Methodological notes

Appendix A

Building a representative sample of EU28 VC-backed start-ups

This appendix details the creation of the representative sample of European start-ups used in this report. Our reference population contains all venture-backed start-ups in Europe, whose initial received investment took place in the period 2007-2015.[1] We further narrow our focus to the 28 Member States of the European Union.[2] This leads to a reference population, approximated through Invest Europe's data, which includes 12,277 early and later stage start-ups (see Table 1 for a definition of VC investment stages).[3]

We collected firm financial accounts, industry activity and patent data from Bureau Van Dijk's Orbis database.[4] Using the identities of invested start-ups and their headquarter locations to match the two data sources, we constructed a sample of start-ups with available performance data. In addition, we incorporated the results from a similar identification exercise carried on the sub-sample of EIF investees to enhance our sample coverage ability. Table 2 illustrates the key financial and innovation indicators used throughout the report, together with a brief description.

Table 1: Venture capital investment stages and their definitions

Seed	Funding provided before the investee company has started mass production/distribution with the aim to complete research, product definition or product design, also including market tests and creating prototypes. This funding will not be used to start mass production/distribution.
Start-up	Funding provided to companies, once the product or service is fully developed, to start mass production/distribution and to cover initial marketing. Companies may be in the process of being set up or may have been in business for a shorter time, but have not sold their product commercially yet. The destination of the capital would be mostly to cover capital expenditures and initial working capital. This stage contains also the investments reported as "Other early stage" which represents funding provided to companies that have initiated commercial manufacturing but require further funds to cover additional capital expenditures and working capital before they reach the break-even point. They will not be generating a profit yet.
Later-stage	Financing provided for an operating company, which may or may not be profitable. Late stage venture tends to be financing into companies already backed by VCs. Typically in C or D rounds.

Source: Invest Europe

[1] Start-ups with follow-on investments in this period, but with initial investment prior to 2007, are excluded from our population, hence our sample.
[2] Belgium, Bulgaria, Czechia, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Croatia, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, United Kingdom.
[3] It is important to note that Invest Europe's population data may not itself be a thorough representation of the underlying EU28 VC ecosystem. For instance, DACH investees tend to be disproportionately better represented in the Invest Europe's dataset. Nevertheless, to our knowledge Invest Europe's population remains the most reliable and accurate representation of the VC ecosystem in Europe.
[4] Orbis is an aggregator of firm-level data gathered from over 75 national and international information providers. Data is sourced from national banks, credit bureaus, business registers, statistical offices and company annual reports.

Table 2: Financial and innovation indicators collected from the Orbis database

Financial/Innovation indicator	Description
Total Assets	Total value of assets.
Number Employees	Total number of employees included in the company's payroll.
(Operating) revenue	Total operating revenues (turnover).
Intangible fixed assets	All intangible assets such as formation expenses, research expenses, goodwill, development expenses and all other expenses with a long term effect.
Cost	All costs directly and not directly related to production of the goods sold (commercial, administrative expenses etc.).
Number of Patents	Number of patent families, that is "a collection of related patent applications covering same or similar technical content".[5]

Source: authors, based on Bureau Van Dijk's Orbis database.

To render financial accounts comparable over time, we deflated all monetary values using harmonised country- and NACE Rev. 2 sector-level producer price indices (collected from Eurostat) with base year 2010. The correspondence between Invest Europe and NACE Rev. 2 classes is illustrated in Table 3.

Table 3: Sectoral classification and concordance with NACE Rev. 2 system

Macro-sector	Sector (full-name)	Nace Rev. 2 classes
ICT	**Business related software**	6201
	Communications	1810; 1811; 1812; 1813; 1820; 2630; 4652; 4742; 5800; 5810; 5811; 5813; 5814; 5819; 5820; 5821; 5829; 5910; 5911; 5912; 5913; 5914; 5920; 6000; 6010; 6020; 6110; 6120; 6190; 6391; 6399; 7310; 7311; 7312; 9512
	Computer & data services	4651; 4741; 6202; 6203; 6209; 6310; 9511
	Computer and consumer electronics	2610; 2611; 2612; 2620; 2640; 2680; 4743
	Internet technologies	6310; 6311; 6312

[5] This is the definition for patent family as described by the European Patent Office (EPO).

Macro-sector	Sector (full-name)	Nace Rev. 2 classes
Life sciences	**Biotechnology**	7210; 7211; 7219
	Healthcare	2100; 2110; 2120; 2660; 3250; 3313; 4774; 8610; 8621; 8622; 8623; 8690; 8710; 8720; 8730; 8810; 8891; 8899
Services	**Business & industrial services**	3311; 3312; 3314; 3315; 3316; 3317; 3319; 3320; 4661; 4662; 4664; 4666; 4669; 4674; 4690; 5210; 5221; 5222; 5223; 5224; 5229; 5320; 6900; 6910; 6920; 7010; 7020; 7021; 7022; 7110; 7111; 7112; 7120; 7320; 7410; 7420; 7430; 7490; 7710; 7711; 7712; 7721; 7722; 7729; 7730; 7732; 7733; 7739; 7740; 7810; 7820; 7830; 8010; 8020; 8110; 8121; 8122; 8130; 8200; 8210; 8211; 8219; 8220; 8230; 8290; 8291; 8292; 8299; 9412
	Consumer goods & retail	1011; 1013; 1020; 1032; 1039; 1041; 1051; 1052; 1061; 1070; 1071; 1073; 1082; 1083; 1084; 1085; 1086; 1089; 1092; 1101; 1102; 1105; 1107; 1300; 1310; 1320; 1330; 1390; 1391; 1392; 1395; 1396; 1399; 1410; 1413; 1419; 1431; 1439; 1511; 1512; 1520; 2219; 2341; 2342; 2349; 2369; 2751; 3102; 3109; 3212; 3213; 3220; 3230; 3240; 3299; 4631; 4632; 4633; 4634; 4636; 4637; 4638; 4639; 4640; 4641; 4642; 4643; 4644; 4645; 4646; 4647; 4648; 4649; 4711; 4719; 4721; 4722; 4723; 4724; 4725; 4729; 4751; 4753; 4754; 4759; 4761; 4764; 4765; 4771; 4772; 4775; 4776; 4777; 4778; 4779; 4781; 4782; 4791; 4799; 9522; 9529; 9600; 9601; 9603
	Consumer services: other	5510; 5520; 5530; 5590; 5610; 5621; 5629; 5630; 7220; 7900; 7910; 7911; 7912; 7990; 8412; 8510; 8520; 8531; 8532; 8542; 8552; 8553; 8559; 8560; 9001; 9002; 9003; 9004; 9200; 9311; 9312; 9313; 9319; 9321; 9329; 9499; 9600; 9602; 9604; 9609
	Financial institutions and services	4610; 4612; 4613; 4614; 4615; 4616; 4617; 4618; 4619; 6400; 6419; 6420; 6430; 6490; 6491; 6492; 6499; 6512; 6610; 6611; 6612; 6619; 6622; 6629; 6630
	Real estate	6800; 6810; 6820; 6831; 6832
	Transport	2910; 2920; 2932; 3011; 3012; 3030; 3090; 3091; 3092; 3099; 4510; 4511; 4519; 4520; 4530; 4531; 4532; 4540; 4910; 4931; 4939; 4940; 4941; 4942; 4950; 5020; 5040; 5100; 5110

The table continues on the next page.

Macro-sector	Sector (full-name)	Nace Rev. 2 classes
Manufacturing	**Business & industrial products**	1610; 1621; 1623; 1624; 1629; 1712; 1721; 1723; 1729; 2211; 2222; 2229; 2319; 2343; 2410; 2420; 2441; 2442; 2451; 2452; 2453; 2454; 2521; 2529; 2530; 2540; 2550; 2561; 2562; 2572; 2573; 2591; 2593; 2594; 2599; 2650; 2651; 2652; 2670; 2710; 2711; 2712; 2720; 2730; 2731; 2732; 2733; 2740; 2790; 2800; 2810; 2811; 2812; 2813; 2814; 2815; 2821; 2822; 2825; 2829; 2830; 2841; 2849; 2890; 2891; 2892; 2893; 2895; 2896; 2899; 3101
	Chemicals & materials	0893; 2000; 2010; 2012; 2013; 2014; 2015; 2016; 2017; 2020; 2030; 2041; 2042; 2051; 2052; 2053; 2059; 2221; 2312; 2314; 4675; 4676; 4773
	Construction	0812; 2223; 2320; 2331; 2332; 2344; 2350; 2361; 2362; 2363; 2364; 2370; 2399; 2511; 2512; 4100; 4110; 4120; 4200; 4211; 4212; 4213; 4221; 4222; 4299; 4313; 4321; 4322; 4329; 4332; 4333; 4334; 4391; 4399; 4663; 4673; 4750; 4752
Green Technologies	**Agriculture & animal production**	0111; 0113; 0126; 0130; 0147; 0149; 0160; 0161; 0162; 0163; 0164; 0210; 0321; 0322; 4622; 4623; 7500
	Energy & environment	0610; 0620; 0729; 0910; 1920; 3500; 3511; 3512; 3513; 3514; 3521; 3522; 3530; 3600; 3700; 3811; 3820; 3821; 3831; 3832; 3900; 4671; 4672; 4677; 4730

Source: Invest Europe (2016).

Identification (and exclusion) of outliers

Our initial sample covers 83% of the initial population. However, preliminary descriptive statistics show that the sample is highly heterogeneous in terms of start-up size and characteristics, beyond what is explained by the differences in investment stages. We deduce that the population (and the sample) must contain a number of outliers that, if not controlled for, are likely to distort the results of our analysis. To identify a restricted sample of companies that qualify for "true" venture capital investments, we treat the formal definitions of Table 1 as a theoretical compass. As a first step, we translate these into data-driven assumptions about the underlying companies. The following assumptions were made for early stage start-ups (at the date of the first VC investment):

E1) less than 10 years of activity,
E2) no positive turnover in the three years preceding the investment date,
E3) less than 250 employees.

The following assumptions were made for later stage ventures (at the date of the first VC investment):

L1) recorded turnover in any of the two years preceding the investment,

L2) active for at least three years and no more than 30 years.

In the absence of relevant financial data, we follow the "benefit-of-the-doubt" approach and keep the existing classification for start-ups in our sample. All firms found non-compliant with the above were discarded from our analytical sample.

According to Table 1, later stage financing tends to back "companies already backed by VCs". As a second step to our strategy to identify later stage outliers, we focus on industries where we do not observe early stage investees, hence we would not expect to find later stage start-ups. The idea is to verify that start-ups classified in the later stage bracket belong to an industry that holds a high (historical) incidence of early stage investments.[6]

In practical terms, we set up the following probit model:

$$y_i = SECTOR\alpha_i + X\beta_i + \epsilon_i$$

where y_i is a dummy variable for the company stage, SECTOR is a categorical variable for the sector, and $X\beta_i$ is a set of controls - firm's age, firm's age squared and country.

We used the model above to estimate the probability of having been an early stage venture investee conditional on the sector and firm's characteristics. We then calculated the average likelihood of each sector[7] to include early stage ventures (p_j), i.e. the average conditional probability of firms in a given sector *j*.

To identify outliers and at the same time reduce the risk of false positives (i.e. true VC investees identified as outliers), we adopt conservative criteria. For sectors with an average probability $p_j < 25\%$, we discarded companies with probability $Pr(y_{ij}) < 20\%$. These outliers had, on average, higher levels of turnover and number of employees at investment date. We are thus reassured that this approach discriminates well between VC- and private equity-backed companies, as the latter are usually larger.

The portion of our initial sample stemming from the EIF investment portfolio also included firms in the so-called "expansion stage", a combination of both later stage and "growth stage" firms.[8] Growth firms are typically more mature and hence are not of interest for our analysis of young and innovative start-ups. In order to detect growth stage firms, we first identified and excluded companies with recorded levels of turnover and employment higher than those of any other observed later stage companies.[9]

Furthermore, we constructed a new probit model including only later stage in the non-EIF sub-sample. This time our dependent variable y_i is 1 if the start-up is a later stage venture. The average conditional probability stemming from this model for each sector (p_j) was much higher than in the previous specification. Therefore, we set a higher threshold $p_j < 60\%$ to identify outlying sectors. Among these sectors, firms with a probability of being later-stage $Pr(y_i)$ lower than 60% were considered growth stage. Once again, this approach discriminates well between later stage and growth firms, which, on average, had higher number of employees, and turnover at investment date.

Overall, we identified and discarded 1,199 firms considered non-compliant with the definitions of early and/or later stage VC investees. As a result, our final sample size for the analysis consists of 8,960 companies.

[6] For this exercise, we employ an extended set of VC investments and relates investees, spanning through the years 1999-2015.
[7] Two-digit NACE code level.
[8] Growth stage investments are a type of private equity investment (often a minority investment) in relatively mature companies that are looking for primary capital to expand and improve operations or enter new markets to accelerate the growth of the business.
[9] Identified as per the procedure described above.

Weighting Procedure

Our data-intensive analyses typically force us to restrict our focus on smaller sub-sets of our sample that hold observable data. This selection process is often non-random, as we encounter large discrepancies in the degree of data usability by e.g., geography, industry, and age.

This implies that, without appropriate adjustment techniques, our results might be influenced by the biased nature of our sub-samples. To address sample representativeness issues, all our analyses employ weights to make each sample more representative of the underlying reference population. This also ensures that our results are more comparable across the report.

To generate our weights, we adopted the so-called Raking approach (Deming and Stephan, 1940). This methodology requires a number of characteristics that can highly predict the existence/absence of data, i.e. the so-called response propensity. Our implementation of the raking algorithm leverages on four key re-weighting dimensions: year of investment, country, sector and stage.

The raking algorithm starts from the unweighted sample and calculates the share of companies in each stratum (analysing one reweighting dimension at a time). It then calibrates the weights so that each sample stratum matches the respective population stratum for the given reweighting dimension, then proceeds to the next reweighting variable in the list (Battaglia et al., 2009). The algorithm iterates until further adjustments do not cause a tangible shift in the weights (Kolenikov, 2014).

Occasionally, we resorted to alternative aggregations of our key re-weighting dimensions. For example, when calculating the weights for the cluster analysis exercise, due to the very small number of observations for a few countries, we aggregated start-ups by macro-regions. This allowed us to improve the data availability in the sample's joint distribution and thus construct more robust weights.

Appendix B

Cluster analysis methods

The objective of "clustering" is to group firms in such a way that between groups, companies would differ substantially in terms of growth trends and, at the same time, they would behave rather similarly within a given group (Everitt et al., 2011). A visual inspection of the distribution of the target variable (i.e. firm growth) would typically be enough to undertake this type of task. However, firm growth is a complex phenomenon that can only be evaluated in a multi-dimensional setting (e.g. turnover growth, staff growth). Against this backdrop, cluster analysis is a convenient approach to classify observations across multiple dimensions.

We evaluate the growth of start-ups along five key dimensions of economic size: total assets (i.e., a measure of economic capital), turnover (measure of output), staff count (measure of labour), intangible assets (a proxy for innovation/productivity) and operating costs (a measure of financial expenditure and a proxy for investments). Growth is measured through the Compound Annual Growth Rate ($CAGR_n$, where n represents the time span, namely 2, 4 or 6 years). For instance, $CAGR_4$ for number of employees is the four-year growth rate of staff starting from the year of investment. We formally calculate $CAGR_n$ using the following formula:

$$CAGR_n = \left(\frac{V_{t_n}}{V_{t_0}}\right)^{\frac{1}{t_n - t_0}} - 1$$

where V_{t_0} is the initial value of the variable under study, V_{t_n} the final value and $t_n - t_0$ is the time horizon in number of years. Our reference time span is four years, and we use the 2- and 6-year time span to compare growth trends over time. As a result, we discard from our cluster analysis all start-ups first invested in the year 2015, as these companies would typically not have enough information to compute four-year growth rates.

To maximise our data coverage, we pool CAGRs by biennia, using earlier period data should the information in the exact period of interest not be available. This approach, based on a relatively mild assumption (e.g., that the three-year growth rate well approximates the four-year growth rate), significantly increases the volume of information at our disposal and reduces our over-reliance on weights to ensure sample representativeness. The exact data rules are as follows:

- If V_{t_0} was missing, we used $V_{t_{-1}}$ instead. In case $V_{t_{-1}}$ was also missing, we took V_{t_1}.
- If V_{t_n} was missing, we used $V_{t_{n-1}}$.

To aggregate companies in profiles, we used a latent class analysis model, also called finite mixture model (Skrondal and Rabe-Hesketh, 2004). This approach proposes a formal statistical model for the sampled data. Specifically, the model assumes that the underlying population is a "collection" of different sub-populations (or clusters), each characterised by its own multivariate normal distribution (i.e., the population has a finite mixture distribution). A drawback of this method, shared with other maximum likelihood strategies, is the considerable number of observations required to obtain robust parameter estimates.

On the one hand, a crucial advantage of formal statistical models is that they allow to hold constant the classification strategy, rendering it "impartial" across samples. This way, we are guaranteed that the same classification approach will hold whether we compare data for 2-, 4- or 6-year growth, or whether we compare VC-backed against non-VC-backed companies. Since data-driven clustering methods (i.e. hierarchical and optimisation clustering) do not allow to hold constant the classification model across samples, this was an important aspect in favour of latent class analysis models.

[10] In the case of $CAGR_2$, when the first year after investment was used as an initial value, it could not also be used as a final value, therefore such firms were discarded from the analysis.

On the other hand, appropriate data transformation was key to the successful application of this model. This is because the high skewness of the distributions of CAGRs and the sometimes-different ranges of variation make it impossible to observe normality in the data. As a result, without adjustments a few variables and observations would disproportionately influence the clustering process, leading to results of poor practical use. Following Signore (2016), we apply a series of data transformation and smoothing techniques to the CAGRs of our five economic size variables.

The clustering approach allows fitting the data under different assumptions about the number of latent classes. Selecting the optimal number of clusters entails the identification of the most "informative" model, i.e. the model that the data fits best. Our final choice for the number of clusters is both data-driven as well as the result of practical considerations. The Bayesian Information Criterion (BIC) indicates that the informational advantage of assuming one additional latent cluster tapers after the fifth cluster. Moreover, the additional growth profiles observed after the fifth cluster pertain to micro-clusters of modest informative power. This drives our final choice of five clusters in the data.

After fitting our final model with five latent classes, we calculated the posterior probability of cluster membership for each cluster and each firm. The posterior probabilities show a highly polarised distribution (i.e. either very high or very low). Against this backdrop, we assigned each firm to the cluster in which its growth profile was most likely to be found. Overall, we were able to classify the growth pattern of 2,160 VC-backed start-ups, invested in the period 2007-2014. Using the weighting approach discussed in Appendix A, we ensure that the aggregate results are representative of the original population under analysis.

Appendix C

Building a counterfactual sample of non-VC-backed start-ups

The counterfactual analysis of early and later stage venture investments tackles the following query: how would VC-backed start-ups perform in the absence of VC? To address this policy question, we exploit the assumptions of Rubin's Causal Model (RCM, Rubin, 1974) to generate a counterfactual group of non-VC-backed firms. If appropriately selected, these control start-ups simulate the (unobserved and unobservable) performance of VC-backed start-ups had they not received the VC investment.

Our identification strategy is largely based on the work of Pavlova and Signore (2019). We provide here a brief overview of their approach: for additional details, the reader is referred to their work. We first make two key assumptions about the data: a) that the Orbis database (our main source to identify counterfactual start-ups) contains a representative sample of EU28 firms, and that b) the sample described in Appendix A is a near-complete representation of the population of VC-backed firms in Europe.[11] These two assumptions allow separating the "treated" (VC-backed) from the "control" population (non-VC-backed).

Based on a thorough analysis of the literature, Pavlova and Signore (2019) construct a treatment assignment model that entails two sets of start-up attributes. The authors call the first set "discriminants" of VC financing, i.e. necessary (but not sufficient) conditions for a VC investment to take place. The second set, called "predictors" of VC financing, includes features that VC investors evaluate in their investment appraisal process. Attributes in this second set can be "traded-off", i.e. one or more characteristics may prevail on others during the VC financing negotiation process.

The theoretical framework above motivates an empirical approach based on a two-step matching process. Pavlova and Signore (2019) first identify appropriate control start-ups by exactly matching on the discriminants of VC financing – country, industry, investment stage, patent ownership, age at investment and degree of innovation. As a second step, the authors construct a propensity score model (Rosenbaum and Rubin, 1983) containing both discriminants and predictors of VC financing. The model's results are further used to select the appropriate counterfactual for each VC-backed start-up.

A significant challenge is brought by the Orbis database, the main source of data for this analysis, which does not cater for the specific information needs of the VC industry. Therefore, we are constrained in the choice of drivers of VC financing that we can actually observe. To offset these limitations, Pavlova and Signore (2019) bring their model to the data by introducing various measures, some original to the VC literature. To predict the degree of innovation of start-ups, the authors use a machine learning algorithm trained to recognise highly innovative business models from short trade descriptions. To measure the "accessibility" of start-ups vis-à-vis active VC firms, the authors use network theory, modelling the European VC ecosystem as a network of VC "hubs" connected by flight routes. Finally, to predict the start-up's access to financing other than VC, the authors construct a proxy for the value of home equity based on satellite imagery analysis. For additional details, the reader is referred to the related work.

Three key distinctions set apart the analysis in this report from the methodology of Pavlova and Signore (2019). The main motivation behind these is the desire to maximise our data coverage and enhance our sample representativeness power.

First, our sample also includes later stage companies, which are outside of the remit and thus excluded from the analysis in Pavlova and Signore (2019). According to the literature, there are some differences in the investment decision process between early and later stage companies. In the case of later stage start-ups, a few (initial) financial metrics

[11] That is, the (conditional) probability for a firm in the Orbis database to be backed by VC, given that it does not belong to our sample, is (approximately) zero.

can be observed, which can shape drastically the views of potential investors. For this reason, we estimate a separate matching model for later stage start-ups, which includes pre-investment financials. We find the existing level of capital and the level of current liabilities to be an important predictor of VC financing.

The second key distinction of this report is that our matching model (both for early and later stage firms) does not include human capital factors – leaving a propensity score model composed of the discriminants of VC financing as well as our "accessibility" index and our proxy for the propensity of the start-up to demand for VC. This choice, significantly advantageous in terms of data coverage, likely introduces some bias in our estimates. The reader is referred to Pavlova and Signore (2019), and in particular appendices G and H, for an analysis of the consequences of such empirical decision in terms of the magnitude of the effects. Our robustness checks indicate that the main findings are maintained (albeit with somewhat smaller average treatment effects) once we further control for the human capital characteristics of start-ups.

The third and final distinction lies in the matching strategy. Once again motivated by the goal to maximise data coverage, we implement the ridge matching estimator of Frölich (2004) to estimate the effects of VC. The ridge matching estimator generates an estimate for the counterfactual mean (i.e. the expected outcome for the treated company had it not received the treatment) that has desirable consistency and efficiency properties in finite samples. The ridge matching estimate for the counterfactual mean can be thought as a "weighted" average of control outcomes. The weight is a function of the distance between the propensity score of the control company and the reference treated propensity score, taking into account the features of the propensity score distribution.

Table 4 provides the list of variables included in our matching model (main effects only, not accounting for interactions and/or higher order effect) complemented by a series of descriptive statistics and the balancing power of our matching method. The second and third column of Table 4 display the matched sample averages of the two evaluated groups. The fourth column displays the P-value of the means test between the groups. Column five displays the percentage bias, i.e. the two samples mean difference as a percentage of the average standard deviation in the treated and non-treated groups.[12] Lastly, column six displays the variance ratio of treated over non-treated. This ratio should equal to one if there is perfect balance. Variables whose post-matching variance ratio exceeds the 2.5th and 97.5th percentiles of the F-distribution are marked with an asterisk in Table 4.

It is worth noting that, similarly to Pavlova and Signore (2019), the ridge matching estimator is constructed separately for each outcome variable. This approach allows flexibility vis-à-vis potential differences in missing patterns across outcome variables, once again benefitting data coverage and representativeness. We evaluate a total of 76,837 candidate control companies in our matching model (both early and later stage). After the matching process, we retain 42,756 control candidates, which are then used to create counterfactual means for 4,039 treated firms.[13]

To carry out our causal analysis of VC on growth patterns, we used the counterfactual means to compute growth rates (and related growth clusters), i.e. comparing counterfactual means across different post-investment periods. Since we are constructing growth rates based on pooled (weighted) counterfactual outcomes, regression to the mean could be an indirect source of bias for this particular exercise. Against this backdrop, more "extreme" results for VC-backed start-ups, i.e. significant under- or out-performance, might be driven to some extent by this phenomenon.

Due to the stringent data requirements (i.e. all financial indicators used for our cluster analysis should be available and the treated companies should be matched), the final sample for this analysis consists of 831 VC-backed start-up and associated counterfactual means. Using the weighting approach discussed in Appendix A, we ensure that the aggregate results are representative of the original population under analysis.[14]

[12] According to the literature, the matching method is considered effective in balancing the distribution of the covariate if it achieves an absolute bias of 5% or below.
[13] This figure pertains to the sample size with available assets data. For other financial figures, sample sizes are typically half this size or lower.
[14] However, because of the significantly reduced sample size vis-à-vis the cluster analysis discussed in Appendix B, it was not possible to obtains perfectly overlapping medians and averages for the matched treated sample.

Table 4: Descriptive statistics of PSM model and balancing checks

Variables	Average		P-value	Percentage bias	V(T)/V(C)
	Treated	Control			
Innovativeness score¤	0.48	0.47	0.17	2.8	0.99
Company accessibility score	0.48	0.47	0.74	0.7	0.97
Company age at inv. Year ¤	2.06	2.03	0.73	0.7	1
Distance from FUA's centroid*	6.68	8.52	0.00	-8.5	0.48
Undevelopable land	0.10	0.10	0.61	-1.1	0.95
Distance from FUA's airport centroid*	36.51	36.97	0.51	-1.3	1.06
Patent at investment year:					
No patent at inv. year ¤	0.74	0.76	0.03	-5.2	n.a.
Has a patent at inv. year ¤	0.26	0.24	0.03	5.2	n.a.
Investment Year:					
2007 ¤	0.18	0.18	0.62	1.1	n.a.
2008¤	0.18	0.18	0.85	-0.4	n.a.
2009 ¤	0.11	0.12	0.46	-1.6	n.a.
2010 ¤	0.10	0.10	0.74	-0.7	n.a.
2011 ¤	0.10	0.10	0.93	0.2	n.a.
2012 ¤	0.10	0.10	0.72	-0.8	n.a.
2013 ¤	0.10	0.11	0.61	-1.1	n.a.
2014 ¤	0.13	0.12	0.16	2.8	n.a.
Macro-sector:					
ICT ¤	0.29	0.3	0.75	-0.7	n.a.
Life Sciences ¤	0.18	0.18	0.85	0.4	n.a.
Manufacturing ¤	0.15	0.15	0.67	0.9	n.a.
Services ¤	0.32	0.33	0.89	-0.3	n.a.
Green Technologies ¤	0.02	0.02	0.98	-0.1	n.a.
Other ¤	0.03	0.03	0.95	-0.1	n.a.
Investment stage:					
Seed ¤	0.71	0.71	0.93	0.2	n.a.
Start-up ¤	0.29	0.29	0.93	-0.2	n.a.
Later stage ¤	0.21	0.22	0.8	-0.5	n.a.
Macro-region:					
DACH ¤	0.29	0.3	0.75	-0.7	n.a.
FR&Benelux ¤	0.18	0.18	0.84	0.4	n.a.
Nordics ¤	0.15	0.15	0.67	0.9	n.a.
Mediterranean ¤	0.32	0.33	0.89	-0.3	n.a.
UK&Ireland ¤	0.02	0.02	0.98	-0.1	n.a.
CEE ¤	0.03	0.03	0.95	-0.1	n.a.

Note: our final matched samples are specific for each outcome variable, with results above pertaining to total assets. Results for other outcomes are qualitatively equivalent. ¤ Exactly matched.

As pointed out in Appendix B, a series of data transformations is rendered necessary in the cluster analysis to ensure normality of the data. One of these transformations is standardisation, i.e. rescaling the data so that their mean is null and their standard deviation is one. To this end, we separately standardise the treatment and control group data. The sub-sample of treated start-ups in the counterfactual analysis is standardised according to the entire cluster analysis distribution, i.e. the 2,160 companies analysed in the second chapter. This ensures that the exact same categorisation of growth rates is maintained for treated firms.

The control group is standardised according to its own distribution. In practical terms, the outcome of using two different distribution on which to rescale the data means that companies will be clustered according to their relative performance in their reference group. This implies that a treatment and a control firm with the same underlying growth rates might fit in two different clusters, due to their different performance relative to the rest of treated and control firms respectively. Start-ups in a given cluster will nevertheless show the same characteristic behaviour, i.e. an overall positive growth with disproportionate intangibles growth for visionaries in both groups. We considered this approach superior to the alternative of standardising both groups according to a common distribution, which would have led to an overwhelming majority of control start-ups being captured by the commoners' group, simply due to the lower intensity of their growth.