Putting the CIS2018 into context

Robbin te Velde
José Cervera
Pim den Hertog

December 2017

devstat

HIGHER SCHOOL OF ECONOMICS

dialogic
innovation • interaction

Project reference:
ESTAT/G/2015/006

Compiled by:
Dialogic (Utrecht, the Netherlands)
Higher School of Economics (Moscow, Russia)
DevStat (València, Spain)

Table of contents

1 Concepts for measuring and understanding innovation

1.1 Introduction

1.1.1 Purpose of the CIS2018 User Manual

This user manual is a supporting document to the Community Innovation Survey (hereafter: CIS). The focus of the document is on the actual use of the CIS data and less so on the implementation of the CIS survey. This is covered in other documentation. This manual particularly deals with many aspect of the analysis of CIS survey results. The primary aim is to show how CIS data can be used, or is already being used, in the practice of innovation research and policy analysis.

The manual follows the structure of CIS in broad lines. In chapters one and two the key concepts of CIS are discussed, respectively innovation and business innovation. Chapter three describes the questions of CIS2018 each and every one. Chapter four deals with the possibilities of combining CIS data with other data sources (i.e., with data linkage). Chapter five elaborates two important subsequent analyses that can be applied to CIS data, namely the calculation of innovation intensity and the profiling of specific business enterprises (i.e., innovation styles). Chapter six covers another important topic, the international dimension of innovation.

1.1.2 The users of CIS

The Community Innovation Survey has many users. The four main types of users are academics, analysts, managers and policy makers. Policy makers and managers are the final users, analysts (research consultants) are intermediaries, and academics are both intermediaries and final users. In practice, there is often an overlap in roles. Nevertheless, each of the roles has specific needs and it are these user needs that drive the construction of a system for measuring and reporting innovation and the subsequent production of innovation data, statistics, indicators, and in-depth analyses of innovative activities.

Ultimately it is the needs of the final users, those of the policy makers and the managers, that should drive the design of the CIS survey and the subsequent processing of the data. Although academics are also final users in their own right eventually their work is also meant to improve the understanding of the aforementioned policy makers and managers of innovation. An important observation is that neither policy makers nor manages are interested in innovation per se. They are primarily interested in the *outcome* of innovation, that is, in its effect on economic development, organisational change, and social transformation.

A marked difference between these two types of users is that policy makers mainly work at the meso and macro level ('the public interest'), and managers mainly at the micro level (their own organisation or benchmark organisations). The (re)use of CIS-data by managers is severely restricted by the fact that micro data cannot be disclosed by entities that collected the information on a confidential basis. This is an additional disadvantage because it is

usually these (innovation) managers who fill out the CIS survey. It is therefore not possible to provide them with tailor-made feedback on the data they have entered in return.[1]

This leaves policy makers as the core target for innovation data. Public policy interest in innovation is indeed extensively reflected in the literature (OECD, 2015). Since innovation is usually not a policy goal in itself, it is important to identify the primary policy goals to ensure that the data that is being collected matches policy needs. So far CIS data has been mainly used in the policy areas of industry and economic affairs. This is probably due to the fact that putting the data to use outside the core policy areas requires the linking of CIS data with existing data sources that cover the target policy areas. Linking CIS data with other data sources (especially administrative data, see chapter 4.1) has a lot of potential but it does require additional investments in existing statistical infrastructures.

1.1.3 Scope and definitions of CIS

As much as user needs drive the collection and analyses of innovation data the opposite also holds. The scope and the method of the data collection and the definition of the core concepts to a large extend determines the actual (re)use of the data. For example, as already been described concerns about privacy of micro data limit the use of CIS data by managers. Somewhat differently, the lack of investments in data linkages from CIS data with other data sources limits the use outside the conventional realm of industrial policy. This is a typical chicken-and-egg situation: there is little demand for data because there is little supply of data.

The actual scope of CIS is also very much determined by the definition of the core concepts. The devil is in the detail here. Seemingly minor changes to the definitions could already give rise to substantial shifts in the scope of the use of the data. Ever since its conception in 1992 CIS has seen quite some changes in this respect. Academic researchers were the key drivers for the first efforts to measure innovation. This has has a strong influence on the first edition of CIS and the underlying Oslo Manual (Arundel & Smith, 2013). Academics have a strong interest in research that can provide predictive and causal interpretations of the effects of innovation, which requires linking innovation data to longitudinal data for variables such as value-added, employment, productivity and user/stakeholder satisfaction, as well as obtaining innovation data through longitudinal surveys. Up until fairly recently most of the linkage was done on the national or sectoral level. Thus the lack of access to micro data was less of a hindrance to academic researchers than to innovation managers. Likewise, the use of *anonymised* micro data is perfectly fit for academic research, but less so for strategic or tactical use in the private sector (e.g., benchmarking with competitors or a precisely defined niche).

[1] However, some national statistical organisations (e.g., Spain and Sweden) give micro-aggregated, *non-confidential* data as an incentive for respondents to provide data, such as comparison of provided data with ratios etceteras.

1.2 The object versus the subjective-based approach

1.2.1 Historical overview

The substantial influence of academic users on the shaping of CIS, in combination with the apparent demand from industry policy makers, has also been decisive in the eventual selection of the unit of analysis in CIS, that is, the choice of the *subject-based* rather than the *object-based* approach. Up until the second revision of the Oslo Manual (OM2), innovations were still defined in terms of 'new products or processes', hence as objects. This fits well with the reference frame of the managers that fill in the survey because they are thinking in terms of (and managing the development of) concrete individual products and services. However although the object approach was very helpful for respondents in order to focus their attention on a few material objects – rather than a vague concept of innovation, it soon become (too) difficult to measure. This is because although most items in CIS referred to all innovations, respondents usually only reported activities aimed at the listed innovations. This rendered data collected on innovations very difficult to be processed and used because of their heterogeneity and technical contents. The latter is a major disadvantage for the use of CIS data in econometric methods that have widely used in the field in innovation sciences (but less so for the more qualitative approaches that are widely used in business and management sciences). As for policy makers, firms (as the 'bearers of innovation') are much more suitable as targets for policies than innovations (as objects) per se.

In the third revision of the Oslo Manual in 2005 (OM3) the subject-based approach was adopted. From CIS2006 on, the firm became the unit of analysis and consequently, the statistical unit of measure. At the same time, the core concept of 'innovation' was defined as a 'process' rather than an outcome. Thus, what CIS actually measures is *the innovation as a process within a firm*. This process encompasses every dimension of an invention, from generation (initiation) to diffusion. In OM2, innovation was still defined as *an object (a new product or service) that is implemented by a firm*. Innovation was explicitly *not* defined as a process. This is mainly to avoid foreseen measurement problems. The innovation process is thought to be more or less continuous, with contant incremental modifications to product rather than distinct events. This makes it difficult to identify these modifications. In OM3 this measurement issue is downplayed by assuming that innovations are in most cases 'events', i.e., the implementation of an object (or method) that has certain novel characteristics for the firm.

The redefinition of the core concept of 'innovation' still remains unclear. In OM2 innovation is not regarded as a process but at the same time the continuous nature of the generation of innovations (as outcomes) is being stressed. In OM3, innovation is defined as a process but is described as a discrete event. In essense, though, innovation has elements of both definitions: it is a continuous process. The confusing is based on an ill-understood concept at a deeper conceptual level, namely the notion of *equilibrium*.

1.2.2 Change as a constant

Equilibrium refers to two different notions. The first one is the conventional conception in economic theory where in the short run exogenous changes might move the system away from the attractor state but where *in the long run the system always returns to its normal condition*. In other words, the statis is primary and the dynamics are only secundary. In the

second notion, *the processes underpinning the continuity of the system as a whole* may be conceived as equilibrating processes. In other words, the dynamics of these processes are primary and the – theoretical – state of rest is secundary (Denis, 2007). 'Theoretical' because the equilibrium towards which the processes are moving is never attained as other processes always intervene. In fact, if the state of rest would ever be reached the system would cease to exist[2].

The definition of innovation as a process in OM3 is still based an a *static* conception of equilibrium. The innovation (as a discrete event) embodies the external shock that temporary moves the economic or social system away from its normal condition. In a dynamic conception of equilibrium, the *constant* introduction of novelty is an essential feature of an economic or social system. Without innovation the system as a whole would sooner of later stop functioning. However, at all levels of the system innovation can only exist by the grace of the simultaneous existence of non-innovation. Thus, at organisational (firm) level innovative activities exist by the grace of non-innovative activities, at industry level innovative firms exist by the grace of non-innovative firms, and at the society level innovative industries exist by the grace of non-innovative industries. The ultimate goal of a manager (micro level) or policy maker (meso and macro level) is then not to promote innovation per se but to maintain the dynamic equilibrium, that is, the right balance between innovation and non-innovation processes. This topic is being covered by CIS2018 in the question on 'innovation activity' (#3.9). This does not only include completed activities (hence 'implemented' innovations, see next chapter) but also activities that are (or were) *intended* to result in an innovation. These are activities that are either still *ongoing* or that have been *abandoned*. The dynamic equilibrium these assumes that these three types of innovation activities (plus a fourth one: *no* innovation activity) should always be in some mind of dynamic balance (i.e., the share of completed activities should neither be too low nor too high).

Text box 1. Question on innovation activities in CIS2018 (CIS 2018 Task Force, 2017)

´**Innovation activity**´ included all developmental, financial and commercial activities, undertaken by a firm, which are intended to or result in an innovatoon.

Research and Development (R&D) comprise creative and systemtic work undertaken in order to increase the stock of knowledge – including knowledge of humankind, culture and society – and to devise new applications of available knowledge.

3.9 DURING THE THREE YEARS 2016 TO 2018, DID YOUR ENTERPRISE HAVE ANY

	Yes	No
Completed activities on product/process innovation*	☐	☐
Ongoing innovation activities at the end of 2018	☐	☐
Abandoned innovation activities	☐	☐
Research and development (R&D) activity** ?	☐	☐

* For all enterprises that reply 'yes' to any category in any of the questions 3.1 or 3.6, the answer to question this category must be preset to 'yes'.

** Internal or external R&D activities leading to expenditure. Please see the annex for definitions of internal and external R&D.

[2] That is, to use a biological analogy, without constant renewal the organism would die.

1.2.3 Change, novelty and improvement

If the constant introduction of novelty is indeed an essential feature of any economic or social system then novelty (the essence of in-*nova*-re) cannot be used as a distinguishing factor. This problem has plagued CIS from the onset. For instance, in CIS2006 the shift in OM3 from object to subject-based approach was adopted but objects were still needed to separate 'innovative firms' from 'non-innovative firms'. The latter are defined as firms which have not implementated any innovation during the observation period. This shifts the burden back again to the definition of innovation as an object. In OM3 an innovation is defined as '[a] product or service that is new to the firm'. However either any change to a product or service (or method for that matter) could be regarded as 'new' or no single change qualifies (for all changes built on earlier changes – genuine novelty does not exist; only '*improvements*'). In OM3 therefore the formulation 'new *or* improved' has been choosen. Obviously, this does not solve the issue because now all changes are still included – and all firms should essentially be regarded as 'innovative firms'. Changing the formulation into 'new *and* improved' does not solve the issue either. Strictly speaking entities cannot be (genuinely) new and improved at the same time, hence the intersection is empty and no single firm would qualify.

In practice, a continuum exists, with different (and dynamic) balances between changes and non-changes. Hence firms are 'more' or 'less' innovative, and improvements are on a continuous scale from 'minor' to 'major'. This moves the measurement of innovation *intensity* (i.e. innovativeness on a continuum) to the foreground (see chapter 5.2). Any attempt to use a dichotomous scale or to introduce a certain treshold level is bound to fail. For instance, earlier versions of OM, the formulation '*significant* improvements' has been used to define a treshold level for innovative firms. However since no clear criteria has been given how to define 'significant' (or 'substantial') the term is rather useless for statistical measurement purposes.

Attempts to place the demarcation outside the firm seem to be more fruitful. First, if 'new to the firm' does not exclude any situation, '*new to the market*' (the next level) might do the trick. Thus, innovations are only regarded as proper innovations if they are 'new to the market' (and not just 'new to the firm'). This does indeed introduce a clear demarcation between 'innovative' and 'non-innovative' firms – provided that a uniform definition of 'market' exists. The underlying assumption is that firms are sufficiently informed about the activities of the few firms which are in direct competition with it. At least, the notion is much clearer than the term 'significant improvement'. However, because the lowest level of 'new to the firm' has policy relevance in its own right it has been decided to maintain this level. From a dynamic point of view this makes sense. For the innovation system as a whole to continue functioning, firms from all sorts of innovation intensity should be present. To cover all firms in CIS, the filtering question on 'non-innovative firms' has therefore been dropped. In subsequent analyses the criterion of 'new to the market' could still be used to distinguish '(more) innovative' from 'non (or rather: less) innovative' firms.

A second attempt is to place the demarcation of innovation further down on the product development process. Now, innovations are only regarded as proper innovations if they are actually *implemented*. Again, this assumes that a unambiguous definition of 'implementation' exists. The problem is that there are different concepts for different types of innovation (e.g., for product innovations that would be: 'introduced on the market', for organisational innovations, 'brought into actual use') and that the individual concepts are still not clear (e.g.,

when is a product really introduced on the market: when it is available at ordering or only when it is actually being sold?).

To arrive at a universal and clearer definition, the concept of 'value creation' could be useful. The underlying idea is that innovations will only generate value once they are properly implemented. The broader notion of 'value creation' (versus the narrow accounting concept of 'generating income') applies to all types of innovations, thus not only to conventional products and processes that are sold on a market. The concept of 'value creation' is also more in line again with the reference framework from managers. It is even regarded as the main objective of a firm and as a key objective of business innovation (Kraaijenbrink, 2015).

Moreover, it could be regarded as the main objective of *any* organization, including public sector organisations. Hence it would extend the scope of CIS, which is now solely on innovation on business enterprises (see chapter 2), to the government sector as well. Having said this, in the absence of generic financial measures (i.e., market prices), in the public sector high quality outcome measures have to be tailor-made, that is they are generally only available for specific innovations (OECD, 2015).

1.3 Innovative activities versus other business activities

Summing up the previous discussions on the definition of 'innovation' (and the derived definition from 'innovative firms') we could say that the key elements are (1) the creation of *value* based on (2) *improvements* to products, processes or practices that already exist in an organisation. Stated differently, the essence of innovation is that of generating value from novelty.

From a dynamic point of view the first requirement ('implementation') would actually signify the *end* of an innovation process – yet the process can never end and one innovation outcome will always build on another. If 'creating value' is the essence of innovation it begs the question what sets innovation apart from 'business as usual' – since *all* activities from a firm (or a government organisation alike) are supposedly aimed as creating value.

The second requirement of 'improvement' seems to have the same limitation. If improvements ('significant positive changes') are defined based on the value that they create, *any* change that creates value is included. Again, if products, processes or practices do not create value they will simply not be implemented, regardless whether they are 'novel' or not.

Defined in this way, an outcome-based measure of innovation then essentially measures *overall* business performance. However the aim of innovation research is to study (the functioning of) innovation processes and their (assumed) contribution to value creation and business performance. As such these processes should be isolated from overall business (or rather: organisational) performance and should thus *not* be defined in terms of 'value created' but rather still in (enlightened) terms of 'novelty'.

2 Concepts and definitions of business innovation for measurement

2.1 Total value created from improvements

2.1.1 Improvements and their contribution to turnover

In the previous chapter the evolution of CIS has been described. The CIS2018 only applied to firms and not to any innovation outside the private sector. The remainer of this manual will therefore focus on business innovation. This permits us to describe the innovation process into more detail. As argued in the previous chapter, one critical issue is how to distinguish the innovation process from other business processes.

The key stage of innovation is the stage in the innovation process when an *improved* product, process, marketing or organisation method starts to generate value for the innovating firm. The challenge is to distinguish 'improved' products, process, and methods from 'non-improved' ones. OM4 rejects the interpretation that any change is an innovation. Instead, only 'significant positive changes' qualify. However, as argued in the previous chapter, without an underlying scale (which is lacking) the term 'significant' is rather useless.

Text box 2. Evolution has no goal

> From an evolutionary point of view there are no such thing as 'good' (positive) or 'bad' (negative) changes. In specific circumstances changes make a species better (or less) fit to its environment. But even if a species would become distinct it has not really disappeared but has rather transformed inself into related forms that are better adapted to their environments. Overall, then,evolution is not necessarily progressive (Dawkins, 1986).

The pragmatic solution is to use a continuous (i.e. innovation *intensity*) rather than a dichotomous scale (i.e. *innovative/non-innovative*). Theoretically, the net impact (i.e., the final value added by the improvement) is then the size of the gross value added *V* times the share of improvement (the innovation intensity I). Because every product, service or method that is (being) implemented will have at least some traces of improvement, the measurement should be at the level of individual products, services or methods.

However, in the subject-based approach that was adopted the measurement is at the aggregate level of the firm. Ideally, the firm should then list all its products, services, and methods and then estimate the gross added value and the innovation intensity for each of these individual entities (the bottom-up approach). Hence for a firm the total value added from innovation would be $\sum_k^n I_k \cdot V_k$. This is obviously a very laborious exercise. A pragmatic approach is to introduce a two-step process: first identity the 'most improved' products or processes and only for these innovation provide individual estimates for I and V. This is basically the approach that has been followed in OM2. This is still only a partial coverage of total nett value as it does not include the residual nett added value from all 'less improved' products or processes.

In CIS2018 all questions on individual entities have been dropped (CIS 2018 Task Force, 2017). It still used a two-step process but top down rather than bottom up to derive total

value created from improved products. Thus it first asks for total turnover and then for the share of 'new or improved' products IP. Thus the share of unchanged ('less improved) products is 1-IP. Total turnover from improved products equals (IP x V(IP) + (1-IP)xV(1-IP).

Text box 3. Question on the total turnover derived from new or improved products in CIS2018 (CIS 2018 Task Force, 2017)

3.3 PLEASE ESTIMATE THE PERCENTAGE OF YOUR ENTERPRISE'S TOTAL TURNOVER[3] IN 2018 FROM PRODUCTS (GOODS AND SERVICES) THAT WERE, IN THE THREE YEARS 2016 TO 2018:[4]

New or improved products		Unchanged products (or with only minor changes)*		*Total turnover in 2018*
_ _ %	+	_ _ %	=	100 %

If possible, separate turnover from new or improved products between products:	
	=
Not previously offered by any of your competitors[6]	_ _ %
	+
Identical or very similar to products already offered by your competitors	_ _ %

* Includes the resale of new products purchased from other enterprises.

[3] Turnover is defined as the market sales of goods and services (Include all taxes except VAT). For Credit institutions: Interests receivable and similar income, for insurance services: Gross premiums written.

[4] This question can be designed according to national needs provided it delivers the described information, in particular the percentages for 'new or improved products' and 'unchanged products'

We have no information on innovation intensity I but since the underlying assumption is that $I_{IP} = 1$ and $I_{1-IP} = 0$ total turnover simply equals V_{IP}. This is not a realistic assumption. Genuinely new (hence 100% improved) products are very rare or even only a theoretical possibility. The same goes for products that have not been unchanged for three years (the period of observation in CIS). In practice, then, $I_{IP}<1$ and $I_{1-IP}>0$. The nett result from a decrease of (I_{IP} x V_{IP}) and an increase in (I_{1-IP} x V_{1-IP}) might not be much different for the *overall* total value created from improved products but this depends on the underlying distribution of I_k and V_k, which are both unknown. In any case, in a subject-based approach it would be better to use the formulation 'due to improvements in general' (which covers all entities) rather than 'due to improved products' (which still leaves out the residual).

2.1.2 On value

The choice for the subject-based approach also affects the scope of the second core element, *value*. Business innovation becomes effective when firms are able to improve the way they are working, and also serving the market, at the point to get an additional value compared to their competitors. As a key function of firms innovation can be assumed to have a rational economic aim. That is, 'improvements' in products, processes or practices will not be implemented by a firm if they do not have a proper return on investment. However, the concept of value is very broad. It encompasses *any* appreciable value of use, not just added functionality and performance but also cultural, symbolic and emotional satisfaction. Moreover, this value could be generated either as a direct effect (i.e., increased turnover from the sales of improved products) but also indirectly (i.e., increased efficiency of business processes).

Table 1. Typology of value creation by improvements

	Functionality and performance	*Cultural, symbolic and emotional satisfaction*
Direct (product innovations) *Indirect* (business process innovations)	Products & services Business processes	Design Image & reputation

In the subject-based approach the firm is the unit of analysis hence it is the firm that reports on the value created *for itself*. Only the value creation for the innovating firm can be identified by survey respondents, i.e., the innovating firm. Although the broad notion of value does not exclude the possibility of spill-overs (i.e., situations in which value partly accrues to other economic or social actors than the original innovating firm) these are *not* reported. To the contrary, in an object-based approach the entire course of an innovation through society could be followed. Much akin social cost-benefit analyses such a description would also include transfers of values between different groups of stakeholders (e.g., an increase in value for one group might result in a destruction of value for another group) (Sartori, et al., 2015).

Indirect effects are also included but only insofar they are linked to direct effects. That is, these improvements should eventually in one way or another lead to an increase in sales volume and/or margins. Innovations in (internal) *business processes* could for instance result in decreasing unit costs of products or deliver and/or increases in the quality or delivery of goods. Improvements in the *reputation* of a firm could increase the perceived quality of the design and of products and services and thus also boost the sales (or prices) of goods.

Such indirect effects can always be identified via direct effects. This is because *in practice it is not possible to improve a product, service or design without improving the underlying business processes*. Theoretically, it is possible to generate novelty 'de novo'. In genetics, for instance, mutation occurs do to random errors in reproduction processes. In some cases, these mutations lead to structural adaptations. However, innovations are supposed to be the wilful outcomes of innovation processes – they are not random products. Secondly, improvements are only regarded as innovations if they are structural in nature – changes must be sustained. This still leaves room for the iconic serendipity but only if this results in products or processes that are being sold an the market. A second, less theoretical, option is create a new product or service by recombining existing (unchanged) components. For example, in

organic chemistry de novo synthesis refers to the synthesis of complex molecules from simple ones. This comes close to the classical Schumpeterian definition of '*Neue Kombinationen'* (Schumpeter, 1934).

2.2 Distinguishing innovation activities from other business activities

The topic of 'new combinations' brings us to the heart of the matter. Could such a re-combination of existing (unchanged) components be regarded as an innovation or not? From a reductionistic point of view there is no novelty involved, hence the simple pooling of existing components will not constitute an innovation. From a systems point of view, the whole can be greater than its parts. In fact, the essence of the innovation *is* the combination of the various existing elements.

Because it is impossible to arrive at an improved product or service without re-organising the existing underlying business processes the theoretical case that a new combination consists solely of existing (intermediate) products and/or business processes is empirically void. Any new combination is at least built on improved business processes. The other way around, any innovation involves new combinations because products or services that are entirely new (which, from an evolutionary point of view is also empirically void) require the establishment of new underlying (business) processes or in any case the embedment into existing business processes. It is not just that a single innovation *can* involve combinations of different types of product and business process innovations, it *will* always involve combinations of (less or more) improved products and business processes. Innovation could thus be defined as a (radical) change in product-market-technology combinations (Buijs, 1987).

For Jack Morton, truly a scholar of innovation practice, *coupling* is the key word (Godin, 2017): "a system is an integrated assembly of specialized parts acting together for a common purpose […] each is dependent for its system effectiveness upon its coupling to the system's other parts and the external world." (Morton, 1971). The 'innovation system' is a subsystem of the firm (as a system), and the firm in turn is a subsystem of an economic system. The 'common purpose' in the innovation subsystem is not to to promote innovation per se but to maintain the dynamic equilibrium, that is, the right balance between innovation and non-innovation processes (see previous chapter). This puts the *management* of innovations at the center. The central question in innovation theory would then be how is it that certain systems (firms) seem to be better able to cope with constant changes (i.e., to successfully couple with the external world) than other systems, and subsequently how these firms have arranged their innovation system, and how they are running it.

In some extreme cases, changes in the coupling with the outside world might force the firm to re-invent itself, that is, to re-organise its system as a whole. This refers to new or redesigned ('improved') *business models*. These business models are the rationale of how a firm creates, delivers and captures value. Changes in its business model, for a firm, is a much more substantial move than just innovating in products, marketing or managerial methods. One might expect to include innovations in business models as well in in the typology of innovations. From a conceptual point of view, the business model innovation refers to the system as a whole, and the other three types of innovation mainly to the innovation subsystem. However in practice the distinction will be difficult to make. In many firms the innovation subsystem will not be formalized and clearly defined. As a subsystem, it is (or should be) strongly interwoven with the whole system which makes it in any case difficult to separate.

This raises the question how to distinghuish the innovation subsystem from the overall firm system. In the absence of a clear empirical demarcation it makes little sense to ask respondents *directly* about the functioning of the innovation subsystem. Instead the survey questions should be on factual processes and results, and from this information theoretical

constructs (e.g., various types of 'innovative firms' or 'innovation styles'; see chapter 5.3 on Enterprise profiling) could be derived. The focus of the survey should be on the coupling of various internal and external innovative and non-innovative components and processes, which is the heart of innovation management (te Velde, 2004). This includes the *combination* of existing products and business processes (managerial and marketing practices) which could be an innovation in its own right.

3 CIS Variables and questions

3.1 Content and structure of CIS

3.1.1 A conceptual model of innovation

Within the definition of CIS, innovation can only happen as a deliberate act of implementation (see chapter 2). Such implementation can refer either to the activities needed to bring new or improved products and processes to the market or to a planned effort for improving internal production processes or the whole organisation of the innovating interprise. The innovation implementation is the evidence of an innovation action (see chapter 2). Thus, in contrast to R&D-statistics, where the focus is on the *input* of the process (R&D expenditure, human resources)[3] in CIS the focus is on the *output* of the innovation process (i.e., the innovation itself).

From the perspective of enterprises innovation only matters insofar it enables them to achieve longer term *outcome* objectives (i.e., to survive, to stay competitive, to be profitable, to increase sales). Firms – as for profit organisations – only invest in innovation because they assume that it boosts (or at least preserves) their long run profitability. The pivotal importance of innovation is based on the assumption that it is central to the growth of output and productivity.

However, the relationship between investments in innovation (input) and profitability (outcome) is only an indirect one. In an innovation process, inputs are first being transformed into new or improved products and services, processes or organisational and marketing innovations. This is the innovation output. It is assumed that only as a further step to the innovation implementation, these outputs will be transformed into inputs of the standard business process that uses the new processes/practices to increase productivity and the new products to increase sales and profitability. This is the outcome. Innovation requires both development of firm resources required to innovate, and the ability to profit from those innovations.

[3] OECD (2015). Frascati Manual 2015. Guidelines for Collecting and Reporting Data on Research and Experimental Development. Paris: OECD.

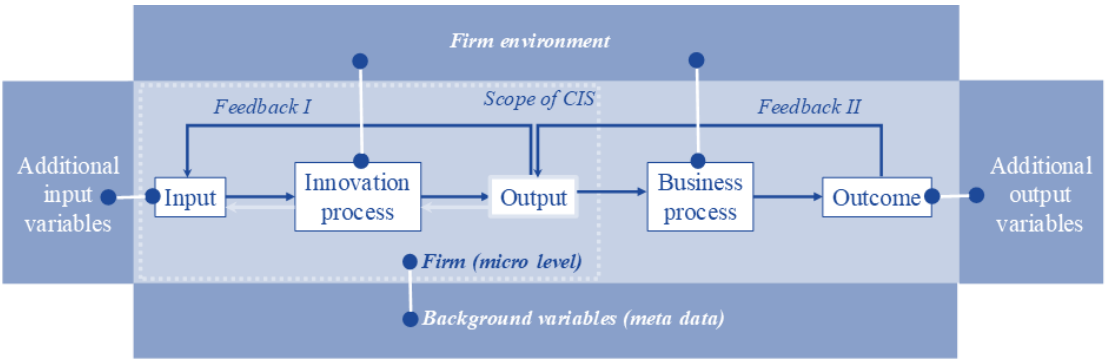


Figure 1. A basic conceptual model to describe the relationships between innovation input, output, and outcome[4]

In the CIS framework, both input and internal processes are defined in terms of innovation (e.g., expenditure *used for innovation*, see Text box 1). It is the quality of the innovation process that determines how and to what extend inputs to the innovation process (e.g., expenditure, human capital, firm-specific knowledge) are being transformed into innovation outputs. In turn, the quality of the innovation process is influenced by several internal (e.g., quality of innovation management) and external factors (e.g., access to sources of information) that assist or hamper innovation.

The innovation process itself is strongly affected by the dynamics and outcome of the business process (the right-hand of Figure 1). The propensity to innovate is linked to the 'propensity to make money'. The basic assumption is that firms who perform relatively well are relatively successful in innovating. Innovation indicators should therefore not only cover the propensity to innovate and innovation intensity (see hereafter, chapter 5.2) but also the ability for firms to achieve (or fail to achieve) innovation *outcomes* of various sorts – such as improved business performance.

3.1.2 Breakdown of CIS2018

CIS208 consists of four modules. The focus of CIS2018 is on the lefthand of Figure 1: the input, process, and output of innovation. Some background variables and items on the firm environment are also included. From the righthand of the figure only one item on business outcome is covered, namely total turnover. The element of business processes is not covered at all. Hence the apparent added value of linking CIS data with external data sources on firm environment and especially on business processes and business outcome (see hereafter, chapter 4).

[4] Dialogic (2016). *Improving the measurement of innovation outcome*. Paper presented at the ESTAT back to back meeting @ Ghent Blue Sky, 22 September 2016.

Text box 4. Overview of CIS questions per module, by type of element

- Module 1 ('Enterprise identification') covers one important
  - *background* variable:
    - the legal structure of the enterprise (#1.1)
- Module 2 ('Strategies and Knowledge Flows') covers two element:
  - *firm environment*
    - overall firm strategy (#2.1, #2.2)
  - *innovation process*, one specific element: knowledge flows:
    - co-creation (#2.3, #2.4)
    - intellectual property rights (#2.5, #2.6, #2.7)
    - external knowledge flows (#2.8, #2.9, #2.10)
    - internal knowledge flows (#2.11)
- Module 3 ('Innovation') is the core of the survey and covers various types of elements:
  - *Innovation input*
    - expenditure (#3.10, #3.11)
    - external funding (#3.12, #3.13)
  - *Innovation process*
    - co-operation (#3.4, #3.7, #3.14, #3.15)
    - status of innovation activities (#3.9)
  - *Innovation output*
    - New or improved produts (#3.1, #3.2, #3.3, #3.5, #3.6, #3.8)
  - *Firm environment*
    - regulation (#3.16),
    - hampering factors (#3.17)
- Module 4 ('Basic information on the enterprise') also covers various types of elements:
  - *Innovation input*
    - employee educational level (#4.2)
    - expenditure (#4.6)
    - internal funding within enterprise groups (#4.9)
  - *Innovation process*
    - innovation co-operation (#4.7),
    - internal knowledge flows wihtin enterprise groups (#4.8)
  - *Firm outcome*
    - firm size (#4.1)
    - total turnover (#4.3, #4.4)
  - *Background variables*
    - firm age (#4.5)

The remainder of chapter 3 more or less follows the structure from CIS2018:

- Paragraph 3.2 ('Information on the enterprise) covers Module 1 and the background variables from Module 4.
- Paragraph 3.3 ('Strategies and knowledge flows') covers the items on firm environment from Module 2 and on knowledge flows from Module 2, and on the innovation process from Module 3 (i.e., 'co-operation) and from Module 4.
- Paragraph 3.4 ('Business and innovation activities and expenditure') covers the items on innovation input (i.e., 'expenditure') from Module 3 and 4, innovation process ('status of innovation activities') from Module 3, innovation output from Modules 3, and firm outcome from Module 4.
- Paragraph 3.5 ('External factors influencing innovation') the items on innovation input from Module 3 (i.e., 'external funding) and Module 4 (i.e., educational level and internal funding) and on firm environment from Module 3.

3.2 Information on the enterprise

Background information (meta-data) on the enterprise is included in Module 1 (legal structure of the enterprise) and Module 4 (firm age). Firm size (Module 4) could also be regarded as a background variable (e.g., as a basis for enterprise profiling, see chapter 5.3) but given the policy relevance that is been attached to job creation the item has been defined here as a business outcome variable instead.

In CIS2018 it is assumed that the enterprise identification is extracted by the NSO from the national Business Register. This required a linkage on a unique business ID. Record linking could either be directly done via the public ID from the Business Register (the Business Register number) or, for the sake of privacy, via a parallel set of unique ID's that is maintained by the NSO (see hereafter, chapter 4, paragraph 4.5.1).

The national business register in turn is linked to the other national business registers in the EU via the EuroGroup Register (EGR), which is part of the FRIBS initiative (see also paragraph 4.3.1). Under FRIBS, enterprise information has been defined in a uniform manner across the EU at three levels:

1. *enterprise groups*: identity, demographic characteristics, the structure of the group, the group head, the country of global decision centre, activity code (NACE), consolidated employment and turnover of the group.

2. *enterprises*: identity and demographic characteristics, activity code (NACE), number of persons employed, turnover, institutional sector;

3. *legal units*: identity, demographic, control and ownership characteristics.

For the framing of the respondent it is of the uttermost important that the answers in the survey are only being answered for the business activities *in the country concerned*. Hence the first question in CIS2018 (no. 1.1) explicitly asks whether the company is part of an enterprise group or not, and if so, then instructs the respondent:

- Only to report about the activities of the enterprise group in the own coutry and
- To exclude all activities of all subsidiaries in parent companies.

Note that both legal structure (ownership and location of branches) and the firm age (date of establishment) could also be directly retrieved from the national Business Register. The use of a single point of registration (the national Business Register) and a consequent re-use of that (unique) registration ensures consistence across data sources (such as CIS), thus avoiding double counting at EU level (see also chapter 6 on globalisation). It also facilitates data collection (e.g., by improving the quality of mailing lists) and reduce response burden (e.g, through prefilling information into an online questionnaire).

3.3 Strategies & knowledge flows

3.3.1 Firm strategy

The way a company starts, launches and implements innovative activities depends very much on its strategy, resources, experience and environment (see hereafter, §0). The *strategy* of the firm one of the elements of

The way a company starts, launches and implements innovative activities depends very much on its strategy, resources, experience and environment (see hereafter, §0). The *strategy* of the firm one of the elements of Porter's classical model of the business environment ((Porter, 1990) see hereafter §3.5.1), together with the *structure* of the firm and *rivalry*. Strategy and structure are internal factors (but external to the innovation process, see §3.5.2), rivalry – the degree of competition in the market in which the firm operates – is an external factor. The management of a firm can use both strategy (in the short run) and structure (in the long run) to respond to changes in the market. Consequently firms that operate on one and the same market can have very different strategies and structures. For a proper comparison across firms it is therefore important to be able to use firm strategy

(and preferably also structure) as a control variable. However, strategy is also relevant to innovation in its own right as some types of strategy are conceptually linked to specific types of innovation. This is most evident for two first pair of items (i.e., a focus on improving existing products refers to process innovations whereas a focus on introducing new products refers to product innovations). The last item could also be used as a control question for the core question #3.1. The link is also present albeit to a lesser in the last item (i.e., a focus on customer specific products to product innovations, vis-à-vis a focus on standardised products to process innovations).

In CIS2018, firm strategy is covered in question #2.1. The items are grouped in opposed pairs of strategic options. The answers are however not coupled hence firms could give scores in similar directions for one pair. This is because in theory firms could pursue two opposed strategies at the same time (e.g., for different products or business units). In practise, an unevocal overall strategy will often be followed – the essence of a strategy is to maintain a certain focus.

2.1 During the three years 2016 to 2018, what describes the strategies of your enterprise to ensure its economic performance?

	This focus described the strategies of your enterprise			
	Very well	Well	Only to some extent	Not at all
Focus on improving your existing products*	☐	☐	☐	☐
Focus on introducing new products*	☐	☐	☐	☐
Focus on low-price (price leadership)	☐	☐	☐	☐
Focus on high-quality (quality leadership)	☐	☐	☐	☐
Focus on a broad range of products*	☐	☐	☐	☐
Focus on one or a small number of key products*	☐	☐	☐	☐
Focus on satisfying established customer groups	☐	☐	☐	☐
Focus on reaching out to new customer groups	☐	☐	☐	☐
Focus on standardised products*	☐	☐	☐	☐
Focus on customer-specific solutions*				

*Goods or services

3.3.2 Knowledge flows: overview

The ongoing sophistication of the innovation strategies implies new ways of leveraging intramural and extramural competences and resources, accessing the most advanced knowledge base and benefiting from the ideas developed in-house. Information on knowledge flows is essential to the understanding the innovation in all existing and prospective theoretical models as well as the contemporary practice of decision making[5].

Module 2 of the CIS2018 presents a systemic approach to capture particular configurations of knowledge flows in the enterprise. Most aspects were already covered in the previous version of CIS but they are now put into context. The module enquires about the internal knowledge sourcing, inbound knowledge flows, the interaction with customers, co-operation

[5] Knowledge flows are generally defined as all processes of transferring knowledge from the place it was created or stored to the place it would be applied (Allen, 1977).

agreements and intellectual property management. These questions can help to understand the different approaches to knowledge sourcing and sharing of enterprises, within their business group or with other enterprises and other organizations. Also, the questions can help to assess their capabilities about new knowledge creation, adoption and dissemination of innovations, and networking.

The module allows measurement of the firms' innovation strategies from the *open innovation* perspective. Introduced by Chesbrough (2003), this framework extends the non-linear (chain-link) model (Kline & Rosenberg, 1986) by specifying extra emphasis on the broader range of knowledge flows across the boundaries of the firm. One of the original definitions (Chesbrough, 2003):

> "open innovation is a paradigm that assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as firms look to advance their technology"

indicates that open innovation accounts for both the inbound and outbound flows of knowledge and technology, on the one hand allowing the company to utilize various external sources and on the other leveraging the benefits from ideas developed in-house and not immediately intended for being launched to the markets. This concept is a convenient umbrella for generalizing the existing and prospective forms of knowledge flows through the porous boundaries of the firm.

Questions in the module follow the perspective on the open innovation proposed by Dahlander and Gann (2010). They distinguish two critical dimensions: the direction of the knowledge flow towards the firm (inbound versus outbound) and the involvement of monetary exchange (pecuniary verus non-pecuniary). As a result, four major types of openness are considered: acquiring, sourcing, selling and revealing:

Table 2. Classification of activities with a company (Dahlander & Gann, 2010)

	Inbound	Outbound
Pecuniary	*Acquiring* Acquiring inventions and input to the innovative process through informal and formal yet money-based relationships	*Selling* Out-licensing or selling products in the marketplace
Non-pecuniary	*Sourcing* Sourcing external ideas and knowledge from suppliers, customers, competitors, consultants, universities, public research organizations, etc.	*Revealing* Revealing internal resources to the external environment

Source: (Dahlander & Gann, 2010)

Sourcing (inbound, non-pecuniary) implies a synergy between in-house processes and open information available without strict financial liabilities. While on the positive side of this type of channeling is cost saving, the literature provides specific evidence on the possible harmful blurring of in-house development activities in attempt to incorporate all the existing information. *Acquiring* (inbound, pecuniary) encompasses all forms of purchase of technologies and R&D effort. The advantage of this channel is the accumulation of competences. A disadvantage is the need to endow resources and the proper governance models of incorporating

(co-creating) new knowledge into the in-house processes. *Selling* (outbound, pecuniary) allows to fully leverage investment in R&D partnering with actors that can bring these results to the market. However, this channel suffers from all the imperfections of the knowledge markets and IP protection mechanisms. *Revealing* (outbound, non-money) implies sharing the knowledge with the network of partners without immediate financial benefit. This channel is efficient when the company faces specific appropriability regimes, and it is overcostly to protect innovation. However, capturing the benefit is subject to a smart design of the strategy.

To capture this diversity of dimensions, CIS-2018 introduces four groups of questions within Module 2:

- *Interaction with customers* (Q2.2, Q2.3, Q2.4) acknowledges the contribution of the end-users into the innovation co-creation effort (often referred as user innovation[6]).
- *Intellectual property management* (Q2.5, Q2.6, Q2.7) provides ground for measuring the flow of knowledge regarding particular types of intangible assets, adding both to the outbound dimension of open innovation and to the domain of innovation management practice.
- *Inbound knowledge flows* (Q2.8, Q2.9, Q2.10) explicitly accounts for the particular channels exploited by the enterprise, including the expanding body of open sources.
- *Internal sourcing* (Q2.11) helps to reveal the configuration of the intramural innovation efforts and capabilities, which is the essential determinant of efficiency for all types of knowledge creation, adoption, and transfer processes.

Co-operation agreements are covered outside Module 2, in Module 3. Since question #3.15 explcitly deals with international co-operation in this Manual it is covered in the chapter 6 on Globalisation (see chapter 6).

- Co-operation agreements (Q3.14, Q3.15) reveal information on the formal mechanisms interaction within the innovation networks.

3.3.3 Rationale

Module 2 provides the facilities to account for the scope of existing and emerging innovation strategies that are highly to rely on complex configurations of knowledge flows within- and across the borders of an enterprise.

A joint analysis of knowledge flows, innovativeness and outcome data is a pillar of studying the mechanics and impact of innovation. Since the first revisions, CIS contributes to understanding the processes of innovation networking, the intensity, and scope of industry-science linkages, the importance of innovation channeling mechanisms to model the impact of particular factors that influence the engagement into the co-operative activities. A growing body of research on the new models of network-driven innovation has expressed a clear need for expanding the set of indicators to account for underexplored components of innovation strategies[7], mainly to overcome the excessive *inbound sourcing bias* of the questionnaire.

---

[6] (Gault & von Hipel, 2009).

[7] CIS-based empirical studies that follow the *open innovation* perspective includes (Laursen & Salter, 2006), UK; (van der Meer, 2007), the Netherlands; (Acha, 2008), UK; (De Backer, Lopez-Bassols, & Martinez, 2008), 26 EU countries; (Barge-Gil, 2010), Spain (Filippetti, 2011); Filippetti (2011), 27 EU Members, Switzerland and Norway; (Drechsler & Natter, 2012), Germany.

A new composition of questions helps to expand the measurement framework of the previous CIS questionnaires and allow for the omnidirectional treatment of the knowledge flows. Likewise complementary dimensions, such as in-house capabilities, intellectual property management, and mechanics of formal co-operative agreements, are also better covered.

The extended data on knowledge channeling is of particular importance to policymakers. Information on the nature and composition of existing and potential linkages between the innovation firms, research organizations, universities and other actors is crucial for the successful design of the efficient policy measures to promote and support all types of intellectual exchange and innovation development.

3.3.4 Interaction with customers

The part on interaction with customers consists of three questions. The first question (Q2.2) deals with the degree of customer involvement. Only the first item refers to *user innovation*. This question contributes both to the analysis at micro-level by allowing to control for the diversity of user-maker relationships and to the macro-level indicators, introducing the formerly missing measure of the individuals' engagement into the innovation activities of the business enterprises.

2.2 In the three years from 2016 to 2018, did your enterprise produce or deliver products (goods or services) in response to specific requirements by users?**

	Yes	No
Your enterprise co-created* products with users**, i.e. the user had an active role in the creation of the idea, design and development of the product (co-creation)	☐	☐
Your enterprise designed and developed* products specifically to meet the needs of particular users** (customisation***)	☐	☐

* A difference between customisation and co-creation is that for 'customisation' the enterprise designed and developed the product alone, whereas for 'co-creation'the enterprise designed and developed the product together with the user**.

** A user can be an end customer or an enterprise which uses a product as an intermediate product.

*** This excludes mass customisation, i.e. customised versions of standard products.

In the follow-up question Q2.3 four types of users are being distinguished. This overcomes the limitations of the existing innovation surveys that propose a unified category for the customers. At the same time this limits the potential descriptive power of the indicators.

2.3 For the products resulting from 'customisation'* or 'co-creation', the users included**

	Yes	No
Private business enterprises	☐	☐
Public sector organisations***	☐	☐
Individuals or households	☐	☐
Non-profit organisations	☐	☐

*This excludes mass customisation, i.e. customised versions of standard products.
** A user can be an end customer or an enterprise which uses a product as an intermediate product.
*** Public sector organisations include government owned organisations such as local, regional and national administrations and
agencies, universities, schools, hospitals, and government providers of services such as security, transport, housing, energy, etc.
Include state-owned enterprises and state-owned corporations.

The last question Q2.4 asks directly for the actual contribution of customer interaction to turnover. However because the two modes of involvement, co-creation and customization are lumped together no split for user innovation can be given. Moreover it remains to be seen to what extent a firm is able to give a reliable estimate of the direct contribution of customer interaction to turnover.

2.4 Please provide an estimate for the percentage of turnover in 2018 from

Products resulting from 'customisation' or 'co-creation'	___ %
Other products	___ %
Total turnover	100 %

3.3.5 Intellectual property management

Intellectual property management is an integral part of the knowledge flow analysis framework. The module proposes three questions concerning IPR.

2.5 In the three years 2016 to 2018, did your enterprise:

	Yes	No
Apply for a patent*	☐	☐
Register an industrial design right	☐	☐
Register a trademark	☐	☐
Claim a copyright	☐	☐
Use trade secrets	☐	☐

*This covers applications filed during the reference period 2016 to 2018 (not the grant of patents). Those countries where 'utility models' are relevant can include a respective category.

Inheriting the design from the previous rounds of CIS, Q2.6 inquires about the engagement of the enterprise into six *formal mechanisms of IPR protection*. The question proved to be

essential to account for the outcomes of the firms' creative efforts as well as understanding the relevance of the IPR frameworks.

2.6 In the three years 2016 to 2018, did your enterprise:

	Yes	No
License out own Intellectual Property Rights (IPRs) to others	☐	☐
Sell own IPRs (or assign IP rights) to others	☐	☐
Exchange IPRs (pooling, cross-licensing, etc.)	☐	☐
Enter into a franchise agreement	☐	☐

Q2.6 proposes the measures for the outbound IPR activities that help to integrate the *open innovation* perspective into the harmonized survey framework.

The new format of IPR questions combined with other variables from the Module 2 will foster the emergence of the empirical evidence to foster the understanding of the network configurations and the corresponding ways to package, deliver and protect knowledge. Related to the information on innovation inputs and outcomes, this would enable the empirical evaluation of the scope of innovation models that have so far largely remained at the theoretical and conceptual stages.

3.3.6 Inbound knowledge flows

The new format of the innovation sourcing questions (Q2.8, Q2.9, Q2.10) proposes means to capture the specificites of the innovation acquisition mechanisms pursued by the enterprise. The conventional channels for innovation (by type of the information provider) reflect the configuration of the innovation sourcing networks. The resulting indicators can bring new insight on the *balance between the codified and embodied knowledge within the innovation networks.*

In earlier versions technical services and goods were taken together but the number of sources of origin (i.e. suppliers) was quite elaborate. In the final version, Q2.8 covers technical services for only two types of suppliers, and Q2.9 goods.

2.8 During the three years 2016 to 2018, did you enterprise buy technical services* from

	Yes	No
Private business enterprises	☐	☐
Public research organisations, universities and other higher education institutions	☐	☐

*'Technical service' includes any consulting activity that involves any kind of technical, scientific or engineering information.

In the second item from Q2.9 an implicit reference to innovation (namely embodied in technology) is being made. In theory, new technology could be implemented without any innovation but in practise but in practice it is almost always accompanied by organisational and/or process innovation.

2.9 During the three years 2016 to 2018, did you enterprise purchase machinery, equipment or software based on

	Yes	No
The same technology used in your enterprise before	☐	☐
New technology that was not used in your enterprise before	☐	☐

Reflecting the growing importance of the open channels, question Q2.10 accentuates the utilization of these sources by the firm. The list of channels includes conventional categories already included in previous CIS questionnaires but is expanded with other prevalent and debated instruments, such as social networks. The explicit question on the reverse engineering practice potentially may lay a new backbone to the technology adoption studies and policymaking practice.

2.10 During the three years 2016 to 2018, did you enterprise use any of the following channels to acquire knowledge?

	Yes	No
Conferences, trade fairs and exhibitions	☐	☐
Scientific/technical journals or trade publications*	☐	☐
Information from professional and industry associations	☐	☐
Information from published patents	☐	☐
Information from standardisation document	☐	☐
Social networks, web-based platforms or crowd-sourcing**	☐	☐
Open platforms or open-source software***	☐	☐
Reverse engineering****	☐	☐

*This can include general business newspapers or branch specific magazines.
** This category mainly relates to 'business to customer' relationships. Crowdsourcing is a specific practice in which enterprises use
contributions from Internet users to obtain needed services or ideas.
*** This category mainly relates to 'business to business' relationships.
**** Processes of extracting knowledge or design information from anything manufactured and reproducing it or reproducing anything
based on the extracted information.

3.3.7 Internal sourcing

Question 2.11 brings insight on the *modality of innovation management culture* of the organization. Practices accounted this question introduce three dimensions of the enterprises' managerial culture. Capturing new ideas and proposals provides an open end for in-house creativity; formalisation is associated with the maturity of the innovation management processes; internal sharing and openness reflects the momentum for organisational learning and development.

This question provides an operationalization for a fuzzy *innovation culture* concept that is acknowledged as one of the crucial determinants of the innovation success.

2.11 During the three years 2016 to 2018, how important to the management of your business and staff were the following methods of organising work ?

	Degree of importance			
	High	Medium	Low	Not used
Planned job rotation of staff across different functional areas	☐	☐	☐	☐
Regular brainstorming sessions for staff to think about improvements that could be made within the business	☐	☐	☐	☐
Cross-functional work groups or teams (combined across different working areas or functions)	☐	☐	☐	☐

3.4 Business and innovation activities and expenditure

3.4.1 Concept

The CIS2018 questionnaire collects information at the firm-level on business and innovation activities expenditure in Module 3 and Module 4. The questionnaire items are described in this section. Chapter 5 examines the statistical analysis that can be carried out to measure innovation intensity (see §5.2).

3.4.2 Rationale

Expenditure in innovation is a main indicator for *innovation input.* It is *directly* measured by breaking down into several types of costs. The notion of direct measurement refers to the possibility of collecting the value of innovation expenditure from the respondent firm, without any need for modelling or any other statistical elaboration. This is because companies usually keep accounting records of their investments and purchases of goods and services. However, in order to be able to allocate expenditures to innovation firms need to be able to recognise the link between the expenditure and any of the innovative activities proposed by the CIS2018.

3.4.3 Implementation of innovation activities

Duration of innovation projects

The implementation of innovation activities is in general a long (and non-linear) process which may take more than one fiscal year – which is the reference period for which economic/financial variables are collected. The innovation project life-cycle (see figure below) shows in a simplified way the time-frame of a project. Knowing the status of implementation of innovation projects in a given reference year (that of the survey) allows user to understand better how the allocation of resources is made in time.

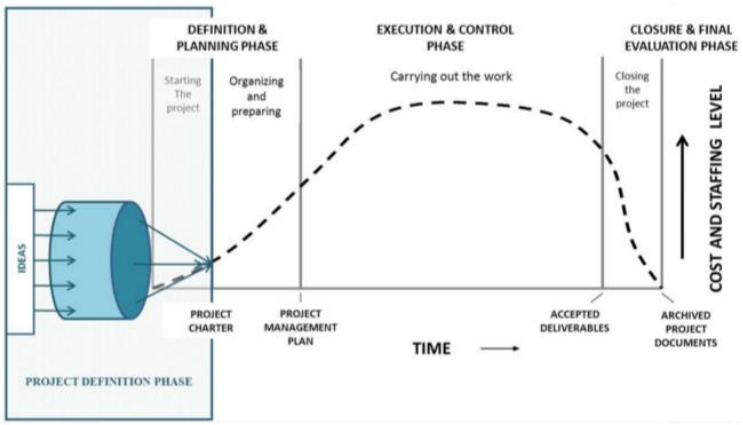


Figure 2. A generic project life cycle model for innovation projects (Marcelino-Sádaba, González-Jaen, & Pérez-Ezcurdia, 2015)

The previously discussed distinction between the *subject-based approach* (i.e. the innovative firm) and the *object-based approach* (i.e. the innovation project within a firm) is important here (see herfore, §2.1.2). When innovation activities are organised as projects, the timing (start-end) is measurable, while innovative activities not organised as projects may present more difficulties in terms of determining their status of completion.

The organisation of innovation within the firm

Internally, companies undertake innovation activities in different degrees of formalisation and flexibilisation (Mattes, 2014). The way a company starts, launches and implements innovative activities depends very much on its strategy, resources, experience and environment. The activities can be undertaken as part of the functional or routine work of the company's staff, or be "packaged" as projects. Low-formalised or highly flexible organisational innovation provides autonomy to firm units or staff, entails informal feedback loops before innovations are adopted, and counts on highly tacit organisational knowledge. Innovation activities in traditional sectors and in the informal economy are rather un-formalised and flexible. A simplified classification is presented in the table below. The shaded cells are relevant for the measurement of innovation intensity.

Table 3. Classification of activities within a company

	Innovative activity	**Non-innovative activity**
Functional	Innovation outside projects	Routine activities
Projects	Innovative projects	Non-innovative projects

Highly formalised, less flexible approaches – more frequent in high-tech, competitive sectors – entail the application of internal rules, binging standards, institutionalised and often requiring written documentation. In general, innovators mix both approaches in what the author calls *'ambidexterity'* and concludes that "[..] while ambidexterity has usually been described at a corporate level, the empirical insights show that the tension is instead resolved to a large extent at a smaller scale, i.e. for individual *projects* […]".

The literature about innovation management recognises an increasing *'projectification'* of the work as enterprises and specialists organise their work in projects rather than on on-going functional basis.[8] Traditional industries that once organised their activities in a functional way are evolving towards project-based forms of organisation. Likewise, emerging industries (ICT, biotechnology) are increasingly adopting project-based forms (Filippov & Mooi, 2010). Tools for project management have been developed from the management of R&D.

There is a risk that the project-based innovation intensity measurement induces a statistical bias. The very fact of organising the innovation activity in projects could already indicate a higher innovation intensity. Bundling innovation activities within a firm in projects is indeed already a sign of certain 'maturity' with respect to innovation. Allocating resources (internal and external) to the implementation of such projects signals that the firm undertakes the innovation activities within a strategy and probably can report on such activities with more detail. Specific innovation project management-related variables may describe the intensity with which the firm undertakes such projects.

Measurement of the status of innovation activities within CIS2018

Taking into consideration the diversity of organisation of innovation activities, the measurement of any innovation effort at the firm level is a statistical simplification of the complex process of innovation within a company. This process is characterised by a mix of activities, implemented in different levels of formalisation, time frames, with different requirements in terms of internal and external resources, and different success probabilities and results. Indeed, aggregating expenditures or any other input or output variable at the firm level does not allow for identifying either the internal allocation of resources to the innovation portfolio, nor for evaluating the contribution of such allocation to the firm's performance. As a results, firm-level statistics may shed little light on optimal innovation strategies for companies (see herefor, §1.2.2).

To partially compensate for this, the CIS2018 questionnaire includes four dichotomic (yes/no) variables on the status of innovation *activities* (questionnaire item #3.9):

3.9 During the three years 2016 to 2018, did your enterprise have any

	Yes	No
Completed activities on product/process innovation	☐	☐
Ongoing innovation activities at the end of 2018	☐	☐
Abandoned innovation activities	☐	☐
Research and development (R&D) activity?	☐	☐

Note that the answers to this question are pre-set to 'yes' when either question #3.1 ("Did your enterprise introduce any new good or service during 2016-2018") or question #3.6 ("'Did your enterprise introduce any new processes during 2016-2018") is answered with 'yes'. Also note that question #3.9 is used as a filter (or essentially as an ex ante control question) to the core question #3.10 on innovation expenditure (see §3.4.4). If none of the answers is a 'yes', question #3.10 is skipped. However, if a firm only has ongoing and/or abandoned innovation activities during the last three years it is still asked about the

[8] An older study showed that 82% of innovators -including business, academic centres, government- use project management techniques (European Commission, 2004).

expenditure on these activities. This is much less restrictive than in the previous edition of CIS (CIS2014) that included firms which had completed innovation activities.[9]

3.4.4 Innovation expenditure

Question #3.10 first separates innovation expenditure from expenditure on R&D and then breaks it down in three broad categories:

3.10 How much did your enterprise spend on innovation and research and development (R&D) in 2018?

	Expenditure on innovation and R&D in 2018	
	Please estimate if you lack precise accounting data	*Please estimate if there were no such expenditure in 2018*
R&D performed in-house	__,___,___,000 €	☐ none
R&D countracted out to others (including enterprises in own enterprise group	__,___,___,000 €	☐ none
All other innovation expenditure (i.e., excluding R&D)	__,___,___,000 €	☐ none
Of which:		
Own personnel working on innovation	__,___,___,000 €	☐ none
Services, materials, supplies purchased from others for innovation	__,___,___,000 €	☐ none
Capital goods (acquisition of machinery, equipment, software, IPRs for innovation	__,___,___,000 €	☐ none

Services may include: product design, service design, preparation of innovation activities other than R&D, training and professional development, marketing, etc. It is important to recall that some expenditures refer to *intangible investment* (e.g. Intellectual Property Rights, IPRs) and that the analysis of such data has received special attention.

Several consistency checks may be developed by analysts of microdata, by relating non-zero expenditure amounts to other questionnaire items. As examples:

- If costs for own personnel working on innovation are non-zero, then answers to questions #3.4 or #3.7 (see hereafter, (see §6.4.1) should be any of the options informing that the innovation has been developed by "your enterprise by itself";
- If question #2.7 (#see herefor, §3.3.5) includes a positive answer on acquisition of IPRs, then it is plausible that the expenditure in capital goods is non-zero.

An exhaustive list of checking procedures (also called "edit rules") is difficult to establish and is generally relevant only to those institutions that are collecting – and accessing – microdata.

[9] See §0 on the impact of censoring and selectivity on statistical analysis.

In addition to current expenditure, CIS2018 intends to collect the *forecast* values for the following years (question #3.11). The precision of such forecast can only be evaluated by analysing the underlying micro data.

3.11 How much do you expect your enterprise's total innovation expenditures* to change in 2019 and 2020?

2019 compared to 2018	**2020 compared to 2019**
☐ Increase *If yes, by* ☐☐☐ % *approximately***	☐ Increase
☐ Stay about the same (+/- 5%)**	☐ Stay about the same (+/- 5%)
☐ Decrease *If yes, by* ☐☐☐ % *approximately***	☐ Decrease
☐ No innovation expenditures expected	☐ No innovation expenditures expected
☐ No innovation expenditures expected	☐ No innovation expenditures expected

*Total innovation expenditures include those for R&D and all other innovation activities.
**If there were no innovation expenditures in 2018 or 2019, pleaseonly indicate if these will increase in 2019 or 2020, respectively.

Innovation expenditure can be compared with overall business expenditure to obtain measures of *innovation intensity*, as well as to elaborate typologies of innovation strategies by analysing the composition of innovation expenditure per categories of expenditure. Consequently, the CIS2018 collects information about general turnover and expenditure in Module 4 (questionnaire items #4.3, #4.4 and #4.6). The analysis of the combination of *innovation expenditure* and other firm-level variables is described in Chapter 5.

The comparison of innovation-related and overall business expenditure as recorded in the CIS2018 questionnaire is shown below, presenting the possible combination of variables.

Table 4. Comparison of CIS2018 questions related to business expenditure in general and to innovation expenditure in particular

Innovation-related (question #3.10)	Overall business expenditure (question #4.6)
R&D performed in-house R&D contracted out to others	*Total expenditure not collected*
Own personnel working on innovation	*Total personnel collected in question #4.1*
Capital goods (acquisition of machinery, equipment, softeware, IPRs for innovation)	Acquisition of machinery, equipment, building and other tangible assets Registering, filing and monitoring own Intellectual Property Rights (IPRs) and purchasing or licensing IPRs from others Marketing, brand building, advertising (including in-house costs and purchased services) Training own staff
Services, materials, supplies purchased from others for innovation	Product design (including in-house costs and purchased services) Software development, database work and data analysis (including in-house costs and purchased services)

Again, chapter 5 describes possible analysis of innovation expenditure data as measures of *innovation intensity*.

3.5 External factors influencing innovation

3.5.1 Concept

Neither innovation nor business processes occur in isolation. Both are affected by a myriad of external factors. The innovation process is thus directly and indirectly (via the feedback loop from the business process) affected by the environment of the firm.

In the classical diamond model from Porter, the business environment consists of five interacting elements and a sixth independent element (‘change’) (Porter, 1990).

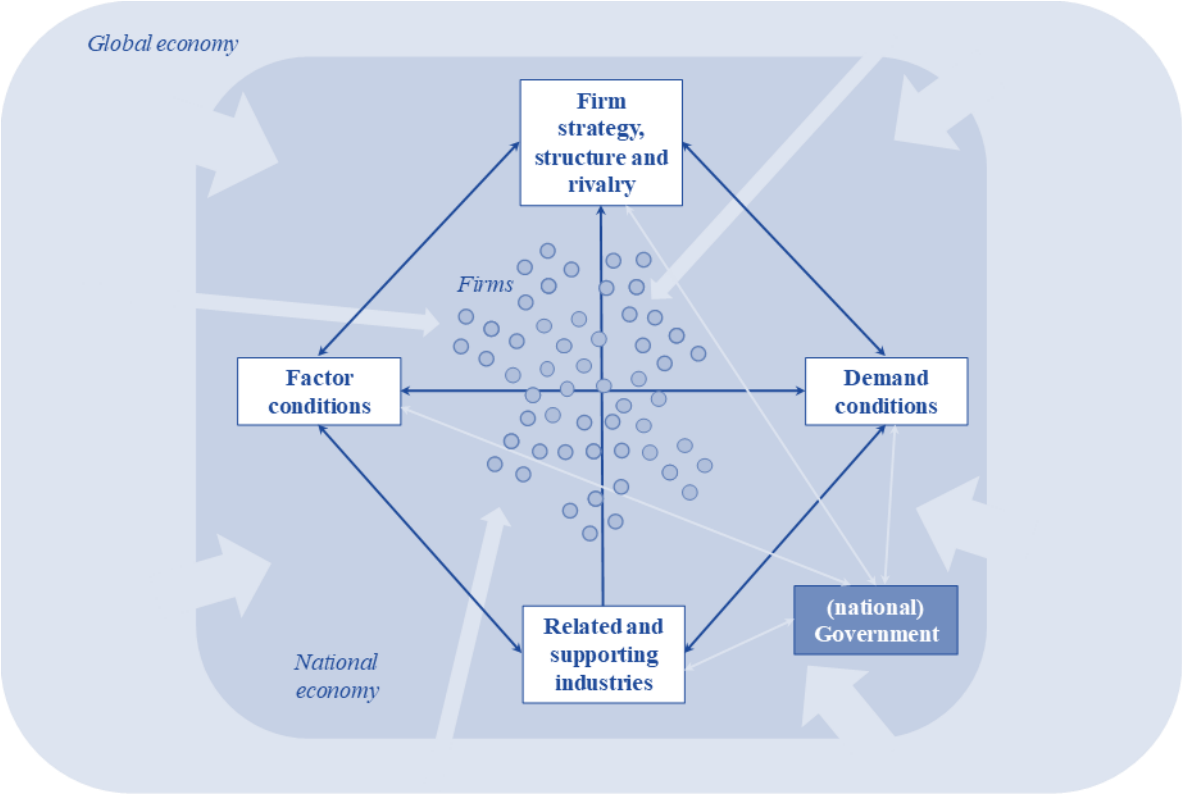


Figure 3. Determinants of national competitive advantage (Porter, 1990), adapted by Dialogic

The aim of the model is to show that a government can improve the competitive situation of a national economy by influencing one or more of the elements in the diamond. These elements in turn determine the competitive situation of firms that operate within the national economy. Contrary to government a firm is not assumed to have any influence on the elements. These are external factors to the firm and it can only adapt to the changes in its environment. Depending on the specific internal strength and weaknesses of the firm and its (dis)ability to change, from a strategic point of view such changes can either be regarded as threats or opportunities.

The business environment is multi-layered. In its daily operations, a firm has the most to with its customers ('demand conditions'), suppliers ('related and supporting industries') and direct competitors ('rivalry'). One layer up is the sector (or sectors) in which the firm is operation. In turn, these sectors are embedded in a national economy (the reference level for Porter's model) and often partly in a global economy. Finally, a national economy is embedded in the global economy. A government always has only a partial influence on the factors in the national economy. They are also (and often largely) determined by developments in the global economy. Likewise, the competitive position of a firm is not only indirectly (via changes in the national economy) but often also directly influenced by changes in the global economy (see chapter 6).

3.5.2 Rationale

To properly compare the performance of firms in terms of the set-up and quality of their innovation processes across sectors or countries one needs to isolate internal from external factors. The external factors can then be used as control variables.

From the perspective of the innovation process (the inside-out view) two types of external factors can be distinghuished:

1. Factors internal to the firm but external to the innovation process;
2. Factors external to the firm

External factors could either directly or indirectly (via the business process) be related to the innovation process. In the latter cases, CIS2018 explictly makes the split between a direct and indirect link to innovation. One example is question #3.13 on financial support:

3.13 During the three years from 2016 to 2018, did your enterprise get financial support?

	Yes	(Part of) this funding was used for R&D and other innovation activities
	Tick all that apply	Tick all that apply
Financial support from local or regional authorities*	☐	☐
Financial support from the national government	☐	☐
Financial support from the EU Horizon 2020 Programme for Research and Innovation	☐	☐
Other financial support from a European Union Institution*	☐	☐
Tax credits and deductions from local, regional or national governments	☐	☐

CIS2018 has some questions on both types of factors and all elements of the Porter model, albeit with a focus on government and limited coverage for the other elements (firm strategy, structure and rivalry; factor conditions; demand conditions; related and supporting industry). Obviously, there are many more external factors that could be considered. For the sake of the length of the survey protocol a selection had to be made. Note that via data linkage a great number of datasources with detailed information on specific factors could be disclosed (see chapter 4.1).

3.5.3 Factors internal to the firm but external to the innovation process

There are two questions in Module 4 on internal factors that refer to the functioning of the firm as a whole but that are also highly relevant to the innovation process. These are respectively the *educational level of employees* (#4.2) and *internal funding* (#4.9). The availability of internal funding (or the lack thereof) is obviously an important input factor to the innovation process – provided that the money is being earmarked for innovation. Question #4.9 explicitly deals with this issue, albeit only for a subset of firms (enterprise groups) and without asking about the actual level of funding.

4.9 During the three years from 2016 to 2018, did your enterprise tried to get or actually gor funding in the form of intra-group loans?

Step 1:	↓ Step 2:	
The enterprise <u>tried to get intra-group loans</u>	**The enterprise <u>actually got intra-group loans</u>**	**(Part of) the <u>intra-group loans were used for R&D and other innovation activities</u>**
Tick all that apply	Please answer only if you re-ponsed Ýes' in Step 1	Please answer only if you re-ponsed Ýes' in Step 2
Yes ☐	Yes ☐	Yes ☐
No ☐	No ☐	No ☐

In principle, *total turnover* (#4.4) could also be included in this category. After all turnover (operating income) is an important external input factor for innovation. Most likely there is a strong feedback loop from business performance to the internal funding of innovation. However total turnover is also an outcome of the business process (and thus indirectly of the innovation process) – it is both cause and effect. Given the critical importance of overall firm performance as a benchmark for innovation performance the variable has been included in Module 3 (and thus in the previous paragraph 3.4 on business and innovation activities and expenditure), and not in Module 4.

3.5.4 Factors external to the firm and directly related to the innovation process

External factors that are directly related to the innovation process and that are included in CIS2018 refer to *external funding* (#3.12 and #3.13), *regulation* (#3.16), and a heterogenous set of other factors (#3.17).

In terms of the Porter model, #3.12 refers to a special set of 'supporting industries' (i.e., financial institutions and venture capitalists) and #3.13 and #3.16 refer to 'government'.

#3.12 has a similar structure than #4.9 but it refers to *all* types of enterprises:

3.12 During the three years from 2016 to 2018, did your enterprise tried to get or actually gor funding in the form of?

	Step 1:	↓ Step 2:	
	The enterprise <u>tried to get</u> funding (e.g. applied for a credit or grant)	**The enterprise <u>actually got</u> funding**	**(Part of) this <u>financial support</u> was <u>used for R&D and other innovation activities</u>**
	Tick all that apply	Tick all that apply, but only if Step 1 is ticked for that item	Tick all that apply, but only if Step 2 is ticked for that item
<u>Equity finance</u> (finance provided in exchange for a share in the ownership of the enterprise)	☐	☐	☐
<u>Debt finance</u> (finance that the enterprise must repay)	☐	☐	☐

Question #3.13 has already been described in the previous section. It is evident that #3.12 and #3.13 are closely related. A relevant question is whether public financial support crowds out private sector funding, or whether it rather acts as a multiplier (i.e. public funding attracts additional private funding – or the other way around; for instance in the case of top-up financing via formal matching requirements).

The focus on #3.16 is on the *nature of the contribution* of government: it could be either positive, negative, or neutral. The unit of analysis is government, not the business enterprise. Note that there are two effects at work here. First, legislation or regulation might be relevant or not. Secondly, if legislation or regulation is felt, even if the direct affect is positive depending on the market conditions and the way the intervention is being implemented the nett effect could be neutral (or even negative) since government interventions inherently disturbs the market and always comes with additional (overhead) costs. Most of the legislation or regulation affect firms via changes in the factor conditions (e.g., salaries are affected by employment laws, prices of raw materials through changes in environmental legislation), and to a lesser extend via changes in demand conditions (e.g., product safety and consumer protection). Often sizeable niches of specialist supporting industries arise on the basis of legislation or regulation (i.e., lawyers, advisors and consultants).

3.16 During the three years from 2016 to 2018, has <u>legislation or regulation affected your enterprises' innovation activities</u> in any way shown in *colums A to C*?

Type of legislation or regulation	**Initiated or facilitated innovation activities**	**Prevented new innovation, or hampered or increased cost for innovation activities**	**Has no effect / not relevant**
	Tick all that apply		
	Column A	*Column B*	*Column C*
Product safety, consumer protection			
Environmental			
Intellectual property			
Tax			
Employment, worker safety or social affairs			

The last question of Module 3 consists of a heterogenous set of items that covers all of the factors from the model (see below, Table #3#). Most of the items are covered in more detail elsewhere in the survey. In these cases #3.17 can be used as a control question.

3.17 During the three years 2016 to 2018, how important were the following <u>factors in hampering your enterprises' decision to start innovation activities*, or its execution of innovation activities</u>?

		Degree of importance			
		High	Medium	Low	Not a constraint
a	Lack of internal finance for innovation	☐	☐	☐	☐
b	Lack of credit or private equity	☐	☐	☐	☐
c	Difficulties in obtaining public grants or subsidies	☐	☐	☐	☐
d	Costs too high	☐	☐	☐	☐
e	Lack of skilled employees within your enterprise	☐	☐	☐	☐
f	Lack of collaboration partners	☐	☐	☐	☐
g	Lack of access to external knowledge	☐	☐	☐	☐
h	Uncertain market demand for your ideas	☐	☐	☐	☐
i	Too much competition in your market	☐	☐	☐	☐
j	Different priorities within your enterprise	☐	☐	☐	☐

The items *d, h* and *i* mention general themes that are do not refer to any other question in CIS2018, hence they are considered to be truly external factors. This items could be used to control for differences in business environments.

Table 5. Cross-references from items #3.17 to other questions in CIS2018 and factors in the business environment model

	Factor	*CIS Theme*	*Refers to question*
a	(Firm strategy)	Finance	4.9
b	Supporting industry	Finance	3.12
c	Government	Finance	3.13
d	Factor conditions	External factors	
e	Factor conditions	Basic information on firm	4.2
f	Related and supporting industry	Knowledge flows	2.3; 2.4; 2.6
g	Related and supporting industry	Knowledge flows	2.8; 2.9; 2.10
h	Demand conditions	External factors	
i	Rivalry	External factors	
j	Firm strategy	Strategy	2.1

4 Data collection

4.1 Linking CIS data with other data sources

4.1.1 Rationale of data linkage

In the previous chapter we have discussed all modules and questions in CIS2018. For obvious reasons, CIS strongly focuses on innovation activities (Module 3) and the knowledge flows that are underlying these activities (Module 2). For the sake of brevity it can only cover the outcome of the business process, and the factors that influence the quality of the innovation process and the business process to a limited extend (Module 4). For instance, only two basic indicators on outcome are included, namely changes over time in total turnover and total number of staff.

Rather than bringing in more variables into CIS, which would dillute the scope of the current targeted framework, the alternative is than to link to external data sources that have a better coverage of the domains that are outside the strict scope of CIS. A more elaborate analysis of firm performance and the role of innovation in this performance would for instance require to include many more intermediate variables about the firm environment (especially linking to the innovation and business process), and more input and outcome variables.

Links could be made from any of the elements in the conceptual model in Figure 4, to any of the components in the fringe of the model. These links (⟷) are already depicted in the figure. For instance, many more additional background variables could be included, such as the geograpical location (e.g., to describe innovation activity at the local level). Information on the enterprise could also derived from third party registers as well. In fact, in CIS2018 it is already assumed that the enterprise identification is extraction from the national business registers (see before, §3.2).

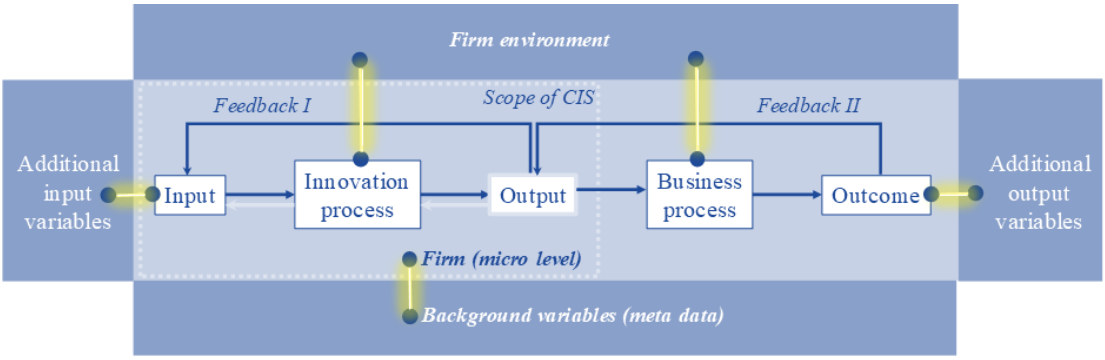


Figure 4. Second display of the basic conceptual model, now with a focus on data linkage

4.1.2 Availability of external data sources

There is a whole range of alternative sources of data that can be used next to the traditional primary data collected in sample surveys such as, but not limited to, innovation surveys. However, the actual availability of data might be an issue. This is because availability has

several layers, and each of these layers can interfere with the eventual use of the data for statistical and scientific purposes.

First, suitable data should exist in the first place. Especially in the realm of business data there is a lot of data available from commercial sources but the difference in nature and/or difference in quality standards could render the data unfit for use in official statistics or scientific research. Also relevant is the stability of the data holder. Private sector data suppliers might change their product portfolio, might be taken over or cease to exist altogether.

The way the data has been collected and processed also affects the methods that could be used for the linkage of the data. In the case of record linkage, individual records have to be discernible. In the case of statistical matching, characteristics of the sample have to be known (see hereafter, §4.4.2).

Secondly, data should also be accessible. That is, existing data always has to be made available by the data holder. The use of data might be bound to various legal conditions. In most countries, for instance, access to personal data is severely restricted. Commercial data holders might also charge high prices for the use of their data and/or their database.

> Even if the third party involved is not the owner of the data at least in Europe the database might still be protected by sui generis rights. Although under the Directive on the the re-use of public sector information [Directive 2013/37/EU] data is in principle free of charge (or at least limited to the marginal costs of the individual request), sometimes private enterprises might have been involved in the (co-)generation, processing or distribution of the data. In these hybrid cases, commercial interests might still be at stake.

Even if the use of data is allowed other conditions might apply to the re-use of the data. With regard to privacy, in many countries only data that do not allow the identification of individuals (or firms) can be made publicly available. Privacy concerns also arise when data matching is being conducted across databases that are held by different organizations, and when the matching requires identifying data to be shared and exchanged between organizations. Different price regimes might also apply to the use (for internal use only) and *re*-use (for wider distribution). The latter might be forbidden altogether by commercial data suppliers. Conditions for (re)use are sometimes also not particularly transparent and/or highly variable. A related issue is that the suppliers of the data that set the conditions should also be accountable for adhering to these conditions.[10]

[10] Finally, if suitable data exists, is accessible, and can be re-used it still has to be comprehensible to enable practical use. In essence, this means that the data should be machine-readable and preferably structured. Moreover, the data should preferably be accompanied by sufficient background information (meta-data). However with the recent rise of data processing techniques that can automatically create machine-processable structures and tags the lack of ex ante structure and meta data is less of a problem.

4.2 Re-using third party data

4.2.1 CIS re-using third party data

The re-use of existing data sources has several advantages. First, it is often more efficient than collecting primary (survey) data. Second, it lowers the administrative burden on respondents (Laux & Radermacher, 2009). Third, as there is a great number of external data sources available, innovation data can be linked to a wide variety of outcome variables (thus not just limited to economic data). Fourth, the data quality (in terms of accuracy, completeness and actuality) from measurements or registrations that are dedicated to the outcome topic at hand is usually high.

However, as described above, the conditions that apply to the data might impede the re-use of the third party data. The use of data from *commercial data holders* might especially be challenging. Re-use might be forbidden altogether (e.g., due to purpose limitation), the prices that are being charged by a commercial data aggregator might be prohibitively high and/or the conditions that apply to the re-use of the data might not be transparent.

The conditions that apply to the re-use of *public sector information* are usually more favourable. We are especially referring here to secondary data that are typically collected in support of some administrative process. As registration is often obliged by law secondary data (such as population of firm registers) often covers the entire population whereas primary (survey) data is based on a sample sources (Buelens, Boonstra, Brakel, & Daas, 2012). Although administrative registers might also contain some measurement errors (e.g., due to administrative errors) the data quality is generally higher than from survey data, which is inherently prone to subjective biases from the respondents.

The linking of innovation input and output data with administrative data is most in line with developments taking place in the international statistical community (e.g., within ESS) and it is in fact already widely practiced. The potential range of administrative sources that could be used for statistical purposes is large and growing (Dias, 2015). Examples of administrative sources that could be relevant to the study of the outcome of innovation processes are tax data, published business accounts, licencing systems, building permits social security data, education records.

Text box 5. Example of linking CIS micro data with administrative data

In their paper on measuring the impact of cultural diversity on innovation, Ozgen, Nijkamp and Poot (2013) combine CIS micro data with micro data from two administrative registers: employee data from the Tax Register (SSB) and demographic background data from the Dutch Municipal Registers (GBA).

Two cross-sections of CIS (3.5 and 4.5) were linked to create a balanced panel of firms that can be followed over four years. Then, the panel of firms was linked to the Tax Register to obtain the actual number of employees per firm and by location. Finally, the new dataset was merged with the GBA to gather the actual number of foreign employees per firm, as well as their country of birth and various other demographic characteristics.

Eventually, this resulted in a dataset with 5,578 observations, consisting of two waves of 2,789 firms. These firms employ about one million workers of whom 11 percent are foreign born.

The use of register data has a long history in statistics. Administrative sources (e.g., census) have been the basis frame from the very beginning of official statistics. However the use of administrative data only really took off from 1980 on. The trend is directly related to the

introduction of integrated information systems within NSOs. Population frames (e.g., business registers) have already been used for decades as a sampling frame in surveys but these frames can now become the backbone of the integrated system ('information warehouses') to which all information could be somehow linked (Kloek & Vâju, 2013).

Table 6. Occurence of CIS micro data linkages and barriers[11]

	CIS data linked (n=31)	Barriers (n=14)			
	yes	legal	lack of resources	Business Register not linked	Linking not explored
Business register data	87%	21%	0%	0%	14%
R&D survey data	77%	7%	0%	7%	36%
ICT survey data	65%	14%	7%	0%	71%
Structural Business Statistics data	71%	21%	7%	0%	43%
data from other sources	48%	n/a	n/a	n/a	n/a

The integration of other data sources in the statistical infrastructure of a NSO would at least require an information system that it would be able to identify and link population units (e.g., individuals and enterprises) across different internal and external data sets. This does not necessarily require a centralized register but at least a series of compatible registers. In the case of enterprises this requires a unique identification numbering system managed by business registers and used for every statistics included in micro-data linking programs (Sturgeon, 2014).

4.2.2 Third parties re-using CIS-data

For obvious reasons this Manual has been written from a CIS-centric view. It should however be noted that for most researchers, analysts and policy makers innovation is not their primary concern. From their point of view, innovation could be one of the potentially relevant background variables, and CIS is one of the data sources on business innovation.

[11] MERIT (2017). *Basic enterprise description variables & Linking CIS data*. CIS 2018 Task Force, Luxembourg, 4-5 april 2017

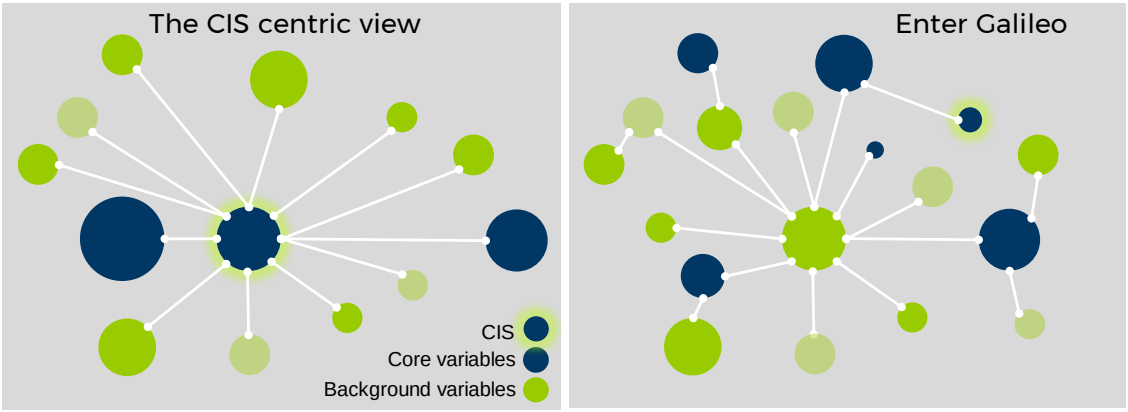
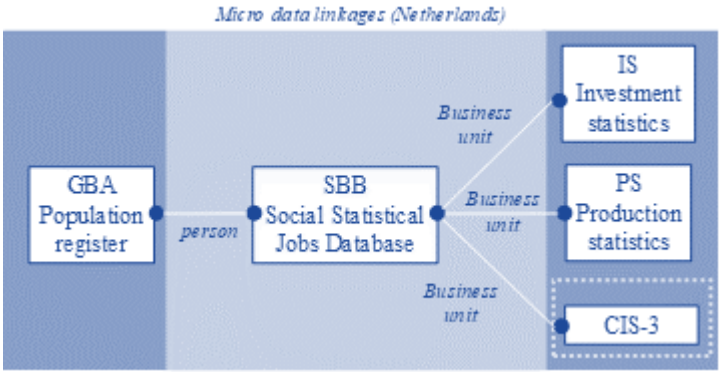


Figure 5. The relative position of CIS in the innovation data landscape

Having said this, CIS is rather unique as a harmonized cross-national data source and through the years it has become the de facto source for studies that use innovation as a background variable. Below is one example of a study on the relationship between skills and wages that used CIS micro data as one of the supporting variables.

Text box 6. Example of using CIS micro data as a background variable

In the cross-country study from Broersma, Koch and Rekveldt (2010) the share of highly educated workers from CIS-3 is an important piece of information, as it enabled the construction of 2-digit industry shares of high educated labour. These industry averages play a crucial role in comparing the industry aggregates of the contructed individual indicators. The CIS-3 share determined where the cut-off point of each distribution is located: above this point employees are assumed to have a high education (or skills). Using CIS-micro data for each employee by age class it could be determined whether she or he has a wage above the reference wage based in the CIS-3 share.



Ironically, to determine whether firms are innovative no core variables from CIS-3 were used. This is because in one of the countries that was compared CIS data was neither part of the panel on employees nor on businesses. Instead, innovativeness was defined by the extent to which a business unit has higher ICT investments than the sectoral average.

4.3 Preconditions for linking data

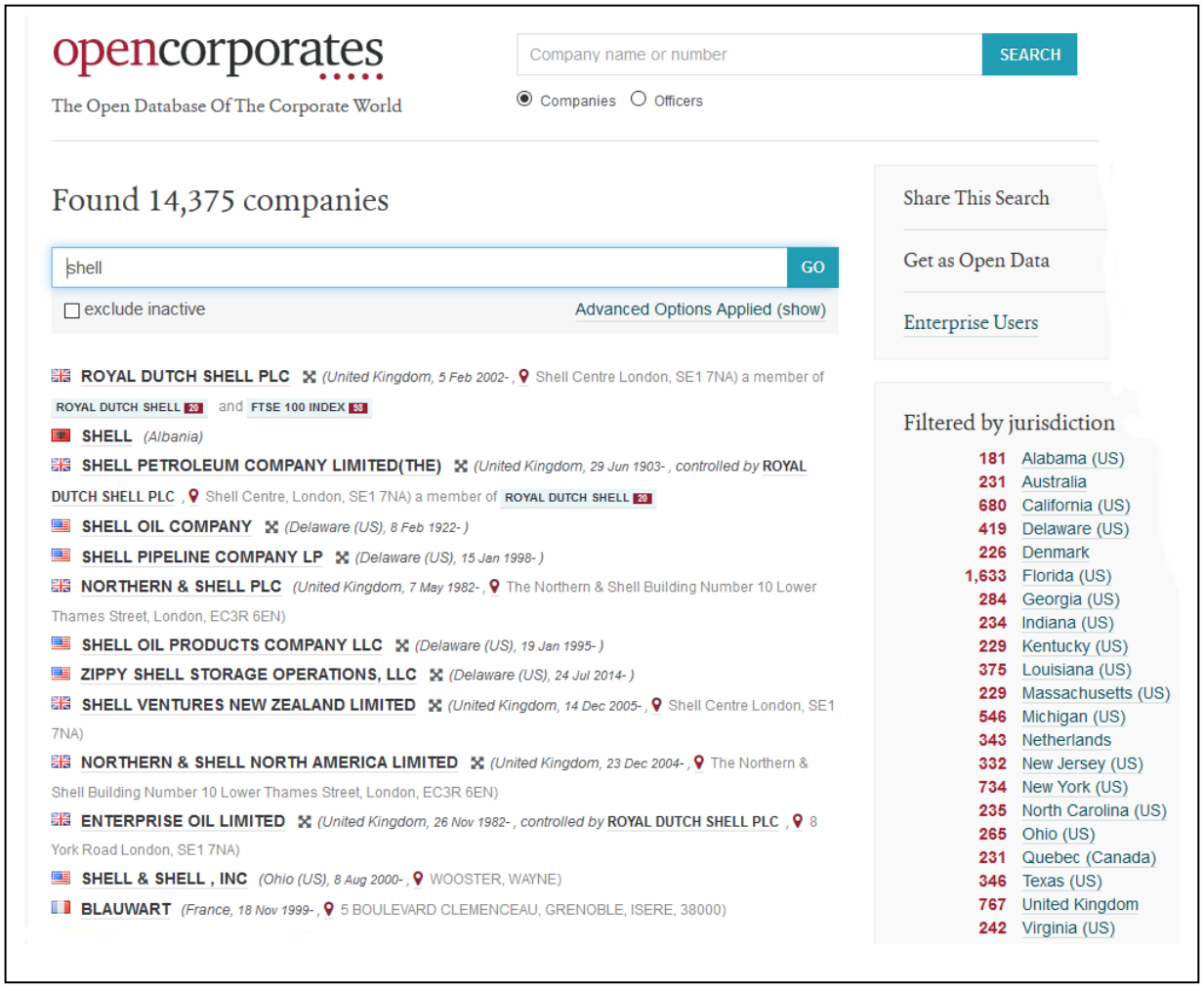
4.3.1 Harmonisation of data

Data integration is defined broadly as the combination of data from different sources about the same or a similar individual or institutional unit. A precondition to data integration is the harmonization of the data sources or dataset integration. Data cleaning and standardisation are crucial preparatory steps to successful data matching. This integration involves the adjustment of data sources at various hierarchical layers.

At the physical layer at the bottom, differences in *technology* need to be overcome. These differences can exist in hardware and software (operating systems, databases' structures and formats). However by using common standards as intermediaries, no further fine-tuning at the technology layer is needed.

The exchange of data subsequently assumes to rely on a common structure. That is, the *syntax* of the records – the way in which entities are represented (e.g., in terms of formats, measurement units, ranges etcetera) need to be harmonized. If there is no common syntax the data needs to be extracted from the data source as raw data and then be cleaned, standardized and eventually parsed into predefined formats and data structures. The objective of parsing is to segment each output field into a single piece of information (e.g., `COMPANY NAME`, `LEGAL FORM`) rather than having several pieces as a single field or attribute). Standardisation is particularly problematic in the case of names. Names are often spelled differently and/or companies (especially large ones) operate under many different names and have many subsidiaries. Much progress has nevertheless been made in name matching. One practical example is the OpenCorporates database that has nearly 100 million companies and that is run by just a couple of database administrators.

Text box 7. Example of name matching by OpenCorporates



At the final layer, differences in the *semantics* of the data need to be overcome. This requires the fine-tuning of meanings, concepts and definitions. Coming to terms with semantics is often the most difficult step in the harmonization process because it requires detailed adjustments and a good understanding of the domain from which the data originates.

At the European level, the EuroGroups Register (EGR)[12], that is part of the FRIBS initiative, works towards uniform definitions of enterprise information at three levels:

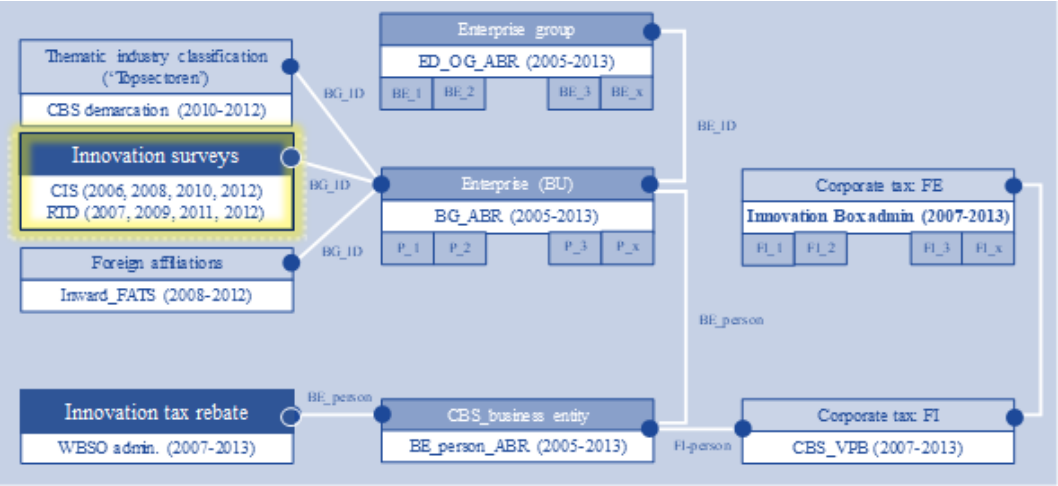
1. **enterprise groups**: identity, demographic characteristics, the structure of the group, the group head, the country of global decision centre, activity code (NACE), consolidated employment and turnover of the group.
2. **enterprises**: identity and demographic characteristics, activity code (NACE), number of persons employed, turnover, institutional sector;
3. **legal units**: identity, demographic, control and ownership characteristics.

[12] see http://ec.europa.eu/eurostat/statistics-explained/index.php/EuroGroups_register

Text box 8. Examble of CIS micro data linkage at three enterprise levels[13]

In the evaluation of a specific innovation policy grant ('Innovatiebox') Dialogic used an elaborate data infrastructure that spanned all three enterprise levels. This nested structure was needed to connect various relevant variables from several registers and databases that were each designed on a different enterprise level.

The major challenge was to link the administration of the actual policy instrument to enterprises. Since the administration was defined in fiscal terms a direct coupling with the core enterprise registers at CBS Dutch Statistics was not possible: fiscal units (FE) and business units (BU) do not exactly overlap. Therefore, the coupling was based on the underlying level of fiscal numbers (FI) and CBS persons (BE_person). Every BU has one or more CBS persons and every FE has one or more FI numbers.



Provided that the statistical infrastructure of a NSO would meet the basic requirements and that a centralized company register with unique IDs would exist, there seem to be no major hurdles to link input data (e.g., CIS data) with outcome data (e.g., on firm performance) from third parties (usually other administrative bodies such as Tax Authorities). In fact, many NSOs already facilitate such linking of data.

The key question for the implementation of the data linking will then be who will actually perform the linkage, and how the confidentiality of the micro data can be safeguarded (see hereafter, §4.5)

Limiting access to micro data is one measure. It means that NSOs have to screen every potential user of its micro data. NSOs then have to make sure that no confidential data is being brought outside the (physical and/or virtual) 'Research Room'. An alternative would be if the NSO provides the linkages and only publishes the aggregated data. However it is often a more practical approach when the end user itself (e.g., a researcher) links the data sets rather than the NSO (this is type 3 in Text box 9 below).

[13] Dialogic (2015). *Evaluatie innovatiebox 2010-2012*. Utrecht: Dialogic.

Text box 9. Current practices at NSO's in granting access to micro data to external researchers (2016).[14]

Type 1. *The NSO makes the linkage in-house and publishes the linked data*. Only one case was found in the sample, namely New Zealand. In the annual Business Operations Survey Statistics New Zealand publishes cross tables on types of innovators (CIS alike) x business performance (income, expenditure, profit). The linkage is based on micro aggregated data, not an a direct linkage between individual records.

Type 2. *The NSO makes the linkage in-house but leaves it to third parties (e.g., external researchers) to publish on the data.* This is also a rare scheme. Israel has a linked set with innovation and value added data. The data set is available for internal use or for the Research Room from the Central Bureau of Statistics.

Type 3. *The NSO provides data sets that can be linked but leaves the actual linking (and publishing) to third parties.* This is the most common scheme. At least the NSOs of Australia, Norway, Sweden and The Netherlands, provide such data sets to external researchers (usually via their Research Rooms). Since firms have a unique anonymous ID across all data sets linking the data is not a major challenge and is, in fact, being done on a frequent albeit ad hoc basis by academic researchers and research consultants who have been granted access to the Research Room.

Type 4. *The NSO does not provide data sets that can be linked*. For a number of practical reasons, there are quite some NSO's who do not provide data sets that can be readily linked. First, relevant data sets might not be available at all. Secondly, linking of data across authorities might not be allowed due to privacy regulation (e.g., Sweden). Thirdly, the country does not have a centralized company register (e.g., Germany) and/or unique IDs are missing (e.g., Belgium, Germany). Obviously, in the latter case matching on partial identifiers could still be used to link records, as for instance OpenCorporates is doing.

The main reasons why type 3 is that most widely found are three: (1) assuming an unique ID exists linking data sets is a trivial operation. It has little added value to offer this service; (2) a NSO has many different data sets hence many possible linkages are possible; (3) often a researcher has specific research questions thus needs a specific selection of variables, industry sectors etcetera. In short, although a NSO could very well offer one big generic linked data set (of which every researcher can make its own cross-section) this is often not the most practical solution.

4.3.2 Micro integration

Micro integration is a complementary part of the overall data integration process. Micro data could in principle be linked without improving the overall data quality of the integrated data set (Al & Thijssen, 2003). The purpose of micro aggregation is to compile *better* information than would be possible by using the sources individually (Bakker, 2011). Micro integration is intended to improve the outcome of record linkage and/or statistical matching. It is applied in situations where variables from different sources may have values that are incompatible or inconsistent with each other at the unit level. Some data sources have a better coverage than others or are just more reliable than others. In many cases there might even be conflicting information between sources at the record level. The basic idea of micro integration is to take the best data quality of several data sources. The final goal is to generate a dataset in which all perceived incompatibility or inconsistency had been removed (Dias, 2015).

[14] Dialogic (2016). *Improving the measurement of innovation outcome*. ESTAT/G/2015/006, Working paper 3.

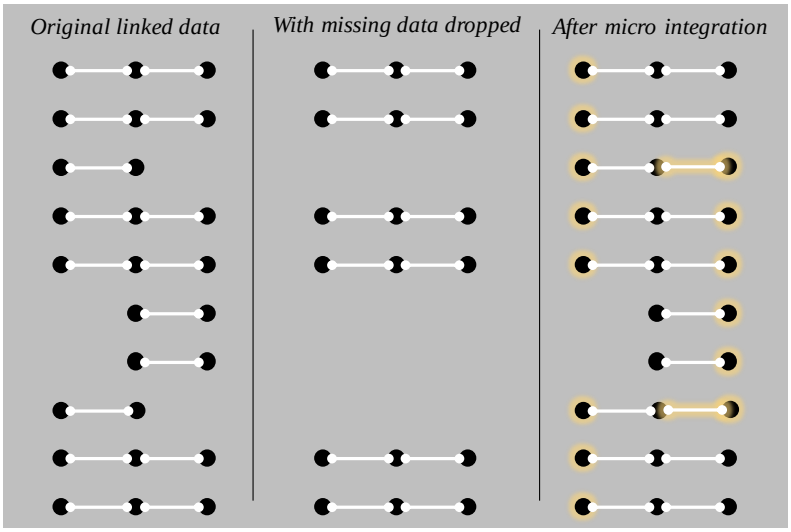


Figure 6. Schematic representation of micro integration

Proper micro integration is only possible in a fully integrated statistical infrastructure. In the ideal case, for a limited number of basic units (e.g., individuals, businesses, buildings) statistics are being complied by matching, editing, imputing and weighting data from the combined set of administrative registers and sample surveys. Since there are inevitably differences between data sources, a micro integration process is needed to check the quality of the data and to adjust for incorrect data. Ideally, the data source with the highest quality for a particular basic unit or variable is used as the overall quality benchmark for the entire system. Hence an important task of a statistical agency is not only to identify the widest range of available data sources but also to assess the strengths and weaknesses of each particular data source that could potentially be added to the system. Eventually, then, micro aggregation could provide far more reliable results because the integrated data are based on an optimal amount of information (Dias, 2015). The coverage of (sub)populations is also better because missing data in one data source can be filled by data from other sources (e.g., by statistical matching). Finally, the consolidation of data sources makes sure that a uniform figure is published for a specific unit or variable.

4.4 Different methods to link micro data

4.4.1 Overview

There are two basic methods to link data:

(1) *Statistical matching*: information on a unit with the same characteristics (hence a *similar* unit).
(2) *Record linkage*: a different set of information on the same unit (hence *identical* records);

Statistical matching can be applied at either the macro level (of groups) or the micro level (of individual units), record linking is by definition solely applied at the micro level. Since the current consensus view is that parameters fed into macro models should be based on solid microeconometric evidence where possible, the availability of micro data is increasingly important in both academic and policy research (Beyer et al., 2013).

The two methods are applied to different types of input data. Record linkage is used when the data sets that are linked have at least a partial overlap in units. In the particular case of

register data, the data set usually covers the entire population. Thus there is a near complete overlap between the two data sets. The units are *directly* linked at the record level. Records can be linked one to one, one to many or many to many. Statistical matching is used when there is no overlap between the units (hence there are two independent samples). In this case, statistical matching is the only possible method to link units.

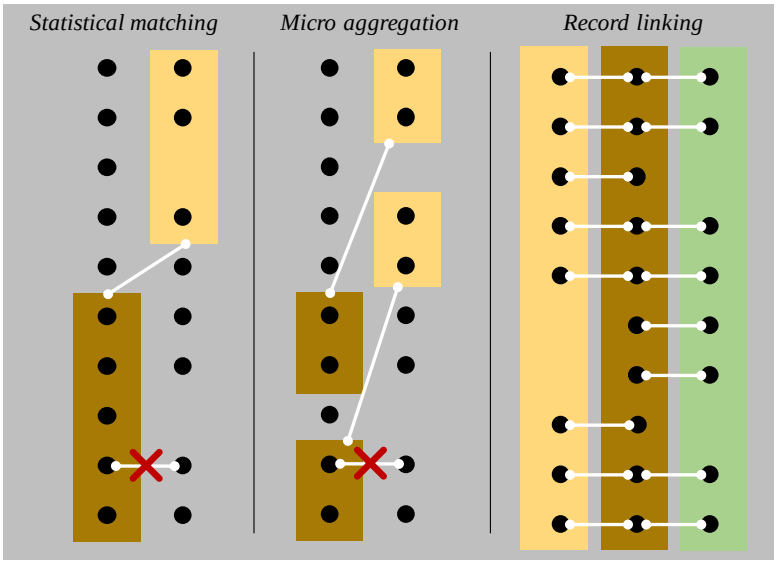


Figure 7. Various ways for an NSO to collect data (Dialogic, 2017)

Record linkage and statistical matching also generate very different types of output data. The units in the output data from record linkage refer to *real-world* entities, that is, individuals or firms that really exist. Obviously, this has severe privacy consequences (see hereafter, §4.5). Statistical matching at the micro level combines two different real-world entities into a new *virtual* unit.[15] This means that there are less privacy concerns. The drawback of statistical matching is that the combination of the data is tailor-made for every analysis. This means that the virtual units that are being generated in the output cannot be re-used for a new analysis – unless the same common variables are being used. Therefore record matching is a more flexible and permanent solution for data integration.

4.4.2 Statistical matching

Statistical matching techniques are used to combine information that is available in distinct data sources (e.g., a CIS dataset and an external data source with outcome data) that refer to the same target population (i.e. firms or a specific subset of firms). An important condition is that the units in the two data sets *do not overlap*, hence direct matching via record linkage is not possible (see herfore, §4.4.3). Note that in the case of public registers the coverage of the population is usually nearly complete, and often unique keys are available as well. Consequently for matching a set of innovation micro data with register data, record linkage is usually the most appropriate method.

Thus, if there is no overlap between the data sets that need to be combined, record linkage cannot be applied but the critical condition for statistical matching ('no overlap') is exactly

[15] hence the alternative label for statistical matching: 'synthetic' matching.

met. The purpose of statistical matching is to study a relationship among *variables* that only occur in either one of the data sets (Y in data set A and Z in data set B). The actual matching is being done on the basis of a variable that occurs in both datasets (the joint variable X). The first condition stipulates that Y and Z are conditional independent given Z. That is, Y and Z should *not* be jointly observed.

In the *micro* approach of statistical matching, a completely new micro-data file is created where data on all the variables is available *for every unit*. Based on the common variable X, for every unit with variable X variable Z is imputed or, the other way around, Y is imputed into units with variable Z as well.

The joint variable X is used to select samples of the two data sets that have the greatest resemblance, that is, as much as possible similar covariate distributions. The strategy is to select those samples for the bias to the covariates is minimized (Stuart, 2010). The matching can be done multiple times and the matched samples with the best balance – that is, the most similar samples of data set A and data set B – are chosen as the final matched samples.

Text box 10. Using propensity scores to calculate the joint variable[16]

> For the calculation of X, propensity scores are often the best available method (Rosenbaum & Rubin, 1983). Propensity scores summarize all of the covariates into one scalar: the probability of being treated. The propensity score is defined as the estimated conditional probability of a unit to belong to data set A (dummy value = 1) or data set B (dummy value = 0). A logit or probit model is estimated with the dummy as dependent value, and the common variables X as independent variable, obviously including the regression constant. Then, for each recipient record a donor unit is searched with the same or the nearest estimated propensity score. Next to mixed methods such as propensity scoring, other types of micro-matching methods are hot deck imputations or regression based models.

The crucial point is that the optimal composition of subsets of units from A and B is dependent on the target variables Z and Y (that are being studied). The choice of the target variables influences all subsequent steps in the matching process (Dias, 2015). This means that the matching should be tailor-made for each specific analysis and furthermore, that the variables Z and Y should be a priori known. This makes statistical matching much less flexible than record linkage. In the latter case, the selection of the subsets can be optimized for the specific purpose of the study and new variables can always be added later on. Data that is being matched on the basis of individual records allows the reuse of existing data sources for new studies. This is usually not to case for data that is combined on the basis of statistical matching.

4.4.3 Record linking

Record linkage is the task of identifying and matching individual records from different databases that refer to the same real-world entities or objects. Records are matched on the basis of a unique unit identifier. In the ideal case, there is already a generic unique ID available (e.g., PIN, business registration number, VAT number) – and it is allowed to use the number as a key to couple data.

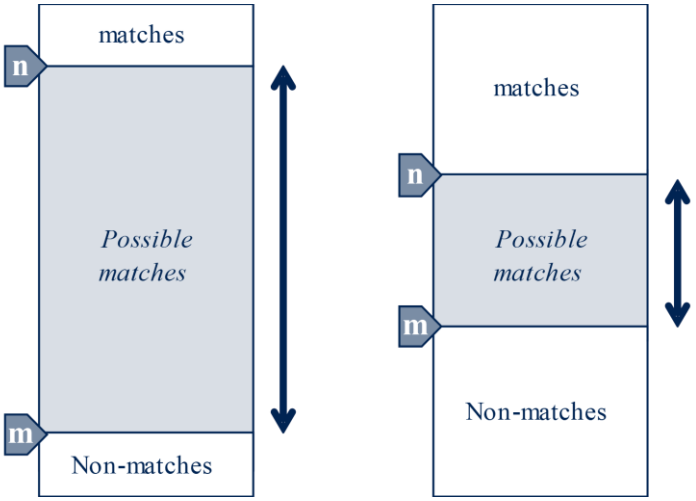
[16] For an overview, see (Eurostat, 2013).

If such a unique key does not exist, records can still be matched by combining several fields of the record, a.k.a. characteristics of the unit.[17] It is the specific *combination* of the supporting fields (*'partial identifiers'*) that needs to be unique, not the fields itself. These partial identifiers should preferably be unique for each unit, available for all records (universal), stable (permanent), recorded easily and without errors (accurate and non-sensible), and simply verifiable (transparent) (Dias, 2015). This obviously assumes a high data quality.

However, in the frequent presence of administrative errors (e.g. due to difficulties with the standardization of names or due to highly dynamic population thus outdatedness of records) it might not be possible to use *deterministic matching*. If data is noisy and contains random errors but there is an array of partial identifiers that could be used for blocking and record matching, one can still resort to *probabilistic matching* (Fellegi & Sunter, 1969). In the latter case, matches (A=A) or non-matches (A≠B) between individual records are not perfect but instead have a probability ('similarity value') between 1 (A=A) and 0 (A≠B). For each candidate record pair *several* attributes are generally compared, resulting in a 'comparison vector' of numerical similarity values for each pair.[18] If the probability is close to 1 there is a *possible* match between the two records (A=a?).

Text box 11. Setting treshold scores in probabilistic matching

In probabilistic matching, the critical problem is to determine the optimal threshold scores for matches (m) and non-matches (n). Obviously, the higher m and the lower m, the larger the number of *possible* matches that need to be further scrutinized. The threshold scores can either be determined by trial and error or by using a model based approach (as proposed by Fellegi and Sunter). The optimal setting depends on the specific characteristics of the data sets hence needs to be tailor-made. The determination of the threshold scores is an iterative process in which the number of type I (false positive, A=B) and type II (false negatives, A≠A) errors are minimized.



[17] In a similar vein, by combining several fields the identity of individuals or individual firms could still be deduced from anonymised data. When multiple high dimensional datasets are being coupled the intersection becomes so small that anynomity can no longer be guaranteed (k-anonymity) (Torra & Navarro-Arribas, 2015).

[18] Compare the use of propensity scores in statistical matching.

To assess the accuracy of the matching, in information retrieval the measures of *precision* and *recall* are often used. Recall is the proportion of positive cases that were correctly identified (or pairs correctly matched, hence recall refers to pairs completeness). Precision (or pairs quality) is the proportion of the predicted positive cases that were correct. Both recall and precision should be high but there is a trade-off between the two measures.[19]

Whereas automatic matching is used by many NSOs as a cost-efficient approach for record matching in bulk, clerical intervention is still needed for the proper resolution of controversial matches. The range of automatic detection could be increased – and thus the deployment of expensive human agents minimized – by optimizing the data preparation (e.g., by additional investments in the data wrangling process) and especially by improving the intelligence of the automated agents.

The drawback of linking records without unique ID is it that it has to be tailor-made every time two or more data sets are being coupled. Moreover, record matching becomes more difficult when the number of data sets increases. The most efficient solution is therefore to assign a unique key to a record once a definitive match has been made. This is usually part of the micro integration process (see §4.3.2).

When a system of registers with unique identifiers has been established, a NSO could in principle combine any register (and census) at any time. This does however require a careful management of the IDs within the statistical infrastructure. For instance, for each of the base registers a *standardized population* or population frame has to be created (Wallgren & Wallgren, 2011). Thus, changes in external registers which could affect the matching precision should be closely monitored. Old and new IDs should for instance be included in a cross reference table together with the reference time when the change occurred.

In the specific case of firms the dynamics in the population are high, thus frequent updates are needed. At the same time, administrative sources for business statistics (such as Chambers of Commerce directories) usually are mainly interested in registering the changes of status and/or in reporting the formal (legal) status of a firm, mostly based on ownership, rather than monitoring its evolution overtime in terms of size and economic activity. For statistical purposes the economic continuity is more important (Kloek & Vâju, 2013). This means that different *continuation rules* will have to be adopted for the base register and the standardized population of firms that is being used as the backbone of an integrated information system of a NSO.[20] The standardized population is therefore an accurate but not an exact copy of the business register.

[19] The full set of measures and derived measures for assessing the accuracy of information retrieval or matching is given by the confusion matrix or error matrix – see Technical annex.

[20] For instance, firms that are still registered as active firms in the business register but that not have submitted VAT declarations for, say, the last two quarters could be dropped from the population frame of firms that is being used by the NSO.

4.5 Privacy and security

4.5.1 Record linking and privacy

In information architecture design the golden standard is to link at the level of individual records that have a universal definition and a globally unique ID.[21] In principle, any new data source or new attribute can then – at any time - be linked to the set of units.

The possibility to link across any data set is also the biggest drawback: this comes with major security and privacy concerns. Linking confidential data at individual level across data sources and data holders is a hazardous venture in terms of privacy and security. This is the price one has to pay for maintaining flexibility.

There are several measures to protect privacy of personal and confidential data but privacy can never be completely guaranteed. At some place in the data infrastructure a coupling with the original unencoded and unencrypted records must be made, thus there is always a theoretical possibility to trace back information to a specific individual or a specific firm.[22] The only fool proof solution would be to use synthetic units (see §4.5.3). The integrated data would no longer be suitable for *administrative* purposes (as it does not refer to real-world entities) but it could still be used for policy making and research purposes, which makes up about 50% of all usage of business statistics (Eurostat, 2015). However one can never be sure to what extent the statistical matching or micro aggregation influences the eventual results of the analyses of the linked datasets. The biggest drawback is the loss of flexibility. The only basis of an integrated statistical infrastructure is record linking.

The two basic measures to protect privacy are to *limit access* to the data or to *conceal* the data. In many countries, only data that do not allow for the identification of individuals (or individual firms) can be made publicly available. At the input side this means that re-use of data sources with personal of confidential data is forbidden altogether (hence the data source is not available) unless an exemption has been made for re-use for official statistical purposes. At the output side this means that data should be aggregated in such a way that it is impossible to reverse engineer the aggregation process (Torra & Navarro-Arribas, 2015).

Ideally, control of the data is at the lowest level. Thus, it is the original data holder (usually the producer or the owner) that keep a full control of what data to release, to whom and in what manner. Thus eventual anonymization of the data should preferably already been done by the data holder, prior to the data exchange with the NSO. However identifiers still have to be known to the NSO otherwise records cannot be linked. In case unique IDs are not available the matching of databases needs to be done on the basis of partial identifiers (see herfore, §4.4.3). However the most suitable (universal, stable, accurate) auxiliary fields to be used for matching usually also contain most confidential data (names, addresses, dates

[21] For example, Universally Unique Identifiers (UUIDs) in software (http://www.ietf.org/rfc/rfc4122.txt)

[22] The NSO typically has a master table in which the original identifiers (company code) is being linked to an anonymous ID that enables the linking of records without revealing the identity of the firm. The actual linking can also be done without revealing the identity of the firm, for instance by using hash tables. Still, even if anonymous IDs are being used users would in principle be able to deduce the identify of individual firms by combining various fields. When multiple high dimensional datasets are being coupled the intersection becomes so small that anynomity can no longer be guaranteed (Torra & Navarro-Arribas, 2015).

of birth or establishment etc.). If there are unique IDs available (social security numbers, business register numbers) in many countries it is not allowed to use these keys to link with external data sources – ironically because they make linkage so easy.

Text box 12. Record linking by parallel IDs

> There is a workaround for NSOs, namely to use a parallel set of unique IDs. In this way, registers can be linked internally, within a NSO, without using the original ID as a key. Thus, in the linking of data the identity of the firm does not have to be revealed. However, at some place in the NSO a linkage table with the original IDs still needs to be kept. Security is maintained by physically limiting access to the data. That is, only a selected number of individuals are authorised to get access. In a similar vein, a NSO could limit the access to the anonymised micro data to on-site use only and require prior screening from the external users.

The best option to preserve the confidentiality of the micro data would be to use *privacy by design*. Again, ideal control of data should be kept at the lowest level. Currently, this is the level of original data holders. But these are still data aggregators. The actual lowest level is that of individuals or firms, i.e. the real-world entities to which the records in the micro data refer.

Text box 13. Use of block chain technology to control data access at lowest level

> There are a number of recent developments that enable to transfer micro-data control to the lowest level of individual users. We are referring here to the emergence of *distributed information systems*, with distributed hash tables and the block chain technology as the most notable implementations. These technologies use the network *as a whole* to verify the legitimacy of data operations rather than some kind of centrally-authorized actor (such as a NSO or a trusted third party, TTP).
>
> The use of distributed ledgers allows citizens and firms to manage the access to their data and to know who has actually accessed the data. The security and accuracy of the ledger is maintained cryptographically (using proof-of-work or proof-of-stake schemes) to enforce strict access control. In essence this allows for the consensual use of personal or confidential micro data in anonymous form for collective intelligence, such as re-use of statistical data for research purposes. Blockchain applications such as Ethereum, for instance, enable to use of so-called 'smart contracts' to create a permanent, publicm transparent ledget system for compiling all sort of personal data (e.g., rights data, digital use data etc.) (Buterin, 2014).
>
> In essence, this allows for the consensual use of personal or confidential micro data in anonymous form for collective intelligence, such as re-use of statistical data for research purposes.

4.5.2 Ensuring confidentiality of administrative data

Confidentiality is a critical issue for the re-use of administrative data. Personal and business data is often protected by the basic principle of purpose limitation. This means that an exemption should be made for statistical purposes. Furthermore, the NSO should ensure that the identity of individuals or firms is protected at all times. Data protection legislation often demands additional steps: anonymisation and dealing with access right (Kloek & Vâju, 2013). Anonymisation should ensure that the data that is being published should never be deducible to a specific individual or firm. Access to secondary data is usually restricted to a known of approved users whom credentials are known (that is, they have been screened) and who (re)use the secondary data is in a controlled environment (e.g., a Research Room). The screening (and subsequent permission to access) could be done by the government agency that holds the administrative data or it could be delegated to the NSO.

4.5.3 Ensuring privacy by micro aggregation

Micro aggregation is a statistical disclosure technique that has been introduced by Eurostat researchers in the early 1990 (Defays, 1997). The basic principle is to split a population in as small as possible groups so that the resulting aggregates cannot be traced back to an individual unit of the population while these composite units still behave similar to the original individual units. That is, a 'virtual person' or a 'virtual firm' is being created that has the same characteristics as the real firms of which the composite unit is constituted. This could be regarded as the statisticians' way of 'privacy by design'. Provided that multivariate (rather than univariate) micro aggregation is being used, confidentiality of the data is ensured because it is embedded into the very design of the statistical technique.[23] That is, the anonymisation of the data is already been done *at the source*, before datasets are being merged. This is a definite advantage over the IT-based method or record linking that uses individual records that refer to real-world entities.

At the same time, similar to micro integration, this is also the biggest drawback of the method. Because the linkage is being done at the level of 'virtual firms' quite a lot of flexibility is been lost. Direct matching with other years for the same datasets (e.g., to facilitate panel designs) or with new data sources (e.g., that contain other types of outcome variables) is no longer possible. The matching is always on a limited number of generic background variables (e.g., firm size, sector, year) that are defined beforehand. However, provided that the size of the micro aggregated cells would be kept reasonably small, there will be a lot of cells (a.k.a. 'virtual firms') available for analysis, hence it will be possible to make a lot of different cross-sections.

Text box 14. Example of micro aggregation (ESSLait)

> Although the micro aggregation can be done on the basis of generic variables this is not the optimal solution. Ideally, the micro aggregation should be tailor-made for the type of analysis that will eventually be conducted on the micro aggregated data (Lamarche & Pérez-Duarte, 2015).
>
> For instance, in the ESSLait project the data has been prepared for the specific purpose of conducting ICT impact analysis (Hagsten, Polder, Bartelsman, & Kotnik, 2013). The input variables have been chosen beforehand and conveniently fixed. This is done because there is a trade-off between the number of variables that are used to compute the Euclidean distance and the errors that occur due to the micro aggregation. One of the main difficulties in micro aggregation is the clustering process of how to select similar units ((i.e. reducing intra-cluster variance as much as possible) while ensuring a sufficient but not too high number of units in each cluster (Lamarche & Pérez-Duarte, 2015). An optimal clustering would require that the variables that are used for the calculations are fine tuned to the specific analyses which will be performed. However, this assumes that the purposes of the analyses are already known beforehand, and this is obviously not always the case.

Micro aggregation has another application besides preserving privacy. A condition for record linking is that the definition of the units is harmonized (see herfore, §4.3.1). In practice, the condition is often not (yet) met. For instance, the definition of the basic unit ('enterprise')

[23] Micro aggregation ensures k-anonymity only when multivariate micro aggregation is applied processing all the variables of the data file at the same time. Otherwise, this is not ensured. In fact, it is often the case that k-anonymity is not ensured. This is so because the set of variables is often partitioned, and micro aggregation is applied independently to each partition element. This is done to achieve a lower information loss (higher data utility) than when applying it to the whole set. In this case, a trade-off has to be found between the information loss and the disclosure risk (Torra & Navarro-Arribas, 2015).

differs across countries because data is collected in different ways.[24] In this case, micro aggregation can be used to group units that are not exactly identical into synthetic units that are 'similar' across data sets.

[24] A relevant development is the changes that have been made to the FRIBS roadmap. In response to the concerns expressed by many National Statistical Institutes, it has been decided to exclude an update of the definitions of statistical units from FRIBS. Instead and in parallel, Eurostat has launched immediate measures for helping Member States in complying better with the existing statistical units definitions in each of the statistical domains and the Business Register (Eurostat, 2015b).

5 Indicators and analysis of innovation data

5.1 Data analysis

5.1.1 Different types of analyses

Few national statistical offices (NSOs) prepare *analytical reports* based on innovation survey data.[25] It is rather done by research centres (such as JRC), universities, think-tanks and reserch consultancies, usually to support policy-making purposes. However, it is useful for NSOs to understand the type of analysis done by the users of innovation surveys' data.

Innovation analysts use a diversity of statistical approaches:

- *descriptive statistics*, based on extrapolation of sample data (using the information of the sample design) and usually disseminated in the form of aggregate tables[26];
- *model-based estimation*, specifying relationships between a set of explanatory (exogenous) and dependent (endogenous) variables: this can be applied to individual firm-level data or to aggregates, and is based on a set of stochastic assumptions on the distribution of variables, errors and on the form of relationships (linear, non-linear, etc.).
- *algorithm-based methods*, such as machine-learning methods (see hereafter, §5.3) which are based on computer-intensive data processing, especially in the case of large or complex files (e.g. patent or bibliography analysis). The use of machine-learning methods is still in its early stage in innovation analysis. A drawback could be that the rather complex underlying methods are less comprehensible to non-technical users.

[25] This is due not only to the mission of NSOs, which may not include the provision of analysis but just the dissemination of tables and microdata for different categories of users, but also due to a general lack of skills in econometric skills.

[26] The Eurostat database accessible at http://ec.europa.eu/eurostat/web/science-technology-innovation/data/database allows users to obtain tabulation of data from all CIS editions, with a number of pre-defined tables customizable for breakdowns (by sectors, countries, size class, etc.) Users can then export the tables (data and metadata) into a variety of formats.

Text box 15. A machine-learning analysis of the determinants of innovativeness of countries (Czyzewska, Szkola, & Pancerz, 2014)

Based on time series of innovation data aggregated at the country level, Czyzweska et al. (2014) propose a clustering method which use an unsupervised machine-learning technique called Self-Organizing Feature Mao (SOM). This method identifies and assesses the correlation among a range of indicators (in this case, time series of the variables used to compute the Innovation Union Scorecard 2011) which are thought to be determinants of innovativeness. Variables related to both the input and the output of the innovation process are considered: investments in R&D, cooperation in the process of innovation introduction and intellectual property rights protection (input), employment in knowledge-intensive activities, exports of high-tech products and services, new to markets and new to firm innovations, revenues from licensing and patenting (outputs).

The method is based on iterative adjustment of clusters grouping countries by similarities in the evolution of each variable, then grouping variables also by similarities. The method is computationally intensive and based on neural networks.

The interpretation of the results of SOM is made in terms of correlations between variables. The capacity of explaining causality relations between variables is however limited.

5.1.2 Level of access to data

The abundant literature on innovation is largely based on the statistical analysis of survey data (Fagerberg & Mowery, 2006), both at the aggregate level (sectoral data, country aggregates presented as summary statistics or indicators) and at the micro data level, thanks to the availability of anonymised datafiles provided by statistical offices (see Text box 16 below).

The type of analysis than can be applied is largely depend on the level of access to the data:

- *access to microdata* allows for adjusting models (such as linear regression and its derived models) with estimation procedures that take into account the observed joint distribution of variables at the individual firm level.[27] The results of the model can be extrapolated to (1) the *sample* only, if the method is not adjusted for the sample design (2) the firm *population* if the method is adjusted for the sample design;
- *access to tabulated data* allows adjusting models at the sector, size class, region, country levels (or other aggregation), but cannot be immediately used to infer innovation behaviour at the firm level (see below, §5.1.3). The international comparability of results depends on the industry composition of the countries.

[27] In the previous chapter 4 the possibilities to link CIS micro data to other data sources are described.

Text box 16. Access to CIS micro data[28]

Access to CIS microdata files for scientific purposes is allowed by Eurostat and the NSOs under the European Statistics Law. This possibility depends on overall microdata availability at Eurostat (CIS microdata provisions are voluntary), Member States' willingness to allow the CIS microdata to be offered for the research use and the permission for using the data for the particular research project. Once the request is approved, the researchers get partially anonymised data in CD-Rom or other supports (CSV or Stata formats), or are authorised to work in the Eurostat's 'Safe Centre' in Luxembourg.

Microdata of the CIS surveys are released normally 2.5 years after the end of the survey reference period, due to the time needed for processing and anonymising the data. The process of anonymisation includes primary and secondary confidentiality:

- *Primary confidentiality* concerns tabular cell data, whose dissemination would permit attribute disclosure. The two main reasons for declaring data to be primary confidential are: too few enterprises in a cell or dominance of one or two enterprises in a cell with respect to the tabulated variable.
- *Secondary confidentiality* concerns data which is not primary disclosive, but whose dissemination, when combined with other data permits the identification of an enterprise or the disclosure of an attribute of the enterprise.

In addition, any statistics (tables, graphs, textual references) on any kind of subpopulation (cell) shall not be published: (1) if they consist of less than 10 enterprises; (2) where one enterprise represents more than 70% of the total sub-population expenditures, employment or turnover; (3) where two enterprises represent more than 85% of the total sub-population expenditures, employment or turnover.

5.1.3 Ecological fallacy

With particular reference to tabular data a core issue is that models which are estimated from *aggregated* (e.g. averages at sector or country level) data cannot be used right away to infer results and draw conclusions that are valid at the *individual* level (e.g. companies). This is due to a statistical issue called 'ecological fallacy' which describes the phenomenon of obtaining correlations of opposite signs between variables when measured at the individual and aggregate levels.

A visual example of this type of fallacy occurs in the X-Y plot below. If the inspection of the data is limited to aggregated data (averages) this would lead to misleading conclusions about the individual units. The example shows simulated data of per capita intake of a certain food product against obesity. Observing only the average values (centroids) of each of each agrgegation seems that more product intake implies more obesity (black dotted line). However, plotting individual data a negative correlation in each subpopulation is observed.

[28] The procedures for requesting access are described in: http://ec.europa.eu/eurostat/documents/203647/771732/How_to_apply_for_microdata_access.pdf

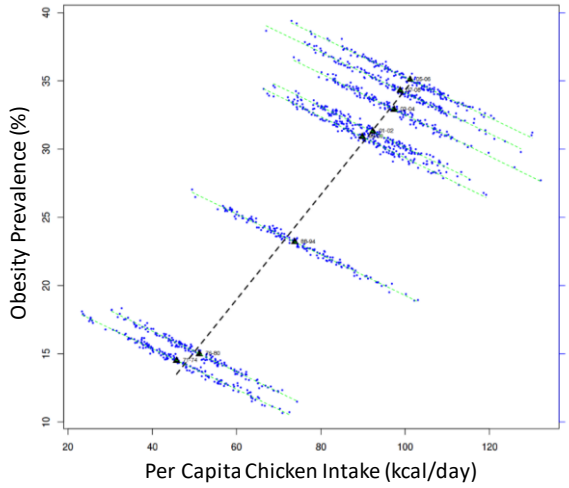


Figure 8. A visual example of ecological fallacy

Ecological fallacy can also be illustrated with a numerical example. In the next tables, fictional aggregate data are tabulated to study the relation between innovativeness and the presence of foreign capital in three industries.

Text box 17. A numerical example of ecological fallacy[29]

Industry A

	NON-INNOV	INNOV	Total
With foreign capital	150	450	600
Local capital	450	450	900
Total	600	900	1500

Industry B

	NON-INNOV	INNOV	Total
With foreign capital	300	450	750
Local capital	450	300	750
Total	750	750	1500

Industry C

	NON-INNOV	INNOV	Total
With foreign capital	450	450	900
Local capital	450	150	600
Total	900	600	1500

The proportion of businesses with foreign capital in industry A is 600/1500 = 40% and the proportion of innovative companies is 900/1500 = 60%. In the same way, prevalence of foreign capital is 50% in industry B and the proportion of innovative firms is also 50%. In industry C, the proportions are respectively 60% and 40%. When observing aggregate data at industry level (table below), an analyst would draw the conclusion that the presence of foreign capital hampers innovativeness *in all industries*, with a perfect correlation of 1.

Industry	Proportion of enterprises with foreign capital	Proportion of innovative enterprises
A	40%	60%
B	50%	50%
C	60%	40%

However, this conclusion is faulty for the individual companies *within industries*. When the odds ratios are calculated for each industry as a measure of association between the two variables ("presence of foreign capital" and "innovativeness") it follows that these is a *positive* assocation. This contradicts the earlier inference that there is a *negative* assocation.

[29] See the Technical annex for a description how to calculate the odds ratio.

5.1.4 Data visualisation

In addition to the traditional phases in policy research of data collection, data analysis and reporting recently the *visualisation* of the results has grown in importance. Although these could still be regarded as a part of the reporting phase the production of concise graphical summaries that appeal to a wider audience is evolving into art in itself. The creation of these so-called 'infographics' not only require statistical analysis but also design and communication skills. Some examples in the realm of STI statistics are shown below.

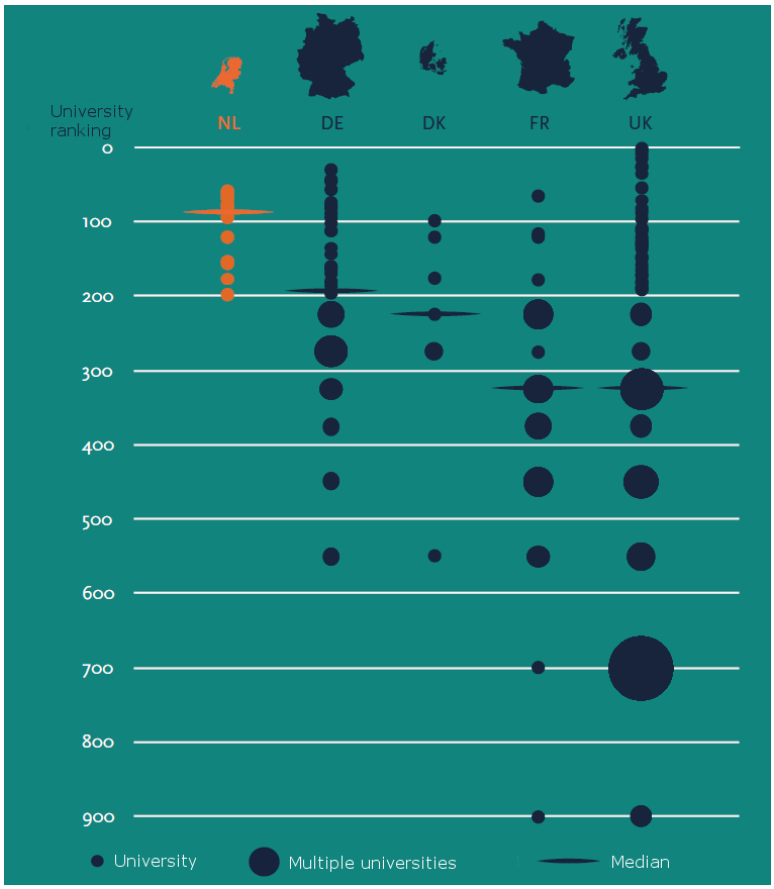


Figure 9. Visualisation of distribution of university rankings by country (KNAW, 2017).

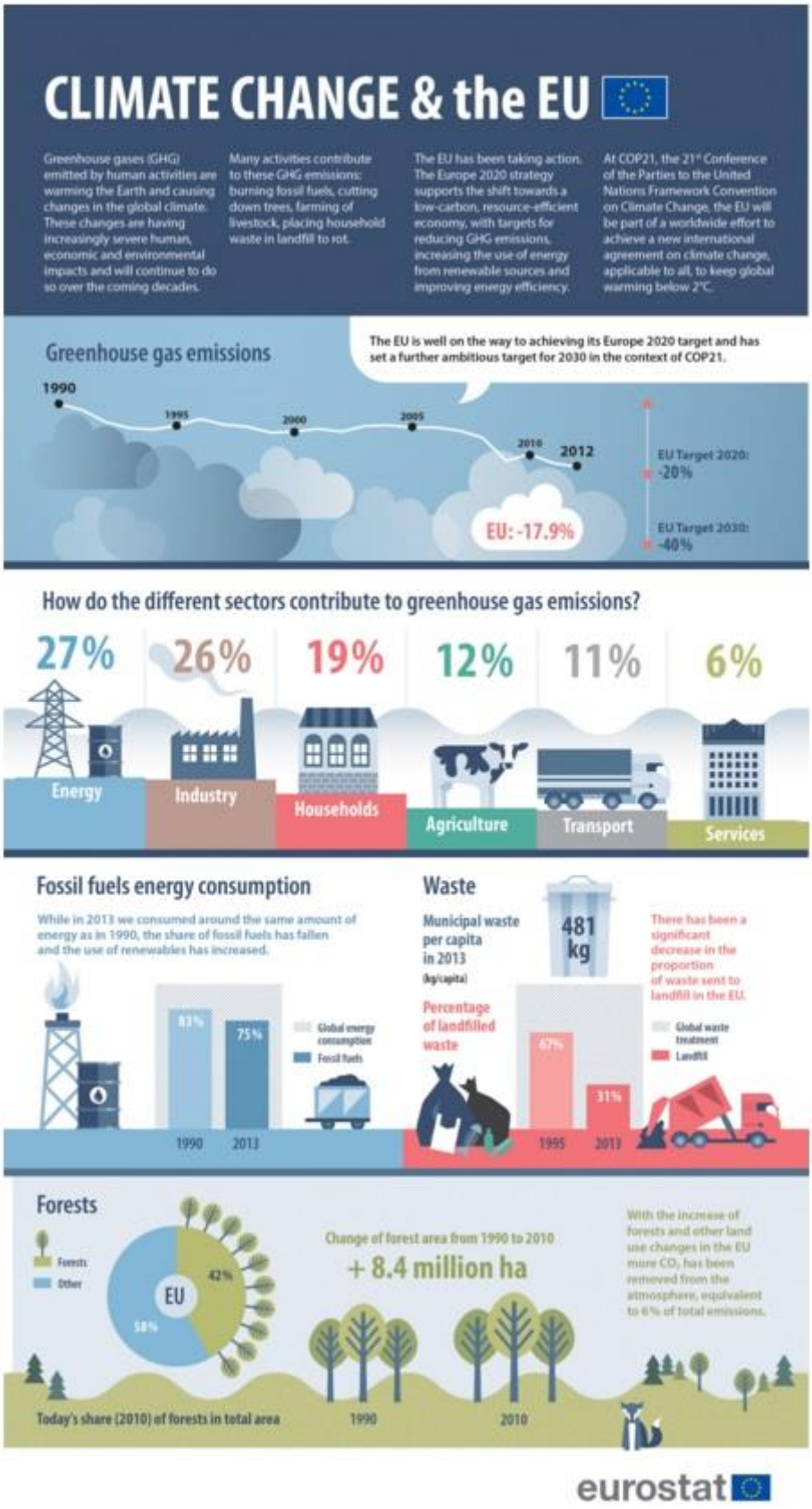


Figure 10. Infographic prepared by Eurostat on Climate Change

5.2 Measuring innovation intensity

5.2.1 Introduction

Innovation intensity is a concept widely discussed in the literature on firm management, microeconomics and growth models, and many of the studies that are based on CIS data use this concept.

There is no consensual definition but at the firm level most analysts distinguish the *propensity to innovate,* i.e. the decision of whether to undertake innovative activities or not –, from the next decision, namely how many resources (financial, human, technological, organizational, etc.) to allocate to innovation – *innovation intensity*. The latter is generally compared with either the total of the activities of the firm or to the average within the sector in which the firm operates.

The measurement of those concepts (propensity and intensity) at the firm level is then made by producing quantitative indicators that reflect the *probability* that a given firm undertakes certain innovation activities (with or without success) given its characteristics and environment (sector structure, government policies, etc.), and the *amount* of resources devoted to innovation.

The CIS2018 questionnaire allows for the identification of the innovation-related activities (e.g., knowledge-based asset creation activities). Simple estimates of the *propensity* to carry out such activities can be obtained by calculating sample percentages of firms (by sector, size, etc.) extrapolated on the basis of the sampling design.

5.2.2 Simple descriptive analysis

NSOs can produce cross-tabulations by firm characteristics from the percentage of firms that indicated to have allocated resources to innovation. Expenditure on innovation can be described using various scales, e.g., allocated or not, absolute amount spent, or as a percentage of total firm expenditure (see Text box 18).

Text box 18. Example of aggregate table with percentages on engagement

Let Exp(i) the expenditure in 2018 for category (i) (i= "machinery, equipment and buildings"; "marketing and branding"; etc.);

Let Share(i) the share of Exp(i) used for innovation;

Calculate ExpInn(i) = Exp(i) x Share(i) to obtain the expenditure in category (i) used for innovation;

Calculate Engagement(i) = (ExpInn(i) > 0), a binary variable that is equal to 1 if the expenditure in category (i) used for innovation is non-zero, and 0 otherwise;

Cross-tabulate Engagement(i), by industry and size, aggregating the individual data as total count or as ratios:

Industry	A: Total number of firms	B: Total number of firms engaged in category (i)	C: Percentage of firms engaged in category (i)
…	…	…	…
Industry *j*	Count of firms with Industry = *j*	Count of firms with industry = *j* and Engagement(i) = 1	(B)/ (A)
…	…	…	…

5.2.3 Modelling the engagement of a company in a certain activity and the resources allocated to it

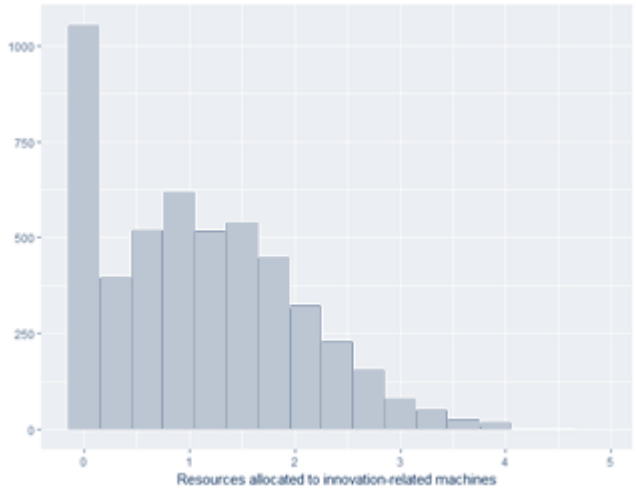
The use of micro data allows for more complex analytical models for the estimation of the probability of a firm to engage in (i.e. to allocate resources to) knowledge-based asset creation activities. This probability can be obtained by econometric methods with binary dependent variables, such as *logit* or *probit* models with exogenous variables. In particular, the analysis of the propensity of engagement in a certain innovation-related activity can be enhanced by linking to variables in the Statistical Business Register or other surveys at the firm-level (see herfore, §4.1.1).

The Tobit model has been used for different applications in the analysis of innovation and technology adoption (see Text box 19). For example, Some econometric studies combine the estimation of the probability of undertaken a certain activity with that of the amount of resources allocated (Mohnen & Röller, 2005) (Mairesse & Mohnen, Working paper series #2010-023, 2010) (Crespi & Zuñiga, 2010). Another example include the comparison of exporters and non-exporters in terms of technology and success in innovation (Nassimbeni, 2001). The author used a Tobit model instead of OLS because the dependent variable had a *censored* distribution (a firm is named as "exporter" if the exports-to-sales ratio was above 0%, all the remaining non-exporters have a ratio of 0%). A similar example is a study to test whether (mangrove rice) farmer perceptions of technology-specific characteristics significantly condition technology adoption decisions (Adesina & Zinnah, 1993). Here, the Tobit model was used to measure not only the probability that a farmer will adopt the new variety but also the intensity of use of the technology once adopted.

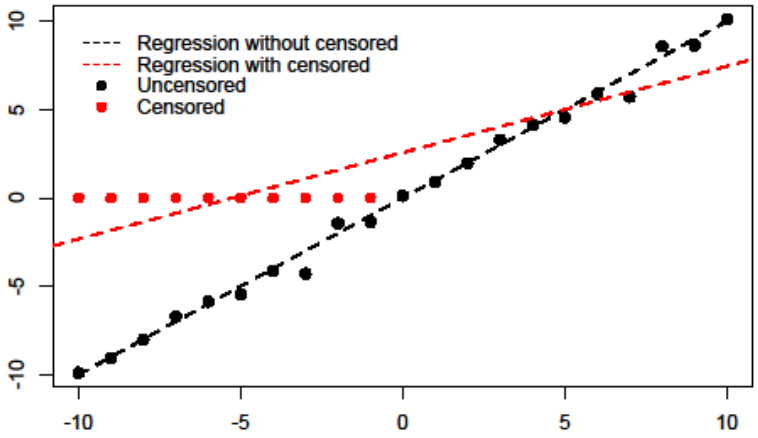
The focus in the CIS2018 questionnaire on knowledge-based activities, organizational practice, etc. provides a richer basis for the segmentation of firms, based on innovation profiles rather than innovation implementation

Text box 19. Measuring the probability of investing in innovation-related IPRs by means of a Tobit model[30]

The Tobit model is used for censored distributions, for example, of the resources allocated to innovated-related IPRs. This variable is only measured for companies that have decided to engage in such activity (it is null for all other companies). Statistically, its distribution would be a mixture of a discrete and a continuous distribution, as in the figure below, where a number of companies show an investment value of 0 and for the rest, there is a skewed distribution.



In these cases, a traditional linear regression model will provide biased results:



The impact of censoring and selectivity on statistical analysis

In previous editions of the CIS, direct measurement was usually carried out on the subsample of firms who had indicated to have having successfully implemented innovations (i.e., self declaration). Mairesse and Mohnen highlight the problematic issue of measuring quantitative variables such as expenditures (or output in innovation) only on innovative firms (Mairesse & Mohnen, Working paper series #2010-023, 2010). The statistical issues of *censoring* and *selectivity* appear in the current practice.

For instance, in the CIS 2014 harmonised questionnaire, firms declaring not having introduced neither new or improved methods of production, logistics or support activities skipped

[30] See Technical annex for an explanation of the Tobit model.

further questions about the description of the process of innovation, with the statistical effect of obtaining a *censored observation of the distribution* (see Text box 20).

Text box 20. Example of a filter question in CIS2014

3. Process innovation

A process innovation is the implementation of a **new** or **significantly** improved production process, distribution method, or supporting activity.

- Process innovations **must be new to your enterprise**, but they **do not need to be new to your market**.
- The innovation could have been originally developed by your enterprise or by other enterprises or organisations.
- Exclude purely organisational innovations – these are covered in section 8.

3.1 During the three years 2012 to 2014, did your enterprise introduce:

	Yes	No	
	1	*0*	
New or significantly improved methods of manufacturing for producing goods or services	☐	☐	*INPSPD*
New or significantly improved logistics, delivery or distribution methods for your inputs, goods or services	☐	☐	*INPSLG*
New or significantly improved supporting activities for your processes, such as maintenance systems or operations for purchasing, accounting, or computing	☐	☐	*INPSSU*

If no to all options, go to section 4

Otherwise, go to question 3.2

3.2 Who developed these process innovations?

In order to avoid potential selection biases censoring should be corrected for. This can be done using sample selection models comprising a regression for the censored variable together with a selection equation (such as the Tobit model). In the absence of additional information about non-innovative firms obtained by merging the innovation survey data with other firm data, there is no possibility to discriminate between innovators and non-innovators and to correct adequately for potential selectivity biases.

In CIS2018 the filtering question on successful implementation has been dropped. As a result, there is no longer a design-based selection bias of the estimates.

5.2.4 *Innovation intensity measured in terms of innovation expenditure and investments in intangible assets*

The concept of *input* innovation intensity relates to the amount of resources devoted by the firm. Studies that break down the allocated resources by categories of activities or characteristics of the firm show that there are significant differences in the innovation patterns about how to allocate resources both in-house (e.g. internal R&D activities) and external (e.g. purchased technology, training, consultancy from external providers).

Innovation expenditure

The most important measures for input innovation intensity are based on innovation expenditure. In CIS2018, *Innovation expenditure* is directly measured at the firm level in absolute terms by the sum of firm's expenditure in acquisition of machinery, equipment, software, other external knowledge, in-house and external R&D, and other innovation activities (including design, training, marketing and other relevant activities).

The notion of direct measurement refers to the possibility of collecting the value of innovation expenditure from the respondent firm, without any need for modelling or any other statistical elaboration. This is because companies usually keep accounting records of their investments

and purchases of goods and services. However in order to be able to allocate expenditures to innovation firms need to be able to recognise the link between the expenditure and any of the innovative activities proposed by the CIS2018.

The Oslo Manual proposes measuring a number of *quantitative variables* that allow for estimating the volume of resources devoted to, and the output of, innovation. These variables could be combined with innovation-related information at the firm- or industry-level from other data sources (e.g., intermediate consumption, investments, sales), to analyse the determinants and role of innovation in business performance and economic growth (see herefor, §4.1). For instance, the data in the table below could be combined with data on engagement in innovation.

Text box 21. Example of aggregate table with percentages on expenditure

Let Exp(i) the expenditure in 2018 for category (i) (i= "machinery, equipment and buildings"; "marketing and branding"; etc.)

Let Share(i) the share of Exp(i) used for innovation

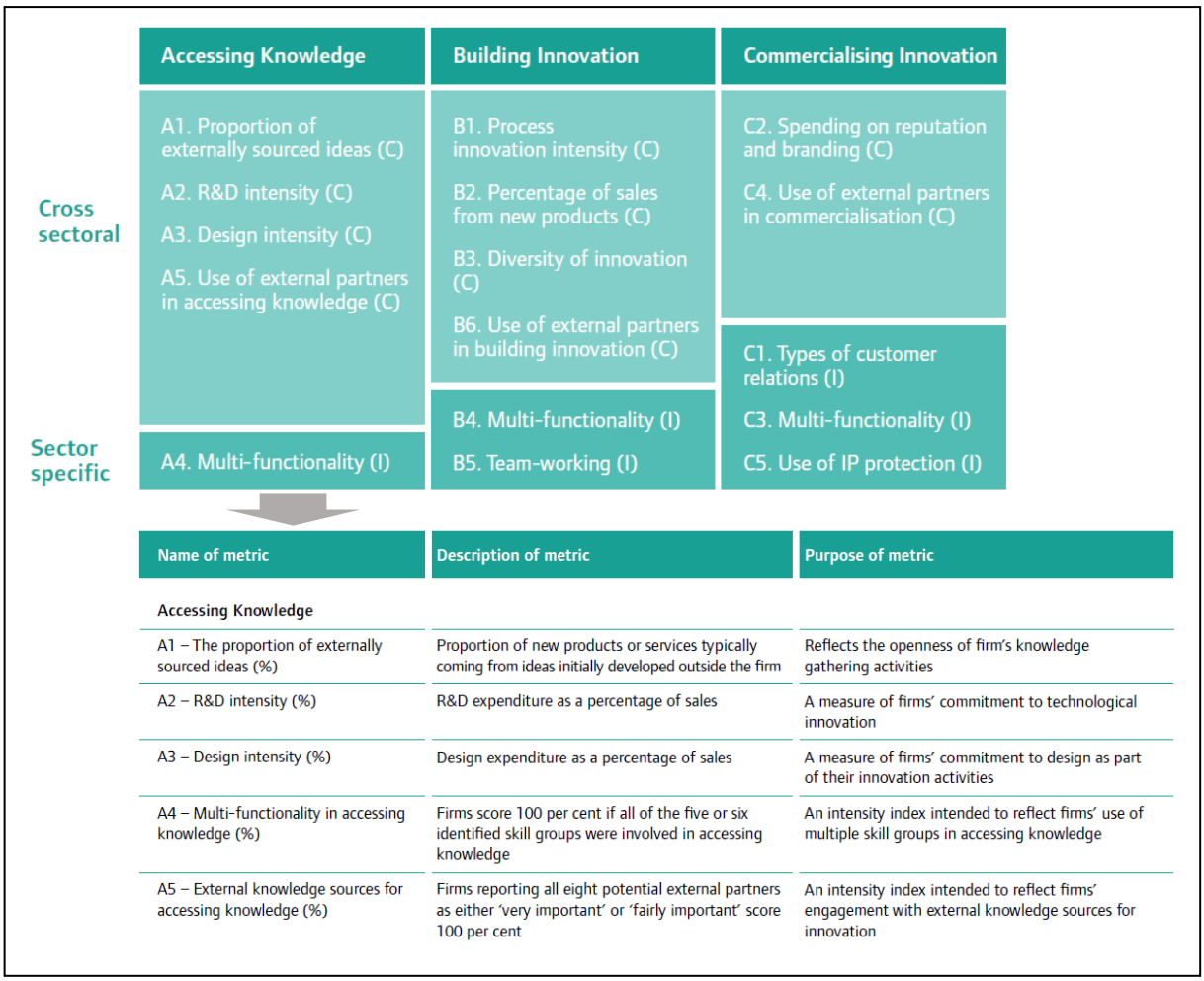
Calculate ExpInn(i) = Exp(i) x Share(i) to obtain the expenditure in category (i) used for innovation

Cross-tabulate Exp(i) and ExpInn(i) by industry and size, aggregating the individual data as totals or as averages:

Industry	D: Total expenditures in category (i)	E: Total expenditures in category (i) used for innovation	F: Share of expenditures in category (i) used for innovation
…	…	…	…
Industry *j*	Sum of Exp(i) for firms with Industry = *j*	Sum of ExpInn(i) for firms with industry = *j*	(E)/(D)
…	…	…	…

A alternative analysis of expenditure-related data is to make use of the *relative* expenditures per categories rather than the absolute values. For example, NESTA (2009) has developed an indicator of *"diversity of innovation activity"* which gives higher scores to those companies engaging (spending) in different types of activities rather than focusing in one component (Roper, Hales, Bryson, & Love, 2009). NESTA isolates the components of innovation expenditure, like R&D expenditure, investments (expenditure) in design, expenditure in process development and branding & marketing expenditure.

Text box 22. Innovation metrics for the Innovation Value Chain (Hansen & Birkinshaw, 2007)



Name of metric	Description of metric	Purpose of metric
Accessing Knowledge		
A1 – The proportion of externally sourced ideas (%)	Proportion of new products or services typically coming from ideas initially developed outside the firm	Reflects the openness of firm's knowledge gathering activities
A2 – R&D intensity (%)	R&D expenditure as a percentage of sales	A measure of firms' commitment to technological innovation
A3 – Design intensity (%)	Design expenditure as a percentage of sales	A measure of firms' commitment to design as part of their innovation activities
A4 – Multi-functionality in accessing knowledge (%)	Firms score 100 per cent if all of the five or six identified skill groups were involved in accessing knowledge	An intensity index intended to reflect firms' use of multiple skill groups in accessing knowledge
A5 – External knowledge sources for accessing knowledge (%)	Firms reporting all eight potential external partners as either 'very important' or 'fairly important' score 100 per cent	An intensity index intended to reflect firms' engagement with external knowledge sources for innovation

Investments in intangible assets

Some of the components of innovation expenditure are intangible in nature and thus rather evasive. Such intangible assets are nevertheless regarded to be an important determinant for firm productivity. These assets resources seem to be a main input to the 'knowledge production function'of firms and contribute to the propensity of a firm to innovate.

The link between investment in intangibles and innovation has been recently studied by Montresor et al. (2014), showing that the intensity of investment in intangibles (with respect to turnover) is higher for marketing and organisational innovators than for product/process innovators (Montresor, Perani, & Vezzani , 2014). The research thus suggests possible mapping between different kinds of innovators and different kinds of intangibles. For instance, some types of intangible assets are easier to incorporate (training, branding/marketing, acquisition of software) than other types that require overcoming entry barriers in terms of internal organisation and level of investment (e.g. establishing a R&D department) (Angotti & Perani, 2015).

The amount and breakdown of investment by type of intangibles could eventually provide insights into the *innovative intensity* of a firm. For instance, investments in 'separable' intangibles such as R&D, software and design can be a proxy for innovation intensity. Further investigating the statistical relationship between intensity and type of intangible investment by typologies of innovators –and even non-innovators – can shed light on the potential of using this quantitative variable to define innovative intensity.

Text box 23. Separable versus non-separable intangibles (European Commission, 2013)

> The Innobarometer 2013 asked through a business survey about expenditure in internal (or developed) and external (acquired) resources for some categories of intangibles: (i) training; (ii) software development, excluding research and development and web design; (iii) research and development; (iv) design of products and services, excluding research and development; (v) company reputation and branding; (vi) organisation or business process improvements. The study concluded that the possibility of 'separating', within the company organisation, activities related to software, R&D and design, from those of a more cross-cutting nature (i.e., branding, training, organisation) showed that companies that declared innovative behaviour scored most for investment in the 'separable' activities. Thus, while investments in 'organisationally separable' intangibles such as R&D and software are more present in product/process innovators while 'non-separable' activities are pivotal for marketing and organisational innovators.

5.2.5 Firm-level innovation intensity calculated as a ratio

To avoid size effects, input variables can be considered as relative to *total expenditure*, *total investment*, *total sales* or *total workforce* (number of employed persons). This allows for comparison of firm-level effort (expressed in terms of expenditure) for innovation with other enterprise activities. Innovation intensity is usually expressed as a ratio between innovation expenditure and some reference variables that describe the volume of activity of the company. These include:

- Turnover (sales), as the most usual way of defining the innovation intensity;
- Total number of employees (FTE) (to obtain innovation expenditure per employee)
- Total value of inputs (consumption);

A frequently used indicator of intensity, especially in studies based on the Community Innovation Survey (CIS), is that of total innovation expenditures / *total turnover (sales)* (Czarnitzki & Lopes Bento, 2011) (Falk & Falk, 2006). In 1997, working on a dataset from the 1992 CIS, Evangelista *et al.* already used this ratio in to analyse differences across industries, size classes, countries and also intra-sectorial concentration of innovation expenditure (Evangelista, Sandven, Sirilli, & Smith, 1997). They mostly observed the uneven distribution of the ratio across different groups of firms (and therefore suggested a log transformation for the application of standard statistical methods requiring at least some distributional symmetry, such as ANOVA).

Considering the similarity with R&D, we may recall the use of ratios to define the *R&D intensity* as *R&D expenditures per employee* (Ebersberger & Lööf, 2004) (Ebersberger & Lööf, 2005). Crespi and Zuñiga use the innovation expenditure per worker as a measure of [input] *innovation intensity* (Crespi & Zuñiga, 2010). Based on the input of human resources, intensity has also been defined as *FTE dedicated to innovation / total workforce.*

When different reference periods are used for the nominator and the denominator several statistical issues arise. For instance, in order to take the time delays between innovation expenditure and its impact on processes and products into account, a reference period of the three last years has been proposed in several innovation surveys. However, the reference period for collecting variables on turnover (sales) and employment is usually annual, which has implications in the definition of intensity as a ratio of innovation expenditure to turnover or expressed per employee. In practice, some assumptions have to be done about the compatibility of difference reference periods for the nominator and denominator.

Hence before considering any ratio of variables, their reference period should always be clarified. Souitaris for instance proposes a mix of reference periods (three and one year), considering as input variables the *innovation input or effort towards innovation* on the one hand, and the expenditure for innovation *in the past three years* over *current* (i.e. for the last year) innovation-related sales on the other hand (Souitaris, 2002).

5.2.6 Model-based definition of innovation intensity

Innovation intensity at firm level

Innovation surveys collect the percentage of sales due to innovative products, thus providing for an *output-based innovation intensity indicator at the firm level*. While this indicator shows the success of innovation, it does not provide any information on why and how the success was achieved. In the absence of a breakdown by being novel or not to the market, it is not possible to distinguish both market and product effects. Particularly in less developed markets this introduces a bias because imitation products may represent a novelty in these markets (Arundel, 2007). Moreover, there is also a time lag between the innovative activity and the launching and commercialisation of the innovative product, which is not reflected in the value of the output indicator.

Mairesse and Mohnen propose a model for an 'innovativeness' factor that is defined at the firm level (Mairesse & Mohnen, 2001). The factor is based on an econometric adjustment of innovative sales after controlling for several structural variables (i.e., sector, size, country). The factor could be regarded as an output-based indicator of innovation intensity. The 'innovativeness' is defined as the difference between the observed percentage of innovative sales and the estimated percentage for a given country, industry and size. The innovativeness factor would complement other measurable input data such as R&D and other current and capital expenditure on innovation, as well as innovative behaviour (R&D activity, continuity of R&D, collaboration and acquisition of technology).

The econometric model may present a selection bias if only innovation input data is collected for firms self-declaring as innovators (see before, Text box 20). The proposed model (generalized tobit) includes therefore two equations: one for the *propensity to innovate* and another for the *percentage of innovative sales*, termed by the authors as 'intensity of innovation'.[31] The equation for the intensity of innovation includes as predictors, in addition to structural effects (industry, size, country) some 'environmental variables' (i.e., the proximity to basic research, the co-operation in innovation) and R&D effects (i.e. the qualitative characteristics of R&D activity and the R&D intensity, as expenditure compared to total turnover).

A major advantage of a model-based approach to determine innovation intensity is that the values can be calculated by the NSO itself, with the in-house use of CIS micro data. This could diminish response burden for respondents. The approach does obviously refer sufficient in-house econometric capabilities. An additional methodological requirement is that an ex post cross-country harmonization of the indicator is needed in order to test the validity of the model in different country samples.

The model-based approach could be extended by combining meaures of input and output intensity, such as *product innovation costs/sales of new products*.

[31] However since the model sets the innovation intensity of non-innovating firms to zero it does not fully overcome the dichotomous classification.

Innovation intensity at sector level

At the aggregate level, the innovative effort is commonly measured by a set of indicators on the proportion (and number) of firms undertaking each type of innovation-related activity and implementing innovations *in a given industry or country* (therefore based on a qualitative/dichotomic variable), as well as by the total expenditure in innovation activities, broken down by relevant business sub-populations (sector, size).

In CIS2018, total innovation expenditure comprises internal and external R&D spending, purchase of machinery and software for innovation projects, purchase of other external knowledge such as patens, licenses and similar intellectually property rights, prototyping and similar preparations for production and market introduction, marketing activities in direct relation with a new product introduction as well as cost for training of employees directly linked to innovation projects. Most companies can provide these data, corresponding to accounting lines or recorded for the purpose of policy-related reporting (e.g. subsidies for innovative activities).

A commonly used model-based approach is to break down input intensity into a number of components, and to uses innovation expenditures as a proxy for innovation intensity (Mas-Verdú, Wensley, Alba, & Garcia Alva, 2011). The model estimates the innovation embodied in an industry separately from the innovation obtained through the domestic purchase and import of intermediate inputs, and investment.

Text box 24. Formal description of total innovation intensity of an industry

Total innovation intensity of an industry j (defined as in_j) can be formally described the sum of five components:

$$in_j = r_j + p^d_j + p^m_j + c^d_j + c^m_j + c^m_j ,$$

where

$r_j = (R_j/X_j)$ is industry j's own innovation intensity (expenditure on innovation activities/output of industry j),

$p^d_j = (P^d_j/X_j)$ is the innovation embodied in domestic intermediate inputs per unit of output of j,

$p^m_j = (P^m_j/X_j)$ is the innovation embodied in imported intermediate inputs per unit of output of j,

$c^d_j = (C^d_j/X_j)$ is the innovation embodied in domestic investment inputs per unit of output of j, and

$c^m_j = (C^m_j/X_j)$ is the innovation embodied in imported investment inputs per unit of output of j,

and these, further broken down by industry.

Although this approach has so far only been applied at industry level it has the potential of highlighting, at *firm-level*, the distinct contribution to innovation intensity of the own efforts of the firm from external sources (i.e., imports, purchases, investments of technology and know-how). Thus, it conceptually provides a basis for assuming that innovation (input) intensity relates to the behaviour of the firm in terms of its own resources to innovate plus the acquisition of 'embodied' knowledge (e.g., in hardware and software or in [consultancy] services).

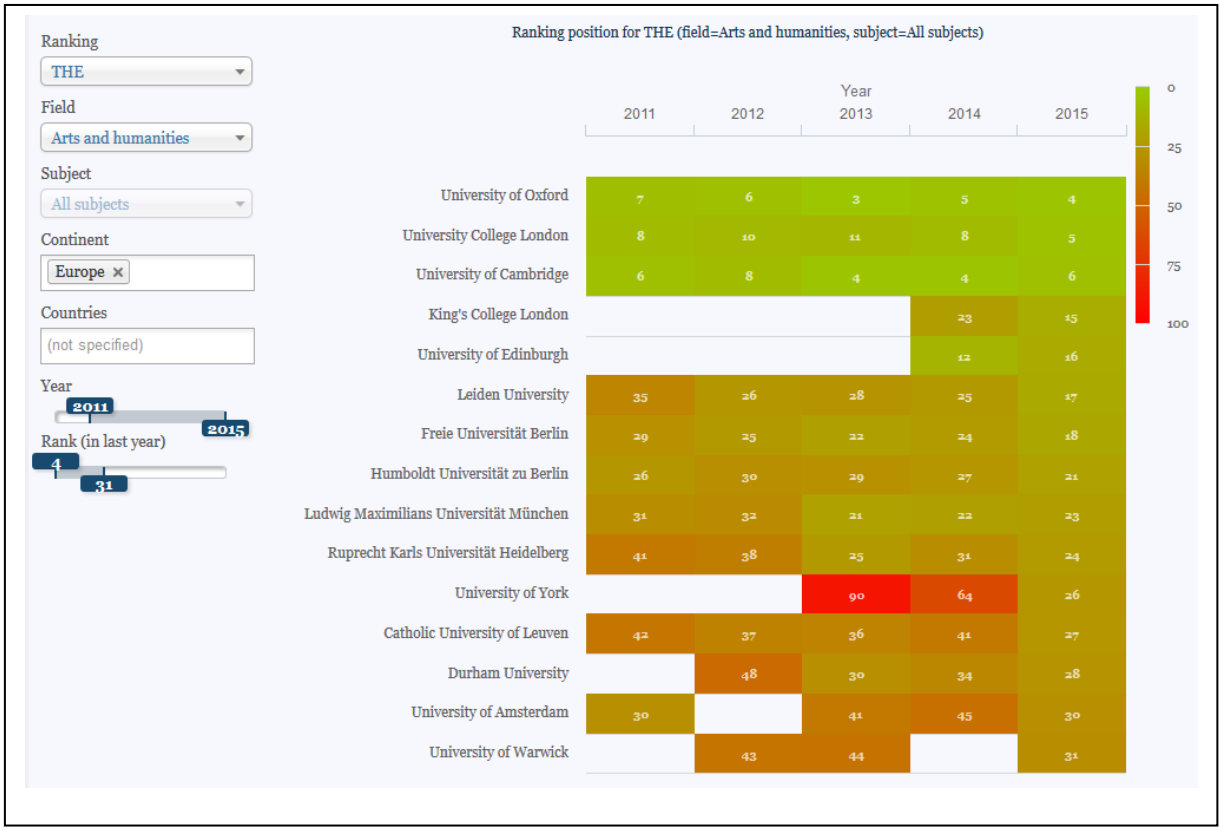
5.2.7 Innovation intensity defined by comparison of a firm with its sector

The classical 1997 paper of Evangelista *et al.* already detected that innovation intensity (defined as innovation expenditure over turnover) was strongly determined by industry and size class, with similarities of patterns across European countries (Evangelista, Sandven, Sirilli, & Smith, 1997). Econometric models of the innovation expenditure at the firm level often use dummies to control for sector and size (Sanguinetti, 2005) (Raymond, Mohnen, Palm, & van der Schim, 2009). Mairesse and Mohnen (2001) compare the percentage of innovative sales of a firm with that of its sector and size interval, to define a level of "innovativeness" (see §5.2.6).

This practice can be considered equivalent, in terms of estimation, to comparing the firm-level values with average values at the sector and/or size interval level, in order to define the firm's innovation intensity.

Similar practice for deriving firm-level results exists in some National Statistical Offices, which provide – as a compensation and incentive for respondents to business surveys – comparisons of their productivity levels (e.g. turnover per employee) with those obtained for the corresponding sector and size. Rankings (in terms of percentiles) are also obtained from the distribution of individual values, to provide with richer information to the respondent.

Text box 25. Example of a dynamic feedback module to compare respondent scores to overall scores[32]



[32] Source: Dialogic/Rotterdam School of Management. https://alluniversitiesranked.com

This type of intensity measurement might also provide insight on the distribution of the innovation effort, distinguishing sectors were most of the innovation expenditure is concentrated in a few leading companies from those in which the innovation is more evenly spread. This suggests that any definition – or analysis – of innovation intensity should consider the industry and size class of the company. The increasing harmonization of industrial classifications and size intervals allows for such cross-country comparability.

5.2.8 Analysis of the innovation output at sector level as a binary variable

The concept of innovation intensity can also be understood as a measure of *success of innovation activities*, i.e. launching to the market new or significantly improved products. A traditional indicator on the *industry level*, whose relevance has however been widely discussed, is the *percentage of innovative firms* obtained on the basis of the self-declaration by respondent firms of the successful implementation of innovations (by launching new or improved products to the market). For example, innovative intensity is more frequently associated with large IT companies that are regularly launching new devices, than to pharmaceutical companies that, despite enormous efforts in R&D, are only able to launch new drugs with a relatively low frequency.

A major conceptual issue is whether this particular measure refers to input or to *output* instead. According to OM2005, an innovation is defined as the (successful) *implementation* of a new or significantly improved product (goods and services), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations. This seems to suggest that the measure is more output than input-oriented (see also Chapter 1).

Text box 26. Example of aggregate table with percentages based on innovation output as a binary variable

Let *InnovNotAvailable* a binary variable that is equal to 1 if the firm declares to have offered one or more new or significantly improved products not yet available from its competitors;

Let *InnovAlreadyAvailable* a binary variable that is equal to 1 if the firm declares to have offered one or more new or significantly improved products already available from its competitors;

Cross-tabulate *InnovNotAvailable* and *InnovAlreadyAvailable* by industry and size, aggregating the individual data as total counts or as ratios:

Industry	A: Total number of firms	Total number of firms having launched new or improved products		Percentage of firms having launched new or improved products	
		B: Not yet available from competitors	C: Not yet available from competitors	D: Not yet available from competitors	E: Not yet available from competitors
…	…	…	…	…	…
Industry *j*	Count of firms with Industry = *j*	Count of firms with industry = *j* and *InnovNotAvailable* = 1	Count of firms with industry = *j* and *InnovAlreadyAvailable* = 1	(B)/(A)	(C)/(A)
…	…	…	…	…	…

5.2.9 Combining qualitative variables to represent innovation intensity

Modelling the relationship between qualitative and quantitative variables is usually more complex than only using quantitative ones. Moreover it remains to be seen to what extent the combination of various qualitative variables (e.g., describing types of innovation activities, degree of novelty of the innovative products or processes, various modes of implementation) can provide robust indicators of innovation intensity.

De Jong uses qualitative variables related to the input, process and output of innovation in a LISREL model to construct *a latent factor* expressed as innovation intensity (de Jong, 2000). Licht and Moch relate several input indicators to the qualitative description of the innovation output (Licht & Moch, 1997). Barlet et al. (2000) provide econometric evidence on the significant impact of the qualitative type of innovation on sales and exports, controlling by industry and size (Barlet, Duguet, Encacoua, & Pradel). Souitaris proposes a list of qualitative variables as possible determinants of a so-called 'innovation rate' of a company, classifying them into (Souitaris, 2002):

- *Contextual variables*: firm's profile (years of operation, growth rate of size, sales, profits and exports), competitive environment (perception of rate of changing demand and intensity of competition);
- *External communication variables*: communication with stakeholders, networking and acquisition of external information, cooperation with external organisations;
- *Strategic variables*: existence and consistency of an innovation budget, business strategy, management of attitudes (towards risks and new technologies), CEO's profile;

- *Organisational competencies*: intensity of R&D and quality control, market competencies, education and education of personnel, training of personnel, internal processes for innovation.

The occurrence of a multitude of interaction effects has been a major challenge to a broader use of analyses that use qualitative variables. However, the recent rise of exploratory machine-learning techniques (i.e., algorithm-based models of statistical interference such as tree-based methods) might improve the ability to investigate the explanatory power of the many qualitative variables that are collected in innovation surveys. Such algorithm-based models do not require the ex ante formulation of any model of the relationship between the explanatory and the endogenous variables. They could be particularly useful in the context of the complex relationships between the variables considered, possibly encompassing non-linear effects and interactions, correlations between predictors and time lags.

5.3 Enterprise profiling

5.3.1 Introduction

Empirical research on typologies of innovation behaviour, using business survey data, is useful to understand how different patterns of innovative activities contribute to the improved firms' performance. This research is the basis for innovation policy and innovation management strategies.

One of the key achievements of the CIS is that it has a harmonized the use of innovation surveys across all member states. Moreover, various countries across the world have also coordinated their national surveys with the latest version of CIS. Harmonization obviously has greatly facilitated the exchange and re-use of innovation survey data.

The unavoidable downside to harmonization is that it makes it seemingly difficult to take local heterogeneity into account – one size just never fits all. However much heterogeneity exists among individual firms within the same industrial branches as well as systemic and significant differences in innovation activities at the level of markets or industries (Peneder, 2010). Adding to this, there are systemic differences between national and regional systems of innovation. Consequently, in every country and in every region different types of innovating enterprises might prevail, and hence there will be a specific demand from local policy makers (e.g., to design targeted policy initiatives). Yet standard indicators cannot take such local differences into account (UNU-MERIT, 2017).

Text box 27. Correcting for systemic differences in innovation profiles across countries

One of the seemingly odd results from CIS3 is that Portugal, which ranks overall number 18 in the European Innovation Scoreboard, has a higher percentage of innovative firms than Finland, which ranks overall number 2. *The outcome can be largely explained by the differences in the prevalance of specific types of innovative firms*. Whereas in Portugal technology modifiers and technology adopters are relatively common, in Finland it are strategic innovators and intermittent innovators (Arundel & Hollanders, Innovation Strengths and Weaknessess, 2005).

Table 7. Share of different type of innovative firms, Finland and Portugal (2004)

	Strategic	Intermittent	Modifiers	Adopters	**Total**
Portugal	3%	15%	16%	13%	**47%**
Finland	13%	19%	10%	3%	**45%**

Standard indicators might fail also to capture certain innovation profiles because they are one-dimensional. For instance, there are no standard indicators that provide a full profile of capabilities or which link capabilities and outputs (e.g. the share of R&D performing firms with new-to-market innovations) (UNU-MERIT, 2017). However among users of CIS data there is an apparent need to be able to distinguish different types of innovating enterprises. For instance, academic researchers often work on highly specialized topics and thus need richer characterisations of innovation capacibilities and performance (e.g., to classify enterprises by their innovative behaviour and specific requirements vis-à-vis their environment). The dichotomic status of innovators vs non-innovators is recognised as too simple for describing business strategies. The one-way classification by type of innovation (product, process, management, marketing, etc.) is recognised as well as limiting, since a large number of companies have mixed strategies (product and process, product and management, etc.).

Text box 28. Profiling enterprises in detail

> The profiling of enterprises beyond basic classifications start with the integration of firm-level and sectoral methodologies (usually building on (Pavitt, 1984)). The actual classification of innovation behavior is done at the micro-level. The clustering of the sectors is based on the occurance of specific types of activities within those sectors (Peneder, 2010). Several studies have built on this framework and have deepened the description of the behaviour of innovative firms. For example, in their study on technological competences of Spanish manufacturing firms Vega-Jurado et al. use Pavitt's taxonomy to control for the impact of the sector on factor importance. This enables them to show that there are significant discrepancies of innovation drivers *within* the sector classification from Pavitt (Vega-Jurado, Gutiérrez-Garcia, & Fernández-de-Lucio, 2009).

In sum, there is a clear need to be able to profile enterprises for at least two reasons. The first one is to correct for systemic differences that exist *across* national or regional populations of firms, or between industries. The second one is to be able to describe the specificities *within* a particular dataset. A combination would be to analyse several datasets in-depth, to identify specific patterns or relationships.

The first aim can be achieved by recombining variables that are already included in CIS. The avoidance of filter questions greatly increases the potential of the recombination of existing *meso data* into new taxonomies (see before, §0). The second aim can be achieved by disaggregating CIS data into *micro data* and then construct new complex indicators.

With regard to meso data is should be noted that there is only an apparent opposition between the rigidity of harmonization and the flexibility of profiling. A proper standardization of data actually greatly facilitates the re-use of the data. The argument is that a small set of standardized elements (e.g., CIS variables) can be *recombined* into many different compositions (e.g., tailor-made composite indicators).[33] For the NSI involved this requires for instance the use of smart tabular ways to present and disseminate large sets of statistical indicators (see for instance (Mazzi, 2015)). This does require a reasonably stable set of variables over time for all participating countries.

With micro data researchers can work directly at the enterprise level. A major advantage vis-à-vis the use of meso data is that data is not (partly) prestructured but researchers are free to create any aggregate category they need, according to adaptable criteria. A

[33] The number of combinations for a composition of m elements from a total set of n elements would be n!/(n-m)!(m!). The maximum number of combinations is m=0.5n. For example, of there are 20 basic variables the maximum number of combinations is at m=10 (i.e., composite indicators consisting of 10 variables), which already gives 20!/(20-10)!10! = 184,756 different combinations.

precondition for the re-use of micro-data by third parties (such as academic researchers) is the establisment of a proper legal and IT infrastructure by the NSI involved that grants access to the data without jeopardizing the strict rules concerning confidentiality, privacy, and security (see before, §4.5)

5.3.2 Enterprise profiling in practice

The rise of the machines

The basic of enterprise profiling is the re-arrangement (regrouping) of a set of enterprises into meaningful groups. There are two basic methods to arrive at such a re-arrangement: *classification* and *clustering*.

In classification, the output is a priori known. That is, the output Y is predicted from the input data X: Y=f(X). Regression is a special case of classification where the output is a continuous, not a discrete value. Applied to enterprise profiling, classification refers to *predefined categories* that are based on a priori knowledge.

In clustering, there is no output data, only input data. All data is unlabeled and the re-arrangement is based on the inherent structure of the (input) data. That is, all observations are assumed to be caused by latent variables (Valpola, 2000). Applied to enterprise profiling, clustering refers to *most distriminating factors* that are derived from the inherent structure of the data.

There are many statistical techniques to classificy or cluster data. The most common technique to define typologies of firms are multivariate analyses such as logistic regression (LOGIT), principical component analysis (PCA) and clustering analysis (e.g., k-Means). With the rise of big data and data science there is a renewed interest in artificial intelligence and machine learning. Both have already been established decades ago but the exponential growth in computational power and storage capacity has enabled the introduction of a plethory of new classification and clustering algorithms. Nevertheless, many machine learning algorithms are in fact classical statistical techniques disguised in the jargon of computer sciences. In terms of machine learning, classification problems refer to *supervised learning* whereas clustering problems refer to *unsupervised learning*.[34] Hence, classical techniques such as linear regression and logistic regression can be regarded as examples of supervised learning and classical techniques such as principal component analysis and factor analysis as examples of unsupervised learning.

[34] There is also a hybrid category of machine learning, namely semi-supervised or *reinforcement learning*. Reinforcement learning is between supervised learning (there is some form of feedback available for each predictive step, i.e. there is some a priori knowledge) and unsupervised learning (there are no precise labels). One example of a reinforcement learning algorithm is Latent Dirichlet allocation, a probabilistic topic modelling technique that is being used in natural language processing to classify texts.

Table 8. First classication of widely used algorithms for regrouping data, based on machine learning task[35]

- Supervised learning
  - Classification (two-class & multi-class)
    - Logistic regression and multinomial regression
    - Artificial Neural networks
    - Decision trees
    - Nearest neighbor methods (e.g., k-NN or k-Nearest Neighbors)
    - Bayesian classifiers (e.g., Naive Bayes)
    - Support vector machine (SVM)
  - Regression
    - Simple and multiple linear regression
    - Ordinal regression
    - Artificial neural networks[36] (e.g., Back-Propagation)
    - Decision tree or forest regression
    - Nearest Neighbor methods (e.g., k-NN or k-Nearest Neighbors)
  - Ensemble methods[37]
    - Random forest
- Unsupervised learning
  - Clustering
    - K-means clustering
    - Hierarchical clustering
    - Expectation Maximization (EM)
    - Deep learning (e.g., Deep Boltzmann Machine, DBM)
  - Dimensionality reduction
    - Factor analysis
    - Singular-Value Decomposition, SVD (e.g., Principal Component Analysis, PCA)
    - Independent Component Analysis (ICA)

Selecting a suitable technique

Which technique or algorithm to use first and foremost depends on the *purpose* of the analyse, and secondly on the specific *characteristics of the dataset* (e.g., structure, size).

With regard to the purpose, we find the supervised pair of classification and regression on the one hand, and the unsupervised of dimensionality reduction pair and clustering on the other hand. Dimensionality reduction and clustering are usually deployed in the initial exploratory and preparatory stages of a research project, respectively to make the dataset more compact (by either selecting a subset of variables or by transforming the data into a space with fewer dimensions) and to partition the dataset into subsets that (ideally) share some common characteristics. Subsequently, the partitioning found in clustering can then be used as input in a (supervised) classification process, or, as one of the measured attributes, in regression to compute new values for a dependent variable for each of the subsets.

Once the aim of the research has been defined, the most suitable technique or algorithm can be selected. Ironically, in the presence of the numerous sophisticated algorithms, this is

---

[35] See also Technical Annex 6.4 for a structured overview of machine learning algorithms.

[36] Note that artificial neural networks can both be used for discrete (classification) and continuous (regression) output. The same goes for decision trees and k-NN.

[37] An ensemble model actually combines the results of several different types of classification models (i.e., they use one of more different types of models for each step in the overall work process of the algorithm (e.g., Bayesian statistics for model averaging and Monte Carlo methods for sampling).

ultimately an empirical inquiry, that is, a matter of trial and error. This is because of the fundamental rule that in a matrix of all problems and all algorithms that the average performance of all algorithms is equivalent, i.e. no one algorithm works best for every problem.[38] There are, however, some basic guidelines.[39]

First of all, unsupervised learning is often used when supervised learning would be more appropriate. When there is robust a priori knowledge on the data, it should be applied in the research project. When such knowledge is lacking, or is disputed, unsupervised learning could be deployed to find new research trajectories. With unsupervised learning it is also possible to learn larger and more complex models than with supervised learning. This is because in supervised learning one is trying to find the connection between two sets of observations. The difficulty of the learning task increases exponentially in the number of steps between the two sets and that is why supervised learning cannot, in practice, learn models with deep hierarchies. If the causal relation between the input and output observations is complex -- in a sense there is a large causal gap – it is often easier to bridge the gap using unsupervised learning instead of supervised learning (Valpola, 2000).

Secondly, however, most sophisticated machine learning algorithms (such as [deep learning] neural networks) require very large amounts of data to train. They perform well on image, audio, and tekst data but are less suitable for relative mundane datasets such as CIS micro data. In general, when the underlying relationships are not all too complex (see before) and the data quality is good (as in the case of CIS data) simpler classical techniques such as linear regression, logistic regression, Naive Bayes and K-means outperform more complex techniques such as deep learning machines, support vector machines and Nearest Neighbors (EliteDataScience, 2017). In general, then, it pays off more to improve data quality than in applying more sophisticated algorithms, for instance by data preprocessing (noise treatment, normalization) and exploratory analysis (sampling, feature extraction). This might explain why we find very few examples of machine learning in innovation research so far. Most of the current projects that use CIS data deploy traditional methods. Nevertheless, there are several machine learning algorithms that seem to be suitable for the profiling of enterprises. Especially Decision trees, Naive Bayes, and hierarchical clustering seem to hold a lot of potential. For exploratory analysis unsupervised learning techniques might be very useful, for instance clustering can be used for sampling, and dimensionality reduction techniques for feature extraction.

In the next paragraphs for each of the possible research stages (dimensionality reduction, clustering, classification, regression) we will describe examples of a traditional and a new (machine learning) technique to profile innovative enterprises.

5.3.3 Dimensionality reduction

Principal Component Analysis (PCA)

Principal component analysis (PCA) has been around for a century and is often used as a default technique in exploratory research to divide a dataset into meaningful subsets. PCA is essentially an unsupervised machine learning algorithm to extract features (i.e., input variables). PCA is a special case of the more sophisticated Singular-Value Decomposition (SVD) algorithm, a generalized technique that has been only been later developed for practical use.

---

[38] The so-called 'no free lunch theorem' (Wolpert & William, 1997).

[39] See also the Machine Learning Algorithms Cheat Sheet (courtesy: SAS) in Technical Annex 6.4.

In contrast to feature selection techniques (such as Genetic Algorithms) PCA (and SVD) do not keep select a subset of the original features but they create *new* features. PCA does this by creating linear combination of the original features. The new features are orthogonal, which means that they are uncorrelated (EliteDataScience, 2017).

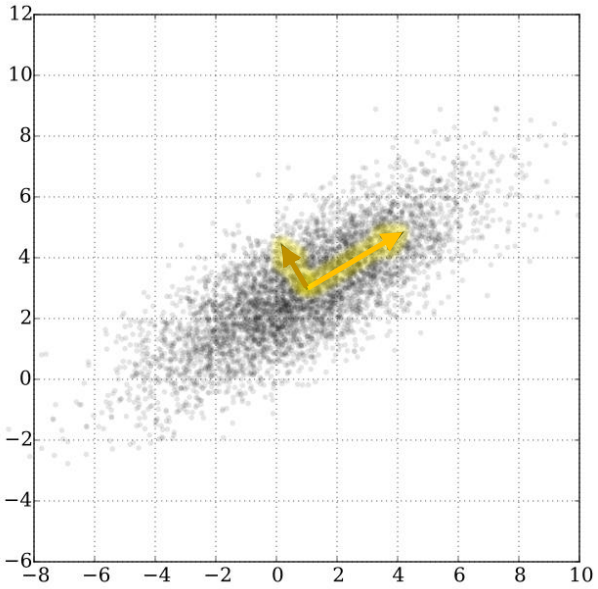


Figure 11. Newly created features by PCA shown as eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue (source: Nicoguaro)

The key advantage of PCA is then its ability to rank these newly created features in order of their 'explained variance'. The trick is now to eliminate the lower ranked features. *This enables the description of the original dataset with a smaller set of (newly created) features*. This is the essence of dimensionality reduction. The reduction of dimensions is often needed in the preparatory data preparation stage for clustering because there are quite many clustering algorithms (e.g., distance-based algorithms) that cannot deal with situations where the number of features (input variables) is very large relative to the number of observations in the dataset (this is known as the 'curse of dimensionality').

Outside the realm of machine learning, though, PCA is often used to 'magically interpret' datasets by autonomously (i.e., unsupervised) generated subdivisions. However, the fact that the new principal components are usually difficult (and theoretically impossible) to interpret is exactly the Achilles Heel of PCA – it is still the research who has to give meaning (label) the components afterwards. Moreover, the researcher also has to define the treshold for cumulative explained variance (compare the weakness of k-Means, where the researcher has to define the number of clusters beforehand).

The limitations of factor analysis are apparent in the description of factors that contrain innovation performance of SMEs in Croatia (Božić & Rajh, 2016). The authors first used a k-Means algorithm to classify the SMEs into three clusters, with minimum within-group and maximum between-group variation. Subsequently, they used factor analysis to distinguish four components with regard to barriers to innovation which they interpreted respectively as 'Organisational contraints' (e.g., "insufficient support from colleagues"), 'Financial constraints' (e.g., "unavailability of bank loans"), 'Market constraints' (e.g., "market dominated by incumbent"), and 'Uncertainty related constaints' (e.g., "perceived risk"). The expectation of the authors was that the three clusters that had been crafted with the k-Means model with coincide with three clusters regarding the intensity of constraining factors, namely: financial

problems, internal constraints, and external constraints. In other words, one cluster from the k-Means model would map to the first component ('organisational constraints'), another cluster to the second component ('financial constraint'), and the third cluster to the third component ('market contraints'), with the fourth component ('perceived risk') as a residual.

Table 9. Cluster means for each of the four main barriers to innovation (Božić & Rajh, 2016)

	Cluster 1		**Cluster 2**		**Cluster 3**	
	Mean	*St.Dev.*	*Mean*	*St.Dev.*	*Mean*	*St.Dev.*
Organisational constraints	3.03	0.79	1.46	0.41	1.54	0.44
Financial constraints	3.52	0.89	1.69	0.59	3.76	0.79
Market constraints	3.31	0.96	2.47	0.78	3.57	0.80
Uncertainty related constraints	3.26	0.86	1.81	0.77	1.87	0.63
Number of employees (mean)	39.4		40.7		21.3	
% firms that report innovation development	57%		79%		87%	
% firms that report radical innovation development	43%		62%		64%	
% firms with no R&D	52%		26%		23%	

The results are not evidently in line with the expectations of the authors. Instead of a clustering by contraint types (financial, internal, external) it rather seems that the population is clustered by firm performance, That is, SMEs in cluster one are reporting high barriers in any of the four discerned factors ('laggards'). In contrast, SMEs in cluster 2 only have relatively high means for external (market) market constraints, but still less than the other two clusters ('leaders'). SME's in cluster 3 are impeded both by financial and external (market) contraints but less so by internal (organisational) or uncertainty related constraints. However, in the intermediate cluster 3 does have the highest percentage of firms that report innovation development. The percentage of firms that report radical innovation development and the absence of R&D is also slightly higher than in the 'leading' cluster 2. The only possible explanation – in this limited set of indicators – for the seemingly underperformance of cluster 3 would then be its significantly lower average number of employees.

The results from the analysis from Hervas Oliver et al. of non-R&D technological innovation seems to be more insightful (Hervas-Oliver, Sempere-Ripoll, Boronat-Moll, & Rojas, 2015). They used a sample of 5.878 non-R&D technological manufacturing and service firms from the Spanish 2006 CIS. A principal component analysis (PCA) was used to construct two dependent variables, 'Production performance' and 'Marketing performence' out of four and three survey items respectively.

Table 10. Composition of dependent variables Production performance and Market performance (Hervas-Oliver, Sempere-Ripoll, Boronat-Moll, & Rojas, 2015)

Production performance	**Market performance**
Explained variance: 63.4%, KMO=0.729	*Explained variance: 72.3%, KMO=0.694*
Reduced unit labour costs	Increasing range of goods or services
Increased capacity	Entering new markets or increased market share
Improved production flexibility	Improving quality of goods or services
Materials and energy saving	

The dependent variables have then be used in a regression model that has 'Organisational innovation' and 'Marketing innovation' as independent variables (together covering the central notion of 'non-technological innovation') and several control variables of which 'external knowledge sources from industry and science' has also been constructed on the basis of a PCA.

The results for Production performance and Market performance are quite similar. Background variables has almost similar scores (e.g., in both cases knowledge from industry 3 to 4 times more relevant than are scientific sources). In both cases, the *joint adoption* of technological process and management innovations has a positive premium effect, albeit the effect is 2-3 times higher for Production performance than for Market performance.

Latent Dirichlet Allocation (LDA)

There are various sophisticated dimensionality reduction algorithms in use in machine learning. For instance, in natural language processing, a subfield within machine learning, the latent Dirichlet allocation (LDA) is widely used to reduce the dimensionality of documents. That is, all words in a text are 'reduced' to one or more topics – LDA basically 'interprets' a piece of text and assigns it to a particular topic. The algorithm is built on the presumptions that documents are characterized by a particular small set of topics, and that these topics use only a small set of words frequently (Blei, Ng, & Jordan, Latent Dirichlet Allocation, 2003). A topic is *not* strongly defined. Instead, it is identified on the basis of automatic detection of the *likelihood of term co-occurrence*. A lexical word may occur in several topics with a different probability, however, with a different typical set of neighboring words in each topic. In the example below, for instance, to topics 'genetics' and 'data' can be described as sectors, with two related topics ('life' and 'brains') as intermediate vectors.

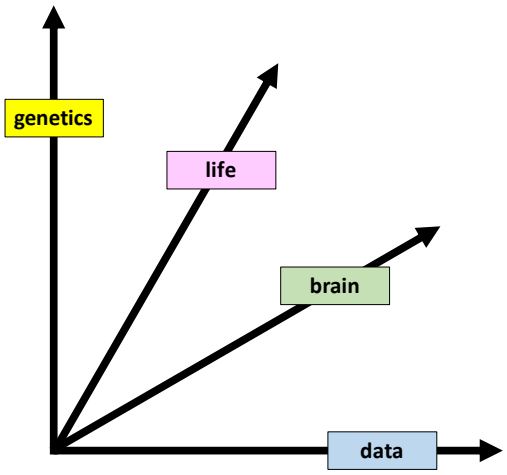


Figure 12. Simplistic Term Vector Model for the topics 'genetics' and 'data'.[40]

Once trained on this space, topics can be classified as *related to* 'genetics' or 'data' and pages can be assigned to these topics. The first instance likewise has probabilities of generating words like 'gene', 'dna', and 'genetics', the second instance words like 'data', 'number' and 'computer'.

[40] Figure inspired by the one found on ttps://moz.com/blog/lda-and-googles-rankings-well-correlated

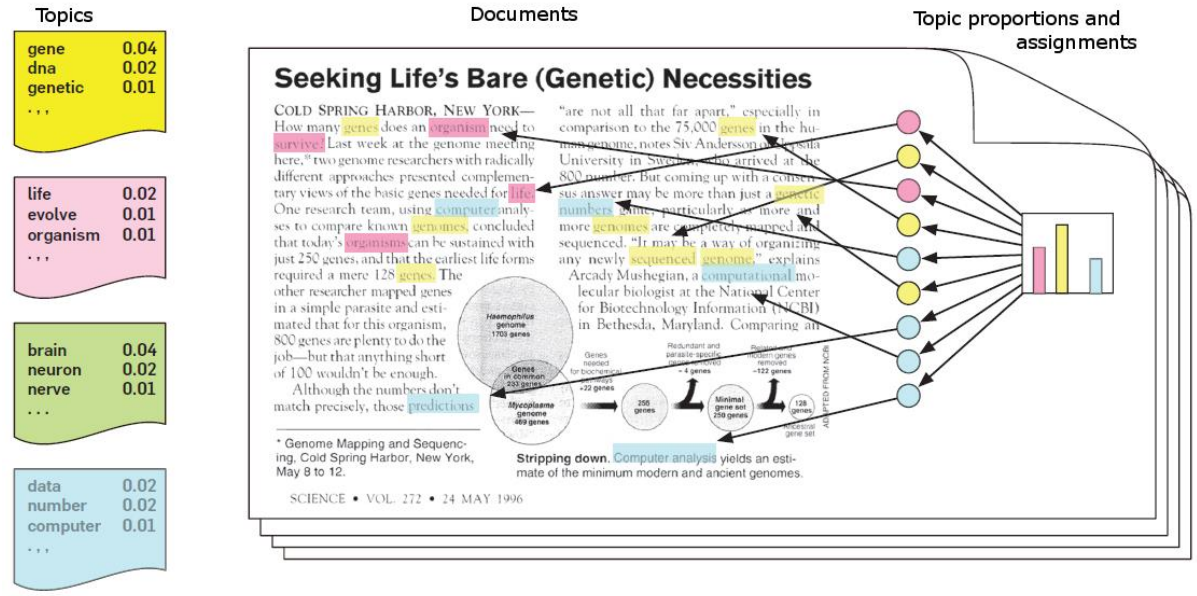


Figure 13. Example of LDA classification of a page on the use of data analysis to determine the number of genes an organism needs to survive (source: (Blei, Probabilistic Topic Models, 2012))

At first sight these sophisticated dimensionality reduction techniques might be less relevant to the profiling of innovative firms. However indirectly these can be of great use. For example, in an ongoing research project on the measurement of innovation in the public sector that is commissioned by Eurostat, the units of observation are webpages on websites of public sector organisations (e.g, municipalities) (Koppers & te Velde, 2017). For specific public services LDA is being used to classify the description of that public service into a specific 'innovation level'. In the table below, for one specific public service (waste collection) four subsequent 'innovation levels' have been defined. Similar to the previous example, Topics can then be classified as *related to* 'door-to-door collection' (level 3) or 'pay-as-you-throw' (level 4). The first instance likewise has probabilities of generating words like 'garbage truck', 'collection schedule', and 'rubbish bin', the second instance words like 'tariff', 'ID card' and 'quota'.

Table 11. Definition of innovation levels for the public service 'waste collection' (Koppers & te Velde, 2017)

Level	Types of waste collection	Description
1	*No separate collection*	Municipal solid waste are trown into common bins in the street
2	*Separate collection*	Separate collection of waste (in common bins in the street) to recycle waste material
3	*Door-to-door collection*	Separate waste streams are collected in separate bins directly at home (periodically)
4	*Pay-as-you-throw*	Separate waste streams are collected in separate bins directly at home (periodically), while a tarif (proportional to the weight) is applied to unsorted waste

5.3.4 Clustering

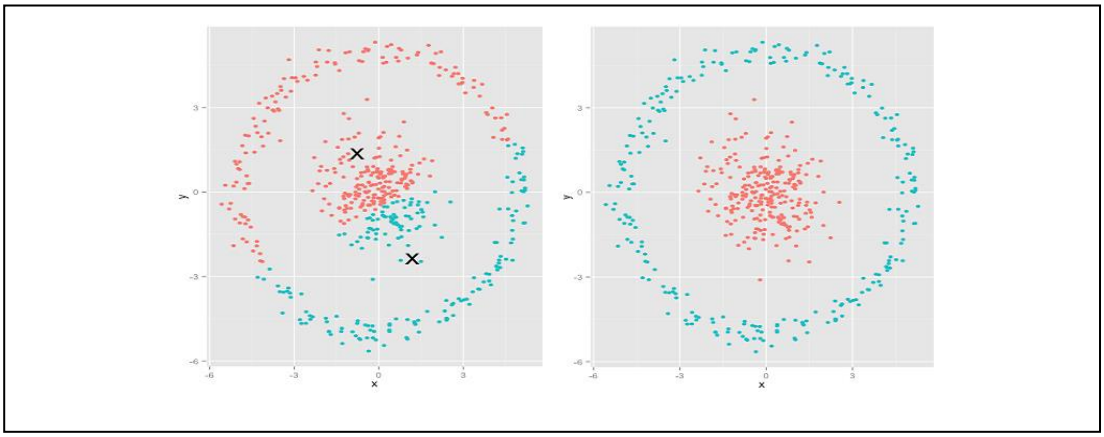
K-Means clustering

K-Means is by far the most used cluster algorithm. It is a relatively simple technique that also works on smaller datasets. With proper datapreparation is is a versatile tool that can be applied to many different types of datasets. There are, however, two limitations. The first (and often overlooked) weakness is that K-Means only works if the underlying clusters in the data are globular. If the underlying clusters are grossly non-spherical, the algorithm produces poor results.

The second limitation is that the algorithm will generate any number of clusters that it is being told to make, that is, this number has to be defined beforehand by the researcher.

The second limitation also applies to hierarchical clustering algorithms. Major advantage from these algorithms over the simpler K-Means is that the underlying clusters do not need to be globular. Hierarchical clustering also scales better than K-Means.

Text box 29. Assignments for non-spherical underlying clusters, K-Means (left) versus Hierarchical clustering (right)



In their exploratory study on the adoption of advanced manufacturing technologies (AMT) by SMEs, Uwizeyemungu et al. used a combination of several types of clustering techniques (see also (Balijepally, Mangalaraj, & Iyengar, 2011)). Data originated from an innovation survey among Canadian manufacturing SMEs. Note that with just over 600 observations this is a small dataset. The assimilation level of AMT was used as a clustering variable. For the organisational performance level, a definition of innovation was used that is close to one in the Community Innovation Survey.

Table 12. Variables measurement AMT clustering study (Uwizeyemungu, Poba-Nzaou, & St-Pierre, 2015)

Category	**Variable**	**Measure**
Clustering variable	Assimilation levels of 20 different AMT adopted	Proficiency in use of each AMT, on a scale of 1 to 5, with score=0 when an AMT is not present
Organizational performance variable	Innovation	Average percentage of sales attributed to new or modified products over the last two financial years
Control variables	Firm size	Average number of employees over the last two periods
	Firm age	Years of existence from the year of creation to the present

Subsector	OECD classification of industrial activities based on technological intensity {low-tech, med-low tech, med-hightech, hightech}

The authors first used a hierarchical (agglomerative) clustering algorithm to determine the optimal number of clusters and to determine the centroids of the clusters (these are the crosses in Text box 29). Four plausible solutions were found (with 2, 3, 4 and 8 clusters respectively). To determine which of the solutions was most stable, the same clustering algorithm was applied to a randomly selected subsample of n=300 and then to a smaller subsample of n=180. The analysis of the dendroids produced with the two subsamples indicated that the solution with 3 clusters was most stable.

Subsequently, a K-Means algorithm that was applied to the complete sample, obviously with K=3 and with the mean values found in the preceding hierarchical clustering exercise. The clustering results already convergent after 11 (out of 100 set) iterations and resulted in the following clusters:

Table 13. Clusters of AMT assimilation patterns by subsector, size, firm age and degree of innovation (expected distribution between brackets)

	Cluster 1	**Cluster 2**	**Cluster 3**
Low-tech	3.9% (7.4%)	8.9% (7,1%)	14.1% (12.4%)
Medium to Low-tech	20.6% (15.7%)	14.5% (15.0%)	21.9% (26.3%)
medium-to High-tech	3,1% (4.4%)	2.9% (4.2%)	10.1% (7.4%)
Firm size (mean)	70.9	60.1	39.9
Firm age (mean)	41.9	40	35.9
Innovation performance (mean)	0.12	0.12	0.10

These results seem rather illusive. There is no clear correlation between innovation intensity (i.e., the distribution of firms in a cluster across the different subsectors) and innovation performance. Only firm size and age (which are most likely correlated) seems to be somewhat related to innovation performance. This shows the general weakness of clustering: results of clustering are very difficult to interpret at face value; a proper interpretation needs additional conceptual development and further research (see next paragraph, (Hollenstein, 2003)).

Hierarchical clustering

Hierarchical clustering techniques can either work *bottom up* ('agglomerative clustering': aggregating individual observations to groups) or *top down* ('divisive clustering': splitting up a set into smaller subsets). In both cases the essence is to measure the dissimilarity between sets of observations, and then use a specific linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Distance-based clustering algorithms use the Euclidean distance between records as a linkage criterion.[41]

[41] In the case of quantitative variables, the Euclidean distance *d* between record given by the n-variables $(X_1, X_2, \ldots, X_n)$ and a record $(Y_1, Y_2, \ldots, Y_n)$ is $d = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$. A modified version knows as the *Mahalanobis distance* takes into account the variances and covariances of the variables.
In the case of qualitative variables, the distance between them can be calculated by coding each

The disadvantage of distance-based clustering is that they soon become very computational intensive once the number of dimensions increases – the aforementioned 'curse of dimensionality'. Therefore, usually dimensionality reduction (such as PCA or factor analysis) are first applied to the dataset.

In his 2002 Hollenstein used hierarchical clustering to profile innovative Swiss service firms (Hollenstein, 2003). Data was taken from the 1991 Swiss Innovation Survey and included 475 firms. Although the Swiss survey does not exactly follow the CIS format the (17) variables[42] that gave been used for classifying the service firms according to their innovative behaviour can be also obtained in the CIS, and include:

- *Input-oriented measures*, including expenditure for research, development, IT and follow-up investments (total and by type);
- *Output-oriented measures*: significance of the innovations in technical terms and in economic terms, IT content of the innovations, patent applications and licences granted;
- *Market-oriented measures* including sales share of new of highly improved services and cost reduction generated by process innovation.

To reduce the number of variables the study first used a factor analysis to collapse the data into five 'factors'. These are uncorrelated variables that contain information which is common to the original variables (see before, §0). Together, the five identified factors explained 56% of the total variance.

Subsequently a hierarchical cluster analysis of the identified "factors" was performed in order to group the firms into a number of categories which are as homogeneous as possible (small within-cluster variance – Ward's criterion (Ward, 1963)) and at the same time as different as possible (large between-cluster analysis). Two additional criteria were taken into account, namely the plausibility of the clusters identified (i.e. "can the clusters convincingly be interpreted as innovation modes?") and the number of firms per cluster. The algorithm found solutions with four, five and six clusters. Based on the criteria, the solution with five clusters was maintained.

Finally, the clusters where examined to see whether they could be interpretated as different *modes of innovation*. As a framework for interpretation, five types of indicators have been used, namely

- innovation indicators
- demand and supply-side determinants of innovative activity
- the firms' position in the knowledge networks
- structural characteristics of the firms
- measures of firm performances.

Based on the interpretation of the five clusters along the five types of indicators the clusters have been labelled as the following five modes of innovation for service firms:

- 'Science-based high-tech firms with full network integration'.
- 'IT-oriented network-integrated developers'
- 'Market-oriented incremental innovators with weak external links'.
- 'Cost-oriented process innovators with strong ex- ternal links along the value chain'.
- 'Low-profile innovators with hardly any external links'.

---

variable with *k* categories into *k-1* dummy variables and applying the Euclidean distance. Other options are based on similarity measures such as $s = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$

[42] Most of the variables are qualitative, either binary (yes/no) or ordinal with five response levels.

The advantage of using the (evolutionary) concept of modes of innovation over a more traditional sectoral distinction in terms of innovation intensity is that these modes can occur across different industries. Indeed, one of the key findings of the study is that firms in most innovation modes are distributed across several industries. Nevertheless, three of out five modes were heavily concentrated in specific industries. The second key finding was that economic performace was only related to the affiliation to a specific innovation mode for 1-5 out of the five modes.

In terms of enterprise profiling, these results suggests that neither the classical classification in industries nor the evolutionary classification in terms of innovation modus are sufficient to properly characterise a set of service firms. Although firms do exercise some degree of freedom in selecting a specific innovation modus their room for manoeuvre is restricted by structural characteristics closely related to the hierarchy of industries in terms of innovation intensity (Hollenstein, 2003). Therefore both types of classification should be used in parallel.

5.3.5 Classification

Logistical regression

Logistic regression (LOGIT, and its close peer PROBIT) is the classification counterpart to linear regression. Predictions are mapped to be between 0 and 1 through the logistic function, which means that predictions can be interpreted as class probabilities. Machine learning algorithms are well suited for data table statistical correlation because they are particularly good at mapping on the functional form of a data distribution without any prior assumption.

LOGIT is the work horse of innovation researchers. The interpretation from the outputs (as propensities) is relatively straightforward, overfitting can be limited (by penalizing coefficients) and the models can be updated with relatively little effort (e.g., using stochastic gradient descent). However, logistic regression has the same inherent weakness as linear regression, namely that they do not work very well when the underlying classes (i.e. decision boundaries) cannot be separated in lines. LOGIT models are therefore not really suitable to capture more complex relationships (EliteDataScience, 2017).

An important question in innovation research is whether innovation *input* (e.g., R&D) or innovation *output* (e.g., novelty of innovation) is the most important driver for firm behaviour. The micro-level model from Tavassoli focuses on the export behaviour of firms (Tavassoli, 2017). The issue here is that most of the micro-level models use R&D as a proxy for innovation. As a consequence they fail to distinguish between innovation input and output. The paper unravels the two variables and shows that actual innovation output (i.e., sales due to innovative processes) drives the exporting performance of a firm much more than innovation input (e.g., R&D). This is because the capacity of a firm to compete internationally (for instance, through introducing new products) involves much more efforts than just innovation, and in turn R&D is only one input factor (and not even a necessary one, as in the case of SMEs).

The data in the study is based on two waves of CIS surveys in Sweden which are merged with administrative data on firm-specific characteristics (e.g., export, productivity, size).[43] Export is measured in two ways: as export propensity (a dummy with exporting firms = 1) and export intensity (the amount of export per employee in the national currency).

[43] On linking CIS data with administrative registers, see before, Chapter 4.

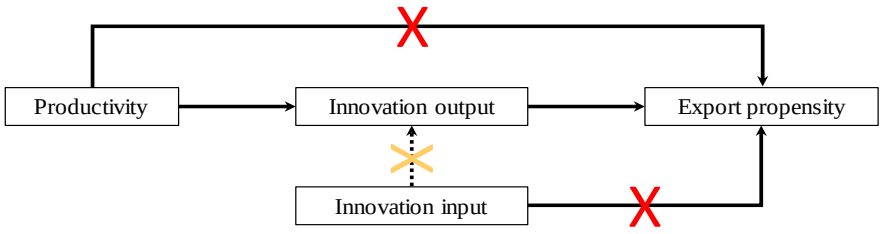
Innovation input is measured as the sum of six categories of innovation expenditures as used in CIS. Innovation output is measured as the amount of sales of innovative products per employee.

The challenge of the study (as in all studies using regression models) is to control for endogeniety between the central variables (here: export and innovation). This is where the aforementioned weakness of logistic regression comes to play: in order to countervail endogeniety one has to conceptualize the causal relations between the variables. i.e. to explicicate beforehand every pathway between the independent and the dependent variable to be able to include it is the model. The number of pathways that can be included is therefore limited. This makes classical regression models ill-suited to cover more complex causal networks with many interdependent relationships.

In the particular case of the study on innovation output and export behaviour there are both theoretical arguments and empirical evidences showing that innovation is endogenous to export. Not only could export and innovation be influenced by the same unknown variable, export and innovation could also reinforce each other (Lilischkis, Abbas, te Velde, & Korlaar, 2016). The first source of potential endogeneity is accomodated by applying panel estimators – which is possible because there are two waves of CIS data available. To deal with the second source of potential endogeniety, the study uses amongst others an alternative measure (i.e., an instrumental variable) of the dependent variable (i.e. export propensity), namely whether firms are new to exports ('export starters') or not (a second dummy with export starter = 1). This alternative specification corroborates the earlier finding that it is innovation output (and not input) that matters because it again shows the positive effect of innovation output on becoming an exporter two years later. This can be interpreted as that the innovation output induces a firm to become an exporter two years later (Tavassoli, 2017).

Text box 30. Overview of most interesting results from micro-level study on relationship between innovation and export (Tavassoli, 2017)

Although inherently limited in nature, the regression analyses shed interesting lights on the causal relationship between innovation and export behaviour. First, the well-established strong association between productivity and export turns out to be an *indirect* one: productivity drives innovation output but it is not directly related to export behaviour. Secondly, innovation output has a strong positive effect on export propensity. On the contrary, innovation input has no (or rather even a slight negative) effect on export propensity. Thirdly, there is only a low level of correlation between innovation input and innovation output. This might be partly explained by the absence of a lag structure (it takes time for innovation input to have effect on innovation output). Still, it is a suprising result that has been empirically observed many times, particularly for Swedish firms (the 'Swedish paradox') (Ejermo & Kander, 2006). Either more detailed data and/or more sophisticated models might be needed in order to explain this particular phenomenon.



***Naive Bayes*[44]**

Naive Bayes is a very simple and scalable algorithm based around conditional probability and counting. To predict a new observation the algorithm looks up the class probabilities in a probability table. The probability table gets updated by training data. The table basically *is* the model. The algorithm assumed that all input features are independent from each other. This is a rather strong assumption that rarely holds in reality (hence the label 'Naive' in the name of the algoritm). Nevertheless, given the fact that the basic assumption often does not hold in practice (i.e., the algorithm is indeed too 'naive') the model turns out to work very well in many cases, although it is often outperformed by more sophisticated models such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) (see before,

[44] Decision trees could also have been included in this section but since they can be used both for discrete (classification) and continuous (regression) output they are included in the section on regression below here.

Table 8).

Tomy and Pardede examine how the analysis and evalution of uncertainty factors with the help of data can predict the success for start-ups (Tomy & Pardede, 2017). The ability to timely identify and select emergent business opportunities is a key characteristic for successfull entrepreneurs (te Velde, 2004). It is vital for nascent entrepreneurs to assess market uncertainty factors which influence business success before making a decision.

Based in literature review the authors first identified the environmental (i.e., political, economical, social, and technological; PEST) indicators that have most influence on the success of risk of a new business. These factors are then used as a scale to predict business success. Secondly, a model was built to predict the success or failure of firms in the pre-start-up phase. The model is used to uncover the frequencies of the relations that links the input uncertainty factors with the success or failure of a firm.

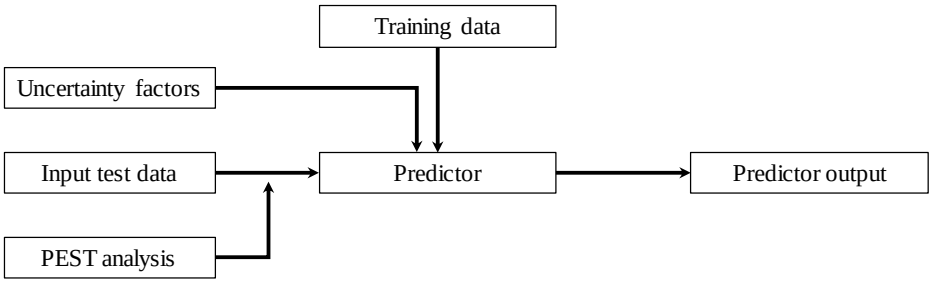


Figure 14. Success Prediction model (Tomy & Pardede, 2017)

To train the model, Naive Bayes, SVM, and k-NN algorithms have been deployed on a dataset of local survey data from Australian ICT companies (260 observations) combined with global data on entrepreneurial activities from the Global Enterprise Monitor (GEM). The two datasets were matched on the aforementioned PEST categories. As a measure of success, profitability and the Global Entrepreneurship Index has respectively been used[45].

Both datasets are small (about 250 ICT firms and 60 economies respectively) but all three machine learning algorithms that have been used work well with even small amounts of training data. All three algorithms were initially trained with 198 sampling units from the ICT survey dataset and 49 records from the GEM dataset. Subsequently, for each of the two datasets the performance of the algorithms has been evaluated in terms of accuracy, recall and precision using 49 training records (ICT survey) and 12 test records (GEM) respectively.[46] The test results have been validated by repeating the tests with other randomly choosen records from the training data sets. In both tests the simple Naive Bayes actually outperforms the other two algorithms on nearly all dimensions.

Table 14. Results of the first test (ICT survey data) (Tomy & Pardede, 2017)

Algorithm	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)	Error rate (%)
Naive Bayes	88%	81%	69%	78%	22%
SVM	87%	75%	69%	73%	27%
k-NN	83%	78%	54%	71%	29%

[45] See https://thegedi.org

[46] See paragraph 5.2 and Technical Annex 6.3 for a more elaborate description of accuracy, recall and precision.

Table 15. Results of the second test (GE data) (Tomy & Pardede, 2017)

Algorithm	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)	Error rate (%)
Naive Bayes	100%	78%	100%	83%	17%
SVM	67%	86%	40%	67%	33%
k-NN	100%	56%	100%	67%	33%

5.3.6 Regression

Linear regression

Linear regression is the most common (and most basic) algorithm for regression. Stating the obvious, linear regression performs poorly when there are non-linear relationships. A remedy is to add interaction terms or polynomials but this can be quite a hazardous and laborious process (see before, §0). There is always the danger of overfitting.

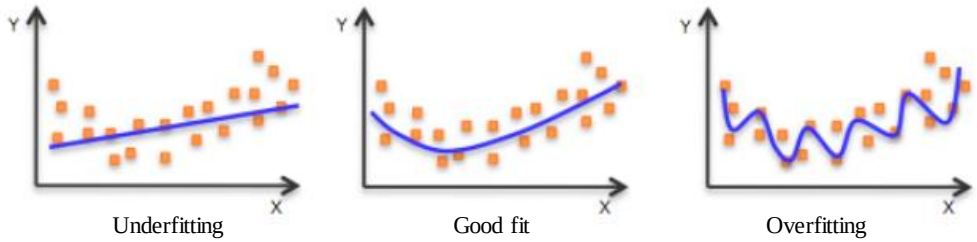


Figure 15. Visualisation of over and underfitting (source: pingax.com)

When linear regression is being used for a comparison between different sub-sets of the same population of firms, the limitations of the method could be less of a problem.

One example is the study from Arora et al. on collaborative innovation and patenting by UK innovations (Arora, Athreye, & Huang, 2016). They use UK CIS6 data to show that both patenting and external sourcing ('openness') are jointly-determined decisions made by firms. Depending on the number of *types* of external partners (one of the items in CIS6), the 329 innovative firms involved were first either classified as 'open' (>=2 types) or 'closed' (<2 types).[47] Next, using k-Mean clustering, a second split was made between 'technology leaders' and 'technology followers'.[48] The clustering was based on two variables that were directly derived from CIS6, namely *R&D intensity* (the log of internal R&D expenditure divived by the number of employees) and the *value of innovation* (the percentage of revenue from product innovation). The 2x2 classes are then described in terms of conditional probabilities (see Table 16). There are clear differences between 'open' and 'closed' firms – in the first class there there are twice as many 'technology leaders' than 'technology followers'. In the latter case, the difference is insignificant (and even slightly negative).

[47] On open innovation and knowledge flows see before, Chapter 3.3.

[48] For a decription of k-Means see before, §0.

Table 16. Percentage of firms patenting focal ('most significant') innovation (Arora, Athreye, & Huang, 2016).[49]

	Leader	Follower	Leader minus follower
Open	25.71 (5.26)	12.90 (3.50)	12.81** (6.09)
Closed	11.67 (4.18)	14.29 (3.43)	−2.62 (5.53)
Open minus Closed	14.05** (6.88)	−1.38 (4.91)	

This simple differences in conditional means obviously do not control for a variety of other factors, such as scale and industry characteristics. To include these factors the authors use a linear regression specification. The choice for a simple model is deliberate: wheres more sophisticated models (e.g., a multinominal logit) yield qualitatively similar results they require considerable more parameters (i.e., in the case of multinominal logit: three times as many). This does not involve involve much more efforts but most importantly, it greatly reduces the statistical power of tests of differences – and these tests are central to this study.

The reason that the usual limitations of linear regression are less of a problem is because when using the coefficients of the regression analysis on the four groups of firms as a measure of conditional mean of patenting, the *statistical significance of the coefficients* is less important for the analysis than the *difference in coefficient values across the groups*. Thus, the equivalence of the conditional mean for patenting is being tested (by means of F-Tests). The "difference in difference" should be positive and significant. As can be seen in the last row of Table 17 this assumption holds (all F-Values are significant), even after controlling for firm and technology characteristics (colums 3-5), and all industry fixed effects (the second column). Thus, it can be concluded that 'open leaders' patent more than 'open followers' and 'closed leaders'. The patent rate of the latter is more or less similar to both 'open followers' and 'closed followers'. In other words, the association between openness and patenting is positive and significant for leaders, and is significantly larger than the association between openness and patenting for followers.

Table 17. F-statistics for difference in estimated coefficients in the OLS model (Arora, Athreye, & Huang, 2016).[50]

		(1)	(2)	(3)	(4)	(5)
H0: coefficient of *open leader* = coefficient of *closed leader*	Difference	14.05	11.26	11.35	10.59	8.39
	F-value	*4.38***	*2.89**	*2.95**	*2.27*	*1.37*
H0: coefficient of *open leader* = coefficient of *open follower*	Difference	12.81	11.93	12.13	13.44	9.92
	F-value	*4.12***	*3.55**	*3.74***	*4.12***	*2.16*
H0: coefficient of *closed leader* = coefficient of *closed follower*	Difference	−2.62	−4.03	−4.5	−4.09	−5.47
	F-value	*0.23*	*0.55*	*0.67*	*0.54*	*1.02*
H0: coefficient of *open follower* = coefficient of *closed follower*	Difference	−1.39	−4.7	−5.28	−6.94	−7.0
	F-value	*0.08*	*0.91*	*1.20*	*1.66*	*1.75*
H0: coefficient of *open leader* − coefficient of *closed leader* = coefficient of *open follower* − coefficient of *closed follower*	Difference	15.43	16.63	16.63	17.53	15.39
	F-value	*3.45**	*3.95***	*4.34***	*4.52***	*3.47**

[49] Numbers in parentheses are standard errors, *** is significance level 1%, ** 5%, and * 10%.

[50] Model (3) includes a dummy for log employment, model (4) adds a dummy for the codification of knowledge and a dummy for the turnover from significant innovation, model (5) adds a dummy for significant innovation = a new good (hence is a product, not a service innovation).

Decision trees and random forest

Classification trees (usually referred to as 'decision trees') are the classification counterparts to regression trees. The algorithm learns in a hierarchical fashion by repeatedly splitting a dataset into separate branches that maximize the information gain of each split. This branching structure allows regression trees to *naturally learn non-linear* relationships. For the classification of datasets with complex relationships (i.e. non-linear decision boundaries) this is a critical advantage over linear regression techniques. They are also robust to outliers. However, when left unconstrained, individual trees are prone to overfitting. We will came back to this issue after the example of the use of classification trees in innovation research. This particular study uses a small tree with few branches and hence overfitting is not an issue.

Using a relatively large set (n=6,855) of firm-level micro data from the 2011 Polish CIS Lewandowska et al. tested the complementarities between product, process, and marketing innovations in the export context. Next, they explored the relationship between innovation cooperation with domestic and international partners and export intensity. Due to the existence of extremely strong asymmetry parametric models could not be used to predict the ratio of new product exports to total new product sales – a key relationship in the study. Therefore, the authors had to resort to a more complex non-parametric approach. They used a classical classification tree algorithm (Automatic Interaction Detection, AID) to evaluate the interaction between the predictors. AID is almost free of parametric assumptions. The algorithm assesses whether interaction effects eventually also occur next to main effects. The result is a neat example of enterprise classification.

The classification tree in Figure 16 shows that strong interaction effects with innovation cooperation only occur for the subsample product-process/product-process-marking innovation. This subsample also has the highest predicted share of new product exports in total new product sales (7.78%). Cooperation with foreign partners has a strong positive effect – it doubles the share (15.54%). Cooperation with domestic partners actually *decreases* the share to 4.00%. The share even drops below the share of the firms that did not undertake any innovation cooperation at all (6.50%).

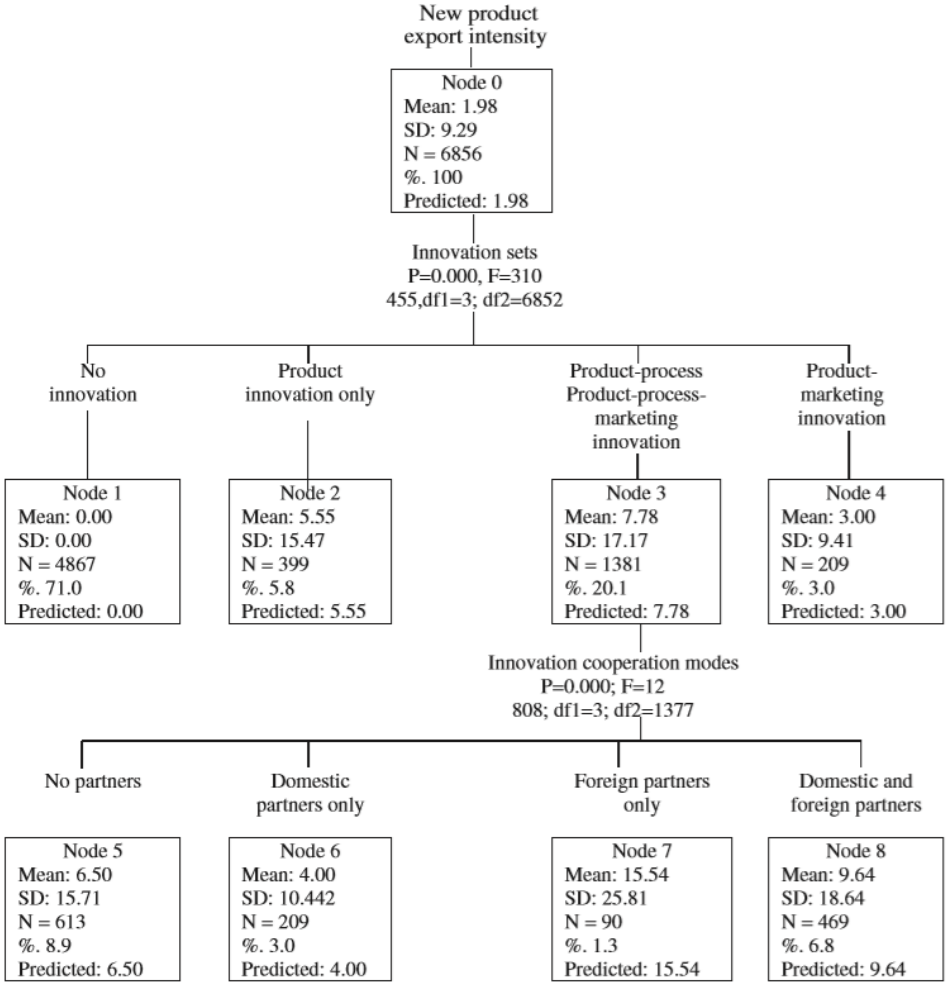


Figure 16. AID Regression tree for the relationship between innovation sets, innovation cooperation modes, and new product export intensity (Lewandowska, Szymura-Tyc, & Golebiowksi, 2016)

The *combination of many individual* trees has proven to be a successfull strategy to deal with the tendendy of classification trees for overfitting. This is indeed what ensemble methods such as *Random Forests* (RF) do. An RF model works by generating ensembles of regression trees built on independent random subsamples of the training data (Breiman, 2001). The model recursively random partitions the data set while minimizing the out-of-sample prediction error of the model. RF models have often outperformed any other classifier and they are widely used in machine learning, bioinformatics, climate science and other natural sciences (Mukherjee, 2015).

To our knowledge the computational intensive RF has not yet been applied to innovation survey data. The algorithm is however already frequently used in the neighbouring field of *innovation management.* One example is the analysis of the product development process for new online services. Hoornaert et al. have applied various sophisticated machine learning algorithms to identify variables that are most useful towards predicting idea implementation in a crowdsourcing community for such an online service (Hoornaert, Ballings, Malthouse, & Van den Poel, 2017). A benchmark of four methods was conducted in predicting whether an idea will be implemented or declined for three different modes to select ideas: *Content-based*, *Contributer-based*, or *Crowd-based.* Data was taken from the Mendeley crowdsourcing community and consisted of 7,046 ideas posted by 5,555 unique contributors during the period 2008-2014. The four methods were the classical linear discriminant analysis

(LDA)[51] and (regularized) logistic regression (LR), and the more recent techniques Stochastic Adaptive Boosting (AB) and Random Forests (RF). Each of these methods can estimate a probability of implementation for a given new idea.

When it comes to the evaluation of the idea selection modes it appears that the combination of the three modes is by far the best solution. This is a stable result across all four methods (see below, Table 18). When analysis the results in more detail it suggests that waiting for crowd data, and especially structured data (i.e., the number of votes and comments that an idea receives per day) may be worthwile: including this information improves idea selection from 18% to 48% over using content and contributor experience. *With regard to the evaluation of methods a highly relevant outcome is dat the non-linear models (AB and RF) substantially outperform the linear models (LDA, LR) when crowd data is incorporated*. This is because the former can capture non-linearities and interactions that are not captured by the latter.

Table 18. Benchmarking Model Performance over Heuristics (Hoornaert, Ballings, Malthouse, & Van den Poel, 2017)

	LDA	Regularized LR	Stochastic AB	RF
Scenario 1: Content + Contributor				
AUC	*.630*	*.629*	*.613*	*.625*
% improvement over crowd vote ranking (*AUC= .564*)	11.7%	11.5%	8.7%	10.8%
% improvement over crowd comment ranking (*AUC= .564*)	-1.1%	-1.3%	-3.8%	-1.9%
% improvement over random idea selection (*AUC= .500*)	26.0%	25.8%	22.6%	25.0%
Scenario 2: Content + Contributor + Crowd				
AUC	*.743*	*.815*	*.908*	*.899*
% improvement over crowd vote ranking (*AUC= .564*)	31.7%	44.5%	61.0%	59.4%
% improvement over crowd comment ranking (*AUC= .564*)	16,6%	27.9%	42.5%	41.1%
% improvement over random idea selection (*AUC= .500*)	48.6%	63.0%	81.6%	79.8%

5.3.7 Conclusions

For the analyst there is a cornucopia of algorithms available to cluster and classify data. However no algorithm works best for every problem. This means that the choice for an appropriate algorithm should be fit the to specific characteristics of the data set at hand. The KISS principle also applies here: although it might be tempting to use more sophisticated methods when (1) the underlying relationships are not all too complex and (2) the data quality is good simpler classical techniques (such as linear regression, logistic regression, Naive Bayes and K-means) outperform more complex techniques (such as deep learning machines, support vector machines and Nearest Neighbors). The latter also (3) require large sets of (training) data and are usually computational intensive.

In the particular case of enterprise profiling based on CIS data, conditions (1) and (2) are probably met but (3) is not hence there is little need to use more sophisticated techniques.

[51] Not to be mistaken with the other LDA, Latent Dirichlet Allocation (see before, §0).

Having said this, next to the various regression techniques that are already widely used for the analysis of CIS data, Decision trees, Naive Bayes, and hierarchical clustering seem to hold a lot of potential.

There is a certain logical order in research within which all techniques can be positioned.

Starting with CIS data, the general rule is that when there is robust a priori knowledge on the data, it should be applied in the research project, i.e. supervised learning should be deployed. Only when such knowledge is lacking, or is disputed, unsupervised learning (clustering and dimensionality reduction) could be deployed to find new research trajectories. In turn, within the branch of unsupervised learning, dimensionality reduction often proceeds clustering, e.g., to reduce the number of features. The result of classification could then be used as an input to classification such as various regression techniques (that are already widely used in the analysis of CIS data).

For exploratory analysis more sophisticated computational intensive algorithms could be used, such as Random Forest and latent Dirichlet Allocation. Both methods require large data sets. Lack of high quality data is less a problem then with traditional methods. In fact, machine learning algorithms were basically devised to deal with noisy data. Thus, other sophisticated algorithms could be used to preprocess data (noise treatment, normalization). Such exploratory analysis is often based on large sets of unstructured data, such as web pages. The outcomes of these analyses then constitute a priori knowledge for the analysis of CIS data (e.g., for the profiling of enterprises). The other way around, reasoning from the side of machine learning, the 'robust' CIS survey data with high data quality could be used to validate the outcomes of machine learning on lower quality data. This is a very important role of traditional statistics in the era of big data.

6 Globalisation and innovation

6.1 Enter globalisation

6.1.1 The scope of international innovation activities

All types of activities in a particular value chain (including design, production, marketing, distribution and support) can be undertaken by one single unified firm or be divided among several firms, in various types of institutional arrangements. Activities in a value chain can be concentrated in one location or be spread over different locations (OECD, 2013). One particular prevalent example are cross-national intrafirm operations within a multinational enterprise (MNE) group.

Much alike research & development, innovation activities follow the global dispersion of production and marketing as well as the expansion of the potential sources for technology and human resources around the world (OECD, 2013) (OECD, 2013). As products and production processes are becoming increasingly complex, firms in most industries become more dependent on using a wide range of network practices to search, access and assimilate knowledge developed outside their own firm boundaries, value chains, sector domains and immediate geographical surroundings. Typically, 15 % to 20 % of enterprises carrying out international sourcing are moving R&D and engineering functions abroad.

However, the majority of international sourcing is still regional rather than global in nature. For instance, over half of the international sourcing of R&D and engineering functions in the EU is being moved to other EU Member States.[52] The crucial point is that globalisation and localization are not opposing forces; they are rather complementary. International R&D and innovation need to tap into the *'local buzz'* and be part of very localized innovation clusters and at the same time need to built *'global pipelines'* to make sure knowledge is distributed and transported (and subsequently absorbed) (Bathelt, Malmberg, & Maskell, 2004).

6.1.2 The changes role of MNEs

There is an emerging strand in literature which emphasizes that the locus of innovation has shifted away from individual firms and their supply chains, towards territorial economies and the '*global innovation networks*' (GINs) by which they are linked (Barnard & Chaminade, 2009) (Cooke, 2013) (Parrilli, Nadvi, & Yeung, 2013). What distinguished GINs from similar notions such as 'Global Value Chains' (GVCs) or 'Global Production Networks' (GPNs) is that they are a particular form of organization at the global scale that is aimed to *solve problems.* The nature of these problems (complex and non-decomposable) requires firms to explore, identify and synthesize globally relevant knowledge (Aslesen, Herstad, & Grillitsch, 2017).

Multinational enterprises (MNEs) are often depicted as the drivers of GINs, by connecting "streams of innovation" taking place in each location (Bathelt, Malmberg, & Maskell, 2004), and reorganizing as they internationalize them (Cantwell, 2009). This is because the global

[52] Source: http://ec.europa.eu/eurostat/statistics-explained/index.php/International_sourcing_and_relocation_of_business_functions

dispersion of innovation has resulted from a scarcity of skilled resources and the need to tap into specialized expertise, and established MNEs with their extensive reach are well-positioned to find and exploit these resources and expertise. Hence MNEs are dominant in both dimensions of leveraging intramural and extramural competences (the increased 'networkedness' of innovation) and in the increasing globalisation of innovation per se.[53]

However, the character of established MNEs has changed and *new types of MNEs* has arisen. Suppliers, vendors, services providers and even buyers of all kinds joined the ranks of MNEs (Sturgeon, 2014). Several mid-sized firms from emerging markets that have been able to develop strong capabilities in the creation and management of global networks are also growing into genuine global players.

The common factor behind these developments is the rise of *external international sourcing as a* third mode of globalisation, next to the traditional modes of international trade between two independent firms (arm-length trade and intra-firm transactions within MNEs (Williamson, 1981). External international sourcing can be regarded as a hybrid. If differs from traditional arm-length trade ('market') because it requires high levels of explicit coordination (Gereffi, 1994). If differs from intra-firm transactions ('hierarchy') because the activities occur between independent firms, although one of these firms could still be a traditional MNE.[54]

GINs could be regarded as a particular appearance of external international sourcing. The distinguishing element between GINs, arm-length trade and intra-firm transactions within MNEs is the *type of coordination mechanism*. Whereas arm-length trade refers to *markets* and intra-firm transactions to *hierarchies*, GINs refer to *networks* (Powell, 1990). In this respect, GINs can be regarded as a specific type of coordination mechanism for the transfer of knowledge which is based on *reciprocal, preferential and long term relations* (Trippl, Tödtling, & Lengauer, 2009) in which *all parties are dependent on resources controlled by others*. Networks may facilitate the exchange of know-how, know-why and know-who, which is crucial for innovation and provide firms with a high degree of organizational flexibility (Powell, 1990). It is this increased networkedness of innovation activities which constitutes a particular challenge for the measurement of global innovation activities.

6.2 The challenge of measuring global innovation activities

6.2.1 The national organisation of data collection

The collection of innovation statistics via CIS is predominantly national. The main jurisdiction for data collection is a country or a region and it is usually the national (or federal) statistical office that is in charge of the actual data collection. Moreover, to ensure that innovation data can be integrated with other statistical sources CIS is aligned with the System of National Accounts (SNA) which provides a globally adopted, generic framework for measuring the economic activities of production, consumption, accumulation and the associated concepts

[53] Evidence for the key role of MNEs in international sourcing is supported by the fact that for most EU countries, 70 % to 80 % of sourcing enterprises are carrying out insourcing (i.e within the same enterprise group), while only 30 % to 40 % were outsourcing (i.e. out of the enterprise group) their business functions abroad (source: http://ec.europa.eu/eurostat/statistics-explained/index.php/International_sourcing_and_relocation_of_business_functions).

[54] In the latter case, external international sourcing has a close conceptual linkage with the notion of Open Innovation (see §3.3).

of income and wealth through the analysis of flows and stocks. The uniform definition of core concepts enables the exchange of data across countries (see §4.3.1).

However this is also where a difficulty arises with regard to the measurement of innovation activities that occur across countries. In the SNA framework, the business enterprise – the fundamental unit of analysis – is defined in terms of *legal ownership*. This is the institutional unit which has legal responsibility for its actions and can own assets, incur liabilities ad engage in the full range of economic transactions. The crucial point is that an institutional unit can *de jure* be independent but de facto be fully or largely controlled by other units. The most common case, obviously, being a domestic subsidiary of a foreign corporation (such as the aforementioned cross-national intrafirm operations within a multinational enterprise group). For CIS the important question then is to what extent the domestic institutional unit (that is under the jurisdiction of the respective NSO) decides upon innovation activities or the foreign unit.

6.2.2 The illusive notion of GINs

To further complicate matters, as described before in GINs there is often no hierarchical (i.e., no ownership) relationship between the collaboration units. The influence goes via the dependency on resources that are controlled by other (foreign) units. Although such global innovation networks are often structural in nature (i.e. they are built on long term relationships) due to the absence of legal ownership GINs are not captured by statistics that are based on the SNA framework.

There is also a mismatch at a deeper conceptual level. What is new to the GIN approach (and what separates it from the GVC and GPN approach) is that it also takes *the network itself as a unit of analysis*. This is important because innovations – understood as processes where *existing* (thus not new, as in R&D) internal and external knowledge and inputs are creatively and efficiently recombined to create new and valuable outputs (Felin & Zenger, 2014) – might only occur at the level of the network and not at the level of individual firms.

6.3 Current approaches to measure global innovation activities

6.3.1 Statistics on international business activities

There are several statistics available which are targeted to the measurement of specific global phenomena. Relevant examples include trade statistics, foreign direct investments (FDI), foreign affiliate statistics (FATS), sourcing and migration statistics. These statistics could potentially also be used to bottom-up describe patterns in international innovation. The presence of international innovation networks might be detected from the occurrence of recurrent, persistent and exclusive (hence preferential) two-way flows in goods, services, money and people between a specific set of enterprises.

Major advantage of these statistics is that the description of location is much more detailed than the (regional) splits that can reasonable be used in innovation surveys. Trade statistics

for instance cover bilateral flows in goods (yet to a lesser degree in services) between all countries in the world. FDI and FATS have a good coverage of corporate ownership.[55]

Alas a major disadvantage of trade statistics is that the unit of analysis is a country or region, not a firm. The identification of relevant interaction patterns between firms would require the avaibility of micro data at firm level. but this data is simple not available because the unit of observation is a particular product or service.

FDI, FATS and sourcing statistics do use the firm as a unit of analysis with a particular focus on ownership relationships and financial transactions. It should be noted that all question that relate to foreign affiliates in CIS2018 should correspond to the FATS related methodology of Eurostat (Eurostat, 2012).[56]A limitation of global FATS and FDI statistics is that they only cover established MNEs. This makes it particularly difficult to detect GINs.

All the aforementioned types of statistics have the disadvantage that they are not originally intended to cover innovation and innovation activities. This makes it difficult if not impossible to distinguish between innovative and non-innovative activities.

The only exception might be *international sourcing statistics* which has most conceptual overlap with the topic of innovation. This data is structured along the lines of business functions (rather than products or services) which is the most suitable classification to introduce innovation dimensions.[57] Furthermore, most sourcing surveys have variables on the underlying motivation to source specific business activities to other firms. Innovation – as a striving for improved quality or introduction of new products – could also be included.[58] Using sourcing data has the additional advantage that the statistics are not by definition limited to specific geographic or corporate boundaries. The sourcing partner can either be domestic or abroad, and can also be – and often is -- outside the own enterprise group. A disadvantage of sourcing statistics is that they are less harmonized than trade and FDI statistics and FATS and often only put out on an ad hoc basis in a limited number of countries[59].

---

[55] The coverage of business activities varies between countries and data sources. For example, the US BEA surveys cover affiliated, intra-group trade but Eurostat's data sets on foreign affiliates (FATS) and other trade in goods and services do not identify intra-group transactions.

[56] See also http://ec.europa.eu/eurostat/web/structural-business-statistics/global-value-chains/foreign-affiliates

[57] For the description of business activities several classifications are in use. The most basic classification has four classes: *Production, Sales & marketing, Transportation, logistics, and distribution, R&D and engineering* (Statistics Canada, Survey of Innovation and Business Strategy 2012). Eurostat's International Sourcing Survey (2012) and NSF's National Organization Survey (NOS 2011) distinguish three additional business activities: *Customer and after sales service, ICT services, Administrative and management functions* (including Facilities maintenance and repair).

[58] The 2012 International Sourcing Survey from Eurostat distinguishes ten motives, of which most are directly related to are relevant to innovation. These are: 'Access to new markets', 'Reduction of labour costs', 'Reduction of other costs than labour costs', 'improved quality or introduction of new products', 'Strategic decisions taken by the group head', 'Focus on core business', 'Access to specialized knowledge/technologies', 'Lack of qualified labour', 'Reduced delivery times', and 'Less regulation affecting the enterprise, e.g., less environmental regulation'. http://ec.europa.eu/eurostat/web/structural-business-statistics/global-value-chains/international-sourcing

[59] In the EU, the International Sourcing Survey has been implemented twice. In the 2007 collection round data has been collected on a voluntary basis for 13 countries: Germany, Czech Republic, Netherlands, Denmark, Spain, Ireland, Italy, Portugal, Finland, Slovenia, Sweden, the United Kingdom and Norway. In the 2011 collection round data has been collected for 15 countries: Belgium, Bulgaria, Denmark, Estonia, Ireland, France, Latvia, Lithuania, Netherlands, Portugal, Romania, Finland, Slovakia, Sweden and Norway.

6.3.2 Innovation surveys

Current measurements of globalisation are either geared towards the market form of coordination (e.g., trade statistics and input-output tables) or the hierarchical form (e.g., statistics on foreign direct investments (FDI) and foreign affiliates (FATS)). These statistics are less suitable to capture the specific network characteristics of GINs. The latter seems to be much closer to the current definition of *innovation co-operation* in CIS2018 (see §3.3).

Qualitative dedicated measurements of innovation (i.e. innovation surveys such as CIS) seem to be most suitable to capture the specific characteristics of GINs. However, most existing surveys have a limited coverage of globalisation and are carried out on samples defined at the national level. Furthermore, although many innovation surveys already have some items that could be used to construct indicators that could in principle partly measure the structure and dynamics of GINs they are not yet fully geared towards that specific goal. Innovation surveys also have some inherent limitations. These could partly be compensated for by linking on firm-level the results of innovation surveys with data from other measurements of globalisation (especially international sourcing surveys)[60].

For the proper typology of innovation networks, it should be described where specific types of innovation-related activities occur. The 'where' refers to the geographic location and the institutional type of the actor where the activity originated from (the source). The 'type of activity' refers to the business activity and to the type of knowledge or technology that is being transferred from the source to the (recipient) enterprise. For the purpose of measuring innovation activity, it is important that the questionnaire items restrict the activities in the context of implementing a new or significantly improved product or process.

For the particular purpose of the measurement of the international dimension, the geographic location of the source is also relevant. This breakdown by location of partner does not allow, however, to understand the nature of innovation cooperation, for which additional classifications on the innovation activities have to be collected. The most basic typology is the split between 'home' and 'abroad' ('rest of the world'). A further refinement would be to ask for the specific country of origin of the source. However, this dichotomy is too crude to distinguish global from international linkages/networks. The latter refers to any cross-national activity whereas *global* specifically refers to activities outside the own region.[61] Obviously, the global scale is a particular trait of GINs, GVCs and GPNs. This breakdown by location of partner per se does not allow, however, to understand the nature of innovation cooperation, for which additional classifications on the innovation activities have to be collected. Obviously, this could significantly increase the response burden in innovation surveys.[62]

Innovation co-operation usually involves the obtaining of information and might involve the acquisition of knowledge and technology. What sets it apart from the other types of transfer

---

[60] For data linkage, see Chapter 4.

[61] Note that in some literature an additional distinguishing feature is added to the notion of 'globalization', namely that it concerns the *functional integration* of geographically dispersed activities (Dicken, 2011). Economic globalization thus requires high levels of explicit coordination that differentiate it from traditional arm-length trade (Gereffi, 1994).

[62] There are basically two ways to reduce the number of options: *Static approach*. Define a limited number of countries in the list of geographic options (see for instance US Department of Commerce 2015 Business R&D and Innovation survey). *Dynamic approach*. The respondent is directly asked to indicate the country of origin of the most important sources, or the most important countries where the firm has conducted the relevant activity at hand (see for instance Statistics Canada 2012 Survey of Innovation and Business Strategy).

is that it requires *active* co-operation with the source of knowledge or technology. Hence co-operation assumes a two-way linkage between the enterprise and the source whereas the other two types are one-way inbound linkages with no further involvement from the source. In surveys, the logical order is to go from the type that requires the most effort/investment from the enterprise – co-operation – via purchase of embodied knowledge to the type that requires the least efforts (obtaining open information).

6.4 The international dimension in CIS2018

6.4.1 Questions sideways related to internationalisation

With regard to the institutional type in questions #1.1 and the more elaborate #4.7# (see below) it is asked whether the enterprise is part of an *enterprise group* or not. The answers can be used as a filtering question for enterprise profiling later on (see §5.3), or to validate enterprise group relationships (i.e., institutional linkages) at EU level that have been recorded in the EuroGroup Register (see §3.2). Question #4.7 also has a geographical component – it uses the dynamic approach to retrieve the country where the group's head office is being located (see footnote 62).

4.7 In 2018, was your enterprise part of…

		Yes	No
(a)	an enterprise group* with the head office** located in [your country]***	☐	☐
	If yes: Are most of the enterprises of that group located in your country	☐	☐
(b)	an enterprise group* with the head office** located abroad	☐	☐
	If yes: Country in which head office is located ……..***		

*A group consists of two or more legally defined enterprises under common ownership. Each enterprise in the group can serve different markets, as with national or regional subsidiaries, or serve different product markets. The head office is also part of an enterprise group.

**'Head office' means the 'Ultimate controlling institutional unit of a foreign affiliate', i.e. the institutional unit, proceeding up a foreign affiliate's chain of control, which is not controlled by another institutional unit. Consistency with the Statistical Business Registers and Statistics on Foreign Affiliates (FATS) should be assured where possible.

***For validation purposes, note that the 1st category (a) and 3rd category (b) of this question are mutually exclusive.

Before the EuroGroup Register (EGR), enterprise level data could only be linked in a limited number of cases, namely when the same enterprise was included in the sample of two or more surveys (which is unlikely except in the exhaustive strata – such as large companies) or when different enterprises are identified as part of a registered group. For international statistics, only if two (or more) enterprises included in business survey samples *and* identified as part of the same MNE group, data could be linked. The establishment of EGR greatly increases the potential of linking methods based on sample surveys (i.e. increasing the sample for analysis), for instance by taking the EuroGroup Register as sampling frame and forcing the inclusion of affiliates.

Both in the case of the use of questions #1.1 and especially #4.7 as a filter by itself or to validate enterprise group relationships across the EU it should be noted that the enterprise is solely defined in terms of legal ownership (i.e., along the lines of the SNA framework).

However, as argued before in the parts on GINs, international innovation activities also frequently occur outside the narrow scope of MNE's.

In Questions #3.4 (product innovation), #3.7 (process innovation), and #3.13 (co-operation) it is asked whether the innovations have been developed together with other enterprises or organisations. However, *no distinction is being made between intra- or extra-group collaboration*. The broad category that is being used includes both independent enterprises and other parts of the same enterprise group (i.e., subsidiaries, sister enterprises, head offices, etc.).

3.4 Who developed these product innovations?

	Tick all that apply
Your enterprise itself	☐
Your enterprise together with other enterprises or organisations*	☐
Your enterprise by adapting or modifying products originally developed by other enterprises or organisations*	☐
Other enterprises or organisations	☐

*Include independent enterprises plus other parts of your enterprise group (subsidiaries, sister enterprises, head office, etc.).
Organisations include universities, research institutes, non-profits, etc.

In a similar vein, in questions #3.9 (on R&D activities) and #3.10 (on innovation and R&D expenditure, see before, §3.4.4) no distinction is being made between intra- and extramural R&D. Note that the last item can be used as a control question to #2.9.

6.4.2 Core questions related to internationalisation

Question #3.15 is the core question on internationalisation. It explicitly covers international innovation co-operation. The focus is on the type of collaboration partner with a split for extra- and intra-group collaboration. Because of this, the geographical segmentation is fairly limited: the crude 'rest of the world' category is split in EU and non-EU. Although *institutional* intra-group linkages within the EU can be traced via EGR it is not possible to see whether these formal linkages are also being used for innovation activities.

3.15 Please indicate the type of innovation co-operation partner by location

Type of co-operation partner	*Tick all that apply* [your country]	Other EU or EFTA	All other countries
Private business enterprises outside your enterprise group	☐	☐	☐
Consultants, commercial labs, private research institutes	☐	☐	☐
Suppliers of equipment, materials, components or software	☐	☐	☐
Enterprises that are your clients or customers	☐	☐	☐
Enterprises that are your competitors	☐	☐	☐
Other enterprises	☐	☐	☐
Enterprises within your enterprise group	☐	☐	☐
Universities or other higher education institutions	☐	☐	☐
Government or public research institutes*	☐	☐	☐
Clients or customers from the public sector	☐	☐	☐
Non-profit organisations			

*but not your clients or customers.

In question #4.8 the innovation activities *within the enterprise group* are further elaborated, with a particular focus on the type of knowledge flow. On a very generic level, the remaining fourth dimension of business activity is also covered (see before, §6.3.2). Contrary to question #3.15 linkages with units *outside* the business group are not covered in this question. Also, with regard to the geographical dimension EU and non-EU are collapsed again into 'rest of the world'. Note that questions #2.8 (receiving technical knowledge) and #4.9 (receiving financial resources) can be used as a control question to #4.8.

4.8 During the three years from 2016 to 2018, did your enterprise engage in any of the following activities with one or more enterprises of your enterprise group?

	Yes, other enterprise in your country	Yes, other enterprise abroad	No
Inflows from other enterprises in your group:	*Tick all that apply*		
Receiving technical knowledge*	☐	☐	☐
Receiving financial resources	☐	☐	☐
Receiving personnel	☐	☐	☐
In-sourcing of business activities	☐	☐	☐
Outflows to other enterprises in your group:	☐	☐	☐
Transferring technical knowledge*	☐	☐	☐
Transferring financial resources	☐	☐	☐
Transferring personnel	☐	☐	☐
Out-sourcing of business activities	☐	☐	☐

* Technical knowledge includes all knowledge needed to solve technical problems in the production process; it excludes all general knowledge not specifically needed to solve particular technical problems.

6.4.3 Overall conclusion on the usefulness of CIS2018 to cover internationalisation

In order to accurately describe the nature and locus of international innovation networks the aforementioned dimensions (*institutional type, geographical location, business activitity, type of knowledge flow*) should be combined. The activities within GINs are then covered by innovation co-operation that take place outside the same business group and outside the own region. To further distinguish GIN's from traditional arm-length trade, two follow-up qualifications can be asked, namely whether the relationship with the specific source is *long-term* and *preferential*.

CIS2018 has questions on all four aforementioned dimensions. In the case of the sideways related questions they are not combined so that splits on for instance geography or intra/extra group collaboration cannot be made (§6.4.1). An exception are questions #3.15 and #4.8 that together make up the core of the coverage of the international dimension of innovation activities (see §6.4.2). However, given the fact that the dimensions are multiplied using all combination in one question would results in a great numbers of combinations. In CIS2018 a pragmatic solution has been adopted by dividing the four dimensions over the two questions #3.15 and #4.8.[63] A drawback is that no precise splits can be made at firm level. This makes it hard to identify GINs or other types of network-based innovation activities. Nevertheless, the establishment of EGR enables a better coverage of firms that are not part of formal enterprise groups. Linking CIS micro data to other data sources (esp. international sourcing surveys) could also improve the identification of cross-national innovation activites.

[63] In the Technical Annex we have proposed a pragmatic concise matrix question which combines all four dimensions.

References

Adesina, A. A., & Zinnah, M. M. (1993). Technology characteristics, farmers' perceptions and adoption decisions: A Tobit model application in Sierra Leone. *Agricultural Economics, 9*, 297-311.

Al, P., & Thijssen, J. (2003). Bespiegelingen over het waarom, de mogelijkhedenen beperkingen van micro-integratie in de sociale statistieken. In J. Nobel, M. Algera, M. Biemans, & P. v. Laan, *Gedacht en gemeten* (pp. 112-122). Voorburg/Heerlen: Statistics Netherlands.

Angotti, R., & Perani, G. (2015). La misurazione degli invesintangibili. In P. Giammarco, F. S. Rota, & C. Casalegno, *La Sfida dell'intangibile: Strumenti, tecniche, trend per una gestione cons apevole nei territori e nelle organizzazioni.* Milan: FrancoAngeli.

Arora, A., Athreye, S., & Huang, C. (2016). The paradox of openness revisited: Collaborative innovation and patenting by UK innovators. *Research Policy, 45*, 1352-1361.

Arundel, A. (2007). Innovation survey indicators: What Impact on Innovation Policy. In OECD, *Science, Technology and Innovation Indicators in a Changing World –Responding to Policy Needs (proceedings of the OECD Blue Sky II Forum).* Ottawa: OECD.

Arundel, A., & Smith, K. H. (2013). History of the Community Innovation Survey. In F. Gault, *Handbook of Innovation Indicators and Measurement* (pp. 60-87). Cheltenham: Edward Elgar.

Aslesen, H. W., Herstad, S. J., & Grillitsch, M. (2017). *Regional innovation systems and global knowledge flows.* LUND: Lund University (CIRCLE).

Bakker, B. (2011). *Micro integration.* The Hague: CBS Statistics Netherlands.

Balijepally, V., Mangalaraj, G., & Iyengar, K. (2011). Are We Wielding this Hammer Correctly? A Reflective Review of the Application of Cluster Analysis in Information Systems Research. *Journal of the Association for Information Systems, 12*(5), 375-413.

Barlet, C., Duguet, E., Encacoua, D., & Pradel, J. (sd). The commercial success of innovation: an Econometric Analysis at the Firm Level. In D. Encaoua , *French Manufacturing. The Economics and Econometrics of Innovation.* Berlin: Springer Verlag.

Barnard, H., & Chaminade, C. (2009). *Global Innovation Networks. What are they and where can we find them?* Brussels: EC (FP7 INGINOUS project).

Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Progress in Human Geography, 28*(1), 31-56.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM, 55*(4).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993-1022.

Božić, L., & Rajh, E. (2016). The factors constraining innovation performance of SMEs in Croatia. *Economic Research-Ekonomska Istraživanja*, 314-324.

Breiman, L. (2001). Random forests. *Machine learning, 41*(1), 5-32.

Buelens, B., Boonstra, H., Brakel, J. v., & Daas, P. (2012). *Shifting paradigms in official statistics. From design-based to model-based to algorithmic inference.* The Hague/Heerlen: CBS Statistics Netherlands.

Buijs, J. A. (1987). *Innovatie en interventie: Een empirisch onderzoek naar de effectiviteit van een procesgeoriënteerde adviesmethodiek voor innovatieprocessen.* Deventer: Kluwer Bedrijfswetenschappen.

Cantwell, J. (2009). Innovation and information technology in the MNE. In A. M. Rugman, *The Oxford handbook of international business* (2nd ed., pp. 417-446). Oxford: Oxford University Press.

Chesbrough, H. W. (2003). The era of open innovation. *MIT Sloan Management Review, 44*(3), 35-38.

CIS 2018 Task Force. (2017). *Harmonised Data Collection for the CIS 2018 (Final Version, 20 December 2017).* Luxembourg: Eurostat Unit G4.

Cooke, P. (2013). Global production networks, territory and service innovation: Stability versus growth. *European Planning Studies, 21*, 1081-1094.

Crespi, G., & Zuñiga, P. (2010). *Innovation and. Productivity: Evidence from Six Latin American. Countries.* Inter-American Development Bank.

Czarnitzki, D., & Lopes Bento, C. (2011). *Innovation subsidies: Does the funding source matter for innovation intensity and performance? Empirical evidence from Germany.* Mannheim: ZEW.

Czyzewska, M., Szkola, J., & Pancerz, K. (2014). Towards Assessment of Innovativeness Economy Determinant Correlation: the Double Self-Organizing Feature Map Approach. *Fundamenta Informaticae, 129*, 37–48.

Dahlander, L., & Gann, D. M. (2010). How open is innovation? *Research Policy, 39*, 699–709.

de Jong, J. P. (2000). *Measuring innovative intensity.* Zoetermeer: EIM.

Defays, D. (1997). Protecting micro data by micro-aggregation: the experience in Eurostat. *Qüestiió, 21*(1), 221-231.

Denis, A. M. (2007). The hypostatisation of the concept of equilibrium in neoclassical economics. In V. Mosini (Red.), *Equilibrium in Economics Scope and Limits* (Vol. Routledge Frontiers of Political Economy, pp. 261-279). Oxford: Routledge.

Dias, C. (2015). Statistical Matching and Data Linking. *presentation at Eurostat, 29-30 November.* Luxembourg.

Ebersberger, B., & Lööf, H. (2004). *Multinational Enterprises, Spillovers, Innovation and Productivity.* Stockholm: KTH Royal Institute of Technology.

Ebersberger, B., & Lööf, H. (2005). Multinational enterprises, spillovers, innovation and productivity. *International Journal of Management Research, 4*(11), 7-37.

EliteDataScience. (2017, May 16). *Modern Machine Learning Algorithms: Strengths and Weaknesses.* Opgehaald van EliteDataScience: https://elitedatascience.com/machine-learning-algorithms

European Commission. (2013). *Investing in Intangibles: Economic Assets and Innovation Drivers for Growth.* Brussels: European Commission, Directorate-General for Enterprise.

Eurostat. (2012). *Foreign AffiliaTes Statistics (FATS). Recommendations Manual.* Luxembourg: Publications Officeof the European Union.

Eurostat. (2015). *Summary report on the open public consultations on FRIBS.* Luxembourg: Eurostat.

Evangelista, R., Sandven, T., Sirilli, G., & Smith, K. (1997). Innovation expenditures in European industry. In European Commission, *Patterns of innovation input, innovation expenditures, non-research and intangible inputs.* Oslo: European Commission, DG-XIII.

Fagerberg, J., & Mowery, D. C. (2006). *The Oxford Handbook of Innovation.* Oxford: Oxford University Press.

Falk, M., & Falk, R. (2006). *Do Foreign-Owned Firms Have a Lower Innovation Intensity Than Domestic Firms?* Vienna: Austrian Institute of Economic Research (WIFO).

Felin, T., & Zenger, T. R. (2014). Closed or Open Innovation? Problem solving and the governance choice. *Research Policy, 43*, 914-925.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*, 1183-1210.

Filippov, S., & Mooi, H. (2010). Innovation project management: a research agenda. *Journal on Innovation and Sustainability, 1*(1).

Gereffi, G. (1994). The organization of buyer-driven global commodity chains: How U.S. retailers shape overseas production networks. In G. Gereffi, & M. Korzeniewicz, *Commodity Chains and Global Capitalism* (pp. 95-122). Westport (CT): Praeger.

Godin, B. (2017). *Models of Innovation: The History of an Idea.* Cambridge MA: The MIT Press.

Hansen, M. T., & Birkinshaw, J. (2007). The Innovation Value Chain. *Harvard Business Review*(June), 121-130.

Hervas-Oliver, J.-F., Sempere-Ripoll, F., Boronat-Moll, C., & Rojas, R. (2015). Technological innovation without R&D: unfolding the extra gains of management innovations on technological performance. *Technology Analysis & Strategic Management, 27*(1), 19-38.

Hollenstein, H. (2003). Innovation modes in the Swiss service sector: a cluster analysis based on firm-level data. *Research Policy, 32*, 845-863.

Hoornaert, S., Ballings, M., Malthouse, E. C., & Van den Poel, D. (2017). Identifying New Product Ideas: Waiting for the Wisdom of the Crowd or Screening Ideas in Real Time. *Journal of Product Management, 34*(5), 580-597.

Kline, S. J., & Rosenberg, N. (1986). An overview of innovation. In N. Rosenberg, & R. Landau, *The positive sum strategy: Harnessing technology for economic growth.* Washington DC: National Academy of Engineering.

Kloek, W., & Vâju, S. (2013). The use of administrative data in integrated statistics. Brussels: NTTS2013.

Koppers, W., & te Velde, R. A. (2017). *Statistical Reporting on Public Innovation. Current status of site-centric measurements to capture public innovation.* Urbino: University of Urbino.

Kraaijenbrink, J. (2015). *The Strategy Handbook.* Doetinchem: Effectual Strategy Press.

Laux, R., & Radermacher, W. (2009). Building Confidence in the Use of Administrative Data for Statistical Purposes. Durban: International Statistical Institute.

Lewandowska, M. S., Szymura-Tyc, M., & Golebiowksi, T. (2016). Innovation complementarity, cooperation partners, and new product export: Evidence from Poland. *Journal of Business Research, 69*, 3673-3681.

Licht, G., & Moch, D. (1997). *Licht, G. and D. Moch. Innovation and Information technologies in services.* Mannheim: ZEW.

Lilischkis, S., Abbas, J., te Velde, R., & Korlaar, L. (2016). *Internationalisation of innovation in SMEs. Case Studies, Exemplary Support Practices and Policy Implications.* Brussels: EC DG Research and Innovation.

Mairesse, J., & Mohnen, P. (2001). *To be or not to be innovative: an exercise in measurement.* Maastricht: MERIT.

Mairesse, J., & Mohnen, P. (2010). *Using Innovation Surveys for Econometric Analysis.* Maastricht: UNU-MERIT.

Marcelino-Sádaba, S., González-Jaen, L. F., & Pérez-Ezcurdia, A. (2015). Using project management as a way to sustainability. From a comprehensive review to a framework definition. *Journal of Cleaner Production, 99*, 1-16.

Mas-Verdú, F., Wensley, A., Alba, M., & Garcia Alva, J.-M. (2011). How much does KIBS contribute to the generation and diffusion of innovation? *Service Business, 5*, 195–212.

Mattes, J. (2014). Formalisation and flexibilisation in organisations. Dynamic and selective approaches in corporate innovation processes. *European Management Journal, 32*(3), 475-486.

Mazzi, G. L. (2015). *Composite indiators, synthetic indicators and scoreboards: how far can be go?* Luxembourg: Eurostat.

Mohnen, P., & Röller, L.-H. (2005). Complementarities in innovation policy. *European Economic Review, 49*, 1431-1450.

Montresor, S., Perani, G., & Vezzani , A. (2014). *How do companies perceive their intangibles? New statistical evidence from INNOBAROMETER.* JRC Technical Reports.

Morton, J. A. (1971). *Organizing for innovation;: A systems approach to technical management (An Innovation book).* New York: McGraw-Hill.

Mukherjee, U. K. (2015). *Managing the Risks and Potential of High-tech Innovations-in-use: Predictive Analytic Modeling with Big Data and a Longitudinal Field Study.* Minneapolis: University of Minnesota.

Nassimbeni, G. (2001). Technology, innovation capacity, and the export attitude of small manufacturing firms: a logit/tobit model. *Research Policy, 30(2)*, 245-262.

OECD. (2013). *Global value chains, global innovation networks and economic performance.* Paris: i4g and OECD.

OECD. (2013). *Knowledge Networks and Markets.* Paris: OECD Publishing. doi:10.1787/5k44wzw9q5zv-en

OECD. (2015). *The Innovation Imperative: Contributing to Productivity, Growth and Well-Being.* Paris: OECD Publishing.

Parrilli, M. D., Nadvi, K., & Yeung, H. W.-C. (2013). Local and Regional Development in Global Value Chains, Production Networks and Innovation Networks: A Comparative Review and the Challenges for Future Research. *European Planning Studies, 21*, 967-988.

Peneder, M. (2010). Technological regimes and the variety of innovation behaviour: Creating integrated taxonomies of firms and sectors. *Research Policy, 39*(3), 323-334.

Porter, M. E. (1990). *The Competitive Advantage of Nations.* New York: Free Press.

Powell, W. W. (1990). Neither Market nor Hierarchy: Network Forms of Organization. *Research in Organizational Behavior*, 295-336.

Raymond, W., Mohnen, P., Palm, F., & van der Schim, S. (2009). *Innovative Sales, R&D and Total Innovation Expenditure: Panel Evidence on their Dynamics.* Montréa;: CIRANO.

Roper, S., Hales, C., Bryson, J. R., & Love, J. H. (2009). *Measuring Sectoral Innovation Capability in Nine Areas of the UK Economy.* London: NESTA.

Sanguinetti, P. (2005). *Innovation and R&D Expenditures in Argentina: Evidence from a firm level survey.* Buenos Aires: Department of Economics. Universidad Torcuato Di Tella.

Sartori, D., Catalano, G., Genco, M., Pancotti, C., Sirtori, E., Vignetti, S., & Del Bo, C. (2015). *Guide to Cost-Benefit Analysis of Investment Projects. Economic appraisal tool for Cohesion Policy 2014-2020.* Brussels: European Union.

Schumpeter, J. A. (1934). *The Theory of Economic Development.* Cambridge MA: Harvard University Press.

Souitaris, V. (2002). Technological trajectories as moderators of firm-level determinants of innovation. *Research Policy, 31*(6), 877-898.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci., 25*(1), 1-21.

Sturgeon, T. J. (2014). *Global Value Chains and Economic Globalization. Towards a New Measurement Framework.* Boston, MA: Massachusetts Institute of Technology.

Tavassoli, S. (2017). The role of product innovation on export hebavior of firms. Is it innovation input or innovation output that matters? *European Journal of Innovation Management*.

te Velde, R. A. (2004). Schumpeter's Theory of Economic Development Revisited. In T. E. Brown, & J. Ulijn, *Innovation, Entrepreneurship and Culture: The Interaction between Technology, Progress and Economic Growth* (pp. 103-129). Cheltenham: Edward Elgar.

Tomy, S., & Pardede, E. (2017). The 5th International Conference on Innovation and Entrepreneurship (ICIE). *Uncertainty Analysis and Success Prediction for Start-ups.* Kuala Lumpur.

Torra, V., & Navarro-Arribas, G. (2015). Data Privacy: A Survey of Results. In G. Navarro-Arribus, & V. Torra (Red.), *Advanced Research in Data Privacy* (pp. 27-37). Cham: Springer International Publishing.

Trippl, M., Tödtling, F., & Lengauer, L. (2009). Knowledge Sourcing Beyond Buzz and Pipelines: Evicence from the Vienna Software Sector. *Economic Geography, 85*, 443-462.

UNU-MERIT. (2017, April 4-5). Profiling innovators (useful variables). *CIS 2018 Task Force.* Maastricht: Maastricht University.

Uwizeyemungu, S., Poba-Nzaou, P., & St-Pierre, J. (2015). Assimilation patterns in the use of Advanced Manufacturing Technologies in SMEs: Exploring their effects on product innovation performance. *Journal of Information Systems and Technology Management, 12*(2), 271-288.

Valpola, H. (2000). *Bayesian Ensemble Learning for Nonlinear Factor Analysis.* Espoo: Finnish Academies of Technology.

Wallgren, A., & Wallgren, B. (2011). To understand the Possibilities of Administrative Data you must change your Statistical Paradigm! *Section on Survey Research Methods* (pp. 357-365). Miami: Joint Statistical Meetings.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*(301), 236–244.

Williamson, O. E. (1981). The Economics of Organization: The Transaction Cost Approach. *The American Journal of Sociology, 87*(3), 548–577.

Technical annex

Odds ratio

The odds-ratio is calculated as follows (for industry A):

$$OR_{IndA} = \frac{INNOV\ and\ Foreign\ Capital_{IndA} \,/\, NON-INNOV\ and\ Foreign\ Capital_{IndA}}{INNOV\ and\ Local\ Capital_{IndA} \,/\, NON-INNOV\ and\ Local\ Capital_{IndA}}$$

For example, given the following matrices:

Industry A

	NON-INNOV	INNOV	Total
With foreign capital	150	450	600
Local capital	450	450	900
Total	600	900	1500

Industry B

	NON-INNOV	INNOV	Total
With foreign capital	300	450	750
Local capital	450	300	750
Total	750	750	1500

Industry C

	NON-INNOV	INNOV	Total
With foreign capital	450	450	900
Local capital	450	150	600
Total	900	600	1500

The odds-ratio of being innovative compared to non-innovative is (450/150) / (450/450) = 3. For industry B, it is (450/300) / (300 /450) = 2.25 and for Industry C is (450/450) / (150 / 450) = 3. These odds-ratios are significatively higher than 1 (asymptotic confidence intervals are [2.77, 3.23] for industry A and C, and [2.04, 2.46] for B), showing that *within each industry,* there is a positive association between the presence of foreign capital and innovativeness, contradicting the earlier inference.

Tobit model

The Tobit model relies on the existence of a latent (i.e. non-observed) variable y_i^*, where

$$y_i^* = X_i\beta + \epsilon_i,$$

and X_i is a vector of independent (exogenous) variables (such as company size, turnover, etc.), β is a vector of unknown coefficients, and ϵ_i is an independently distributed error term, assumed to follow a Normal (Gaussian) distribution with zero mean and constant variance σ^2. Then the observed variable y (in our case, the investment in innovation-related IPRs) is defined as

$$y_i = \begin{cases} y^* \; if \; y^* > \tau \\ \tau_y \; if \; y^* \leq \tau \end{cases}$$

which is censored from below at τ (in our case, 0). Usually τ_y and τ are the same value (in this example, for simplicity, they are assumed equal).

If the latent variable y_i^* is distributed as a Normal distribution with mean μ and variance σ^2, then the probability of an observation being censored is a mixture between a continuous (Normal) and a categorical distribution where all the probability in the censored area is assigned to the point τ.

$$P(censored) = P(y^* \leq \tau) = P\left(\frac{y^* - \mu}{\sigma} \leq \frac{\tau - \mu}{\sigma}\right) = \Phi\left(\frac{\tau - \mu}{\sigma}\right),$$

where Φ is the cumulative distribution function of a standard normal distribution. In addition, the probability of an uncensored observation is

$$P(uncensored) = P(y^* > \tau) = 1 - \Phi\left(\frac{\tau - \mu}{\sigma}\right) = \Phi\left(\frac{\mu - \tau}{\sigma}\right)$$

Then, the probability function of the censored variable is

$$f(y_i) = \begin{cases} f(y^*) \; if \; y^* > \tau \\ P(y = \tau) = P(y^* \leq \tau) = \Phi\left(\frac{\tau - \mu}{\sigma}\right) \; if \; y^* \leq \tau \end{cases}$$

This probability can also be expressed as

$$f(y_i) = [f(y^*)]^{d_i}\left[\Phi\left(\frac{\tau - \mu}{\sigma}\right)\right]^{1-d_i},$$

where d_i determines if the observation is uncensored (1) or censored (0).

From the definition of the latent variables is derived that its expected value is $E(y^*) = X_i\beta$. Also, the expected value of the censored variable is

$$E(y) = \big(P(uncensored) \times E(y|y > \tau)\big) + \big(P(censored) \times E(y|y = \tau)\big)$$

With ordinary least squares (OLS) there will be clearly a biased estimation.

Confusion matrix

A confusion matrix is a table that is used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known:

	Predicted	
Actually observed	No	Yes
No	True negative	False positive
Yes	False negative	True positive

From these values, various measures can be derived that are used for assessing the accuracy of information retrieval or matching. Some widely used measures are:

$$Accuracy = \frac{True\ positive + True\ negative}{Total}$$

$$Recall = \frac{True\ positive}{False\ negative + True\ positive}$$

$$Precision = \frac{True\ positive}{False\ positive + True\ positive}$$

For example, given the following matrix:

	Predicted	
Actually observed	No	Yes
No	80	15
Yes	5	145

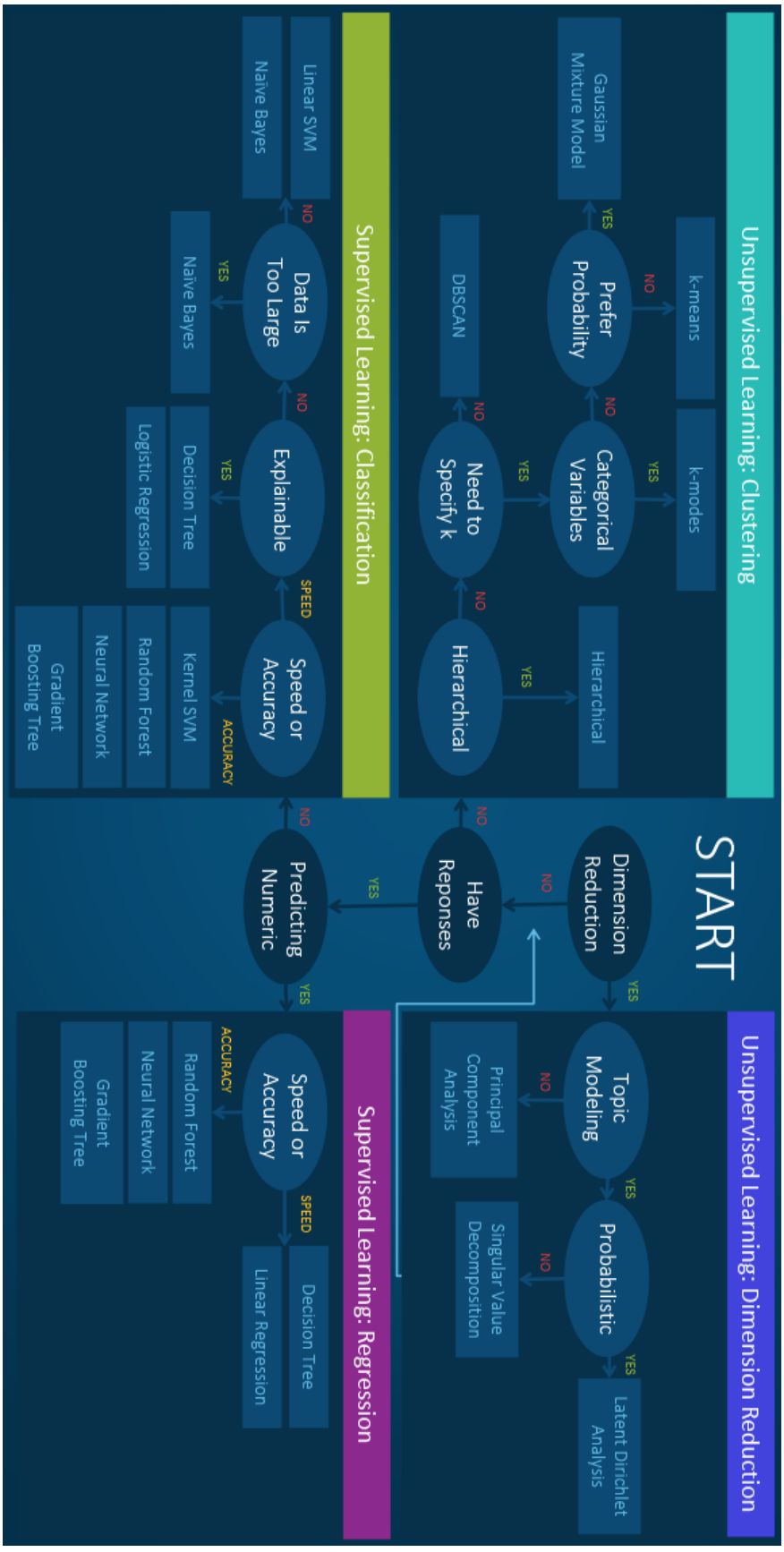
Accuracy is (145+80)/245 = 0.918

Recall is 145/(5+145) = 0.967

Precision is 145/(15+145) = 0.901

Machine Learning Algorithms Cheat Sheet

Courtesy of SAS (Hui Li)
https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use



Globalisation matrix question

- A matrix should be compiled that has the institutional types x the basic geographic classification (with three classes) as rows, and the basic classification of four or eight business activities as columns (see next page for an example)
- The matrix should be filled (with YES/NO answers) in for each of the three types of transfer of knowledge and technology, in the proposed order.
- The options in the matrix are dichotomous items: the combination in the specific cell has been an input to the innovation processes of the firm or it has not.
- The question in the first loop could be formulated as: **"For each of the four business activity mentioned below, in the context of implementing a new or significantly improved product or process, with which kind of enterprise or institution did you actively co-operate during the last three years?"**
- The question in the second loop could be formulated as: **"For each of the four business activity mentioned below, in the context of implementing a new or significantly improved product or process, from which kind of enterprise or institution did you purchase external knowledge and/or knowledge and technology embodied in capital goods (e.g., machinery, equipment, software) and services (e.g., hiring of experts) during the last three years?"**
- The question in the third loop could be formulated as: **"For each of the four business activity mentioned below, in the context of implementing a new or significantly improved product or process, from which kind of enterprise or institution did you obtain open information during the last three years?"**

1) **For each of the four business activity mentioned below, *in the context of implementing a new or significantly improved product or process* , with which kind of enterprise or institution did you actively co-operate during the last three years?**

Within the same enterprise group	R&D	Production	Marketing	Distribution
Parent company				
home				
same region				
rest of the world				
Other enterprise				
home				
same region				
rest of the world				
Controlled affiliate				
home				
same region				
rest of the world				
Outside the enterprise group, private sector				
Customer				
home				
same region				
rest of the world				
Supplier				
home				
same region				
rest of the world				
Competitor				
home				
same region				
rest of the world				
Other enterprise in the same industry				
home				
same region				
rest of the world				
Consultancy				
home				
same region				
rest of the world				
…				
Outside the enterprise group, public sector				
University or other higher education institution				
home				
same region				
rest of the world				
Public research organisation				
home				
same region				
rest of the world				
…				

Contact:
Dialogic innovatie & interactie
Hooghiemstraplein 33-36
3514 AX Utrecht
Tel. +31 (0)30 215 05 80
www.dialogic.nl