeek (2008), Tobit models are used to study processes where the dependent variable is continuous, but its range may be constrained and "most commonly this occurs when the dependent variable is zero for a substantial part of the population, but positive (with many different outcomes) for the rest of the population". In the case of Gibrat's law, Tobit models are used to model the survival of firms in order to test and correct for sample selection bias[15]. Intuitively this means that in the Tobit model the impact of the non-surviving firms is attempted to estimate more analytically rather than setting bluntly the growth rate at -1. According to Dougherty (2002) a Tobit model consists of two components, which are a probit model and a standard regression.

For the sake of simplicity, the probit model is explained first as a standalone component. In a probit model, a dependent binary choice / variable that takes the value of either 1 or 0 is studied. The aim is to estimate what factors have an impact on the occurrence of the binary dependable variable. Equation 5 illustrates, where $y*$ is the binary variable and β are the estimated coefficients[16].

5)
$$\begin{gathered} y_i^* = \beta x + \varepsilon \\ y_i = 1 \quad if\ y^* > 0 \\ y_i = 0 \quad if\ y^* < 0 \end{gathered}$$

In the case of Gibrat's law, the binary variable would naturally be firm survival. The starting point is to estimate a linear function of the variables that determine the probability of the choice. An example of such a function could be equation 3. The binary variable is then estimated by fitting the data to a standardized cumulative normal distribution[17]. According to Dougherty (2002) "maximum likelihood analysis is used to obtain estimates of the parameters".

In a Tobit model, the probit analysis is used to analyze a latent variable y_i^*. The difference between a Tobit model and a probit model is that the latent variable (i.e. the dependent variable estimated by the probit model) can also take other values than 1 and more importantly the latent variable is used in further analysis according to Verbeek (2008). As stated above, there are many different variants of the Tobit model. The key difference between a standard Tobit model and a sample selection model is on what basis an individual observation is included into the sample. In the former variant, sometimes denoted a type I Tobit model, the selection is based on the same principle as the regression itself is done. In the latter model, the sample selection model, the selection, as the name implies, is (partially) separate from the regression model. The sample selection model is the one used in the studies performed by Hall (1987) and Evans (1987 a, b) as they assume that the survival

[15] According to Verbeek (2008) there seems to be a strong belief that a Tobit model could eliminate sample selection bias, but this is certainly not generally true.
[16] $y*$ also known as the latent variable,
[17] In a logit model, the linear function is fitted to a logistic function.

process is different from the growth process. Therefore, it is discussed in more detail. The model is presented formally in equation 6.

6)
$$\begin{aligned} y_i^* &= \beta_0 + \beta_1 x_{1i} + \varepsilon_{1i} \\ z_i^* &= \beta_2 + \beta_3 x_{2i} + \varepsilon_{2i} \\ y_i &= y_i^*, z_i = 1 \qquad if\ z^* > 0 \\ y_i\ is\ not\ observed, z_i &= 0 \qquad if\ z^* \leq 0 \end{aligned}$$

In equation 6, the latent variable is represented by z_i^*. This is unobserved and as stated it is estimated with the probit model. In Hall (1987) the latent variable is estimated using the function of firm characteristics such as industry or beginning of period size. Evans' (1987 a, b) latent variable can be thought of as the value of remaining in business in excess of opportunity cost. The latent variable is thus used to select the appropriate sample. The observation is included into the sample if the latent variable is estimated to be over 0. Y_i is then the actual variable that is studied and in both Hall and Evans y is the growth rate. Y_i is the observable growth rate and z_i^* is the estimated growth rate[18].

There are two possible ways to perform a type II Tobit regression. In the Heckman two-step procedure, first the latent variable is estimated with a probit model in order to study whether the observations should be included or excluded into the sample. The following step is to perform a regression on the remaining sample (i.e. for those that $z^* > 0$). However, if the regression is simply done on the selected observations, then the estimates will be inconsistent, because according to Dougherty (2002) the expected value of ε_{1i} is nonzero for observations in the selected sample if ε_{1i} and ε_{2i} are correlated. Fortunately, the exact expected value can be deduced analytically and it is presented in equation 7, where $\sigma_{\varepsilon 1i \varepsilon 2i}$ is the population covariance between ε_{1i} and ε_{2i}, $\sigma_{\varepsilon 1i}$ is the standard deviation of ε_{1i} and λ_i is the inverse Mill's ratio[19]. According to Dougherty (2002), the sample selection bias using only the selected observations can be thought of as an omitted variable bias, with the λ being the omitted variable. Naturally, this will lead to that λ will appear in the disturbance term creating a correlation between the disturbance term and the independent variable.

[18] The standard Tobit Mode / type I Tobit model is a special case of the type II Tobit model where $y_i = z_i$.

[19] The inverse Mill's ratio, λ_i, is

$$\lambda_i = \frac{f(v_{i)}}{F(v_i)}$$

7) $$E\ (\varepsilon_{1i}|\varepsilon_{2i} > -\beta_2 - \beta_3 x_{2i}) = \frac{\sigma_{\varepsilon_{1i}\varepsilon_{2i}}}{\varepsilon_{1i}}\lambda_i$$

The correct way to take this omitted variable bias into account is by adding the result of equation 7 as an explanatory variable to the regression equation $y_i^* = \beta_0 + \beta_1 x_{1i} + \varepsilon_{1i}$. It is good to note that the omitted variable can always be calculated as the needed inputs to calculate the variable depend only on the selection process. The Heckman two-step procedure will lead to consistent estimates.

The second possible way to estimate a type II Tobit model is to perform a maximum likelihood analysis. According Verbeek (2008) both methods will produce a consistent estimator, but the two-step procedure will not be efficient. However, some caution should be used when using a type II Tobit model. Verbeek (2008) notes that "routinely computed OLS standard errors are incorrect, unless the covariance is 0". Another issue is that if the independent variables would be the same both in the selection equation and in the regression model then there would be a problem. Dougherty (2002) states that at least one selection variable should not be in the regression model. Finally, he also points that if the selection variable is unjustly added in the OLS regression it can have a significant impact even if it shouldn't[20].

In addition to introducing a new framework to test Gibrat's law, Evans (1987 a, b) also used a new linear function in the Tobit estimation instead of the one presented in equation 3. The novel factor was to introduce age as a possible explanatory variable for the dependent variable[21]. The theoretical justification that age could have an impact is given by Jovanovic (1982) whose model is presented in section 4.1. Evans's linear function used in the Tobit estimation is presented in equation 8 where A_t stands for age, S_t for size, B_t for number of plants and μ is disturbance term[22]. Evans estimates the growth function g by taking a second-order expansion in the logs. Hall (1987) used size as the independent variable. Another noteworthy point is that instead of estimating the relationship between the current periods and previous period size, Evans (1987 a, b) and Hall (1987) regressed the linear function directly on growth.

8) $$\frac{[Ln\ S_{t\prime} -\ Ln\ S_t\ ]}{d} = \ln g(A_t, S_t, B_t) + \mu_t$$

[20] I.e. the coefficient would be nonzero when it should be zero.

[21] Number of plants was also considered as a factor, but it was dropped because preliminary results found it insignificant.

[22] This was the planned equation, before the preliminary results.

Evans (1987 a) studied 20,000 manufacturing firms and Evans (1987 b) studied the firms in the Small Business Data Base "which was constructed by the Office of Advocacy of the U.S. Small Business Administration (SBA) from information originally collected by Dun and Bradstreet for its credit reports". Both of Evans' studies start from 1976 and ends at 1982. Hall's (1987) two datasets consist of publicly traded manufacturing firms. The first dataset contains all firms with employment data from 1972 to 1979 and the second dataset extends from 1976 to 1983. In total 962 firms were in both samples.

All three studies clearly reject Gibrat's law. Evans (1987 a) clarifies that "the negative relationship between growth and size holds for 89 percent of the industries and the negative relationship between growth and age holds for 76 percent of the industries". Hall (1987) states that "with respect to the size-growth relationship, we have negative results in the sense that neither measurement error (serial correlation) in employment nor sample attrition can account for the negative coefficient on firm size in the growth rate equation." Both of Evans' studies found that "the departures from Gibrat's law tend to decrease with firm size."

Hall (1987) didn't find evidence that would suggest serial correlation. Evans didn't treat the matter in either of the studies. Also, heteroscedasticity was studied and Evans (1987 b) found results that would imply of heteroscedasticity, but controlled it with White's test[23]. Evans (1987 a) and Hall (1987) didn't find the heteroscedasticity to be significant, but nevertheless controlled it with White's test. As stated above, all the studies used the type II Tobit model to correct for sample selection bias. All three studies found that the firm size had an impact on the survival.

Multivariate models have become a standard after Hall's (1987) and Evans' (1987 a, b) studies. Also other potential explanatory variables have been introduced. For example Hamilton et al. (2002) studied a multivariate model where legal form was one of the additional variables[24]. Dunne & Hughes (1994), Wing & Yiu (1996) and Rodriguez et al. (2003) found a negative relationship between age and firm size, but in Hamilton et al. (2002) the results were more mixed. Bechetti & Trovatto (2002) found some evidence that finance has an impact, but the magnitude of the results

[23] White test checks that the residual variance of a variable in a regression model is homoscedasticity.

[24] Hamilton et al. (2002) found that legal form had impact. For example publicly traded firms grew faster than private firms. Hamilton et al. (2002 suggest that "this result is the outcome of capital constraints faced by private firms, as well as their bias against high-risk initiatives where they face unlimited liability."

vary depending on whether non-surviving firms are included or not. Johansson (2005) found that the effect of government ownership is insignificantly negative in all regressions, but Wing & Yiu (1996) found that government ownership had a role.

To sum up, Evans (1987 a, b) and Hall (1987) were the first studies that took sample selection bias into account. The studies have had a significant impact on the study of Gibrat's Law. Evans (1987 a) states himself that "the major contribution made here is to test Gibrat's law for theoretically relevant samples of firms after controlling for sample censoring." In addition, new independent variables were introduced. Naturally, new testing methodology has been introduced after the three studies. One example is the different panel tests. However, this will study will not go deeper into these new methodologies.

2.3 A critical look on the testing and results

In general, the empirical studies reject Gibrat's law, but the rejection is not as unanimous as one would expect. In other words, there are also studies that contradict the general rejection of the Gibrat's law, at least for some subsamples. One example of such studies is done by Vander Vennet (2001) who studies the growth of national banking sectors in the OECD countries. Although Gibrat's law relates to firm size, Vander Vennet (2001) argues "that the evolution of individual banks will to a large extent be determined by the economic and regulatory environment of their home country."[25]

The study is performed using the OLS method following the specification of Mansfield-Chesher studies. In other words, equation 4 is used to test and therefore serial-correlation is taken into account. There is no need to take sample selection bias into account as all the national banking sectors are included. The study starts from 1985 and ends in 1994. For this period, Gibrat's law is rejected. The study is then split into two sub-periods. The first one covers the years from 1985 to 1989 and the second study covers years 1990 – 1994. For the former, Gibrat's law is rejected, but during the second period the law is accepted.

There are other studies on the financial sector such as Goddard et al. (2002) that studied the credit union sector in US and found rather diverse results. In general they reject Gibrat's law, except when

[25] Vander Vennet (2001) continues that "this is especially true in terms of the available strategic options in areas such as functional de-specialization, degree of internationalization, access to funding and capital. Therefore the evolution of aggregate national bank sectors in the OECD area is analyzed"

they estimate the growth of total memberships using panel testing techniques. Nevertheless, Goddard et al. (2002) found that for univariate models Gibrat's law is accepted in some periods. Actually for the multivariate models Goddard et al. (2002) found that "in general larger credit unions grew faster than their smaller counterparts, and that there is a positive relationship between size and age." irrespective of whether the estimation was done using OLS or panel techniques. Interestingly, Das (1995) found a negative relationship between growth and size, but a positive one between age and growth when studying the computer hardware industry in India.

It is interesting how the result for the same industry can fluctuate with different time periods. Vander Vennet (2001) noted that he studied the different time periods separately because a series of major deregulation initiatives were implemented at the end of the 1980s. It is likely that this changed the results[26]. The result would imply that institutions and regulations could have an impact on whether firm size has impact on the growth rate. Findings by Audretsch et al. (1999) support this view as they report that "there is virtually no evidence to link firm size with survival" in a study on Italian manufacturing. Their result contradicts earlier studies for other countries such as Germany, United Kingdom and the United States. The reason they present is that the underdeveloped and highly imperfect Italian capital market entails barriers to entry which lead to "a pre-entry selection process which selects only those characterized by the choice of more capital intensive production techniques, techniques, larger availability of internal finance and easier access to outside financing." To sum up, institutions, regulations and other tacit factors have impact on the results.

Also, business cycles could be a possible explanation why the results vary over time. Hardwick & Adams (2002) report that "for example, small firms may tend to grow faster than larger firms during an economic boom (as experienced by the UK at the end of the 1980s) in response to greater consumer confidence and higher spending". With respect to the business cycle and institution aspect, it is good to keep in mind that Gibrat didn't specify for what time period the law is applicable and hence there are a variety of different time periods to be used. In general, a ten year span is used, but there are also studies that have a longer time span than this. For example Mazzucato (2003) studies Gibrat's law over a 30 year time span and found that the law describes the statistical process of firm growth better in the early phase of industry evolution in the auto and

[26] Vander Vennet (2001) continues "examples include interest rate deregulation, liberalization of capital flows and a harmonization of bank capital requirements initiated by the Basle Committee".

PC industry[27]. Furthermore, he finds that in the later phases the firm growth rates tend to be more stable and structured, which would imply that the law should be rejected.

In addition to the three major challenges in testing Gibrat's law presented earlier, there are still challenges that need to be taken into account. Size is one such example. The original study performed by Gibrat used the number of employees as a measure for size. Naturally, there is quite a variety of different size measures and it doesn't seem that one leading measure would have emerged. Rodriguez et al. (2003) for example test the law using total net assets, operating income, added value and equity as measures of size. However, the measure of size also seems to depend on the industry in question. The studies on the financial sector, such as Vander Vennet (2001), seem to favor the total assets as a proxy for the size. Johnasson's (2004) study on Swedish IT follows the original line of Gibrat by using employment as the measure. Wing & Yiu (1996) also used employment as a measure in a study on Chinese manufacturing firms.

Johansson (2004) rejects Gibrat's law as null hypothesis, but Wing & Yiu (1996) actually fail to reject the null hypothesis that β_1 in equation 3's is 1 and the null hypothesis that there is serial correlation when measuring firm size in terms of number of employee. They state that "this suggests that the employment growth of firms in the current period is on average independent of the employment growth in the last period." What is interesting is that Wing & Yu (1996) actually reject the same hypothesis when they use output as their measure of size. However, they reject the Gibrat's law in whole because both measures show that there is heteroscedasticity. But it is interesting that a simple change of measure can alter the results in such a drastic manner.

Unfortunately, size is a non-trivial matter. For example, the third version of Gibrat's law states that the law is only valid for a subsample of firms that have a surpassed a level of minimum efficient scale (MES). The reason for this could be in the relationship between survival and firm size. Audretsch et al (2004) state that "as long as the likelihood of survival is also independent of firm size, Gibrat's law would be expected to hold for a reasonably large sample". One could argue that for firms that have surpassed the MES-level the link between survival and size doesn't exist anymore which will lead to a Gibrat-like growth pattern. This would imply that small firms are mean-reverting if they are under the MES-level. However, if the likelihood of survival would be linked to a certain specific measure, then the testing of Gibrat's law could go wrong simply by

[27] Mazzucato (2003) defines the first 30 years as the early phase of industry evolution.

choosing the wrong measure. Furthermore, in study on farm size, Weiss (1998) found that there could actually be two scale-thresholds. Therefore, it can be the case that it is not enough to only study the subsample of firms that are over the MES-threshold as “medium-sized” firms converge to the second threshold.

Heshmati (2001) studied Swedish micro and small firms in the region of Gävleborg during 1993 to 1998[28]. The data was not split into sub-periods, but what makes the study interesting is that the results are very sensitive to functional form and estimation methods of the panel technique. Sample selection bias is taken into account in all different methods by adding the inverse Mill’s ratio as suggested in section 2.2.3. Gibrat’s law is rejected using the OLS method. In addition, Heshmati (2001) rejects OLS as a method as there is considerable firm-level heterogeneity and it would lead to either over- or understatements. Instead a multi-step generalized least square (GLS) estimation procedure is used. He finds that the growth-size relationship is negative when using the “employment model” and positive when using the “sales model”, respectively. Heshmati (2001) notes that, in general, employment growth model are found to be more sensitive to the choice of functional form and estimation method of the panel technique[29].

In other words, there are a quite many ways to test Gibrat’s law and it is not 100 % clear that what the correct way is. It is likely that sample selection bias is present and this would be in favor of Tobit models. However, Verbeek (2008) noted that Tobit models may not be adequate to correct for sample selection bias. On the other hand, according to Goddard et al. (2002) panel techniques should be in favor of cross-sectional regression because the cross sectional regression suffers from a loss of power. As stated above, Heshmati (2001) also favors panel techniques. However, the panel techniques also have their limitations, namely the lack of proper data. This means that there is a lack of longitudinal data sets tracking the evolution of firms. In addition, there is some ambiguity regarding which panel test actually should be used as Heshmati (2001) reports different results for the fixed effects model and the random effects model. Given all the challenges presented in this section, the following section presents a new way to look at the matter.

[28] According to Heshmati (2001) these firms have employees between 1and 100.

[29] Difference can be created by treating the disturbance term differently. Heshmati (2001) elaborates that there are two types of model which are the fixed effects (FE) model, where the disturbance term is assumed to be fixed and correlated with the explanatory variables, and second, the random effects (RE) model, where disturbance term is assumed to be random and not correlated with the explanatory variables.

2.4 The next step?

As stated before, the literature on testing Gibrat's law is rather extensive as it covers various time periods and different industries. In addition, the law has been tested with various methods and different datasets. The general result is that Gibrat's law can't be a "law" in general or in a strict sense, but there also some mixed evidence. There have been time periods and industries when Gibrat's law is accepted. Audretsch et al. (2004) sum up that in general for larger firms Gibrat's law tend to be valid and Evans (1987 a, b) similarly reports that deviations from the law are smaller for larger firms. It is not only the size of the firm, but also the general business cycle can have some impact on the results. The question is not whether Gibrat's law is valid or not, but rather when and with what restrictions is it valid.

As reported in the previous section, for example, Vander Vennet (2001) found that Gibrat's law was valid in a sub-period after capital markets were deregulated and as stated Vandet Vennet's study is not the only one indicating that sometimes the law should be accepted. Also, the sample selection bias and the exclusion of the very smallest firms can have an impact. As discussed in section 2.3, the actual methodology of the empirical test can also have an impact on the results. Given all these challenges in empirical testing and the fact that there is already a substantial amount of literature, this thesis proposes that rather than empirically testing Gibrat's law, a step backwards should be taken and the theoretical framework and mindset on firm growth should be studied in more detail.

It is the hope that by analyzing theoretical models and searching for theoretical justifications, cases and scenarios could be found and pointed out that would support Gibrat's law. A concrete aim of revising the firm growth's theoretical framework is that more specific testable hypotheses could be formulated rather than the traditional null hypotheses. The different and more specific hypotheses could then be one by one falsified by further testing with preferably different empirical methods to ensure more robust results. Also, there is the additional benefit that other variables that have an impact on growth could identify as by-product of this structured analysis.

2.4.1 Possible reasons to reject Gibrat's law

The empirical studies have presented various theoretical reasons to reject Gibrat's law. One of the most popular explanations is that smaller firms are better at innovating. Hamilton et al. (2002) sum up that "one explanation commonly put forward for the finding of faster growth of small firms is that they have a greater capacity to innovate at least in specific technological environments".

Basically, this is the Schumpeter Mark 1 argument, which argues that SME's are more likely to provide the bulk part of innovations. Also, as discussed the size question is interdependent with the age question, as young firms tend to be smaller. Moreno & Casillas (2007) present the argument that in general young, and hence small firms, are not only better to innovate, but also more proactive and less risk averse than older firms. Their argument is that risk taking and pro-activity is the very essence of young firms as it provides the opportunity to exist and thrive.

Calvo (2006) finds that both product and process innovating firms have grown more than non-innovating firms. If it is indeed that innovations leads to growth and small firms innovate, then this would explain why Gibrat's law should always be rejected. It also would mean that if a subsample were studied, consist of non-innovating and large firms, it could be the case that Gibrat's law would be valid for this sample. On the other hand, according to Schumpeter Mark 2 small firms wouldn't involve themselves in many R&D projects as it is very costly for SMEs to finance it. In addition, it has been noted that large firms have the alternavite to diversify risk over many R&D projects. Consistent with this view, Evangelista et al. (1998) found that innovation is more prevalent in large firms in a study on large European firms. However, the statistics don't tell everything as according to Ortega-Argilés et al. (2010) small firms carry out informal R&D. In addition, Van Dijk et al. (1997) have found that small firms tend to produce more patents and innovations than larger firms by unit of input invested in R&D. To sum up, there is seems to be ambiguity what is the role of innovation for different types of firms.

Moreno & Casillas (2007) further continue their argument that in general young and small firms are more flexible and thus they have less rigid routines. It is the flexibility that enables the firm to find and to create new growth opportunities. This relates to another topic that explains why large firms grow at a slower pace. The theory is that larger firms have higher agency costs. This idea is not new as according to Hamilton et al. (2002) it was already Penrose (1959) that argued that difference in growth rates were present due to differences in internal resources and notably the existent of diseconomies of scale in managerial coordination as the organizations grow. This means that the ability to allocate resources within the firm decreases with the size and the complexity of the organizations. Similarly, Wing & Yiu (1996) presents the idea of structural inertia that explains "how the internal organizational structure interacts with the environment." This leads to that older and larger firms are slower to change, because these changes can undermine their accountability and reliability. In other words, large firms want to manage their reputation more carefully as they have more to lose from a possible downfall.

Dunne & Hughes (1994) argue that large firms grow slower, because they are more diversified and thus more prone to more variable growth rates leading to decreased overall growth rate. On the other hand, reasons why small firms would grow slower can also be identified. Moreno & Casillas (2007) state that most of the studies on growth report that "small and medium-sized firms consider that the accessibility to sufficient financial sources is either a handicap or a brake to growth". Finally, Hamilton et al. (2002) note that large firms require large absolute growth in order to keep their growth rates constant. They give an example that "10% growth for a $4 million company requires $400,000 in expanded markets, while a $4 billion firm would need to secure a new market of $400 million – an exponentially more difficult task".

2.4.2 The introduction of the holistic approach

The above reasons have certainly an impact on the growth and are valid reasons why small firms could grow faster. However, they are in a sense scattered as they can't be (easily) extrapolated into general rules that would more strictly define when Gibrat's law is valid. The role of innovation is a good example as there were many factors both in favor and against that small firms benefit from innovation and hence grow faster. It was argued in section 2.3 that Gibrat's law should be accepted for larger firms that have surpassed the MES threshold. However, this is not always the case as there are many studies done solely on large firms with different methodologies that reject the law. The results vary from industry to industry. A tractable and structured framework would help to identify all the relevant factors and clarify Gibrat's law in more detail. One idea could be that if certain industry characteristics that supports the law could be indentify with the help of a theoretical framework, then one could test only these industries? The finance sector could be an example industry.

Also it is good to remember that although there seems to be a statistical relationship between firm growth and size, it doesn't necessarily mean that small firms grow fast, because they are small. It could be very well the case that (small) size is simply a proxy for something else. Stam (2010) notes that the statistical relationship "does not necessarily improve our insight into the role of growth processes and strategies for firm growth, as firm size and firm age can be indicators for multiple mechanisms (e.g., economies of scale, learning effects, reputation effects)". This encourages studying the growth of the firm in a more analytical manner as one could more easily identify cases when size is only a proxy.

There are certainly many different frameworks and models to analyze Gibrat’s law and growth, but a logical step would be to analyze the problem in a more traditional economics framework. As Weiss (1998) notes that “two facts about this model are remarkable: its parsimony and its lack of economics”. In other words, a traditional maximizing framework should be presented. Actually, this is what happened as new maximizing models were introduced. According to Sutton (1997) “the aim was to move instead to a program of introducing stochastic elements into conventional maximizing models”[30]. Jovanovic (1982), Hopenhayn (1992) and Ericson & Pakes (1995) represent these new types of industry dynamic models. This strand of literature looks at firms, and subsequently at growth, in a more holistic manner by incorporating the growth question into the profit maximization problem. In other words, the evolution of firms’ size is not a standalone process or phenomena, but rather a result of firms optimizing their production.

Firm-level heterogeneity, entry and exit are other common features that have an important role in these models. Stochasticity is used to impose the firm-level heterogeneity. In other words, the randomness that was central in Gibrat’s law is now only a part of a larger model. Selection is a common theme in these models as according to Armington & Acs (2004) all these models suggest that firm growth rates results from the effects of “noisy” selection and incomplete information.

Although there are a lot in common between the new type models of industry dynamics, Hopenhayn’s (1992) model is able to capture many of those essentials improvements in a tractable manner. The modeling of the random element is also close to the mindset of Gibrat’s law although there are some different assumptions namely on the persistence of the growth. Nonetheless, Hopenhayn’s model will be presented as a base case for the new models of industry dynamics in the next chapter. Extensions to Hopenhayn’s model and other similar models are then presented in chapter 4.

[30] However, Sutton (1997) argues that stating that the earlier models were "not maximizing models" is misplaced. He continues that “what is striking about the "stochastic growth models" is not their lack of "optimizing agents," but their reliance on Gibrat's Law”. Nonetheless, this thesis makes a clear distinction between the model similar to Gibrat’s and the new type of models of industry dynamics. There are two reasons. This thesis’ focus has been on the original Gibrat’s law and not on the extensions that are more complicated. Secondly, as will be seen in chapters 3 and 4, there are large structural differences between the two model types such as non-existing production optimization in the strand of stochastic models.

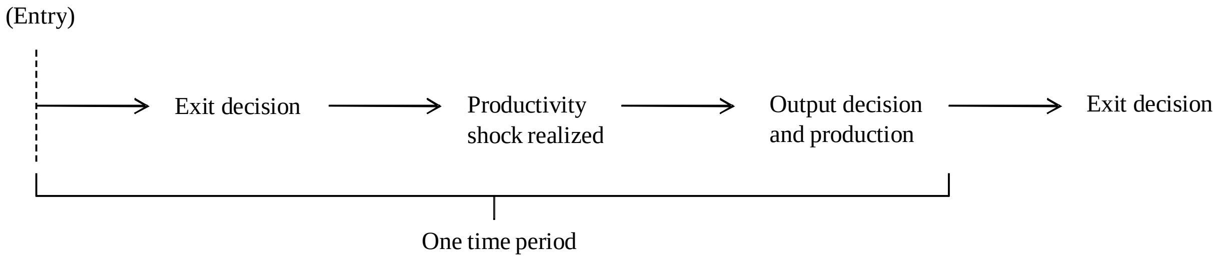
3 Firm-level heterogeneity, entry and exit

The aim of this chapter is to give a comprehensive presentation on Hopenhayn's model in "Entry, exit, and firm dynamics in the long run equilibrium" in order to justify why Gibrat's law can't be valid and further elaborate what creates growth. This chapter is divided in the following manner. First, the assumptions and the model are presented. After this, the competitive and stationary equilibriums are discussed and what are the implications of the model. Finally, a critical view on the model and a path for the future analysis are presented.

3.1 The description of the model

In Hopenhayn's (1992) industry dynamics model, there is a continuum of firms that produce a homogenous product in a competitive market. It is assumed that labor is the only input of the firm although the model can be easily extended to cover multiple inputs. In addition, an exogenous shock, φ, defines how productive the firm is. The shock is the only source of uncertainty in the model and it follows a Markov process independent across firms with conditional distribution $F(\varphi'|\varphi)$. The shock has a crucial role in the model and it will be discussed in more detail in the coming paragraphs.

Time is discrete and the incumbent firm has two possible actions in each period. Before observing the productivity shock, the incumbent firm can either exit the industry or it can decide to continue and produce. It is good to note that the output decision is made after the shock is revealed and thus the uncertainty is resolved for that period. The inverse demand function gives the aggregate demand and it is assumed to satisfy general conditions (e.g. demand is strictly decreasing, continuous). Picture 1 illustrates what actions and events take place during one time period and in what order.



Picture 1 - The incumbent firm's actions and events during one period

3.1.1 Exogenous productivity shock as the source of firm-level heterogeneity

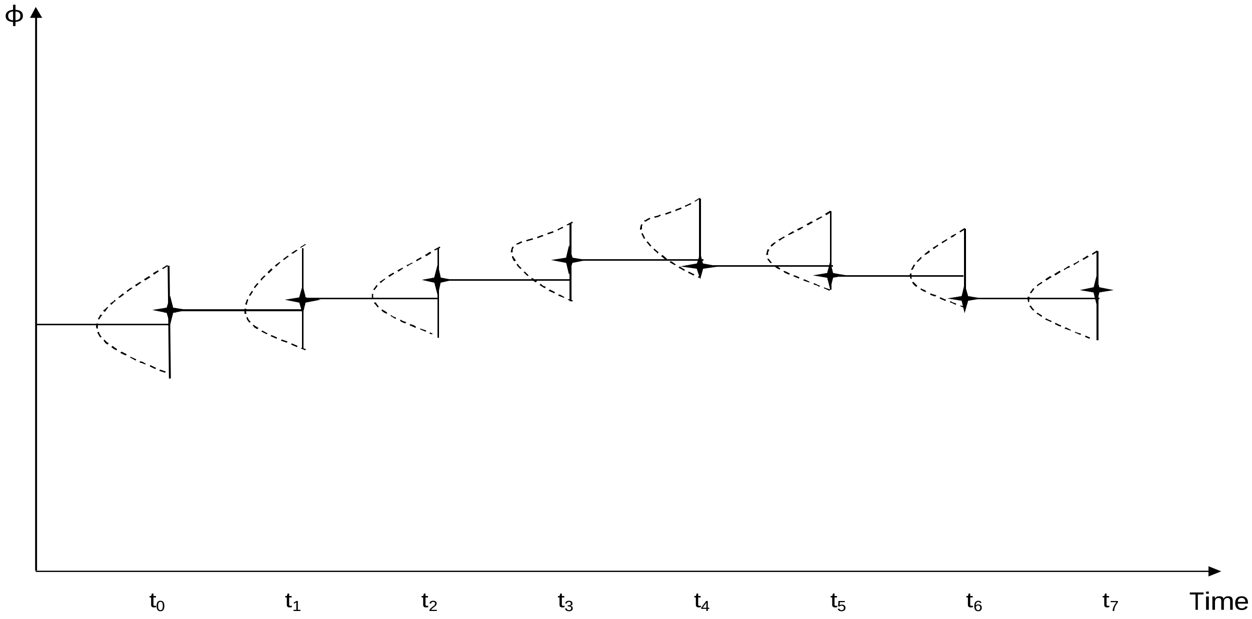
The size of the productivity shock is normalized to be on an interval between zero and one. Naturally, a higher productivity shocks is better for the firm. As stated above, the productivity shock will have a significant role in the model as the shock will be the main factor that explains why firm's size fluctuates. An intuitive interpretation would be that the shock reflects different skill levels. What does it then mean that the shock follows a Markov process independent across firms? First of all, the shocks are firm specific meaning that firm's shock has only direct impact on the firm itself and consequently firm's realized shock doesn't alter others' productivity. A common definition of a Markov process is that it is a process that has the property that given the current realizations, future realizations are independent of the past[31]. This basically means that current information / situation / level is the only that matters when forming the expectations. However, it doesn't mean that history wouldn't have an impact. The correct interpretation is that Markov processes are path independent meaning that it doesn't matter how the productivity level is reached if the levels are the same. Usually, the Markov process is assumed because it is much easier to treat from a mathematical point of view and it is a fairly good generalization.

Hopenhayn further assumes that F, the conditional distribution of the productivity shock, is strictly decreasing in φ. Hopenhayn interpret that this means that the higher the productivity shock in period t the more likely are higher shocks in period $t + 1$. According to Harris & Li (2010) this implies that productivity levels are persistent. In other words, there is some level of serial-correlation. This assumption is a direct contradiction to the implied assumption in Gibrat's law, where there was no serial correlation. That said, the level of serial correlation is not specified and Hopenhayn's model does not assume that the productivity shock will stay high / low for certain. There are fluctuations and the shock can evolve almost without boundaries[32]. Hopenhayn simply assumes that is more likely that the shock stays roughly the same. The persistence of the productivity shock will have a considerable role as will be discussed in the results-section.

Picture 2 presents a possible path for the productivity shock φ. In the example, the firm gets a good start and the productivity start to increase. This development continues for few periods, but then something happens and in period 4 for some unknown reason the productivity is lower than in the previous period. How could one justify this sort of productivity path? One possibility is learning by doing, where the firm gets better the more it produces. Some firms are simply better at learning and

[31] See for example Stokey & Lucas (1989).
[32] The shock is restricted to be between zero and one.

evolve faster. But what would explain the drop in the productivity, that occurred at t = 4 in the example below? Learning by doing can’t explain this and Hopenhayn itself doesn’t provide an intuitive reasoning for productivity shock. The exogenous productivity shock just exists.



Picture 2 - An example of productivity shock evolution

Normally, a Markov process implies that the future would be impossible to predict. This is true on firm-level, but Hopenhayn argues that on aggregate level this is not true. To the contrary, the aggregate output, the employment, the prices, and the frequency distribution will be deterministic, because “the frequency distribution for the idiosyncratic shocks each period coincides with the probability distribution dictated by the initial distribution, the conditional distribution function, and the entry and exit rules” and in addition there is no other source for aggregate uncertainty[33]. In other words, the path for example for the aggregate output and the employment are known. This is also true for both the input and the output price and firms can therefore make decision based on perfect foresight.

3.1.2 The rationale of exit

Although the productivity shock is persistent, eventually the productivity will fall below such a level that firm has to carefully scrutinize whether is wise to continue or not. This fall will be

[33] According to Hopenhayn (1992) the fact that “there are a large number of firms and since the conditional distribution function F and the probability measure v over initial states are the same for all firms” guarantees this.

inevitable, because it is assumed that the life span of the firm is almost surely finite. The assumption is certainly realistic as it hard to come up with examples of firms that would have operated forever. The exit decision will be made before observing the next period’s productivity shock and it is optimal for the firm to exit when the productivity shock goes below the reservation value for the first time. According to Hopenhayn, firms form the reservation rule in the same manner, but each firm will have their own optimal reservation value. The reservation value is the value of the productivity shock when the expected value of the firm is zero and thus the firm is indifferent between continuing and exiting. In other words, the exit decision is an endogenous result as the firm will exit when expected value of the firm will turn negative for the first time. Therefore, according to Aw et al. (2001) the exits are concentrated among the firms with the lowest productivity.

The reservation rule is not static, but instead firms constantly update their expectations when new information is revealed (i.e. when they observe their new productivity shock). It is good to note that the value of the firm is the expected sum of all the returns (profits and losses) from the periods when the firm operates. In other words, there can be periods when the firm makes a loss and still continues to operate in industry. Hence, a loss in one period is not a sufficient condition to exit. This result seems to be realistic as many firms can have consecutive years of losses and still continue to operate. The reason behind this result will be discussed in section 3.3.3.

One final note on exit, the firm has three different cost types, which are entry, variable and fixed costs. The variable costs are costs that arise from used labor, but the link between entry and fixed costs is more fascinating. The entry cost is nonrecoverable and thus sunk after entry. It won’t have an impact on the production decisions. The firm has to pay the fixed cost each period and Hopenhayn justify the fixed costs by the fixed outside opportunity cost for some resources such as managerial ability. The fixed costs are significant for the model as Hopenhayn notes that they are a prerequisite for exit. If there were no fixed costs, the firm could continue as long as the marginal cost would be lower than prices and stop production for those periods when it isn’t. Fixed costs in each period create the pressure to perform and if the profit can’t cover the fixed costs, the firm will exit.

3.1.3 Firm’s objective and entry

Given the assumptions on productivity, optimal exit and different costs, the firm’s objective is then to maximize the expected discounted profits. All the different components are represented in equation 9. It is the functional or bellman equation that is an essential part of dynamic programming problems[34].

9)
$$v_t(\varphi, z) = max\left[\pi(\varphi, p_t, w_t) + \beta max\left\{0, \int v_{t+1}(\varphi', z)F(d\varphi', \varphi)\right\}\right]$$

In the equation, π stands for profit and it is determined by the productivity shock φ, the input and the output prices in the given period, w_t and p_t. The second component of equation 9 is the expected value function, where z_t is the price sequence (p_t, w_t). The value function is a concept that simplifies infinite-time horizon models. It is a representation of all the future profits functions after the current period[35]. The function is calculated given the next period’s shock, φ’, and the deterministic price sequence z. The agent in question has to decide only how much is allocated between the current and the next period. The discount factor, β, represents the preferences between different time periods. In Hopenhayn’s model, the value function is an expectation and not an absolute value, because the productivity shock is a random variable. Therefore, one has to integrate over the whole conditional distribution F.

The maximization between zero and the expected firm’s value simply represent that the firm has the option to exit. The zero value is the normalized outside opportunity cost. If one relates this normalization back to the discussion on optimal exit, the link between the reservation value, exit and outside opportunity (i.e. fixed costs) becomes clearer. The reservation rule was the value of the productivity shock when the firm’s expected value is zero. The reason for this formalization is that the opportunity cost is normalized to zero. If the opportunity cost would be something different, this also would be reflected on the reservation value. For example if the opportunity cost increases, the reservation value would also increase. Intuitively this means that the firm evaluates between two alternatives and chooses the one that have a higher pay-off.

The firm tries to maximize the expected discounted profits, but as said it will inevitably exit. An incumbent’s exit will not necessarily lead to that the aggregate output level decreases as entry is a

[34] See for example Stokey & Lucas (1989) for more information on dynamic programming and recursive methods.

[35] One can envisage the value function as a sum of the all profit functions after the current period. In other words, it gives the future value of the firm as the name implies.

possibility in the model in each period. Similarly, replacing failing incumbents is not the only reason for new firms to enter the market. In other words, actual amount of incumbents can fluctuate. One interesting result in Hopenhayn's model is that the entry and exit will be equal in the steady state. It is good to note that entry and exit are not an exogenous process. To the contrary, they are endogenous. Aw, Chung & Roberts (2003) states that "the endogenous variables produced by the model are the flow of entrants into the market in each period and the minimum level of productivity required for incumbents to stay in the market"

If the firm wants to enter the industry, it has to pay the nonrecoverable entry cost. After paying the cost, the initial productivity shock will be revealed for the entrant. Put differently, the entry will happen before the actual productivity shock is known represented by the dotted line in picture 1. After the entry, the new firm behaves like an incumbent firm. According to Hopenhayn new firms will enter the market until expected discounted profit net of the entry costs is zero. The entry decision is not static as entry is a possibility in each period.

The evolution and the state of the industry can be described by the measure μ over firms' shocks. It describes how the industry has evolved and how many firms there are given a measure. Equation 10 describe the industry evolution measure in more detail.

10)
$$\mu_{t+1}([0,\varphi']) = \int\limits_{\varphi \geq x_t} F(\varphi'|\varphi)\mu_t(d\varphi) + M_{t+1}G(\varphi')$$

The first component on the right hand side describes the state for the incumbent firms. The idea is to calculate how many firms are in a given productivity shock range in the next period and sum them up. M_{t+1} is the mass of new entrants and G is the distribution for the initial productivity shock. Therefore, the second component describes the state for entrants. This concludes the description on structure of the Hopenhayn's model.

It goes without saying that Hopenhayn's model is far more elaborate than the one presented by Gibrat. The whole starting point is entirely different as the main action is the individual firm's profit maximization. Gibrat's model was justly criticized due its lack of economics, but presenting a set of nice assumptions and a model does not equal economics. Proving that the equilibrium exist is just as essential as presenting the model. Therefore to complete the model from an economics perspective and to highlight the difference between Hopenhayn and Gibrat, the equilibrium's existence is discussed rigorously in the following section.

3.2 The equilibrium

As stated in the previous section, the incumbent firm has to decide whether to continue or not in the start of the each period. If the firm decides to continue it also has to decide how much to produce given the aggregate demand and input prices. The problem repeats itself in each period the firm continues to operate. In other words, the model is formed in a recursive manner. As an extra layer to the exit and production decisions, there is the aspect of possible new entry and the infinite time-horizon. Given the problems complexity, a good starting to point is to define the conditions for the different equilibriums.

3.2.1 The conditions for competitive and stationary equilibrium

Hopenhayn defines four conditions for the industry's competitive equilibrium that consists of bounded sequences of all variables, formally $(p_t^*, w_t^*, Q_t^*, N_t^*, x_t^*, M_t^*, \mu_t^*)$. The competitive equilibrium is the traditional concept of equilibrium where there exists a market clearing price. Evidently, the first condition for a competitive equilibrium is that there exist such prices that clear the market. This should be true both for the input and the output markets. The second condition stipulates that exit rule is chosen optimally. Intuitively these conditions transforms to the idea that the incumbent firms behave rationally. In concrete terms this would mean that the firms don't waste resources. For example the firms produce the correct amount of goods compared to the given demand. Another example could be that the incumbent wouldn't prolong its stay in industry, but rather exit as soon as the reservation value is hit. The smallness of each firm and the fact that the idiosyncratic shocks aggregates to deterministic variables are critical background assumptions that ensures that equilibrium exists.

The third condition is that the possible entrants behave rationally and that there are no further incentives to enter the industry. Put differently, this means that new firms will continue to enter as long as the expected discounted profit net of the entry costs is zero and after this there is no further entry in that period. To sum up the incumbent firms' and entrants' optimal behavior, the fourth and final condition is that the industry state measure should behave consistently with respect to the optimal entry and exit rules and initial firm distribution. The industry state measure was described in equation 10. The last condition is a sort of a sanity check condition.

It goes without saying that the conditions described above have to be valid in each period. It can't be the case that the firm behaves irrationally in one period and in one period it is rational. In other words, both the incumbent and the entrant have to be consistent. The period-to-period optimality will then form the competitive equilibrium. The competitive equilibrium is a necessary condition for the stationary equilibrium which is the "equilibrium" for the industry evolution. This means that the four conditions specified above must also hold in the stationary equilibrium, but in addition Hopenhayn defines the stationary equilibrium as a vector ($p*$, $w*$, $Q*$, $N*$, $x*$, $M*$, μ*) where the different elements are the steady states for the bounded sequence in the competitive equilibrium. It is good to note that time-subscripts have been dropped in the definition of stationary equilibrium.

Put differently, stationary equilibrium is the state where the firm distribution is constant and steady. For example if there is no variation in the amount firms in the industry or in the quantity produced, then the equilibrium is stationary. That said, it doesn't mean that the industry is populated by the same firms after the steady state is reached. Actually, what Hopenhayn shows is that it is possible to have stationary equilibrium with positive entry and exit given certain conditions. The amount of firms is unchanged, but the actual firms changes. Evidently, this must be true as Hopenhayn has assumed that the firm's life-span is finite and hence the productivity will fall for all firms at some point below the reservation rule.

The problem is complex from a mathematical point of view, but fortunately there are solution methods to study the model in more tractable form. One of the applicable methods is dynamic programming as the model is formed in a recursive manner and the time-horizon is infinite. In the following sections the equilibrium will be characterized both from the social planner's and the individual firms' point of view. Looking at the model from different aspects will be important in the results section.

3.2.2 The social planner as the agent

The sequence problem is described in equation 11. According to Hopenhayn the equilibria in the model maximize the net discounted surplus. The objective function, that represents the surplus, consists of four different components. The two first components are revenue, $R(Q_t)$, and variable costs, $C(N_t)$[36]. The other components are the entry and the fixed costs adjusted by the amount of relevant agents. Surplus is then discounted and summed up. The feasibility correspondence $\Gamma(\mu)$ restricts alternatives. The correspondence is defined as the set of all sequences $\{N_t, Q_y\}$ that comply with the optimal entry and exit rules and the firms' production capabilities. The purpose of the feasibility constraint is to restrict the social planner so that she can't choose such a sequence that the firm's couldn't execute. It is assumed that the feasibility correspondence $\Gamma(\mu)$ is closed, convex and non-empty.

11)
$$V(\mu) = \max_{N_t, Q_t} \sum_{t=0}^{\infty} \beta^t \left[R(Q_t) - C(N_t) - c_e M_t - c_f \mu_t(S)\right]$$
$$s.t.$$
$$\{N_t, Q_t\} \in \Gamma(\mu_t)$$

Equation 9 presented the firm's problem in a functional equation form. As equation 11 presents the sequence problem from the social planner's perspective a new functional equation is needed. Hence, equation 12 presents the problem in functional equation from the social planner's perspective. The additional benefit of introducing the functional equation from the social planner's perspective is that it creates extra clarity and thus makes the equilibriums' analysis more structured.

12)
$$v(\mu) = \max_{\{N,Q\} \in \Gamma(\mu)} F(N, Q, \mu) + \beta v(\mu')$$

The return function $F(N, Q, \mu)$ represents the objective function in equation 11 for one period. Intuitively, this is the net consumer surplus. The industry state measure μ is the state variable and μ' is the next period's industry state. It can be assumed that the measure is well-defined. The control variables are the input and output quantities, N and Q respectively. The value function is represented by $v(\mu)$[37]. The time subscripts are dropped, because the variables refer only to one time-period. It is good to note that from the social planner's point of view there is no uncertainty as the

[36] The revenue function is defined as R(Q$_t$) = $\int_0^{Q_t} D(x)dx$ and the variable cost function is defined as C(Nt) = $\int_0^{N_t} W(x)dx$.

[37] Similar to section 3.1.3, the value function is the sum of all return functions after the first periods. In other words, it is a representation of the total surplus after the current period.

productivity shock is independent and identical-distributed across firms. Hence, all the aggregate variables are deterministic as explained in the section 3.1.1.

The return function is bounded and continuous and therefore it is natural to assume that the function belongs to space of continuous bounded functions. Then according to Stokey & Lucas (1989) it is enough to show that correspondence is nonempty, compact-valued and continuous and that the return function is bounded and continuous. Evidently, the return function satisfies the requirements given the topology of the correspondence. Hopenhayn further assumes that correspondence is compact and he shows that the feasibility is nonempty. Finally, it can be assumed that feasibility constraint is continuous as the output and input variables, n_t and q_t, for the individual firms are continuous[38]. There is no reason to assume that the variables will be discontinuous when they will be aggregated. Hence by Stokey & Lucas' (1989) theorem 4.6 the dynamic programming problem will have a unique solution. According to Hopenhayn this "implies a dynamical system on the space of bounded positive measures given by $\mu_{t+1} = H(\mu_t)$ where H is a nonlinear map".

3.2.3 Equilibrium and the firms as decision makers

The above proves that a unique equilibrium exists and it was proven from the social planner's perspective. However, the existence of a unique equilibrium can also be proved when the individual firm is the agent. The two approaches will have the same solution and in section 3.3.1 it is discussed what actually this implies. The second approach is preferred by Hopenhayn as it more useful in characterizing the results.

The four conditions for a competitive equilibrium are utilized when proving the stationary equilibrium from the individual firm's perspective. The reason is that the conditions have to be valid also in the stationary equilibrium as explained in section 3.2.1. The first step is to prove that there exists market clearing prices that are constant. It can then be showed that the incumbent's problem presented in equation 9 has a stationary solution given the constant market prices. In other words, incumbents solve the following equation[39].

[38] In addition, Hopenhayn's lemma 1 should be valid. This is a condition that the state measure and price vector converge to steady values.

[39] Equation 13 is a variant of equation 9 where the time subscripts are dropped due to the stationary equilibrium and where the price vector is presented through a function of μ in order to impose the stationary solution.

13)

$$v(\varphi,\mu) = \tilde{\mu}(\varphi,\mu) + max\left\{0, \beta \int v(\varphi',\mu)F(d\varphi'|\varphi)\right\}$$

It can be shown that Equation 13 has a unique solution by applying standard dynamic programming arguments[40]. Once it has been established that there are market clearing prices and that the value of the firm can be calculated, the following step is to show that the three other conditions are met. This is done by introducing three equations that satisfy the conditions and the value function, which ensures the market clearing prices.

14)

$$\int v(\varphi',\mu)F(d\varphi'|x) = 0$$

15)

$$\int v(\varphi,\mu)v(d\varphi) = c_e$$

16)

$$\mu = m(x,M) = P_x\mu + Mv$$

Equation 14 and 15 represent the exit and entry conditions, respectively. Equation 16 is a variant of equation 10 where $m(x,M)$ is an invariant measure for exit rule x and entry mass M. P_x is a bounded linear operator that can be thought as an indicator function that specifies which productivity shock levels should included in the calculation of the measure[41]. The subscripts define what the actual cut-off level is and in equation 16 the cut-off point is defined as the optimal exit threshold. Put differently, P_x is another way to state that only those firms should be included in the calculation of the state of the industry that has a productivity shock above the optimal exit threshold level.

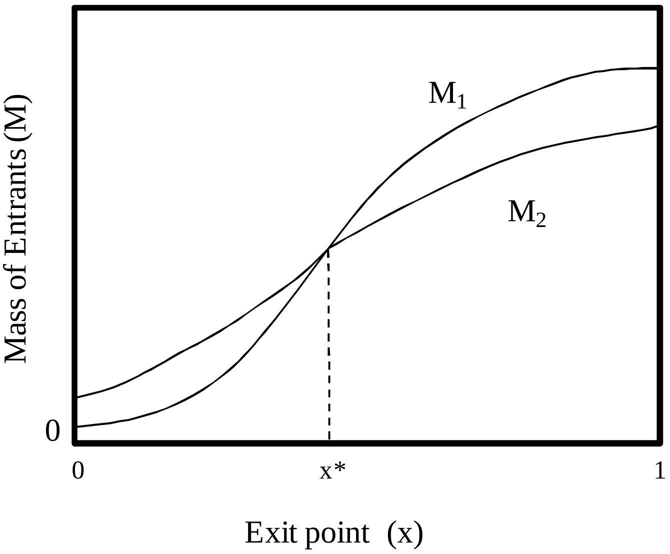
The invariant measure m is well-defined and jointly continuous. In addition, the measure is decreasing in x and increasing in M. Intuitively this means that the more there is the entry (i.e. higher M) and the lower exit rule x is, the more firms should there be in the industry. $M_1(x)$ is defined as the mass of entrants when the exit rule is optimal given for a fixed exit rule. Correspondingly $M_2(x)$ is defined as the mass of entrants when there are no further incentives to

[40] It is good to note that the equation has a stochastic element as the productivity shock is a random variable. Therefore, the line of arguments presented for the deterministic case in section 3.2.2 can't be replicated as such.

[41] Formally, the Px is defined as

$$P_x(\varphi,A) = \begin{cases} \int_A F(ds|\varphi) & if \varphi \geq x \\ 0 & \text{ot□erwise} \end{cases}$$

enter the industry given an exit rule. Both $M_1$and M_2 are well-defined and they are deduced from equation 14, 15 and 16. In other words, the measures M_1 and M_2 contain only those points that satisfy the equations above[42]. According to Hopenhayn a stationary equilibrium exists with positive entry and exits if there is an x* ∈ (0, 1] such that $M_1(x) = M_2(x)$. In order to clarify the above, picture 3 presents the measures in a graphical manner.



Picture 3 - Existence of stationary equilibrium taken from Hopenhayn (1992)

At x = 1, M_1 will always be higher than M_2. M_1 should be thought as the amount of entrants needed to replace the incumbents that exit in order to hold the exit rule optimal (i.e. to ensure that the distribution of firms stays stationary). At x = 1 all incumbents will exit the industry and there are not enough entrants to replace the mass departure, because the cost of entry is too high with the respect to the productivity requirement. Put differently, there is a mismatch between the possible optimal entry and the needed entry to hold the firm size constant.

Hence there are two possibilities i) M_1 and M_2 will cross at some point and ii) M_1 is always above M_2. In alternative i) a stationary equilibrium exists albeit it may not be unique. In the second alternative there will be no stationary equilibrium if there is positive entry. However, it can be shown that if there is no entry then a stationary equilibrium exists. Naturally this means that there is zero entry and exit and the incumbents remain in the industry forever. In case i) there will be a

[42] As an example $M_1(x)$ is defined as the x and M points that satisfy the following equation

$$\int v(\varphi', m(x, M_2(x))F(d\varphi'|x) = 0$$

unique stationary equilibrium if the functions crossed just once. According to Hopenhayn, M_1 is strictly increasing and M_2 is nondecreasing meaning that if the cost of entry is low enough it can only be the case that the measures crosses only once. These attributes rely on the past assumption and proposition made in the model.

Hopenhayn also shows that if the profit function is not increasing in φ, then there will be multiple equilibriums. This is Hopenhayn's assumption A.2b, but he further argues that one of the following conditions is necessary for it to hold. The first condition is that that if the industry is a price taker in the input markets (condition U.1). The second condition is that if the profit function is separable into productivity shock component and price component (condition U.2)[43]. The first condition guarantees that the firm can always benefit from an increase in productivity. The following example could be considered. If the output prices were not fixed then it could result in a situation where the firm would like to increase production, but it will not do so, because the input prices could increase to be too high. According to Hopenhayn the condition U.2 will be satisfied if the production function is for example homogenous of degree one in the vector of inputs and shocks. So, the condition U.2 basically means that if firms can choose the production input ratios independent of the productivity then the production will be increasing in φ. In addition to condition U.1 and U.2, there has to be an upper limit on the cost of entry. If there were no such limit, a situation could exist where the cost of entry would be too high which effectively deters entry.

Hopenhayn sums up that "the existence of a stationary equilibrium with positive entry and exit is equivalent to the existence of a stopping rule with finite expectations and a mass of entrants such that for the (stationary) prices correspond to the associated invariant distribution this stopping rule is optimal and the expected discounted profits of entrants are equal to the cost of entry". In other words, from the firm's perspective the problem is an optimal stopping problem. If it can stop optimally and there are enough entrants to replace the incumbents that exit, then there will be a stationary equilibrium. One of the results is that the firm's growth rate will be decreasing in size. This negative relationship between firm size and growth rate is a contradiction to Gibrat's law. This and other results will be discussed in the next section.

[43] Formally condition U.2 is $\pi(\varphi, p, w) = h(\varphi)g(p, w)$.

3.3 Results

Few words should be said on the comparison between Hopenhayn and Gibrat as it now has been verified that the Hopenhayn's model is a functioning economics model. First of all, the main difference is that in Hopenhayn's model firms make decisions on production, entry and exit and the different outcomes are a result of these decisions while in Gibrat's model the firms just existed and, bluntly said, something just occurred and this result in growth or not. With Hopenhayn's model, it can be verified that the outcomes are result of firms making decisions and this is a nontrivial result. The power of the Hopenhayn's model is considerably larger than Gibrat, because of it can take a stand in multiple issues. One such issue is the role of social welfare that couldn't be verified in the models that had the mindset of Gibrat's law. Welfare is discussed in the following section among other things.

Also, Hopenhayn's model shows that there exists a stationary equilibrium with positive entry and exit. What makes Hopenhayn's work different and robust from the preceding literature is that the entry and exit are part of the limiting behavior of the industry given that certain conditions are met. For example in Jovanovic (1982) exit and entry is only part of the adjustment to a steady sate. The results are obtained by assuming firm-level heterogeneity through serially-correlated productivity shocks. Similarly, it is a nontrivial result that entry and exit is part of a stationary equilibrium. It has interesting results on for example used resources.

3.3.1 Resources reallocation and price taking firms

Positive entry in stationary equilibrium existed when entrants had the incentives to replace all the incumbents that exit. This occurred at point $x*$ in picture 3. As explained in section 3.1.2, it is the firms with low productivity that exit the industry. To the contrary, the new firms' actual productivity is not known a priori to entrance. In some case, some entrant's realized productivity is higher than the incumbents that exit and similarly for some entrants the realized productivity is lower. It is only known that the expected value of the entrant is equal to the cost of entry (for the marginal entrant). Naturally, the low productivity entrants will be replaced in the next period, but it is good to highlight that the high productivity entrants do not change the aggregate productivity. The reason is that in a stationary equilibrium in addition to firm size distribution also aggregate output, input, prices are constant.

Although entry doesn't directly change the aggregate productivity, it still has an impact on the aggregate productivity. The comparative statistics of the entry costs illustrates the importance of entry. Hopenhayn (1992) shows that if the cost of entry increases, it implies that the marginal reservation rule $x*$ and the mass of entrants will decrease. Formally, this means that the M_2-curve shifts downwards as an increase in entry costs will raise the level of discounted profits needed to make entry profitable for each reservation value level. It goes without saying that this discourages entry. For the incumbent, the cost of entry is sunk and therefore M_1-curve will remain unchanged. However, this doesn't mean that the changes in cost of entry don't have an impact on the incumbents. To the opposite, an increase in cost of entry implies less selection which means that low-productivity firms will continue longer in the industry.

According to Aw et al. (2003) decreased entry "makes it easier for low-productivity incumbents to survive, reduces the amount of exit, and results in an industry characterized by a higher proportion of low-productivity producers". Assuming that the input prices are constant, then output prices increases with cost of entry resulting in higher employment and output for each φ[44]. This is the price effect according Hopenhayn (1992) which ensures that high productive firms can produce more. However, this higher price level has an opposite effect. Balasubramanian & Sivadasan (2009) elaborates "more specifically, the larger the sunk entry costs, the greater should the expected value function be, which requires a higher average price level to prevail in equilibrium. The higher average price level allows some relatively inefficient firms to cover their fixed costs". This is defined as the selection effect.

According to Hopenhayn (1992) the "strength of each of these effects depends on properties of the stochastic process for the shocks and the production function". What is clear that the required productivity for the marginal incumbent will be lower if cost of entry increases. The expected lifetime of firms will be also higher. Higher cost of entry will also lead to a lower turnover rate, the rate between entrants and total number of firms. The level of aggregate productivity is ambiguous depending how the different components are specified. Thus, the level of entry has an indirect impact on the level of aggregate productivity.

[44] The reason is that cost of entry decreases leads to a decrease in the mass of entry. This in turn will lead to that the output prices will increase, because if it doesn't then the aggregate output will grow without bound.

However, it doesn't mean that the production levels, given a cost of entry, were not socially optimal. In other words, there is no welfare loss. The social planner couldn't do no better than the "invisible hand". The solution method presented in 3.2.2 where the social planner maximizes net discounted surplus will end up to the same result as when the incumbents or entrants independent from any central guidance maximized their own expected discounted profits. The latter case was presented in section 3.2.3. According to Bergin & Bernhardt (2008) the standard approach characterizing industry dynamics is to show first that the competitive equilibrium corresponds to the solution of a social planner's problem, and then solve the social planner's problem. The stationary equilibrium is then a special case of a competitive equilibrium in the sense that the different variables are constants. Stationary equilibrium thus preserves the properties of a competitive equilibrium.

What is truly remarkable is that there is no loss from a welfare perspective although there is firm-level heterogeneity, entry and exit. Evidently, information is imperfect as this is a requirement to have firm-level heterogeneity. If there were no uncertainty, only the best would enter resulting in identical firms. The reason why entry is needed is that it resolves the uncertainty. As Cabral (2007) states "the only way to determine a firm's efficiency is to actually enter the industry". Although there is a cost of entry, the industry is still competitive as the "extra costs" of arising from entry are not transferred to customers. The reason is that each firm is so small that they don't have market power and therefore they are forced to be price taker. The cost of entry is then just spread over to all participants in the industry or as Cabral (2007) sums up that "the basic idea is the same as in the model of perfect competition: A very small firm has a negligible impact on other firms and on price. It follows that it internalizes all of the costs and benefits from entering or exiting the industry: What is good for the firm is good for society." What is true for entrant is also true for the incumbent. Incumbents don't have market power and hence they are price takers. Hence, price is equal to expected marginal costs and as a result the market is efficient from a welfare point of view. As stated in section 3.2.1, the smallness of each firm is a crucial assumption in the model.

To sum up, the main reason for competitive equilibrium is that the agents are price takers due to their small size. According to Cabral (2007) Hopenhayn shows "that the market equilibrium is efficient if firms are price takers —even if efficiency varies across firms and across periods". Entry is important, because it preserve the smallness of firm and heterogeneity. Similarly, exit is needed so that low productive firms can cease their operations. The possibility to exit gives each firm the same alternative to test their efficiency. In other words, resource reallocation (i.e. exit and entry) is one of the factors that keep the industry competitive in a world where there is imperfect information. The existence of a competitive and stationary equilibrium has also implications on the relationship between firm size and growth, which is the topic of the following section.

3.3.2 The relationship between firm's size and growth rate

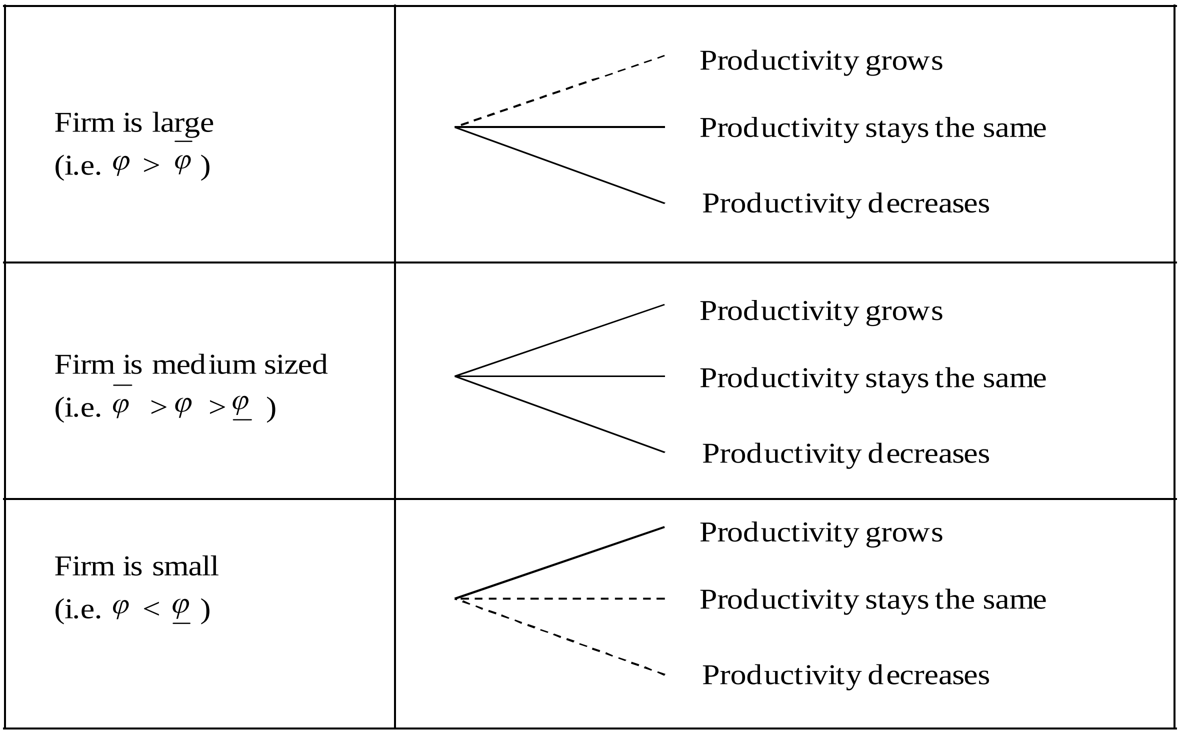
The evolution of the productivity shock will dictate the path of the industry, but it is the possibility to exit that creates a negative relationship between firm's size and growth rate. The following example illustrates the reason. In the example, it is assumed, for the sake of simplicity, that there are only three different size classes – small, medium-sized and large firms. There is a one-to-one relationship between firm size and the productivity shock and hence one can use size and productivity interchangeably. Similarly for the sake of simplicity, one could assume that the productivity shock could only have three distinct paths[45]. In the future periods, productivity shock could increase, stay the same or decrease. These simplifications are then presented in picture 4.

Firms will exit when the expected value turns negative for the first time. This basically means that low-productive firms (i.e. small firms) that expect that their productivity will stay the same or even decrease will exit. The exit is represented by the dotted lines in lowest row in picture 4. Therefore, the only small firms that remain in the industry are small firms that grow fast. This is not true for medium-sized or large firms, because selection doesn't matter as much for these classes. If an equivalent negative productivity shock would both hit a small firm and a large firm, it is more likely that the large firm will continue to operate, because it doesn't want to forfeit the option to operate in the industry. Therefore, medium-size firms' productivity has the possibility to evolve in all directions. It is good to note that the largest firm's productivity can't grow beyond a certain limit, because the size of the productivity shock is restricted from above. Hence, the dotted line for the large firms.

It is good to note that much can't say on the evolution of the medium size firms, but actually this doesn't matter. The selection of small firm's is the only thing needed to reject Gibrat's law. In other words, because the small firms are selected there is a negative relationship between firm size and growth rate. The evolution of medium size firms can't change this. If Gibrat's law would be valid in its purest form it would mean that all the different size classes would grow at an equal rate. This is not the case in Hopenhayn's model if entry cost is sufficiently low. However, the situation is

[45] The path of the productivity shock in Hopenhayn's model was continuous.

different when the cost of entry is high enough to deter entry completely. Then growth will be purely random. The reason is that if no entry occurs, incumbents remain in the industry forever, as explained in section 3.2.3, and then evolution of industry is dictated only by the productivity shock which was by definition random.


Firm is large
(i.e. $\varphi > \bar{\varphi}$)
Productivity grows
Productivity stays the same
Productivity decreases
Firm is medium sized
(i.e. $\bar{\varphi} > \varphi > \underline{\varphi}$)
Productivity grows
Productivity stays the same
Productivity decreases
Firm is small
(i.e. $\varphi < \underline{\varphi}$)
Productivity grows
Productivity stays the same
Productivity decreases

Picture 4 - The path of productivity shock evolution for different size classes

The productivity shock follows a different process in Hopenhayn’s and in Gibrat’s model, but if one could assume for the sake of the argument that the processes were sufficiently close, the result would be interesting. As stated many times in chapter 2, there are subsamples that fail to reject Gibrat’s law. Could it be the case that these industries that don’t have rejected the law are similar to the industry that Hopenhayn’s model would define? If this is the case then it would mean that Gibrat’s law can never be valid, because growth wouldn’t be random, because growth is a result of firms maximizing profits and the visible “stochasticity” is only a result of this. Put differently, it is the classical argument that Gibrat’s law can be replicated with Hopenhayn’s model, but not the other way around.

Gibrat’s law could be tested for an industry that satisfies the Hopenhayn’s industry assumptions and where there has been a large increase in cost of entry, preferably over a shorter period time. The “new” null hypothesis should then be that the deviations from Gibrat’s law should be smaller after the increase in cost of entry. The problem is then of course to find an industry that would fit these requirements. Another downside with this hypothesis is that sample selection bias would be still present.

Age itself alone will not have an extra predicative role in Hopenhayn's model meaning that two firms with identical size, but with different ages would have same path. Therefore, from a selection point of view it doesn't matter if the firms are incumbents or new entrants. However, there is a difference between the different age cohorts. According to Farinas & Ruano (2005) any cohort of surviving firms at time t stochastically dominates the cohort of entering firms at $t + 1$. The reason is simple. The survived cohort contains older firms that have been more exposed to the selection process leading to a higher threshold for failure. To sum up, according to Cooley & Quadrini (2001) one of the model's primary results is that, conditional on age, the dynamics of firms (growth, volatility of growth, job creation, job destruction and exit) are negatively related to the size firms.

3.3.3 Profits and the option of staying in the industry

The cost of entry has another important implication for the incumbent. It separates the entrant and the incumbent firm. The entrant has to pay the cost of entry when entering the industry in addition to the each period's fixed cost. For the incumbent it is sufficient only to pay the fixed cost. In effect, the incumbent firms have an option that gives the possibility to stay in the industry without the need of pay a new entry cost. This option is naturally valuable and because of the option, it can be reasonable to endure periods with losses. In other words, the incumbent firm stays in the industry although it is making a loss, because it expects a turnaround in the future. The incumbent will not exit the industry for few periods, because if the business environment would turn once again positive it would have to pay the cost of entry. So, although the current cost of entry is sunk, the cost of entry from the possible re-entry is still relevant for the incumbent. This would explain why loss-making firms would stay in the industry. In addition, Hopenhayn shows that there is a positive lower bound for average industry profits.

Balasubramanian & Sivadasan (2009) notes that "these firms [the firms in the industry] may not necessarily make a good return on their entry costs, which in this model they incur on entry, before they know their true productivity levels. However, having already incurred these sunk costs of entry, the inefficient firms will choose to remain in the market as they are able to cover their recurring costs at the prevailing price level". It should be added that incumbents de facto earns more than the opportunity cost, so in this sense firms that remain in industry are not doing anything wrong.

The results are impressive and indeed Hopenhayn's article is a seminal paper within the industry dynamics and evolution's research. However, the model is built on certain key assumptions such as the productivity shock. Are the assumptions valid and realistic? Does it really explain why the industry evolves? Are there other variables that have been omitted from the analysis for the sake simplicity and tractability and are these omitted variables relevant and important? These questions are discussed in the next section where a critical view of the Hopenhayn model is presented.

3.4 A critical view

Although Hopenhayn's results are impressive, there are some drawbacks. The competitive equilibrium is only a partial equilibrium as Sleet (2001) notes. Basically, this means that Hopenhayn has assumed that the household (i.e. the demand side) acts optimally not matter what happens. Of course, the model is a simplification focusing on the dynamics of heterogeneous firm, but the partial equilibrium approach reduces the robustness of the solution if the intention is to show a truly competitive equilibrium. In other words, a truly competitive equilibrium would require that households behave optimally as a by-product of their own utility maximization.

Also, what guarantees that an individual firm wouldn't grow to have such a significant role that it would have market power? The equilibrium was competitive, because the individual firm was so small that it was essentially a price taker. The new flow of entrants reduces this pressure as explained in section 3.3.1, but the following example could be considered. Hopenhayn states that changes in size distribution are ambiguous after an increase in cost of entry, because there was both a price and a selection effect. The exact results depend on the exact form of the underlying distribution. Couldn't it be case that for some distributions, the price effects dominates so much that there would be few firms that would dominate everybody else? In other words, what would happen if few firms would be able to exercise market power? This is unclear and can't be answer without additional assumptions made on the underlying distributions. The question is not trivial as it could be the case the welfare implications could change if an incumbent or entrant would grow so much that it effectively would have market power.

In the models, the firms that are not productive will exit the industry. This is realistic, but exit is not always synonymous to failure as Plehn-Dujowich (2009) notes that "existing theories of industry dynamics focus exclusively on new entrants and assume that firm exit is synonymous with failure". This is also true for Hopenhayn (1992), but the matter is not as straightforward as one would expect. There are two aspects that should be considered.

The first is that although a firm ceases their operations in one industry, it doesn't mean that the resources are salvaged. According to Dunne et al. (2005) on average 22 % of firms exit an industry or market in order to start producing a new product, but the importance of realigning the production varies from industry to industry. Dunne et al. (2005) continue that the respective percentage is 10 % for bakeries and ready-mix concrete, but for concrete block and brick sector the percentage is 33 %. Bernard et al. (2006) found that 66 % of surviving firms change their product mix every 5 years. Therefore, Plehn-Dujowich (2009) concludes that firms relocating across industries or product lines are empirically relevant in industry dynamics.

It is good to note that the critique is not entirely correct. Firms will exit when the value of staying in the industry is less than the opportunity cost and Hopenhayn don't specify that the resource couldn't be used again. So there is the possibility that firms could reallocate their operations, but it is important to remember that the firms will consider this alternative only then when their productivity falls below the industry threshold. The reason is that the opportunity cost is normalized to zero for each firm. In other words, firms will exit only when they are unproductive in their current industry and not because they would be more productive in other industries. As seen above, firms quite often change their market and so this seems somewhat unrealistic.

The second aspect is mergers and acquisitions. As stated in section 2.2.2, Dunne & Hughes (1994) argued that survival function is non-linear. Small firms have a high death rate, but the rate is not considerable smaller for medium sized firms. The reason is that these medium sized firms are taken-over rather than simply ceasing their operations. In Dunne & Hughes (1994) report that 10 – 12 % of medium-sized and large firms are taken over. For small firms the figures are similar, but the takeovers represent a considerable larger share of industry exit for medium-sized and large firms. This aspect has not been taken into account in Hopenhayn's model. A good question is that is there actually any benefit of merging two units as they would have limited market power. Nevertheless, the important question is that does the possibility of mergers and acquisition change the analysis on the growth-size relationship presented in section 3.3.2. This should be further studied.

Another point is that there is no aggregate uncertainty, meaning that the only source of the uncertainty is the firm-specific shocks. Certainly, aggregate shocks can play a role in real life as seen by the latest financial crises in 2008. Although the points above are all valid, there is one problem that should be focused on. This is the role and assumptions around the productivity shock. For example it is assumed that the firms did know the productivity shock's distribution. This is certainly unrealistic. In addition, the model's many results follow directly from the shock's evolution. One could say that instead of showing impressive results, Hopenhayn just assumed them.

3.4.1 The problem of exogenous shocks and the omitted variables

Hopenhayn's primary result is that the productivity shock defines and drives the evolution of the market. One has to ask that does this really explain anything. Hopenhayn's contribution to the evolution of industry dynamics can be compared with progress between the basic Solow model and the general Solow model. Adding an exogenous growth component to the basic Solow model enabled continuous growth. Although the extension is not by any means trivial, it didn't explain why the growth occurred in more detail. Similarly, the exogenous productivity supplements earlier models, but it doesn't explain why firms are different in their internal efficiencies. In other words, the shock is in a sense a good explanation, but it isn't sufficient. The inevitable question is that what drives then the productivity?

This leads to the next and final step, which are the omitted factors. As any model, Hopenhayn's model is a simplification in order to highlight one certain aspect or relationship. In this case the highlighted relationship is the one between productivity, profitability and selection. As Foster et al. (2008) points out that "productivity is only one of several possible idiosyncratic factors that determine profits. However, other idiosyncratic factors may affect survival as well". Productivity is certainly an important relationship, but one can't deny for example the role of finance as a constraint. Of course one could argue that Hopenhayn's productivity shock entails all the different aspect of firm's productivity ranging from efficient manufacturing to superior marketing, also finance.

Also, the model was very supply driven, so one extension could be to treat demand as stochastic and see whether the results are still valid and robust. Foster et al. (2008) continue along this strand and show how selection can created with also demand factors. Finally, a natural extension to

Hopenhayn's model is to treat the productivity shock as an endogenous variable. In other words, R&D should be introduced as a possible action for the firms. Different extensions are the theme of the following chapter. The aim is to explain in more detail what could explain the productivity shocks evolution and supplement the picture on the firm's behavior.

For example by adding a finance constraint, Cooley & Quadrini (2001) shows how age dependency can be introduced to the model. For example the size of the firms is more persistent for older firms. As Hopenhayn acknowledges age for the individual firm does not have an extra predicative role in the model. Is this a drawback? It was already shown by Dunne et al. (1989a) and Evans (1987) that age has a role. However, one could argue that size and the age of the firm move together. If size could be used as a good proxy for age, then it would unnecessary to analyze the age's roles. Nevertheless, introducing the age component could be a good extension for the Hopenhayn model and actually Hopenhayn himself suggests this.

4 Alternative explanations for variability of firm size

The previous chapter explained in detail Hopenhayn's seminal model on industry dynamics and firm-level heterogeneity. The main result was the existence of a stationary equilibrium with positive entry and exit. In addition, as a by-product it showed how small firms grow faster than larger firms due to the selection-effect. However, the model had some limitations. One of the elemental problems in the model was that it didn't really explain that much, as all the results were built upon the assumption of the productivity shock. In others words, the model didn't explain how the productivity shock was formed and it omitted several interesting variables such as finance constraint or the demand stochasticity.

The purpose of this chapter is to extend the model to cover more variables with different assumptions. The extension will not be analyzed with same vigor as Hopehayn's model, but rather the focus is to present how the models' assumptions and results are different. The chapter consists of three sub-sections. The first contemplates the role of Bayesian learning. After this, the role of finance is studied in two different sections.

4.1 Passive learning (and fittest survives)

In Hopenhayn, the firm knew the distribution of the productivity shocks and merely reacted to each period's updates on the shock. The firm didn't engage in any learning activities as explained in section 3.4. The purpose of this and the following section is to discuss what impact does learning have on the results. This section focuses on the concept of passive learning. This is done by introducing the classical article by Jovanovic (1982) “Selection and the evolution of Industry”. The models have very similar premises and in fact they base the analysis and the model on common observations, although Jovanovic focuses more on the individual firm’s behavior in addition to dynamics of the industry itself.

The general structure is very similar to Hopenhayn and other industry evolution models[46]. Jovanovic presents a model where there is infinite amount of firms that are price takers just as in Hopenhayn. The incumbents manufacture homogenous products in a small industry and the firms’ objective is to maximize expected profits. The input prices were deterministic and dynamic in Hopenhayn, but in Jovanovic’s model input prices are assumed to be constant. Each period the incumbents have to decide whether to continue and if they do so, then how much to produce. The decision on production happens before the random variable is revealed. However, the significant difference between Jovanovic’s and Hopenhayn’s models is in the assumptions around uncertainty.

Both models presented so far have had firm-level heterogeneity. In Jovanovic, firm’s costs create the heterogeneity as they have different cost efficiency levels that are drawn from an initial distribution. This stochasticity in costs will lead to that the firms are different. It is a good question that is productivity that much of a different concept from cost efficiency. Therefore in this section core efficiency will be used as a term to define both productivity and cost efficiency. The separating factor between the two models is the assumptions on the distribution of core efficiency. In Jovanovic, the true costs are constant, but unknown. The firm only knows the distribution and the variance of costs. In Hopenhayn, the distribution of productivity was not specified in more detail, but one can assume that the agents know the distribution, mean and variance of the productivity. Ericson & Pakes (1995) validates this assumption by stating that competitive firms “doesn’t engage in Bayesian learning as they know the distribution of those shocks”.

Equation 17 illustrates the composition of costs in Jovanovic. The variable x_t is a random variable independent across firms and it consists of two components. The first is the true costs, θ, that is constant and it defines the core efficiency of the firm as larger values of θ will generate larger costs

[46] Actually, Jovanovic’s model precedes Hopenhayn’s model and according to Hopenhayn (1992) Jovanovic’s model was one of the first models of industry dynamics with firm-level heterogeneity.

at all levels of output. As stated above, true costs are unknown to the firms. The second factor is the firm-specific shock or noise variable that is i.i.d. The firm knows the mean and variance of ε_t. It is good to note that as the uncertainty is only on firm-level it means that the aggregate variables are deterministic just as in Hopenhayn. In other words, the incumbents treat the price path with perfect foresight. This does not mean the prices are constant over time.

17) $x_t = \xi(\eta_t) \qquad \eta_t = \theta + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)\ i.i.d.$

It is the incumbent firms' mission to find out the true nature of the core efficiency. This is done by observing the fluctuating costs and using the knowledge of the distribution and variance to narrow down the true costs. As the firm receive more information the precision of the estimate increases. As mentioned earlier, this process is called Bayesian learning. Another term used to describe the activity is passive learning as the firm learns as a by-product of its operations. Intuitively this means that the firms enters or continues in an industry, because it wants to obtain evidence if it is profitable or not.

In case the incumbent judges that it is inefficient (i.e. it has a high true costs), it will exit. The exit decision will be made by comparing the opportunity cost with the value of continuing in the industry. Once the opportunity cost is higher than the firm's value the firm will exit. The value of the firm is calculated as dynamic programming problem similar to Hopenhayn. The dynamic programming is a suitable calculation method as there is an infinite-time horizon and the structure of the problem is recursive.

4.1.1 The elemental role of information

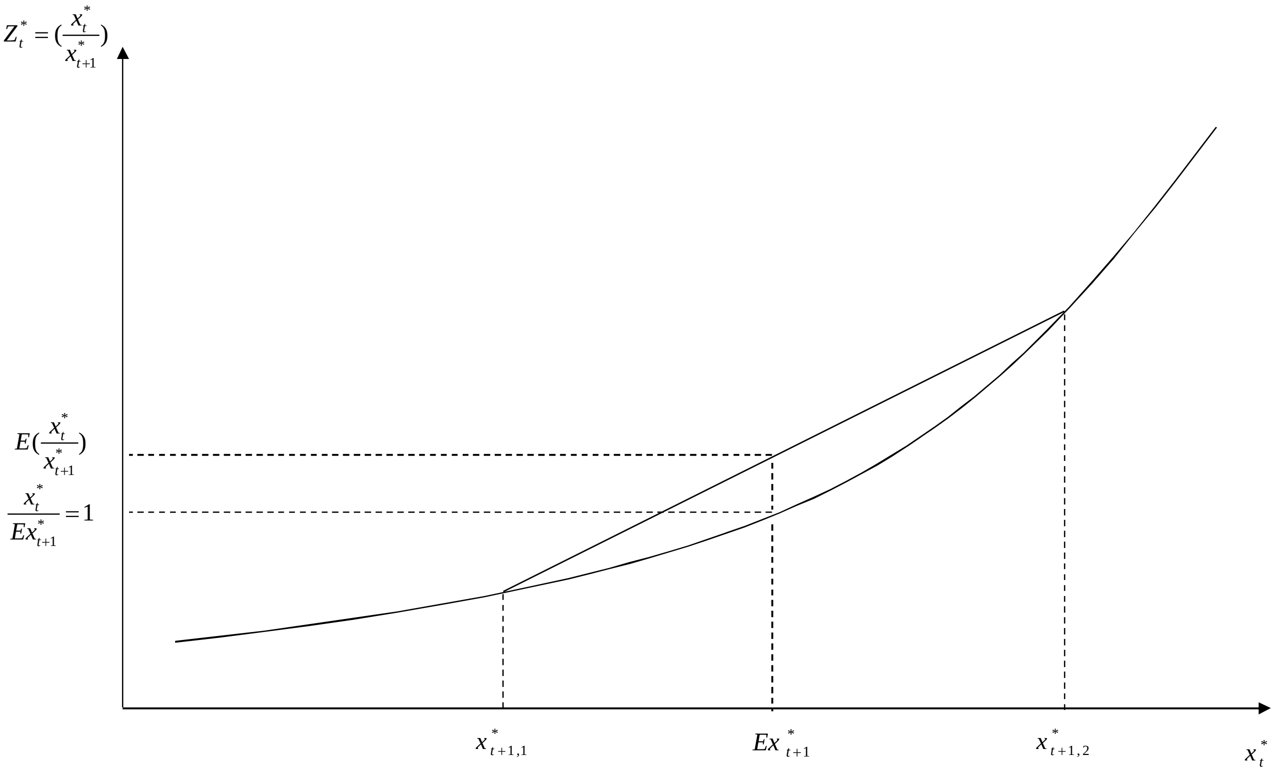
The exit process and role of information in Jovanovic is different from Hopenhayn. Of course, the firms react to new information in both models, but in Jovanovic the information is utilized to determine the full nature of the firm. One prerequisite for this is that, the core efficiency of the firm is constant, because if core efficiency would fluctuate the firm couldn't narrow it down. To the contrary, in Hopenhayn this core efficiency is continuously changing. This has an impact on the reason why firms exit. In Hopenhayn, the incumbent firm will exit only then when it believes that core efficiency has reached an unprofitable level for a sufficient long enough time while in Jovanovic the exit will happen when the firm is sure that it is an inefficient operator. The difference is subtle, but important, because it crystallizes why the models are different. In practice, this means that in Hopenhayn an incumbent can have the worst core efficiency in the industry, but still remain in the market if it believes that the productivity will (soon) surge. To the contrary, in Jovanovic if the incumbent would realize that it is the worst player it will definitely exit.

Entry is also a possibility in Jovanovic. This happens by paying a cost of entry. The cost is sunken and nonrecoverable and according to Jovanovic it can be thought for example as a cost of establishing a particular location. Once the possible entrant has decided to enter, it will be treated as an incumbent. This is standard. Before entering, the possible entrant does not know the actual true cost parallel to the incumbent. The entrant knows the distribution, the mean and the variance of the potential true costs. This is very similar to Hopenhayn, but role of entry is not as significant.

In Hopenhayn, the role of entry and the exit was that it ensured resource reallocation. New entrants replaced incumbents that's productivity shock was falling and hence were effectively unprofitable. In Jovanovic, the firm will exit if the observations reveal that the firm is inefficient as stated above. One by one, the inefficient drops down from the industry. This leads to that eventually only the most efficient survive the competitive pressure. In other words, after industry has reached the stationary equilibrium only the truly efficient remains in the industry. There is no need to enter, because all agents know that the incumbents have a superior cost advantage and therefore the information gathering process is redundant and entry ceases. It is interesting how the role of information has an impact also on the entry process.

Only if the production q is concave in x_t, then there will be entry and exit in a stationary equilibrium. The reason is that if the production function q is concave in x_t, then each period the remaining firms will produce less and less. The new entrants replace this decrease. Naturally, before the stationary equilibrium entry can occur under both assumptions. New firms would enter to gather information and try their luck. To sum up on entry, in Hopenhayn entry occurred in stationary equilibrium, because they replaced exiting firms. In Jovanovic, there is not a similar need because the only reason to enter is to acquire information and once the market has saturated there is no room for entry.

Jovanovic's model can also show that younger, and hence smaller firms, will have a faster growth rate than larger firms. Jovanovic first argues that the growth rate can't be the same for all firms, because it would require that the entire distribution of Ez_t be the same for all[47]. If Gibrat's Law would hold, it would require that two firms with the same estimate of the true efficiency, but different precisions, would have the same distribution. This cannot obviously hold. In other words, the first version of Gibrat's law will not hold in the model. Jovanovic's second argument states that the growth rate can't be same within a single age cohort. The reason is that the smaller firms will have a higher variance in their growth rates as they are more sensitive to information. Brock & Evans (1986) states that this is, because of young firms have less precise estimates of their true abilities. Young firms will also grow faster and this is the consequence of simple statistical principle, the Jensen's inequality.


$Z_t^* = (\frac{x_t^*}{x_{t+1}^*})$
$E(\frac{x_t^*}{x_{t+1}^*})$
$\frac{x_t^*}{Ex_{t+1}^*} = 1$
$x^*_{t+1,1}$
Ex^*_{t+1}
$x^*_{t+1,2}$
x_t^*

Picture 5- Jensen's inequality in Jovanovic's model

According to Jovanovic (1982), change in the costs z_t (i.e. the increase or decrease of the true efficiency) is increasing in x_t as can be seen above in picture 5. Basically, this can be translated into that the higher costs are the higher will be the benefit from a decrease in costs. The expected change in costs will always be higher than the actual growth rate as implied by the Jensen's inequality[48].

[47] $Ezt = \frac{x_t^*}{x_{t+1}^*}$, the expected rate of decrease in costs (i.e. the expected increase in efficiency)

[48] i.e. $E_t\left(\frac{x_t^*}{x_{t+1}^*}\right) > \frac{x_t^*}{Ex_{t+1}^*}$

The link between costs, x_t, and size is similar to the one the relationship between productivity and size in Hopenhayn's (1992) model. Those firms that have a high costs will also be small. Therefore, small firms will grow faster than the large firms and Gibrat's law can't hold in environment that is specified by Jovanovic model.

However, there is a more interesting finding regarding the growth-question. Namely, according to Jovanovic the true efficiency converges to a constant. Therefore, for the mature (and hence larger firms) the growth rates should be equal. The reason is lies in selection and the fact that the precision increase as time lapses. In other words, as the industry matures the firms' growth rate converges and decreases as there is less need to revise one's production. This is because the surviving firms have already gathered enough information and they have deducted that they are efficient. It is good to note that in Jovanovic's model there is also negative relationship between age and growth. This could imply that Gibrat's law could be valid for matured industries or at least the deviations from the law should be smaller for larger and more matured firms as found by Evans (1987 a, b). With regards the new and more refined null hypothesis, a possible new test could be to choose two similar industries, but with different maturities or phases in the product life cycle. It should be expected that the more matured industry follows more closely Gibrat's law.

Jovanovic's model fortifies the view that Gibrat's law can't be valid and there are no clear cut special cases where the growth would be stochastic as in Hopenhayn's model. Further, it validates that the age is an important factor when discussing growth. The age-dependency emerges, because firms don't know their distribution and this is a realistic assumption. Nevertheless, there are aspects that the model doesn't take into account. One is active learning, but the theme of the following section is to combine the age and size dependency. This is done by introducing capital and finance as possible factors.

4.2 Introducing finance constraints

The previous section discussed the role of (passive) learning for industry evolution. The purpose of this section is to discuss impact of finance. The interest is in environments where there is a friction in finance for example due to high debt or equity costs. Cooley & Quadrini (2001) presents a simplified model of Hopenhayn where finance is an issue. What makes the approach of Cooley & Quadrini more robust is that their most important contribution is to show the relationship of simultaneous dependence of industry dynamics on size and age by assuming persistent shocks and frictions in finance. This simultaneous dependence means that similar size companies will grow at different rates if they have different ages. In addition, they present more detailed predictions on behavior of the firm for example on the size of dividend it pays and debt it takes.

4.2.1 Frictions in finance and persistent shocks

Cooley & Quadrini build their firm dynamics on a simplified version of Hopenhayn. Basically, it means that there is a continuum of firms that produce a homogenous product. Instead of maximizing expected profits the firm's objective is to maximize expected dividends over an infinite-time horizon[49]. There is no outside pressure from other industries. So the "competition" occurs only within the industry. The incumbent firm can use both capital and labor as inputs for production and they are perfect complements. According to Cooley & Quadrini this means that the capital-labor ratio is constant. Depreciation needs to be taken into account when capital is a possible input variable. Machines and equipment wear out as they are used. In Cooley & Quadrini's model it is assumed that capital depreciates with a constant rate. The production technology has decreasing returns to scale due to concavity of the production function. The intuition given in the article is that limited managerial or organizational resources lead to that output will increase less than the increased inputs.

The model is built on the assumption of firm-level heterogeneity. There are two sources of heterogeneity. In addition to the standard productivity shock, frictions in finance will make firms (even more) different. As in Hopenhayn, a productivity shock stipulates how efficiently the inputs are used in production. The productivity shock itself is defined more elaborately than in Hopenhayn as it is composed of two different parts. The first part is the persistent shock, similar to the serially correlated productivity shock and it can be interpreted as technological differences. The persistent shock follows a first-order Markov process, meaning history doesn't have a role as discussed earlier. The second component represents the pure accidentally or lucky events and they are modeled as nonpersistent and i.i.d shocks with zero mean. The distinction between two types of shocks was not present in Hopenhayn and therefore Hopenhayn's model can be seen as a special case of Cooley & Quadrini's model in this case. The reason is simple. In Hopenhayn's model the pure accidental random variable exists, but it has zero mean and variance and hence no impact.

[49] If all dividends were to be paid to shareholders, then it wouldn't make a difference if the firm value or dividends were maximized. However, in Cooley & Quadrini some of the earnings are retained in the firm and this could have implications on the results.

There are also differences in timing of the different shocks. The persistent shock is revealed one time period before the production while the “luck” shock is revealed in the same period as production occurs. The amount of capital and labor is decided in the same period as the persistent shock is revealed. Cooley & Quadrini states “in the absence of financial frictions, the efficiency level of the firm fully determines its size”. Put differently, the results would be very similar to Hopenhayn.

The question of finance was irrelevant in the previous models. It was assumed implicitly that firms had ample resources and access to the capital markets without any extra costs. This is not the case in Cooley & Quadrini as finance is the extra layer that makes the approach and results different. Generally, new investments’ financing can be categorized into internal and external financing. Internal finance consists of retained earnings and external finance is understood as new equity and debt. All three forms of finance are possible in Cooley & Quadrini. There is a wide range of articles and studies on the relative advantages and disadvantages of the different forms of finance. This thesis will not go deeper into the finer details, but a general stylized fact is that external finance is more expensive than internal.

This stylized fact is reflected in Cooley & Quadrini’s model as they state “the financial frictions arise because of the following assumptions: (a) there is a cost or premium associated with increasing equity by issuing new shares, compared to reinvesting profits; (b) defaulting on the debt is costly”. Equity finance can be expensive because it dilutes the value of existing stocks. Another reason could be the information effects related to the information asymmetries between managers and investors as Myers & Majluf (1984) showed. However, also the debt has to be expensive, in order to have frictions in the financial market. Cooley & Quadrini’s (2001) states if debt were costless, all firms would prefer it over expensive equity. It is good to note that to the contrary to other models so far, the firm has to decide whether it wants to pay dividends or not. After all, the firm’s objective is to maximize dividends. Therefore, there exists a threshold level on equity and when this threshold is surpassed the firm will start to pay dividends. Dividends and the payout-ratio stipulate the level of available internal finance.

In the model, the debt contract lasts for one period. The borrowed money should be paid back together with interest at the end of the period. The loan is acquired from a financial intermediary. The primary reason that debt is expensive is that it is not risk-free as there is a possibility of default and hence the firm wouldn’t be able to repay the borrowed funds. To compensate for the possible bankruptcy, the financial intermediary will charge a higher interest rate which depends on the probability of default. The firm will decide on default after productivity shock and amount of revenues have been revealed. It will default if its net worth is equal to or less than zero. The financial intermediary verifies the bankruptcy. It is important to note that the cost of verification is transferred back to the firm by embedding the cost into the interest rate. Interestingly, default does not lead to liquidation and exit as it is not in the interest of the financial intermediary and therefore the debt is renegotiated on new terms.

Although the default decision seems to be similar concept to the exit decision in Hopenhayn and the other models, it is not so. To the contrary, the firms continue after the debt has been renegotiated and hence no exit occurs due to defaults. Exit will take place exogenously and the firm will exit when it turns unproductive. Hence, it is only the productivity shocks that drive the exit behavior. Why would any firm want use equity if default doesn't lead to exit? The reason is that debt is expensive and after a certain threshold it reduces the firm's value as will be seen below. After the default and renegotiations, the firm returns back to business. To sum up, verification and the possibility of bankruptcy make the debt financing expensive.

But what would be the reasons to use external finance? The most imminent reason is that it gives the possibility to expand the firm's asset and hence production without the immediate need to generate extra revenue. If external financing is introduced, the firm can expand its production beyond a level that only internal finance would permit. Naturally, this will lead to increased expected profits. The downside of debt is that it is expensive and that debt amplify the volatility of firm's value, which has an adverse effect on firm value due concavity. New equity doesn't have this amplification effect and to the contrary it reduces the stress from default risk. Cooley & Quadrini notes that the problem on equity expansion is that due to decreasing returns to scale the increase in production is not proportional to the increase in equity. The right combination of debt and equity is a trade-off between the advantages and disadvantage of the different forms of finance. It goes without saying that finance will have a considerable effect on dynamics of the firm and industry.

As in all the industry models covered so far, entry is a possibility. Entry happens by paying a fixed cost, which can be assumed to be sunk, nonrecoverable, and the additional cost of issuing new equity. Naturally, the firm enters only if the expected value of the firm is larger than or equal to cost of entry. According to Cooley & Quadrini all new firms will be of the highest efficiency, because high productivity entrants have a higher firm value than entrants with small productivity. The assumption is justified by the observations that in general new firms are more efficient than incumbents due to better technology. In each period, entry will occur as long as the surplus from entry is nonnegative. The arbitrage conditions will ensure that entry is optimal[50]. The role of entry and exit can be seen as part of the resource reallocation process just as in the other models.

4.2.2 Simultaneous dependencies between growth and age

It goes without saying that the model can produce more diverse predictions than Hopenhayn's model as finance is an additional component and source of heterogeneity. Cooley & Quadrini can for example show that small firms take on more debt or that small firms face higher rates of profits only conditioning on the frictions in finance and surviving. The first result is obtained, because the firm becomes more alarmed of the profits' volatility as their size increases and therefore borrows less in proportion to its size[51]. Hence, the share of debt is decreasing in size of the equity. Cooley & Quadrini notes that "as a consequence of higher borrowing, small firms face higher probability of default". The second outcome is as a by-product of the capital structure decision. As the larger firms use more equity, they will be less profitable also due to diminish returns on scale.

The outcomes above are interesting as they supplement the picture on the behavior of the firm, but Cooley & Quadrini can also comment on the growth-size relationship. They firsts study the scenario, where there are frictions in finance, but no variations in the productivity shock. They are able to show that growth-size relationship is negative. This is done by assuming that the productivity shock is binary in the sense that in only takes values zero or one. The firms with a zero productivity shock will exit and hence the remaining firms have identical internal efficiencies. The negative relationship can be traced back to financial results presented above. Smaller firms will have higher rate of earnings and the dividend payout ratio will be lower for smaller firms. In other words, small firms plowback more money into the firm and have relatively more internal financing. These retained earnings are then used to finance investments and which will lead effectively to faster growth. The standard deviation for growth is also higher for smaller firms expect for very small firms.

[50] Arbitrage conditions means changes in output and input prices which lead to changes in firm value

[51] Cooley & Quadrini argue "that the firm compares the marginal increase in the expected profits with the marginal increase in its volatility". The reason that volatility is used is that the increasing volatility decrease firm's value due to objective function's concavity. Basically, volatility is one of finance's costs.

According to Cooley & Quadrini "the model generates an unconditional age dependence of firm dynamics", but this dependence is ostensible. Cooley & Quadrini continues that "young firms are small, which in turn derives from the small size of new entrants." In other words, the reason why there seems to be relationship between age and growth is that age is only a proxy for size. The relationship can be shown to be robust in case also the productivity shock can vary. In order to create a simultaneous dependence both on age and size two sources of firm-level heterogeneity is needed. In this variant, an interesting finding is that higher productivity firms have higher default rates because they predispose themselves to more risk.

Cooley & Quadrini show that how the growth rate, default rate and job creations have a negative relationship with both age and size with the exception of job creation for very small firms. The results are received after the age and size, respectively, have been controlled. The reason for size dependencies is similar to the case presented above, small firms plowback more revenues than large firms and this results in growth. The age dependence outcome results from the assumption that young firms are highly productive. This leads to that young firms have higher rates of profits and as explained above, it will eventually transform into faster growth. It is good to note that if the shocks were not persistent there would be no age dependency, because the difference between the firms would fade away quickly. Similarly, if the new entrants would have low productivity, the age-dependency wouldn't exist.

To sum up, Cooley & Quadrini's main result is that firm's size will not anymore depends on its internal efficiency. The size-dependency existed, because small firms plowed back more money in in order to avoid using costly external finance. In Cooley & Quadrini's model, the size and growth relationship will always be negative, simply because the entrants had a high productivity. In other words, Gibrat's law will not be valid when studied with Cooley & Quadrini's model. However, the age-dependency can vary and a testable hypothesis could be to look at industries were the productivity doesn't fluctuate anymore. This implies that for more mature industries the age-dependencies should be more stochastic than for young industries.

The results are interesting when compared to the ones obtained in Hopenhayn. After all, Hopenhayn didn't have any age dependence even though there were persistent shocks. The difference seems to be in the role of new entrants. In Hopenhayn, the entrants didn't know what type of firms they are, but in Cooley & Quadrini the role is known and hence only high productive firms enter. In the following section, another variant of Hopenhayn's model is presented with liquidity constraints. Although the assumptions are similar, the results are somewhat different from Hopenhayn.

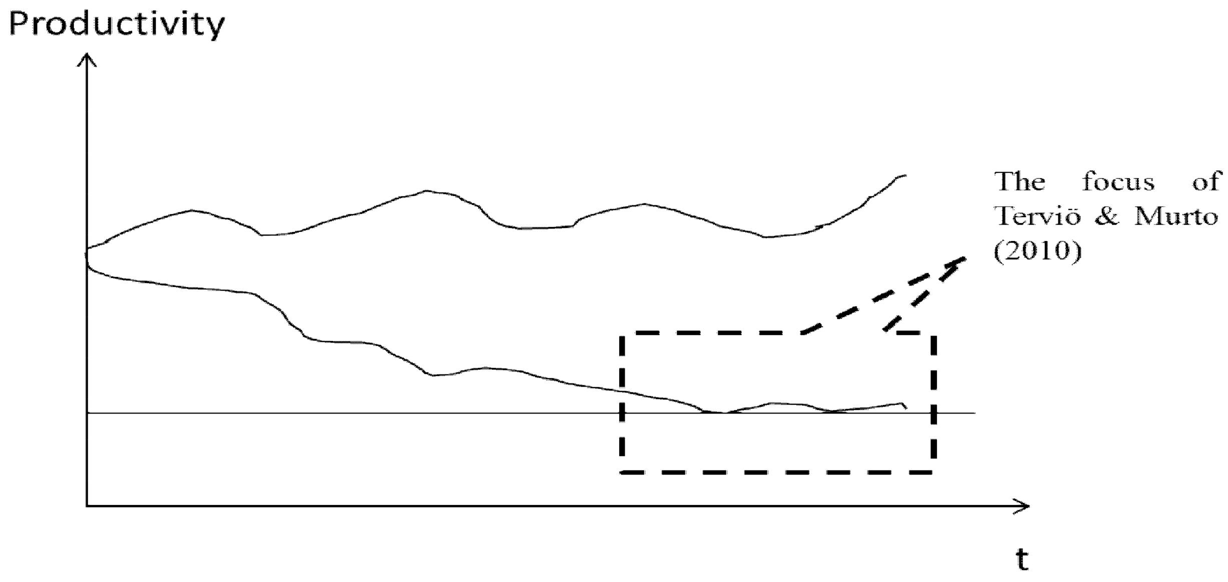
4.3 Survival of the fattest

Also Murto & Terviö (2010) has studied the impact of finance on the firms' evolution and behavior. The main difference with Cooley & Quadrini's (2001) model is that the inability to finance its operations leads to exit when in Cooley & Quadrini's model this only leads to renegotiation of debt. As in other models, the profit-maximizing firm is the primary agent and the objective of the firm is to maximize the expected present value of the income to the shareholders[52]. The fear of premature exit, and hence the loss of business, encourage firms to hoard cash as a safety measure. This is costly, because cash pays less interest than dividends. Murto & Terviö explains the liquidity constraint as the inability to raise new funds[53].

The other and more important result is that the model leads to the "survival of the fattest". The result is rather different when compared to the other models. For example Jovanovic's (1982) model leads to the "survival of the fittest" and in Hopenhayn (1992) it was always the most productive that survived in the industry. So, in a sense introducing a liquidity constraint, changes the results drastically. However, the results of the different models are of course not directly comparable. The focus in Murto & Terviö's model is not the same as in the ones presented earlier. The key conceptual difference is illustrated by picture 6 which is an adaption from Jovanovic's (1982) article. Basically, Murto & Terviö are interested in what happens once the firm's productivity has decreased sufficiently low that exit is a valid alternative. They argue that sometimes it is rational policy to exit the industry precautionary.

[52] The model is solved numerically with dynamic programming as there is no close-ended solution.

[53] The source of the inability to raise new funds is not modeled. Murto & Terviö discuss that their model is a special case of a more general model where raising news funds is expensive. In their case, the cost of raising new cash is so high that it is never optimal to do so. Murto & Terviö presents also the model where it is possible to raise new cash as a lump sum. Finally, there are no other sources of imperfections.



Picture 6 – Illustration of the conceptual differences between Murto & Terviö (2010) and Jovanovic's (1982) model.

4.3.1 The importance of cash

According Murto & Terviö (2010), their model is similar to the ones in Hopenhayn (1992) and Dixit & Pindyck (1994, Ch 8.4), but naturally there are differences. The firms' revenue consists of two components as the revenue is the product of price and the firm specific productivity. However, the productivity shock follows a different path. The productivity shock follows a Brownian motion in Murto & Terviö (2010). In other words, the productivity shock evolves continuously while in Hopenhayn (1992) and the other models the productivity moved in discrete jumps[54]. The random walk is presented in equation 18 where dw is the increment of a standardized wiener process and μ is the expected mean of productivity.[55]

18) $$dz = \mu zdt + \sigma zdw$$

Naturally, the price is determined by the demand curve that is everywhere strictly downward sloping. The industry is competitive so each individual firm doesn't have market power. Hence, the marginal revenue is constant, p. In order operate the firm has to pay a fixed cost. Hence, the profit is difference between revenue and the fixed cost. Additionally, the firm earns interest on the cash it

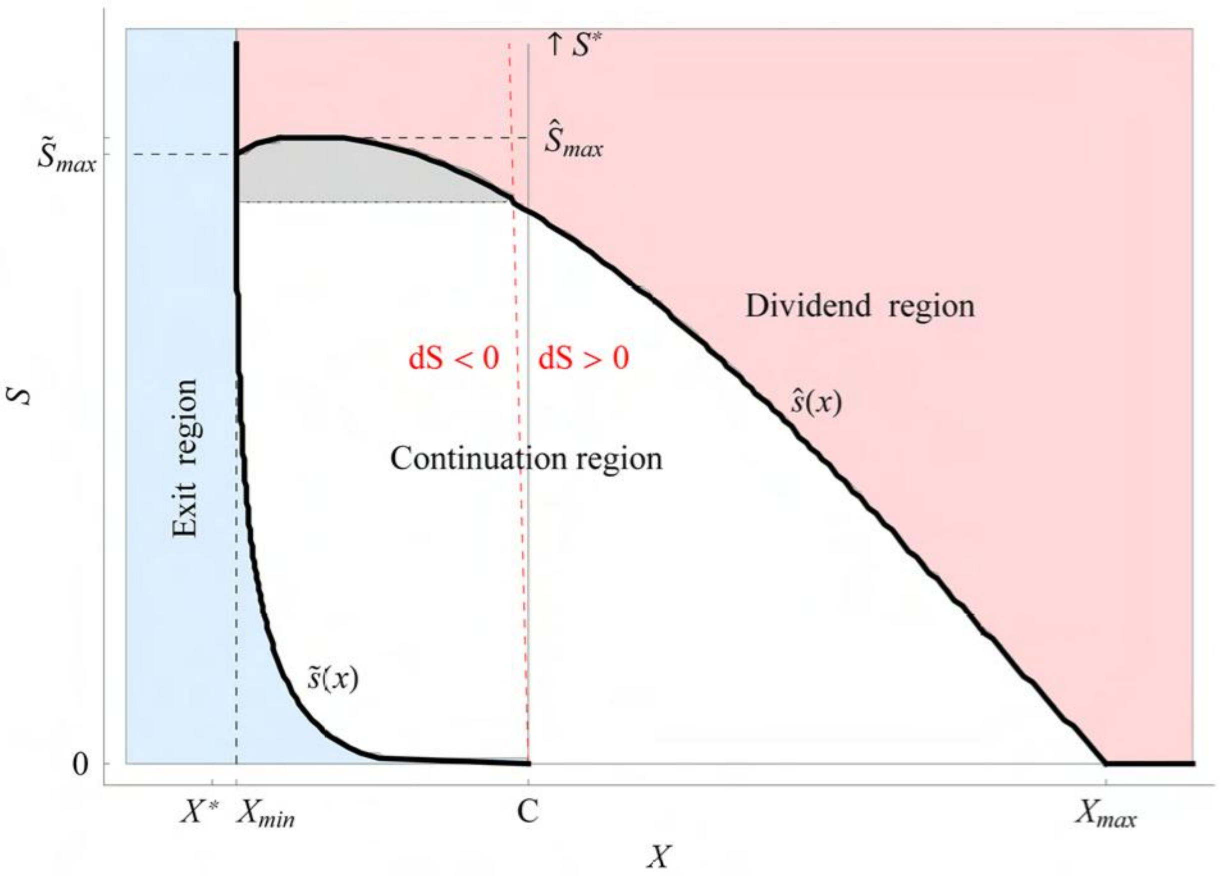
[54] There are also similarities. For example the Brownian motion is also a Markov process as was the productivity shock in Hopenhayn (1992).

[55] The shocks, dw, are independent across firms.

accumulated. The profit can be paid out as dividends or retained as cash in the firm. As stated earlier, cash has an important role in Murto & Terviö's model and to further understand its role one should look at the exit process[56].

Exit is irreversible and without an additional exit cost. As in Hopenhayn's (1992) model, the firm will exit if the productivity decreases under the critical threshold level. Similarly, this doesn't mean that making a loss is a sufficient reason to exit as discussed in section 3.3.3. In other words, there can be periods of losses and in Murto & Terviö's model. These losses have to be covered with cash. This is the major difference with other models as they implicitly assumed that firms can always cover their losses. In Murto & Terviö's model, firms have an initial stock of cash and this can only change by retaining some of the earnings. As stated, retaining some of the earnings is costly as cash pays less interest than dividends. Nevertheless, hoarding cash can be optimal as in case that the firm doesn't have enough cash to cover its negative cash flow it is forced to exit immediately irrespective of what the productivity is. If firm decide to exit as a precautionary measure, the remaining cash will be paid back to the shareholders. Obviously nothing will be paid to the shareholders if the firm is forced to exit.

[56] Murto & Terviö presents two different cases, one where the firm is unconstrained and one where the firm is constrained. The assumption and model presented are for the constrained case.



Picture 7 - Optimal policy regions of a liquidity constraint firm from Murto & Terviö (2010)

Picture 7 illustrates the impact of the liquidity constraint on the firm's behavior. The picture is taken from Murto & Terviö's article where x is the productivity of the firm, S is the amount of cash the firm has and c is the fixed costs. The firm has three different alternatives and they are exiting the industry, continuing in the industry and paying dividend and finally continuing in the industry and not paying dividends. The firm will always pay dividends when it is in the dividend region, but it will never stay there as the firm is always moved to the continuation region after the cash been distributed to shareholders[57]. When the productivity is sufficiently high or the firm has ample finance, the probability of exit is so small that the firm can transfer the cash to the shareholders. According to Murto & Terviö (2010), the liquidity constraint has no direct impact on the dividend policy, only on the exit policy.

The firm will always exit when revenue falls below $x*$ irrespective if there is liquidity constraint or not. The liquidity constrained firm is forced to exit when revenue, x, is smaller than costs, c, and it

[57] Naturally, no cash is paid in the continuation region.

has no cash. The most interesting area is the one where the firm has a very little cash (small s) and revenue (x) is just above the minimum required revenue (x*). According to Murto & Terviö, this kind of firm could in principle continue, but they further elaborate that "for sufficiently small *s* the firm is so unlikely to bounce back to a positive cash flow before *s* hits zero that it is better exiting immediately and just taking the remaining *s*." In other words, the firm exits as precautionary action in order to ensure that the shareholders receive at least something. This precautionary exit is illustrated by the small area between continuation region and the dotted $x_{\min}$ line. The result of precautionary exit is that some marginally productive firms that would have survived a temporally loss will exit due to insufficient funds (or more accurately in order to preempt forced exit)[58].

In addition to the forced a precautionary exit, Murto & Terviö has assumed an exogenous death rate λ "at which firms are forced to exit with their cash holdings as the exit value". This will not drastically change the behavior of the firms as the difference according to Murto & Terviö is that firms take probability of exogenous death into account in their discounting. The assumption is made to guarantee a steady state. Firms can enter the industry by paying the cost of entry, φ. The productivity of the entering firms is known and all the new firms have initial cash holdings s_0. The entry is endogenous and it must fulfill the zero-profit condition[59]. Similar to Hopenhayn (1992) there is no aggregate uncertainty in the steady state. According to Murto & Terviö "all firms follow the same optimal policy, which in turn results in a stationary distribution of [productivity] z".

4.3.2 Another look on the growth and welfare question

Regarding Gibrat's law the results are similar to Hopenhayn (1992) if production (i.e. productivity) is a measure of size. There are few exceptions. Namely, the selection effect is much stricter as there are precautionary exits. On the other hand, if the firm has ample cash it can sustain longer periods of loss and therefore is not forced to exit as often as in Hopenhayn's model. The total impact of these two contradictory forces is ambiguous without additional assumptions. Murto & Terviö analyzed the impact of the liquidity constraint with numerical analysis and found that "the liquidity constraint has a negative impact on mean productivity at low levels of [cost of entry] φ". So, when the cost of entry is low, the "fat" firms survive and one should expect that there should be fewer deviations from Gibrat's law. This could be a new testable hypothesis if an industry that fits the

[58] One implication of this is that all exits are basically precautionary.

[59] Formally this that the following equation must be true

$$V(pz_o, s_0) = \varphi + s_0$$

given assumptions can be identified. Murto & Terviö has estimated that when the cost of entry is sufficiently low the mean productivity decreases up to 15 %. The impact of the model's second parameter initial cash, s_0, has also an impact. Murto & Terviö states that "the liquidity constraint is harsher when s_0 is small, so the relative distortion is always decreasing in s_0 as the constraint becomes milder".

Liquidity constraint has also an impact on welfare and Murto & Terviö has identified three different sources of distortion. They are higher aggregate entry cost (due to higher turnover), lower average productivity, and higher liquidity costs. They further state that "the only component of welfare that can be affected by the liquidity constraint is consumer surplus, which varies in the opposite direction as [price] p". The distortion of from the lower average productivity is due to the same contradictory forces as stated above. To sum up, liquidity constraints can have a clear impact both on welfare and the growth-size relationship. The exact magnitude depends on the further assumptions, but these findings further strengthen the view that Gibrat's law can't be valid.

5 Conclusions and discussion

This thesis has studied Gibrat's law and the firm dynamics with the help of conventional industry dynamics model. Gibrat's law is an old theory and it has been studied extensively for past 60 years. Although the initial studies accepted the law, further studies have shown that the law should be rejected. The development of empirical methods is one of the reasons why the stance on Gibrat has changed. However, the results have not always been clear-cut and as explained in section 2.3, there have been many studies that have accepted the law in whole or for some subsample. Therefore, this thesis states that the question is not whether Gibrat's law is valid or not, but rather when and with what restrictions is it valid. Another important matter to explore is that why Gibrat's law should be valid.

This thesis argued that the Gibrat's law can't be a law in a strict sense although it can be possible to observe Gibrat-like firm growth. The earlier studies have presented plenty of reasons why the law should be rejected. A popular explanation was that small firms are better at innovating. Yet, at the same time it have been acknowledged that it is more likely that large firm invests more in R&D, because of the larger risks for small firms. Another explanation was that small firms are more agile than large firms because of risk of losing reputation or diseconomies of scale. These are fine reasons, but there are more fundamental and comprehensive explanations to indicate that Gibrat's law can't be valid.

Namely, it can be shown that in Hopenhayn's (1992) industry dynamics model firm's growth rate can be both stochastic and proportional to firm size. Hopenhayn's model is a traditional industry dynamics model where there is a continuum of firms that maximize the expected firm value. Both entry and exit is endogenous. In addition, the firm's productivity shock is stochastic making the firms heterogenetic. Therefore, it is truly amazing that Hopenhayn can show that the stationary equilibrium is competitive and hence there is no welfare loss. If the cost of entry is not too high and some auxiliary assumptions are satisfied, then small firms will grow faster than larger firms. The reason is selection as small firms that don't have a bright future prospect don't have the incentive to continue in the industry and thus only the most efficient small firms are "selected" to continue. However, if the cost of entry is too high, then there is no selection and the industry will continue as such forever implying that the observed growth is stochastic.

However, this doesn't mean that the actual growth is stochastic, because growth is a result of firms maximizing profits. In other words, both incumbents and entrants make different decisions based on the available information and these decisions leads eventually to growth that sometimes can be observe as random patterns. For example if the entry cost is too high, then the possible entrants decide not to join the industry. This decision helps creating the environment, where one could observe stochastic growth. Therefore, growth is never purely stochastic as there is an underlying process that is deterministic. These findings could explain why sometimes studies reject the law and sometimes they accept it. Of course Hopenhayn's model is a stand-alone model, but it opens the possibility for other models were Gibrat's law could be a special case. Three other models were presented that were similar to Hopehayn's model and they all more or less confirmed that Gibrat's law can't valid.

This thesis also touched upon on what induces growth. In Gibrat's model growth was random, at least compared to the firm size, but the conventional economics models offered more intriguing explanations for growth. Hopenhayn (1992) explained that growth arises from superior productivity meaning that those firms that are more productive will also grow faster. A similar explanation where given by Jovanovic (1982). The difference between these two models is that in Hopenhayn's model the firms reacted to an exogenous productivity shock and in Jovanovic the firms attempted to deduct their own true productivity. Jovanovic defined productivity as cost efficiency while Hopenhayn didn't make this distinction.

One of the past research’s problems is that the law has been studied in a generic way meaning that the focus has solely been on testing if the law is valid or not. One of the aims of this study was to identify to testable hypothesis that would more accurate and hence they would further clarify the role of Gibrat’s law. Fortunately, it was possible to find new testable hypothesis. For example Gibrat’s law could be tested for an industry where there has been a large increase in cost of entry, preferably over a shorter period time[60]. The expected result is that after the increase in cost of entry, there should be less deviation from Gibrat’s law. One possible idea for further studies is to test these new hypotheses.

The models used in this study are not designed specifically to study Gibrat’s law and many times the models’ focus was on industry-level rather than firm-level. The results from these industry models can be seen as more robust as the growth-size relationship results are by-products of other studies. Nevertheless, one possibility for future study is to design models that are more focused on the growth-size relationship in order to create new testable hypothesis. There are a plethora of models available, so these could be reused somehow.

Another possible area that should be further studied is the role of the demand side. All the models presented in the thesis are supply-driven and they assume that the demand side behaves optimally. In other words, these are partial equilibrium models and of course they are not as robust as general equilibrium models. Bergin & Bernhardt (2008) has already started to study the role of demand. In their model “the dynamics of an industry is subject to aggregate demand shocks where the productivity of a firm's technology evolves stochastically over time”. This strand of literature could be extended for example looking at both frictions in finance and aggregate demand shocks at the same time.

The conclusion of this study is that in majority of the cases small firms indeed grow faster than large firms. This is supported both by theoretical and empirical evidence. It can be case that sometimes the growth is observed as stochastic, but it would seem that underlying process is indeed deterministic as there are profit-maximizing firms that act and make decisions. The actions are not random, but sometimes this leads to growth that is observed as random. In other words, this study concludes that Tekes should continue to target small and medium size firms with their subsidies.

[60] The industry should satisfy the assumptions made by Hopenhayn (1992)

References

Almus, M. & Nerlinger, E.A. 2000, "Testing 'Gibrat's Law' for Young Firms - Empirical Results for West Germany", *Small Business Economics,* vol. 15, no. 1, pp. 1-12.

Armington, C. & Acs, Z.J. 2004, "Job creation and persistence in services and manufacturing", *Journal of Evolutionary Economics,* vol. 14, no. 3, pp. 309-325.

Audretsch, D.B., Klomp, L., Santarelli, E. & Thurik, A.R. 2004, "Gibrat's Law: Are the services different?", *Review of Industrial Organization,* vol. 24, no. 3, pp. 301-324.

Audretsch, D.B. & Santarelli, E. 1999, "Start-up size and industrial dynamics: some evidence from Italian manufacturing", *International Journal of Industrial Organization,* vol. 17, no. 7, pp. 965 - 983.

Aw, B.Y., Chen, X.M. & Roberts, M.J. 2001, "Firm-level evidence on productivity differentials and turnover in Taiwanese manufacturing", *Journal of Development Economics,* vol. 66, no. 1, pp. 51-86.

Aw, B.Y., Chung, S. & Roberts, M.J. 2003, "Productivity, output, and failure: A comparison of Taiwanese and Korean manufacturers", *Economic Journal,* vol. 113, no. 491, pp. F485-F510.

Balasubramanian, N. & Sivadasan, J. 2009, "Capital resaleability, productivity dispersion and market sturcture", *Review of Economics and Statistics,* vol. 91, no. 3, pp. 547-557.

Becchetti, L. & Trovato, G. 2002, "The determinants of growth for small and medium sized firms. The role of the availability of external finance", *Small Business Economics,* vol. 19, no. 4, pp. 291-306.

Bergin, J. & Bernhardt D., 2008, “Industry dynamics with stochastic demand”, *Rand Journal of Economics,* vol. 39, no. 1, pp. 41-68.

Bernard, A.B., Redding, S., Schott, P. K., 2006. “Multi-product firms and product switching”. *Centre for Economic Performance (CEP) Discussion Paper No. 736.*

Brook, W. A. & Evans, D.S., 1986, ‘The economics of small businesses, their role and regulation in the U.S. economy”, *Holmes & Meier Publishers,* New York

Cabral, L.M.B. 2007, "Small firms in Portugal: a selective survey of stylized facts, economic analysis, and policy implications", *Portuguese Economic Journal,* vol. 6, no. 1, pp. 65-88.

Calvo, J. 2006, "Testing Gibrat’s Law for Small, Young and Innovating Firms", *Small Business Economics,* vol. 26, no. 2, pp. 117-123.

Chesher, A. 1979, "Testing the Law of Proportionate Effect", *The Journal of Industrial Economics,* vol. 27, no. 4, pp. 403-411.

Cooley, T.F. & Quadrini, V. 2001, "Financial markets and firm dynamics", *American Economic Review,* vol. 91, no. 5, pp. 1286-1310.

Das, S. 1995, "Size, age and firm growth in an infant industry: The computer hardware industry in India", *International Journal of Industrial Organization,* vol. 13, no. 1, pp. 111-126.

Dixit, K. A. & Pindyck, R. S, 1994, “Investment under Uncertainty”, *Princeton University Press*

Dougherty,   2002, “Introduction to econometrics, second edition”, *Oxford university press,* Oxford

Dunne, P. & Hughes, A. 1994, "Age, Size, Growth and Survival: UK Companies in the 1980s”, *Journal of Industrial Economics,* vol. 42, no. 2, pp. 115-140.

Dunne, T., Klimek, S.D. & Roberts, M.J. 2005, "Exit from regional manufacturing markets: The role of entrant experience", *International Journal of Industrial Organization,* vol. 23, no. 5, pp. 399-421.

Dunne, T., Roberts, M.J. & Samuelson, L. 1989, "The Growth and Failure of U. S. Manufacturing Plants", *The Quarterly Journal of Economics,* vol. 104, no. 4, pp. 671-698.

Ericson, R. & Pakes, A. 1995, "Markov-Perfect Industry Dynamics: A Framework for Empirical Work", *The Review of Economic Studies,* vol. 62, no. 1, pp. 53-82.

Evangelista, R., Sandven, T., Sirilli, G. & Smith, K. 1998, "Measuring Innovation in European Industry", *International Journal of the Economics of Business,* vol. 5, no. 3, pp. 311-333.

Evans, D.S. 1987, "The Relationship Between Firm Growth, Size, and Age: Estimates for 100 Manufacturing Industries", *The Journal of Industrial Economics,* vol. 35, no. 4, The Empirical Renaissance in Industrial Economics, pp. 567-581.

Evans, D.S. 1987, "Tests of Alternative Theories of Firm Growth", *The Journal of Political Economy,* vol. 95, no. 4, pp. 657-674.

Fariñas, J.C. & Ruano, S. 2005, "Firm productivity, heterogeneity, sunk costs and market selection", *International Journal of Industrial Organization,* vol. 23, no. 7, pp. 505-534.

Gibrat, R. 1931, “Les inegalites economiques; aux inegalite's des richesses, a la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel, *Librairie du Recueil Sirey*

Goddard, J.A., McKillop, D.G. & Wilson, J.O.S. 2002, "The growth of US credit unions", *Journal of Banking & Finance,* vol. 26, no. 12, pp. 2327-2356.

Goddard, J., McMillan, D. & Wilson, J.O.S. 2006, "Do firm sizes and profit rates converge? Evidence on Gibrat's Law and the persistence of profits in the long run", *Applied Economics,* vol. 38, no. 3, pp. 267-278.

Goddard, J., Wilson, J. & Blandon, P. 2002, "Panel tests of Gibrat’s Law for Japanese manufacturing", *International Journal of Industrial Organization,* vol. 20, no. 3, pp. 415-433.

Greenaway, D., Gullstrand, J. & Kneller, R. 2009, "Live or Let Die? Alternative Routes to Industry Exit", *Open Economies Review,* vol. 20, no. 3, pp. 317-337.

Hall, B.H. 1987, "The Relationship Between Firm Size and Firm Growth in the US Manufacturing Sector", *The Journal of Industrial Economics,* vol. 35, no. 4, The Empirical Renaissance in Industrial Economics, pp. pp. 583-606.

Hamilton, O., Shapiro, D. & Vining, A. 2002, "The growth patterns of Canadian high-tech firms", *International Journal of Technology Management,* vol. 24, no. 4, pp. 458-472.

Hardwick, P. & Adams, M. 2002, "Firm Size and Growth in the United Kingdom Life Insurance Industry", *Journal of Risk & Insurance,* vol. 69, no. 4, pp. 577-593.

Harris, R.I.D. & Li, Q.C. 2010, "Export-market dynamics and the proability of firm closure: Evidence for the United Kingdom ", *Scottish Journal of Political Economy,* vol. 57, no. 2, pp. 145-168.

Heshmati, A. 2001, "On the Growth of Micro and Small Firms: Evidence from Sweden", *Small Business Economics,* vol. 17, no. 3, pp. 213-228.

Hopenhayn, H.A. 1992, "Entry, Exit, and firm Dynamics in Long Run Equilibrium", *Econometrica,* vol. 60, no. 5, pp. 1127-1150.

Johansson, D. 2004, "Is small beautiful? The case of the Swedish IT industry", *Entrepreneurship & Regional Development,* vol. 16, no. 4, pp. 271-287.

Jovanovic, B. 1982, "Selection and the Evolution of Industry", *Econometrica,* vol. 50, no. 3, pp. 649-670.

Kalecki, M. 1945, "On the Gibrat Distribution", *Econometrica,* vol. 13, no. 2, pp. 161-170.

Kumar, M.S. 1985, "Growth, Acquisition Activity and Firm Size: Evidence from the United Kingdom", *The Journal of Industrial Economics,* vol. 33, no. 3, pp. 327-338.

Laitinen, E. K. 1999, **“**Stochastic growth processes in large Finnish companies: test of Gibrat’s law of proportionate effect”, *Liiketaloudellinen Aikakauskirja*, no.1 pp. 27-49

Lotti, F., Santarelli, E. & Vivarelli, M. 2003, "Does Gibrat's Law hold among young, small firms?", *Journal of Evolutionary Economics,* vol. 13, no. 3, pp. 213-235.

Mansfield, E. 1962, "Entry, Gibrat's Law, Innovation, and the Growth of Firms", *The American Economic Review,* vol. 52, no. 5, pp. 1023-1051.

Moreno, A.M. & Casillas, J.C. 2007, "High-growth SMEs versus non-high-growth SMEs: a discriminant analysis", *Entrepreneurship & Regional Development,* vol. 19, no. 1, pp. 69-88.

Murto, P. & Terviö, M, 2010, “ Exit Options and Dividend Policy under Liquidity Constraints”, Working paper, http://www.hse-econ.fi/tervio/

Myers, S.C. & Majluf, N.S. 1984, "Corporate Financing and Investment Decisions when Firms have Information that Investors do Not have", *Journal of Financial Economics,* vol. 13, no. 2, pp. 187-221.

Ortega-Argilés, R., Vivarelli, M. & Voigt, P. 2009, "R&D in SMEs: a paradox?", *Small Business Economics,* vol. 33, no. 1, pp. 3-11.

Penrose, E., 1959, "The Theory of the Growth of the Firm", *Basil Blackwell,* Oxford.

Plehn-Dujowich, J.M. 2009, "Entry and exit by new versus existing firms", *International Journal of Industrial Organization,* vol. 27, no. 2, pp. 214-222.

Reichstein, T. & Jensen, M.B. 2005, "Firm size and firm growth rate distributions--The case of Denmark", *Industrial & Corporate Change,* vol. 14, no. 6, pp. 1145-1166.

Rodriguez, A.C., Molina, M.A., Perez, A.L.G. & Hernandez, U.M. 2003, "Size, age and activity sector on the growth of the small and medium firm size", *Small Business Economics,* vol. 21, no. 3, pp. 289-307.

Simon, H.A. & Bonini, C.P. 1958, "The Size Distribution of Business Firms", *The American Economic Review,* vol. 48, no. 4, pp. 607-617.

Stam, E. 2010, "Growth beyond Gibrat: firm growth processes and strategies", *Small Business Economics,* vol. 35, no. 2, pp. 129-135

Stokey, N. L. & Lucas, R. E. 1989, "Recursive Methods in Economic Dynamics". Harvard University Press

Sutton, J. 1997, "Gibrat's legacy", *Journal of Economic Literature,* vol. 35, no. 1, pp. 40-59.

Toivanen, O., Takalo, T., Tanayama T. 2010, "Expected benefits of innovation policy", *HECER working paper*

van Dijk, B., Hertog, R.D., Menkveld, B. & Thurik, R. 1997, "Some New Evidence on the Determinants of Large- and Small-Firm Innovation", *Small Business Economics,* vol. 9, no. 4, pp. 335-343.

Vander Vennet, R. 2001, "The law of proportionate effect and OECD bank sectors", *Applied Economics,* vol. 33, no. 4, pp. 539-546.

Wagner, J. 1992, "Firm Size, Firm Growth, and Persistence of Chance: Testing GIBRAT's Law with Establishment Data from Lower Saxony, 1978-1989", *Small Business Economics,* vol. 4, no. 2, pp. 125-131.

Weiss, C.R. 1998, "Size, growth, and survival in the upper Austrian farm sector", *Small Business Economics,* vol. 10, no. 4, pp. 305 -312.

Verbeek, M. 2008, 'A Guide to Modern Econometrics, Third Edition", *John Wiley & Sons*

Wing, C.C.K. & Yiu, M.F.K. 1996, "Firm dynamics and industrialization in the Chinese economy in transition: Implications for small..", *Journal of Business Venturing,* vol. 11, no. 6, pp. 489-505.