Discussion Paper

Measuring the internet economy in The Netherlands: a big data analysis

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

2016 | 14

Lotte Oostrom
Adam N. Walker
Bart Staats
Magda Slootbeek-Van Laar
Shirley Ortega Azurduy
Bastiaan Rooijakkers

Content

Key messages

- This research, a three-way partnership between Statistics Netherlands, Google and Dataprovider, is the first to combine big data with regular statistics to study the Dutch internet economy.
- As part of this research, a new set of operational definitions was developed to describe the way businesses use the internet.
- A sophisticated matching algorithm resulted in a comprehensive set of Dutch businesses and their websites, suitable for extensive further analysis.
- In 2015 550,000 businesses (36% of all businesses) had a website. Of the businesses which do not have a website, 83% represent self-employed persons.
- The core of the internet economy (online stores, online services and internet related ICT) consists of 50,000 businesses and provides 345,000 jobs (4.4% of the total) and a turnover of € 104 billion (7.7% of the total). In terms of size, the core of the internet economy is roughly comparable to industries like 'construction' or 'accommodation and food service activities' or 'transportation and storage'.
- The results show that only half of all online stores belong to the retail industry according to the Standard Industrial Classification. This indicates that industries other than retail are using e-commerce to sell their products directly to consumers.
- This study is the first to present results of online services as an separate category of businesses. This relatively young category consists of 5,700 businesses, provides 26,000 jobs and has an annual turnover of € 10 billion in 2015.
- Certain regions are more prominent in the internet economy than others. Online services are most prevalent in the regions around Amsterdam and Groningen. Internet related ICT businesses are more often based around Amsterdam and Rotterdam, and in the province of Flevoland.
- There are many opportunities for further research, which principally result from the richness of the big data source used in this study, in combination with the analytical opportunities offered by regular statistics. In particular, future research could focus on creating a time-series of data or repeating the study in other countries.

1. Executive summary

The internet is becoming progressively more important to many aspects of our lives, our societies and our businesses. Moreover, The Netherlands has a strong international position in terms of connectivity and internet usage. Demand is growing to better understand its nature and effects, while traditional statistics do not capture very well many of the specific aspects of the internet economy. As an important player in the internet economy, Google has approached Statistics Netherlands to carry out a study to deepen the understanding of the internet economy using an innovative approach, similar to a previous study done in the UK (NIESR & Growth Intelligence, 2013). This research report is the result of a first study into combining web-based (big data) sources and more classical statistical sources to get a more complete feeling of the internet economy and its impact.

As the Bean Review (2016) demonstrates, the internet has complicated our economic systems, but it is also a source of vast amounts of data with which the internet economy can be studied. With this in mind, a three-way partnership between Google, Statistics Netherlands and Dataprovider has been set up to study the internet economy. Our partner Dataprovider has extensive experience in crawling the internet to collect data on websites on a regular basis, in particular companies' websites. In the current research project, their data are made available to Statistics Netherlands. This data source constitutes the innovative 'big data' aspect of the internet economy. In addition, Statistics Netherlands possesses a large amount of statistical data on the businesses within the economy. It remains then to link Dutch websites to Dutch businesses. Doing so facilitates a deeper analysis of the internet economy.

Another important reason for this research is that there is not yet a broadly accepted definition of the internet economy. This research contributes to this debate by constructing a pragmatic definition within the context of the available web-based source. Importantly, the definition was formulated in cooperation with stakeholders from Dutch government, business-world, academia, Google and Dataprovider. The resulting definition classifies businesses with websites into various categories depending on how a business makes use of the internet. These categories are:

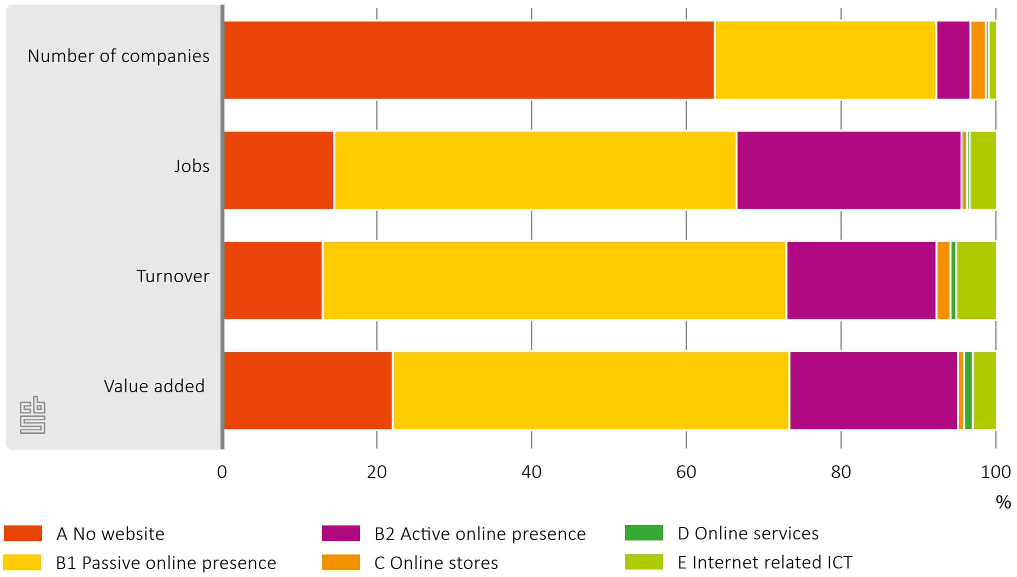
- A: Businesses without websites
- B: Businesses with a passive (category B1) or active online presence (B2)
- C: Online stores
- D: Online services
- E: Internet related ICT

Websites are allocated to these categories predominantly according to the information available from Dataprovider. Further, Categories C, D and E as a group constitute the 'core' of the internet economy. The core consists of online stores, online services such as dating sites, price comparison sites, or online entertainment, and of internet related ICT such as app developers, web-hosting and internet marketing. Outside of the core we distinguish two further types of online presence for businesses: active and passive. Active online presence means that businesses provide a manner to interact with them directly, such as making a reservation or ordering a brochure. Passive online presence means that businesses purely use the internet to provide information about their activities and to publicise their organisation.

To analyse the internet economy characteristics in a coherent way, the characteristics of the websites need to be linked to statistical information on the businesses behind the website. This implies a nontrivial methodological challenge which is dealt with using two key pieces of information. Firstly, Statistics Netherlands records the websites of business in its General Business Register (GBR). Secondly, businesses often report their Chamber of Commerce (CoC) number on their website. These identifiers provide the basis upon which websites can be linked to businesses. The successful linking to the GBR facilitates further links to a variety of Statistics Netherlands data sources. These data sources allow us to build an understanding of the characteristics of the internet economy from a variety of perspectives including, turnover, employment, and geography.

Our analysis identifies circa 550,000 businesses which are in some way present on the internet. This constitutes 36% of all businesses. Of the businesses which do not have a website, 83% represent self-employed persons. Of all self-employed persons, we find that almost 70% do not have a website. The characteristics of the internet economy are summarised in the following figure. Using four economic indicators, the figure shows both the share of the internet economy of the whole economy as well as the distribution of the internet economy across the categories.

1.1 Relative distribution of number of companies, jobs, turnover and value added by Internet categories, 2015



This shows that the majority of business with websites fall in the category of passive online presence. Many business thus use the internet predominantly to share information about their business online. Active online presence is the next largest category in terms of the number of business, followed by the categories within the core: online stores, internet related ICT and online services. We find that the core constitutes a modest but appreciable proportion of the economy as a whole. The core consists of 50,000 business (3.3% of the total). In 2015, the core of the internet economy provided 345,000 jobs (4.4% of the total) and a turnover of € 104 billion (7.7% of the total). In terms of magnitude, the core of the internet economy is roughly the same as the sectors ‘construction’ or ‘accommodation and food service activities’ or ‘transportation and storage’.

We also identified some interesting patterns in the results, even though we can make no inference regarding causality. We find that businesses without a website, account for proportionally much less turnover than those with websites. We also see that businesses with an active online presence, account for proportionally more employment. Again, without inference of causality, we see that businesses with websites generally (with the exception of online stores) proportionately account for more value added in the economy. Geographically, the results show that certain areas are more prominent in the internet economy than others. Groningen, in the north of the Netherlands, is important with respect to many categories of the internet economy and Amsterdam constitutes an important hub for online services and internet related ICT.

Given the exploratory nature of this study, we carefully consider the limitations and strengths of our data and methodology. Here we quickly mention some of the more intuitive issues.

Regarding the data:

- It is estimated that 95% of all Dutch websites are included in the Dataprovider data so the coverage is not complete.
- Many, especially small, businesses use social media (Facebook for example) for their online presence and we have no data about these pages.

Regarding the linking process:

- Not all businesses report their CoC number on the website and not all businesses record their website in the GBR.
- A business may have many websites which fall under different categories. This complicates allocating a business to a given category.

The strengths of this research lie in the innovative approach and the use of big data in combination with classical statistical sources. The rich data set opens up many opportunities for further research, which could not be covered in the context of the current project. For example, creating a time-series of data on the internet economy opens up many possibilities for analysis of trends and for gaining deeper insights in evolutions. We could for example look at how the different categories of the internet economy are developing over time. We can also look at the changes in businesses within different categories over time. This would allow us to gain insights into the extent to which businesses in particular categories grow at different rates, or indeed close down. The approach developed in this project of linking website information with classical statistics could also be used in other countries. Finally, we consider the possibility for broader methodological advances which could facilitate more accurate and/or more detailed study of the internet economy in the future. Machine learning approaches show much promise for this kind of analysis, and could potentially be used to simultaneously allocate websites to categories and link them to General Business Register.

2. Introduction

Over the past few decades the internet, ICT and digital products and services have provided growth and/or start-up opportunities for many businesses. It is commonly understood that the economy and the internet are now inseparably linked. The recent proliferation of online stores for example has fundamentally altered the nature of retail and consumer spending habits. New key industries have sprung up around ICT and software creation which fundamentally depend on the existence of the internet to do business. It is in fact hard to think of many significant economic activity which does not make use of the internet in some way. We can therefore speak in general terms of the internet economy and wish to gain a better understanding of precisely what it is and what its characteristics are. This report is an experimental step towards defining and understanding the internet economy, using the internet economy in the Netherlands as a case study.

The Bean Review (2016) makes an important argument regarding the internet economy. Namely, the report analyses the challenges resulting from the expansion of the internet economy for our ability to measure economic activity. Going a step beyond this, the report also demonstrates how the internet provides a plethora of data with which can be employed to better understand the economy. One example is the use of online data on job vacancies to improve our understanding of the labour market. This research can be considered within the broader context of the Bean Review as moving towards both a better understanding of the internet economy and also a better understanding of how the internet can be employed to derive new statistics and insights into social and economic phenomena.

This research is conducted in cooperation with Google and Dataprovider. Due to the leading position of the Netherlands in terms of connectivity, internet usage and innovation on the internet, Google has approached Statistics Netherlands to engage in research on the internet economy and to use big data in this analysis. The integration of this big-data source within our research is of fundamental importance to fulfilling the research remit. This research also fits the strategic priorities within Statistics Netherlands. Statistics Netherlands aims for innovation in statistics. This project necessitates the development of new methods, use of new data sources and new partnerships. Also, this project builds on Statistics Netherlands' focus on big data. Accordingly, Statistics Netherlands is uniquely positioned to address the following central research aim:

To explore the possibilities to deepen our understanding of the importance of the internet economy.

More specifically, the main research tasks are:

1. Construct a definition of the internet economy that: a) reflects the beliefs on what the internet economy is and b) pragmatically considers the possibilities of big data analyses.
2. Show the importance and size of the internet economy as part of the Dutch economy.
3. Show, by proof of concept, the possibilities of new measurement methods for producing statistics.
4. Explain differences from standard statistical concepts/classifications and existing statistics.

Explicit in the research remit is the use of big-data in pursuing these research tasks. Dataprovider is a Dutch company which crawls the web and structures and provides the data. This data source gives a unique insight into the structure and contents of the internet and provides many variables which are useful to study the link between the economy and the internet.

The results of this work feed into policy discussions. The contribution of the internet economy depends fundamentally on the available infrastructure. The availability of high-speed internet facilitates the internet economy, but is also costly. Understanding the size and workings of the internet economy therefore constitutes important policy related knowledge. This report is especially relevant in this context because we present results spatially. This research is therefore also relevant for spatial economic policy.

In this report, we begin in chapter 3 by looking at the existing literature on the internet economy in order to produce a conceptual definition of the internet economy. We then consider the data to which the definition can be applied in chapter 4. In chapter 5, we explain how the data is processed and linked together in addition to explaining how the conceptual definition of the internet economy is operationalized. Chapter 6 presents the results in terms of the demographics of the businesses in the internet economy, the regional distribution of the internet economy, and other economic indicators such as employment and turnover. Chapter 6 discusses these results, with a particular focus on evaluating the methodology, as befits an experimental study such as this.

3. Approaches to defining the internet economy

There is currently no broadly accepted definition of the internet economy. It is however, broadly accepted that it is difficult to construct a unambiguous definition of the internet economy. This ambiguity is highlighted by NIESR and Growth Intelligence (2013) who state that 'The 'digital economy' is not straightforward to define, as it is variously used to refer to a set of sectors, a set of outputs (products and services), and a set of inputs (production and distribution tools, underpinned by information and communication technologies)'.
One potential approach to deal with the challenge of conceptually defining the internet economy is to be very broad. This approach has been adopted by the OECD, who define the internet economy as:

'the full range of our economic, social and cultural activities supported by the internet and related information and communications technologies' (OECD, 2013).

This definition is broad, in the first instance, in its inclusion of social and cultural activities, and in the second instance in that these activities need only be 'supported' (thus not entirely facilitated) by the internet. Further, the inclusion of 'related….technologies' exacerbates the ambiguity/broadness. This definition was however created as part of a 'vision for the internet economy'. It is therefore not a definition which was determined in order that it be applied to data. Nonetheless, this definition demonstrates recognition of the interconnectedness of the internet and the economy.

A more practical definition is provided by the Boston Consulting Group:

'The share of GDP that consists of online consumption and purchases, of investments in internet capacity and of net exports of internet services and products.' (Boston Consulting Group, 2011).

This definition is clearly constructed in order to determine the size of the internet economy by allocating specific activities to the internet economy. Importantly, this definition represents a macro approach to the definition of the internet economy. Specifically, it focuses on disaggregating GDP and then re-aggregating the components which can be allocated to the internet economy. We see then that this is a definition constructed with the pragmatic task of delimitating the economy in mind.

The McKinsey Global Institute (2011) defines internet related activities according to the following classification:

- Internet related services (e-commerce, content and other utilization of internet)
- Telecommunication related to internet (e.g. Broadband)
- Software and services (e.g. IT consulting and software development
- Hardware (e.g. computers or Smartphones)

This definition is much more specific than the other definitions which we have considered up to now and hints towards a more micro-orientated approach to defining the internet economy. To illustrate what is meant with a micro-data approach let us consider a definition constructed by Statistics Netherlands in related work (Statistics Netherlands, 2015). The topic of Statistics Netherlands (2016) is not the internet economy but the ICT economy, which is defined as:

'The businesses which according to the Standard Industrial Classification (SIC) belong to category 26: ICT industry, 46: wholesale ICT equipment and 582, 61, 62, 63 and 95: collectively, the ICT service sector.'

The approach in this definition fundamentally differs from that of the Boston Consulting Group, because it is a micro, not a macro approach. While Boston Consulting Group focuses on disaggregating aggregate indicators, Statistics Netherlands uses its array of micro-data (at businesses level).

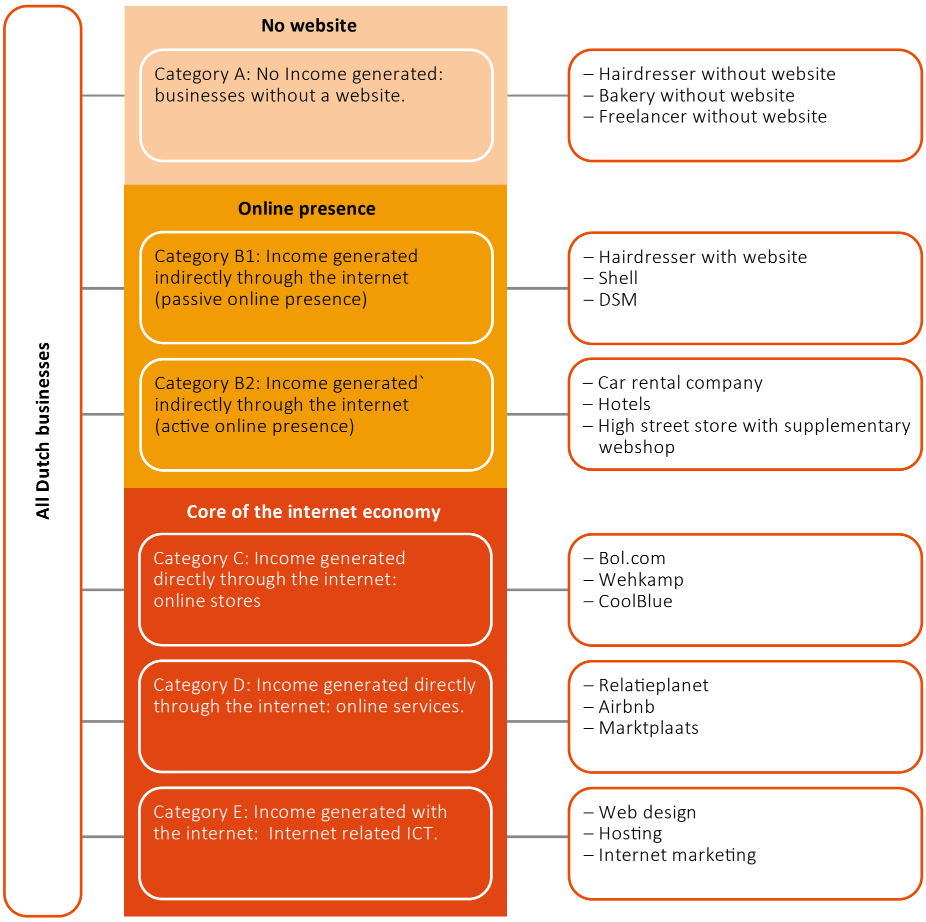
The internet economy could be defined by making use of the SIC by adding the SIC codes for retail sales via internet (4791) to the definition of the ICT economy of Statistics Netherlands (2015). There are several reasons why we do not adopt the SIC categorisation for the definition of the internet economy in this discussion paper. Firstly, there are doubts about the ability of the SIC categorisations to deliminate the internet economy in concordance with the developing ideas regarding what the internet economy is (NIESR and Growth Intelligence, 2013). This is principally because of the interconnectedness of the internet within many facets of the economy. The internet economy clearly exists outside of the SIC categorisations which, at first glance, would be used to deliminate it. An example is the proliferation of online stores among businesses whose principal activity falls into non-retail SIC codes. This problem is particularly prominent because of the speed at which businesses are incorporating the internet into their business activities while the SIC code of a business is predominantly determined by the business activity at the time of registration at the Chamber of Commerce. Further, the SIC categorisations have not been updated since 2008. Consider for example, the concept of an 'app', which was much less prominent in 2008. For these reasons, it is desirable to consider alternative methods for deliminating the internet economy than the application of the SIC categorisation.

This discussion paper adopts a micro-data approach for studying the internet economy but employs an alternative method to the traditional method of employing the SIC categorisation. Micro-data approaches in general provide the most detail and the greatest breadth of possibilities for analysis. Further, the Dataprovider data also exist at a micro-level (the website level). We can therefore build our study of the internet economy by combining data sources at the micro level in order to maximise the analytical possibilities.

In this discussion paper, instead of looking at a categorisation of the economy and considering which aspects of it can be attributed to the internet economy, we begin with a categorisation of businesses in terms of their relationship with the internet. This categorisation was derived through analysis of Dataprovider data and by consultation with stakeholders (a steering group consisting of representatives the Dutch government, business-world, academia, Google and Dataprovider) and is shown in Figure 1.

The categorization shown in figure 1 can be considered as a 'micro-data' approach. We thus began with the smallest unit (the business) and considered how these businesses, in terms of how they use the internet, could best be grouped together into diverse categories. The easiest category to define is Category A, which consists of businesses without a website.

3.1 A categorization of the businesses according to their use of the internet


All Dutch businesses
No website
Category A: No Income generated: businesses without a website.
– Hairdresser without website
– Bakery without website
– Freelancer without website
Online presence
Category B1: Income generated indirectly through the internet (passive online presence)
– Hairdresser with website
– Shell
– DSM
Category B2: Income generated` indirectly through the internet (active online presence)
– Car rental company
– Hotels
– High street store with supplementary webshop
Core of the internet economy
Category C: Income generated directly through the internet: online stores
– Bol.com
– Wehkamp
– CoolBlue
Category D: Income generated directly through the internet: online services.
– Relatieplanet
– Airbnb
– Marktplaats
Category E: Income generated with the internet: Internet related ICT.
– Web design
– Hosting
– Internet marketing

These business are as such not considered part of the internet economy. In this way, our definition of the internet economy deviates from other definitions, such as that of the 'digital economy' as employed by NIESR and Growth Intelligence (2013). In that study, a business can be considered part of the digital economy even if it does not have a website.

Categories B through E are all businesses with websites and can therefore, to varying extents, be considered part of the internet economy. The differences between these four categories revolve around how the business generates income in relation to the internet. Category E consists of businesses which make the internet possible. They are the web designers, the hosting companies and the internet marketers. Cloud services and app design among other services fall into this category. If the internet did not exist then these business could fundamentally not exist. This category is therefore referred to as 'Internet related ICT'.

The distinction between Category D and E can be loosely understood by considering the nature of the services involved. The provision of cloud services is inextricably linked with the internet and thus cloud services belong in Category E. Dating services existed before the internet. Dating services now make extensive use of the internet: they have in fact become inextricably linked to the internet, but they would still exist without the internet. Another example is the housing market, which is now facilitated greatly by the internet, but would still very much exist without it. This category is therefore referred to as 'online services'.

Category C consists of online stores, which is defined as businesses with e-commerce activities. While this category seems simple, it is also the category for which the problem of internet dependence for revenue creation is most prominent. For example, amazon.com generates all of its sales via internet: it has no physical/high-street shops. We can therefore attribute all of the revenue of amazon to the internet economy. The Dutch department store Bijenkorf however sells both through its traditional shops and through the internet. Therefore, while many businesses have an online store, they can only be considered partially part of the internet economy. We will explain later how allowance is made for this.

Together, categories C, D and E are considered to be the 'core' of the internet economy. Businesses which have a website but do not fall into the core of the internet economy are considered business with an 'internet presence'. Category B consists of businesses that only make indirect use of the internet to generate revenue. This category is therefore referred to as the 'internet presence' category. Generally, these websites provide information about non-internet related business activities. For example, a consultancy firm uses a website to provide information about its services, publicise it work, place job advertisements and to display information for potential and existing employees.

Given this categorisation of the internet economy, we can now be clear about what is not included in the definition of the internet economy. What is not included is predominantly determined by the available data. Dataprovider can only provide information on the publically available internet. This means that many business-to-business uses of the internet cannot be included in this study. Consumer-to-consumer economic activity can also not be measured. The best example of this is marktplaats.nl (the Dutch equivalent of Craigslist or eBay). Marktplaats as a business is included, but it is not possible to consider the transactions between consumers that are facilitated by Marktplaats. As such, our definition is limited to economic use of the publically available internet and excludes consumer-to-consumer transactions.

In the following section, we will introduce the data which allowed us to allocate businesses to a given category and to derive indicators to analyse the properties of the internet economy according to the above definition.

4. Data sources

Obviously, Statistics Netherlands possesses a great deal of data on the economy at the level of businesses. In order to study the internet economy, data on the internet is also necessary. This data is provided by the (aptly named) Dutch business Dataprovider[1)]. In short, Dataprovider uses web-crawling to index the internet. This data is then structured and made available to clients. Dataprovider shares this data with Statistics Netherlands. Vitally, this data often contains the Chamber of Commerce (CoC) number of the business which is the subject of the website. The CoC allows the internet data to be linked to data about the business connected to the website. In this way, a dataset is built which combines information on the economy and the internet, and as such provides insight into the internet economy.

[1)] https://www.dataprovider.com/.

4.1 Internet data: Dataprovider

The resulting Dataprovider database which Statistics Netherlands has been provided with is in principal a list of all Dutch websites. Each website is described according to a set of variables including business names, chamber of commerce numbers, shopping cart systems and site traffic estimation, among many others. The data is updated monthly.

Dataprovider deals with the important question of whether or not a website is a Dutch website. Dataprovider treats all websites with a .nl Top Level Domain (TLD) as a Dutch website. If a website is hosted on a .com TLD then the following approach is taken. If a .com websites uses the Dutch language it is recorded as a Dutch website. If the websites uses a .com TLD and makes no use of the Dutch language then it is a Dutch website if it is hosted in the Netherlands and displays either a Dutch address or telephone number. According to these decision rules, the Dataprovider web-crawler finds approximately 2.5 million websites on the publically available internet in the Netherlands in March 2016. This is estimated to represent 95% of Dutch websites[2)]. There are three principal reasons why a website of a business can be missed by Dataprovider. Firstly, a website with no links to it cannot be found. Secondly, a business which uses Facebook (or some such website) for its internet presence will not be identified[3)]. Thirdly, Dutch websites which use a .com TLD, are hosted from abroad and do not use the Dutch language or have a Dutch address or phone number will not be identified as Dutch. For a detailed list of a the variables in this Database see Appendix A.

Dataprovider also provided us with an additional dataset referred to as the Call To Action (CTA) database, which we use to complement to the main Dataprovider database. The CTA database contains all Dutch websites for which there is at least one way in which a user can, loosely speaking, ‘interact’ with the website. There are six ways in which a user can interact with a website: order, buy, view the shopping cart, make a reservation/booking, subscribe or register. If an such an interaction is facilitated by a hyperlink or a button, then this website, and the methods of interaction, are recorded in the CTA database. We use these variables in this research to refine our understanding of the activities of websites.

4.2 Economic data

Statistics Netherlands makes use of diverse internally available data sources to provide data on the economics of the internet economy. The General Business Register (GBR) provides the backbone and structure of the dataset. All the other datasets are used to complement and enrich this backbone. The last step in our methodology is to link all of the enriched datasets to the GBR in order to provide as much information as possible on the nature of the internet economy. This section first describes the GBR and then proceeds to describe the various datasets which are used to enrich the GBR.

General Business Register (GBR)
The GBR is a database which structures businesses in the Netherlands. A given ‘business’, roughly speaking, can consist of many smaller businesses, or be subsumed into a larger businesses. When ordinary people think of a business, they are generally thinking of, in

[2)] This estimate is derived using the database of Stichting Internet Domeinregistratie Nederland, which registers all Dutch domain names (Dataprovider 2016).

[3)] Dataprovider could not provide data from Facebook because Facebook does not allow third party web crawlers or indexing to take place on its site.

Statistics Netherlands terminology, the Enterprise Group (EG). The EG is the top level of business aggregation. The best example is the Dutch business Phillips, which consists of many smaller parts that operate independently on a day-to-day basis. These separate parts of the EG are 'Business Units' (BUs)[4]. Alternatively, smaller EGs may only consist of one BU. Many of the statistics which Statistics Netherlands possesses are at the BU level. It is important to note that the relationship between the CoC number and the BU is not one-on-one. A BU may have more than one CoC number, and this relationship is made clear in the GBR. The business unit may also consist of multiple Local Business Units (LBUs). Business units in the retail industry often have multiple LBUs, which in that case are simply the multiple shops owned by the same business.

The GBR provides information on the 1) size, 2) sector and 3) age, of BUs. The size (point 1) of a BU is determined by the number of employed persons. The number of employed persons includes:

- all employees and managers on the payroll,
- employees on the payroll of other companies or institutions but employed by their own company or institution and as such in fact belonging to the staff (hired staff),
- employed owners, partners, partnership members and participating family members,
- temporary workers,

but does not include employees seconded out to other BUs. In this study we group BUs into the following 4 size classes: 1 employed person, 2–49 employed persons, 50–249 employed persons and 250 or more employed persons. Sectors (point 2) are groupings of businesses according to their main activity. Statistics Netherlands uses the SIC to classify enterprises by their main activity. The SIC is a hierarchic classification of economic activities[5]. As such, there are up to 5 digits in a SIC code, but even the first digit alone provides some information on the nature of the sector. In this study, we focus mainly on the first two digits so as not to get caught up in unnecessary detail. Lastly, age (point 3) is determined by the date given at registration. This date is corrected for mergers, take-overs, spin-offs etc.

Production Statistics (PS)

PS provide a picture of employment in, and the financial position of, businesses in the Netherlands. Statistics Netherlands compiles production statistics for the following sectors of industry: mining and quarrying, manufacturing and construction, production and distribution of energy and water, repair of consumer goods, wholesale and retail trade, hotels and restaurants, transport, storage and communication, business and personal services, environmental services and health and welfare. PS are derived from surveying a stratified random sample of businesses which are part of the 'business economy' (see box 1). For small businesses (fewer than ten employees) data for the PS are taken from tax data as much as possible. Businesses with fewer than 50 employees receive a questionnaire on a sample basis and businesses with 50 employees or more are all included in the survey. The sample size varies strongly per sector of industry, as the number of businesses also varies per sector. Overall more than 80,000 enterprises are invited to participate in the survey, about 10% of businesses in the Netherlands.

[4] A Business Unit is a statistical unit that groups all the parts of an enterprise contributing to the performance of an activity at class level (4-digits) of NACE Rev. 1 and corresponds to one or more operational subdivisions of the enterprise. We choose the term Business Unit because it is the most accurate translation of the Dutch term used. The appropriate term according to eurostat is 'Enterprise Unit', which we avoid because of the confusing abbreviation (EU).

[5] The SIC is based on the classification of the European Union (Nomenclature statistique des activités économiques dans la Communauté Européenne, or NACE) and on the classification of the United Nations (International Standard Classification of All Economic Activities or ISIC).

The business economy

The ‘business economy’ is a subset of economic activity which relates directly to commerce. An example of economic activity which is not part of the business economy is government spending (education, policing), which contributes to Gross Domestic Product, but not directly to the business economy. In this study, we present some results for the business economy and some results for the economy as a whole. Specifically, results in terms of turnover, value added and production[6] are presented for the business economy. One important reason for this is that only businesses can create value added (in the economic sense of the word). Government activities on the other hand, while being very valuable, do not create value added in a way that is observable in micro-data, and therefore we do not include turnover outside of the business economy. Other indicators such as the number of businesses and jobs are presented for the entire economy. Jobs are considered good for the economy regardless of whether they are in the business economy or not. Accordingly, we measure all businesses and jobs within the internet economy. An overview of which indicators are presented for the business economy and which are presented for the whole economy is shown below.

4.1. Overview of population per indicator

Indicator	Population
Number of businesses	Whole economy (NACE codes A-U)
Age	Whole economy (NACE codes A-U)
Size	Whole economy (NACE codes A-U)
Sector	Whole economy (NACE codes A-U)
Employees	Whole economy (NACE codes A-U)
Jobs of employees	Whole economy (NACE codes A-U)
Province and COROP	Whole economy (NACE codes A-U)
Turnover	Only business economy (NACE codes B-J, L-N)
Value added	Only business economy (NACE codes B-J, L-N)
Production value	Only business economy (NACE codes B-J, L-N)
Employed persons	Only business economy (NACE codes B-J, L-N)
Full-time equivalents	Only business economy (NACE codes B-J, L-N)

Baseline

Baseline complements the GBR with data from the tax office regarding the Value Added Tax (VAT) and profit declarations received from the BU. In combination with the PS, this allows us to determine the following variables at BU level:

- Production value; the value of the goods and services produced, valued at basic prices[7].
- Value added; the value of all goods and services produced (‘production value’ or ‘output’), minus the value of the goods used as inputs to production.
- Employed persons; these are all persons who are working in one or several jobs as employees or as self-employed for a resident institutional unit (company, institution or household). Employed persons include all persons who have a paid job for at least one hour a week.

[6] We present the indicators employed persons and full-time equivalents for the businesses economy only. These variables come from a data source which is only concerned with the business economy.

[7] Basic prices are defined as the prices experienced by the producer. As such product related taxes have been subtracted from the original prices, and subsidies haven been added to them.

- Full-Time Equivalents (FTE); the full-time equivalent is obtained by dividing the annual contractual hours of the job by the annual contractual hours considered full-time (in the same company). Two half-time jobs add up to one full-time equivalent.

Because Baseline extracts the data from VAT and profit declarations, only businesses who are obliged to provide this information are in the dataset. For example, companies that work in the supply of water are exempted from taxes on their profits, therefore they are not captured with Baseline. Due to this incompleteness, we only use the data to describe the business economy (see box 1). In addition, the VAT and profit declarations from large and complicated businesses are difficult to merge with the GBR. As a consequence they are occasionally not included in the Baseline dataset. Given the relative importance of some of these companies, we will use data from the PS to include their production, value added and employment.

Turnover Statistics (TS)
Because data in the PS only contains a sample of businesses, the TS is employed to improve coverage. Statistics Netherlands has developed a mixed-source production system that uses VAT data for the smaller units and sample survey data for the largest units to produce quarterly and yearly revenue statistics. This way, turnover is available for very nearly the entire population. The turnover statistics exclude VAT and include returns from both primary and secondary business activities.

Regiobase
Regiobase is a database designed to provide insight into the geography of economic activity at the regional level. Regiobase contains all the LBUs in the GBR. Additionally, Regiobase contains variables which allow data at the LBU level to be derived from data at the BU level. Regiobase also provides insight into the business activities of the LBUs, as these can differ between the business activity of the BU as a whole. For each LBU, Regiobase contains a postcode, and this allows for data to be analysed and presented spatially. In this study we disaggregate The Netherlands into provinces and COROP areas[8]. It is important to note that the majority of the statistics in this study are presented at BU level whereas Regiobase employs LBUs.

Policy Record Administration
The Policy Record Administration is used in this research to derive the number of employees for given BUs. The Policy Record Administration is a record of the employment history of all workers in the Netherlands. The data is collected via income tax returns, and is subsequently processed by the state unemployment insurance provider in order to determine rights to claim money from the state in the case of unemployment. As part of this record, the BUs at which a worker has been employed are also recorded. This allows, at any given time, to derive the total number of employees at a given BU. From this database we derive the number of employees and number of jobs of employees at a BU, whereby an employee is defined as a person who has a contract with an economic unit to carry out work in return for financial remuneration.

[8] A COROP region is a spatial area within The Netherlands of which there are 40. The abbreviation stands for Coördinatiecommissie Regionaal Onderzoeksprogramma, literally the Coordination Commission Regional Research Programme.

Retail Survey
The retail survey provides data on a sample of businesses with more than 10 employees in the retail sector. The size of the sample is 7,456 BUs and the data refers to 2015. If the business operates an online store then the turnover from that online store is reported separately, next to the turnover for the whole business. However, turnover from online stores is not reported for businesses of a certain size classes and for certain SIC codes (retail and wholesale). This data in the retail survey is used in the categorisation of online stores.

ICT Use Survey
This survey contains annual data on automation and the use of information and communication technology (ICT) in companies in the Netherlands. The results describe, among other things, the use of computers, the internet, electronic buying and selling, software and ICT applications and show the trends in these phenomena for the period since 2003. The survey is carried out on a sample of roughly 11,000 businesses from the population of businesses with at least 10 employees (which consists of approximately 60,000 businesses). We used data from this survey to cross-validate our results and improve our methodology.

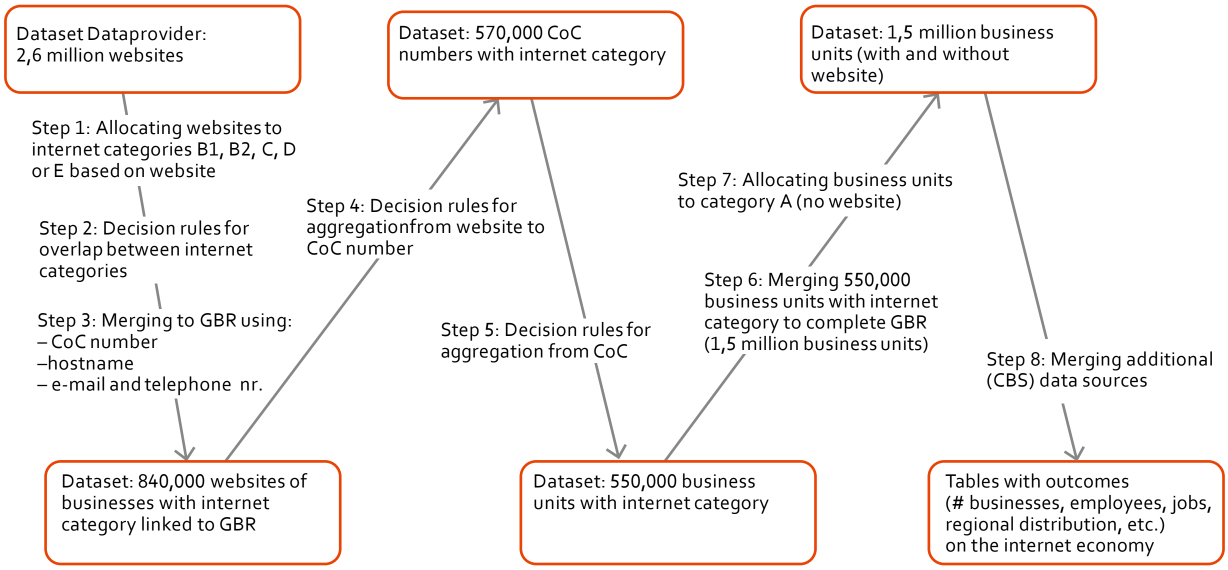
Other sources
Finally, we make use of several other publicly available sources of data from the internet. For example, where lists of online stores are available online, we use this to check that our method to identify websites is capturing all the most important online stores (see section 5.1.1). We also use lists of ICT businesses for a similar purpose (see section 5.1.2). These sources allow us to check the plausibility of our results and to complement the other data sources as necessary.

5. Methodology

This section describes the steps to create a database from which all the results are derived. Figure 5.1 shows a schematic summary of all the research steps.

We begin with the Dataprovider data on Dutch websites and allocate each website to a category. We explain our method for this in section 5.1. After websites have been allocated to given categories, we link the websites to the GBR. The method for linking to the GBR is quite complex and employs diverse methods. For some linking methods more than others, we can be more confident that the link between the website and the BU is correct. Therefore, each linking method is described in detail in section 5.2. At this stage, we have a database in which a BU can have 1) no website, 2) one website or 3) multiple websites allocated to it. This means that the database is not unique at the BU level. In order to accurately represent the economy, this database needs to be unique at the BU level. Further, we need to translate the categorisation of the websites attached to each BU to the classification of the BU. Thus if, for example, one BU has two websites, one belonging to category B and the other belonging to category C, then the BU could be categorised as B or C. To deal with this, we develop a series of decision rules which allow us to create a database which is unique at the BU level and for which all BUs are allocated to a category. These decision rules are explained in

5.1 Overview methodology in research steps


Dataset Dataprovider: 2,6 million websites
Step 1: Allocating websites to internet categories B1, B2, C, D or E based on website
Step 2: Decision rules for overlap between internet categories
Step 3: Merging to GBR using:
– CoC number
–hostname
– e-mail and telephone nr.
Dataset: 840,000 websites of businesses with internet category linked to GBR
Step 4: Decision rules for aggregationfrom website to CoC number
Dataset: 570,000 CoC numbers with internet category
Step 5: Decision rules for aggregation from CoC
Dataset: 550,000 business units with internet category
Step 6: Merging 550,000 business units with internet category to complete GBR (1,5 million business units)
Step 7: Allocating business units to category A (no website)
Dataset: 1,5 million business units (with and without website)
Step 8: Merging additional (CBS) data sources
Tables with outcomes (# businesses, employees, jobs, regional distribution, etc.) on the internet economy

section 5.3. If no website can be allocated or linked to a given BU then that BU is classed as Category A: business without a website. Finally, we link the database to several additional Statistics Netherlands data sources. In this way, we enrich the backbone of BU's with as much information possible, with the aim of maximising the insight into the internet economy.

5.1 Allocating websites to categories

In this section, we describe the methods used to allocate websites to categories. These allocation methods can result in a website being allocated to more than one category. We refer to this as 'overlap'. The last subsection explains the rules used to deal with overlap.

5.1.1 Category C: Online stores

Firstly, a point of terminology. Websites can have the functionality generally associated with the term e-commerce. A business with e-commerce functionality may be referred to as an 'online store'. Of course, many businesses which engage in e-commerce also engage in other activities which would better describe their core business activities. Whether a business which engages in e-commerce is best classified as an online store is a question which is dealt with in sections 5.3 and 5.4.

Determining whether a website has e-commerce functionality is a relatively simple process because many variables in the Dataprovider database pertain to e-commerce. For commercial reasons, Dataprovider has collected data on, for example, the presence of shopping carts and the presence of different payment methods. Further, Dataprovider constructs a concrete e-commerce indicator. This indicator was derived using a machine-learning algorithm. The input for the machine-learning algorithm was a list of websites with concordant relevant variables, which are known with certainty to be e-commerce website and a list of websites which are known with certainty not to be online stores. The machine-learning algorithm takes this knowledge and applies it to websites for which it is not known whether the website is an online store. The machine learning algorithm then assigns a probability that a website is an e-commerce website. Thus, some websites have low probabilities (5% for example) and are thus most likely not e-commerce website, while some have high probabilities (95% for example) and are thus most likely e-commerce websites. It remains then to choose a cut-off point. Dataprovider chooses a cut-off point of 85%. During this research, we analysed this choice by looking at the websites either side of the cut-off point. On this basis, the performance of the machine-learning algorithm and the appropriateness of the 85% cut-off point appeared to be performing satisfactorily. Below the cut-off point most websites don't seem to be an online store and above the cut-off point most websites were.

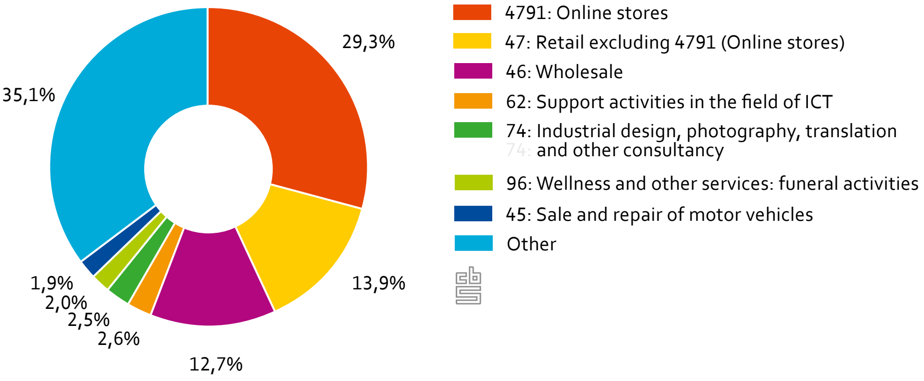
It is, however, possible that websites which are actually online stores can be missed using this method. We therefore searched for other data with which we could test whether the Dataprovider machine learning algorithm had identified at least all of the most important online stores. The Dutch website jouwaanbieding.nl contains an up to date list containing the most popular online stores. Within this list were several online stores which were better placed in other categories, so we manually removed these from the list. From the remaining website approximately 250 additional online stores were identified and added to category C.

Finally, we employed the Call To Action (CTA) dataset to further refine the category. Of the six calls to action in this dataset (order, buy, view the shopping cart, make a reservation/booking, subscribe or register), the following three are most closely associated with online stores: order, buy and view the shopping cart. We analysed different combinations of these 3 calls to action in order to find a combination which added a significant number of online stores to the category without adding any websites which were not online stores. The best performing set of calls to action was 'buy' and 'view the shopping cart'. This choice of calls to action identified around 5,400 online stores which were not yet categorised as such.

Comparison to SIC codes for online stores

The use of big data within this project facilitates a different perspective on the nature of the Dutch businesses. It is particularly interesting to make comparisons between the nature of businesses according to the SIC codes and the categorisation of businesses in this study. In this box we analyse the SIC codes of all the businesses which are classified as online stores according to our definition. The results are shown below.

5.1.1.1 SIC codes of businesses classified as online stores



The results show that there is a significant overlap between the SIC code for online stores and our online store categorisation (29.3%). We also find that there are many businesses in SIC codes other than 4791 (online store). The figure shows the largest 9 SIC codes not belonging to 4791. Unsurprisingly, a large number of retailers who are not online stores according to their SIC code, are online stores according to our definition. This can easily occur if a retailer decides to diversify into online sales. Several of the other SIC codes can also have retail activities (62, 45, 90, 96) alongside their other business activities. Art is an interesting example. An artist who sells her own paintings online would probably strongly disagree with the proposition that she is an online retailer and therefore would be unlikely to classify herself as such. Overall, the data show that the SIC codes are not capturing the extent to which online stores are an important part of the economy. On the other hand, the SIC code 4791 also identifies businesses as online stores that are not part of our definition. For example our study misses businesses that do not have their own e-commerce website but sell their products through websites such as amazon.com or bol.com. Further research is needed to fully understand these discrepancies.

5.1.2 Category D and E: Online services and internet related ICT

Category D (Online services) consist of websites which are used as a means to generate revenue without selling goods through what is generally understood to be a online store. Consider for example the case of marktplaats.nl. In this case, the service is a market facilitation: connecting buyers sellers and facilitating transactions. Dating websites are a similar example, as are price-comparison sites, or websites such as booking.com which facilitate the booking of holiday accommodation and hotels. News websites are included in this category because they provide an online service.

Online stores and online services naturally cannot exist without the internet and its related industries, the category E. Traditionally captured by a selection of SIC codes, the supporting industries of internet were often part of the ‘digital economy’ as described in previous papers. With the big data approach in this publication, we aim for a more stringent definition of this type of website. Computer repair is for example not included in our definition of the internet economy. Examples of businesses in category E are host and cloud services, website and app developers and internet consultancy and marketing. Note that many of the services provided by category D could still be provided without the internet. However, category E is entirely dependent on the internet. Without the internet there can be no cloud services or app developers.

The method for allocating websites to categories D and E is fundamentally different from the method used for category C. This is because there are many variables in the Dataprovider dataset which relate to online stores, most importantly the variable e-Commerce certainty. There are however no variables which can so directly be employed to allocate websites to categories D and E. This is especially the case because categories D and E include many different kinds of services and products. Online services range from dating services to online car auctions, while internet related ICT ranges from hosting service to online consultancy. This makes a straightforward selection on the basis of one or more Dataprovider variables impractical. A more pragmatic approach was therefore needed. The method that has been developed applies a combination of steps which are described as follows:

1. Keyword selection: the indicator 'keyword' in the dataset from Dataprovider captures the words that appear most frequently on a given website. The keywords therefore provide insight into the type of website and the content. As such, if keywords can be identified which relate to a particular category, then the presence of these words in the keywords for a given website can be used to allocate the website to the appropriate category. We started with separating the categories D and E into subcategories with specific topics. For each of these subcategories lists of keywords were created, primarily on the basis of keywords from prominent websites in the subcategories. Combinations of keywords were also used, i.e. the presence of one of two words would be not count as evidence that a website belongs to a given category, but that keyword in combination with another keyword would count. In Appendix B , the full lists of subcategories and keywords are presented. Using these keyword lists, websites were allocated to the various subcategories with categories D and E.

2. Refinement of selection: while the keyword selection succeeded in capturing relevant websites, it also captured many websites which better fitted into categories other than D and E. This occurred because website keywords do not necessarily correctly describe the nature of the website. Therefore, other variables from the Dataprovider dataset and the GBR were employed to refine the selection. From the Dataprovider dataset, we mainly used variables regarding the topic (indicator 'Category') and type of website (indicator 'Websitetype'). Additionally, we made a preliminary link to the GBR at the level of website in order to provide extra information for this refinement. Later, as described in section 5.2, we made the definitive link to the GBR at BU level. However, at this stage, we wanted to obtain more information on the business behind the website in order to better allocate websites to categories. For the preliminary link, we employed the same methodology as described in section 5.2. Making the preliminary link allowed us to employ information about the sector (the SIC codes) and the size of the business (size classes, based on number of employees). The result was a more precise categorisation of websites thanks to the exclusion of websites which were better placed in other categories.

3. Manual adjustments: this is the final stage in the process. Firstly, the websites in the different subcategories were ordered in terms of their 'importance' (turnover, number of employees and 'economic footprint[9]'). The 100 most important websites for every category were then manually inspected to check that no websites were mistakenly included during the first two steps, and also to check for the presence of the larger well-known websites that should fall into given categories. This identified several areas for improvement which result from incompleteness of datasets and inaccurate links between websites and the GBR (as would be expected from a preliminary linking). In some cases, the categorisation of a website was reallocated manually. This often involved visiting the specific websites in order to judge into which category they belonged. In other cases several external sources have been used to determine the completeness of our categorisation[10]. At this stage, we were satisfied with the categorisation for category D, but found evidence that websites belonging to category E were not being identified with sufficient completeness. We dealt with this by identifying SIC codes which fell under the subcategories (shown in table 5.1.2.1). Websites which fell under these codes according to the preliminary link with the GBR were then allocated to category E.

[9] The economic footprint is simply an indication of how big the website is on the web. It is measured by looking at the technology and, popularity.

[10] The website appspecialisten.nl was used for its list of software developers who create apps. We also acquired a list of ICT businesses from the 'MKB Innovatie Top 100', which list the top 100 best ICT business.

On the basis of the subcategories determined for the keyword selection, we have created aggregated subcategories for the final categories. The following subcategories, including the topics included, are distinguished:

5.1.2.1 Subcategories of categories D and E

Internet category	Subcategories	Description
Category D: Online services	Leisure	hotels, holidays, flights, food
	News and entertainment	news, blogs, vlogs, games, videos, music, gambling, adult, e-learning
	Business	finance, advertising, jobs
	Retail	price comparisons, markets, tickets, auctions, car sales, housing
	General services	dating, transport, visualisations
Category E: Internet related ICT	Hosting and cloud	webhosting, servers, datacentres, cloud services
	Websites and apps	website and app designers and developers
	Software	developers and suppliers of software products and services
	Marketing and consultancy	marketing and consultancy services related to websites
	Infrastructure and security	infrastructure and security (firewalls, etc) for IT
	Datamining & Big Data	website crawlers, big data services, machine learning

5.1.3 Category B: online presence

Categories C, D and E constitute the 'core' of the internet economy. The remaining websites are conceptualised as having an 'online presence'. This means that the website's principal function is to provide information on the business behind the website. The website may also provide services associated with categories C, D and E but these services are in principal additional services to the central purpose of the website, which is to provide an internet presence to the business in question. For example, Royal Dutch Shell has a website which provides information on the business but clearly the business should not fall into categories C, D and E.

Other websites do not fall into categories C, D or E but still provide somewhat more functionality than simply information provision. In the simplest case, one may be able to subscribe to a newsletter or to emails which provide information on new products. Another example is large companies which also manage a small online store. Consider the case BMW, who have a small online store in the Netherlands selling BMW merchandise, but this in no way means that BMW can be considered an online store. We therefore split category B into Categories B1 and B2, whereby B1 is termed 'passive online presence' en B2 is termed 'active online presence'.

To determine whether a website is categorised as active online presence (B2) we make use of the CTA dataset. If a website has calls to action (as do all websites in the CTA dataset) then we can consider the website as having an active online presence if it has not yet been allocated to categories C,D or E. Category B1 is not dealt with at website level. Instead we determine Category B1 at the BU level. As such, we first link the categorised lists of websites to the GBR (as described in section 5.2). All websites which linked to businesses which had not yet been assigned a category, were then assigned the category B1. The definition of B1 is thus in practise: websites which are not allocated to any other category. These are thus websites outside of the core of the economy which provide purely an online presence to the businesses behind the websites.

Category B1 was also used to deal with problems associated with the businesses with the most complicated structures. These businesses are referred to as TopX and consist of businesses where the Enterprise Group (EG) consists of multiple BU's. These complications result from the fact that websites are linked to businesses via the CoC number, and the CoC number is generally only linked to one BU. However, the website for that single BU, was often found to be more closely related to the whole EG rather than the specific BU. Conceptually then, all the BUs under the EG make use of the website even if the website is formally only linked to one BU. We therefore do not wish to assume that all the other BUs are not part of the internet economy. In order to include these other BUs in the internet economy, we allocate them to Category B1. The categorisation of the BU which links directly to the website is determined by the methods described above. Thus, for all TopX EGs, one BU is allocated to a category and all other BUs are automatically allocated to B1.

5.1.4 Dealing with overlap

Of all the websites, 4% were allocated to more than category. This occurs because these websites have characteristics which are associated with multiple categories. Because we wish for a website to be allocated to only one category, we formulated a series of decision rules to remove the overlap. Overlap can easily occur for example between categories C and D because price comparison sites and house buying websites share many similar characteristics with online stores in terms of the language used on the website. A website can have an online store and simultaneously offer other online services for example. In general, the reliance on keywords in the methodology for categories D and E results in overlap simply because websites contain keywords which are associated with both categories .[11)]

In order to remove the overlap, the overlapped websites were analysed at the level of subcategory in order to understand the principal causes of the overlap. By analysing the principal causes, appropriate decision rules were formulated to allocate overlapped websites to a unique category in the most appropriate way. Often additional Dataprovider or GBR variables were employed in the specification of the decision rule. As an example, Dataprovider construct a variable named 'category' that provides some general information on the type of website. Some websites have the category 'IT-services and Telecom'. We found that websites which were allocated to both C and E could be best placed in category E if the website was categorised as 'IT-services and Telecom' according to Dataprovider. Otherwise, we allocated the website to category C.

5.1.4.1 Number of websites per type of overlap

	Category C Online stores	**Category D Online services**	**Category E Internet related ICT**	**Total**
Category C Online stores	.	490	1,360	64,580
Category D Online services	**490**	.	1,290	9,690
Category E Internet related ICT	**1,360**	**1,290**	.	37,580
Total (excluding overlap)				111,850
Total (including overlap)				**108,740**

[11)] Because only the websites unallocated to C, D or E were allocated to category B, there is no risk of overlap with category B.

5.2 Merging to General Business Register (GBR)

By this stage, every website is allocated to a category. The next step is to link websites to a BU. This allows the economic data available in Statistics Netherlands and the internet data from Dataprovider to be combined and thus analysed together at the BU level. In order to link two datasets, a key is required. A key is a variable which is presents in both datasets to be linked. In this project we make use of various combinations of keys into order to maximise the number of successful links which can be made. The degree of confidence whether the correct link has been made varies between key combinations. We therefore describe in detail the different key combination which we have used.

Key combination 1
The first key combination consists of website names and CoC numbers. Many websites in the Dataprovider data have a corresponding CoC number which was obtained by web-crawling. Additionally, many BUs in the GBR record the website of the BUs. The fact that the link is made on the basis of both CoC numbers and website names, indicates a very high linking accuracy. This is referred to as Key Combination 1. Of all the websites which could be linked to the GBR 32% of these were linked using Key Combination 1.

Key combination 2
Key combination 2 is used when the website names match in the GBR and in the Dataprovider data but the CoC number in the Dataprovider data does not concur with the CoC number in the GBR which is indicated by the website name. To solve this problem, we look at the telephone and email address data from Dataprovider and with the GBR. When the analysis of the telephone and email address indicates that the CoC in the GBR is the correct one, we assume that the website links to the business corresponding to the CoC number in the GBR. Of all the websites which could be linked to the GBR 8% of these were linked using Key Combination 2.

Key combination 3
Key combination 3 is the same as key combination 2 in terms of the problem but not the solution. Key combination 3 is thus used when the website names match in the GBR and in the Dataprovider data but the CoC number in the Dataprovider data does not concur with the CoC number in the GBR which is indicated by the website name. The solution is different because is in this case, the analysis of telephone and email addresses indicates that the CoC from Dataprovider corresponds to the correct businesses in the GBR. In this case then, the website is linked to the business in the GBR corresponding to the CoC number from Dataprovider. Of all the websites which could be linked to the GBR, 35% were linked using Key Combination 3.

Key combination 4
For some cases there are no logical links between the CoC number and hostname from Dataprovider and that from the GBR. Examples are when there is no hostname in the GBR available and/or there is no CoC present in the Dataprovider data. In this case, the best solution is to use the CoC number derived from the email and telephone details in the Dataprovider data. These email and telephone details are matched to those registered in the GBR to allocate the correct CoC to the website. Of all the websites which could be linked the GBR, 9% were linked using Key Combination 4.

Key combination 5

The final category consists of cases where more than 10 websites have the same CoC number. Some businesses may indeed have more than one website, for example, there are several online stores which each have several websites selling different products. In many cases however, more than 10 websites link to the same CoC number because the CoC number on a given website does not correspond to the business behind the website. For example, a website may contain a list of businesses who supply a particular product, which includes the CoC number for each business. A given CoC number can show up multiple times over different websites. Another example is of hosting/web-design companies which display their CoC on the websites they host or designed. For these cases we incorporated a special decision rule. If there were more than 10 websites that belong to one CoC number and one of these websites was a hosting/web-design company than this CoC number was allocated to category E. Of all the websites which could be linked to the ABR 16% belongs to this category.

All websites which could not be linked according to one of the 5 key combinations are from this point on excluded from the analysis. The websites which could not be linked will include all the websites of private individuals which fall outside of the definition of the internet economy, as well as websites where there was not sufficient information to make a link to the GBR. The different key combinations are summarised in the following table.

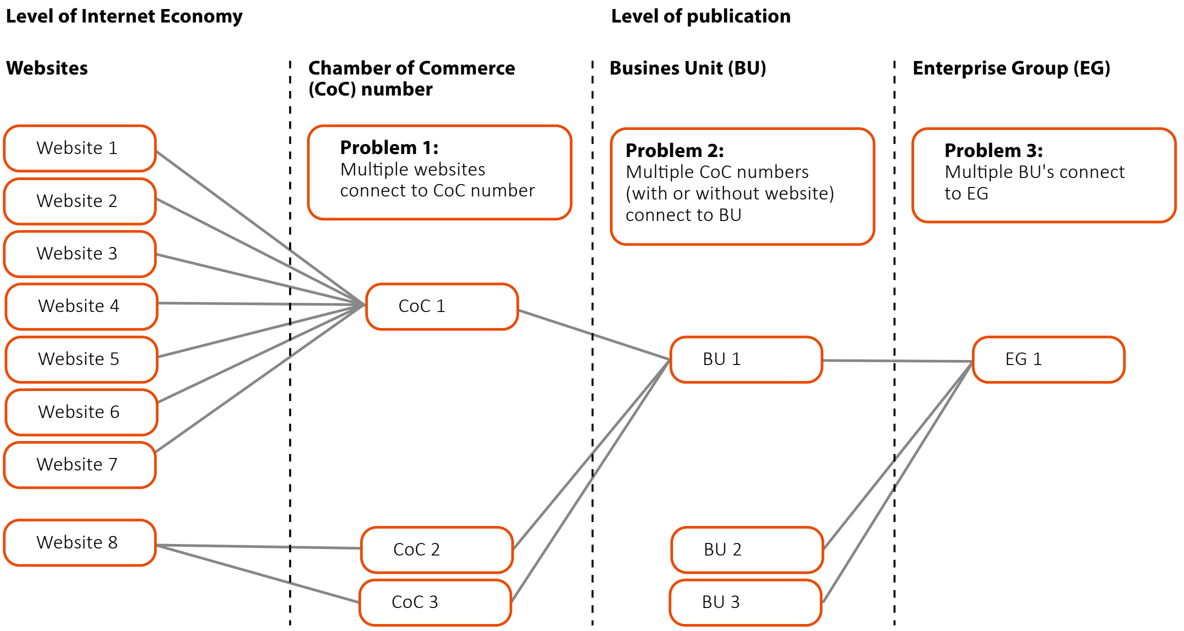
5.2.1 Results of the merging process

Key combination	Description	Count	Percentage
1	hostname (Dataprovider) = hostname (GBR) and CoC (Dataprovider) = COC (GBR)	272,000	**32**
2	hostname (Dataprovider) = hostname (GBR) and CoC (Dataprovider) <> CoC (GBR) -> CoC (GBR)	66,000	**8**
3	hostname (Dataprovider) = hostname (GBR) and CoC (Dataprovider) <> CoC (GBR) -> CoC (Dataprovider)	294,000	**35**
4	hostname (Dataprovider) <> hostname (GBR) and CoC (Dataprovider) <> CoC -> CoC through e-mail or telephone number	76,000	**9**
5	CoC (Dataprovider) > 10 hostnames	132,000	**16**
Total matched		**840,000**	**100**
Total not matched		**1,948,000**	

5.3 Decision rules

Decision rules are formulated to determine the role of BUs in the internet economy based on the information which we have gained about the website(s) of the BU. Decision rules would not be necessary if every business had only one website, only one CoC number and only one BU. The situation can be more complicated because businesses often have multiple websites. Consider for example a large business with separate websites for customers, businesses, careers and sales. It is likely that these websites will not appear in the same category of the internet economy. Additionally, due to the structure of GBR, multiple CoC numbers often link to a single BU. Subsequently, multiple BU's can link to one EG when large companies have diverse activities. The following figure presents these different problems.

5.3.1 From Internet Economy to level of publication; problems


Level of Internet Economy
Level of publication
Websites
Chamber of Commerce (CoC) number
Busines Unit (BU)
Enterprise Group (EG)
Problem 1: Multiple websites connect to CoC number
Problem 2: Multiple CoC numbers (with or without website) connect to BU
Problem 3: Multiple BU's connect to EG
Website 1
Website 2
Website 3
Website 4
Website 5
Website 6
Website 7
Website 8
CoC 1
CoC 2
CoC 3
BU 1
BU 2
BU 3
EG 1

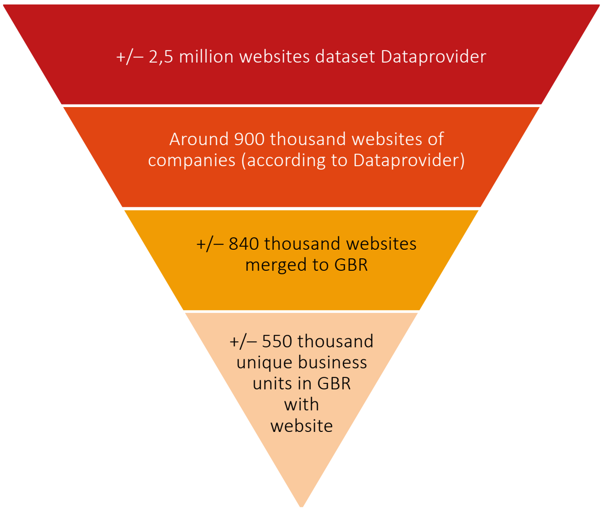
Our aim was to create a list of all the businesses in the GBR and allocate every BU to a category of the internet economy according to the website(s) belonging to that BU. It was therefore necessary to develop a set of rules to counter the problems regarding the translation from website to BU level. The rules applied for this can be divided in two groups:

1. *Hierarchy of categories*: the categories within core of the internet economy (C, D and E) are always preferred to the category Online Presence (B). We choose this hierarchy in order to maximize the information provide by allocating a website to a category. Consider for example a BU that could be allocated to D or B. If we allocate it to B then we know less about that BU because category B is more general.
2. *Importance of activities*: for the remaining categories we selected the website with the highest average 'economic footprint' (a Dataprovider variable) or the CoC number with largest number of employees. Both economic footprint and the number of employees were chosen because they should indicate which activity of the business is the most important activity.

An additional problem is that some BUs which fall under large and complex EGs may not be allocated to a category even though other BUs under the EG have been allocated to a category. This is a problem because the BUs which have not been allocated are likely to make use of the website of another BU within the EG. Consider the large Dutch company Phillips. A given BU may make use of the website of another BU to recruit staff. On this case, we therefore change the BU's with no website (category A) into online presence (category B1). The absence of a website for these BU's is the result of our link between websites and CoC's, i.e. it is not possible to link one website to multiple CoC's. A website linked to another BU in the EG is therefore most likely to represent the business activities of the remaining BUs.

Finally, the largest 100 business in each category (according to turnover) were manually inspected to confirm that the decision rules were functioning appropriately. The process up to now is summarized in the following figure.

5.3.2 Summary of method


+/– 2,5 million websites dataset Dataprovider
Around 900 thousand websites of companies (according to Dataprovider)
+/– 840 thousand websites merged to GBR
+/– 550 thousand unique business units in GBR with website

5.4 Turnover of online stores

At this stage, a list of online stores (at BU level) existed. This list however included businesses with an online store which may only produce a small amount of their turnover from their online store. It was therefore necessary to ask whether such businesses can really best be classified as online stores. A logical approach, which we adopt in this study, is to classify online stores as such if more than 50% of their turnover comes from the online store. Applying this decision rule is however challenging because the percentage of turnover generated by e-commerce is only available for the small sample of business in the Retail Survey. Further, this is not a representative sample of businesses because it excludes businesses with less than 10 employees. We therefore wish to estimate the revenue from e-commerce for businesses which have an online store, and for which these data are not available from the Retail Survey.

To do this, we adopt regression analysis. This method is employed to predict the turnover from e-commerce for businesses which are not in the retail survey. The method attempts to explain variations in turnover from e-commerce with other variables about the business. The first variable which we consider is the total turnover of the business. In general, if total turnover increases, then supposing a fixed portion of turnover from e-commerce, the e-commerce turnover will increase also. We therefore include the variable total turnover to control for this effect. We then experimented with many other available variables (from Dataprovider and the GBR among others) to try to explain as much of the variation in turnover from websites as possible. We identified the size of the businesses behind the e-commerce website, the average number of pages on the websites of the business, the presence of shopping carts to all be significant determinants of turnover from e-commerce.

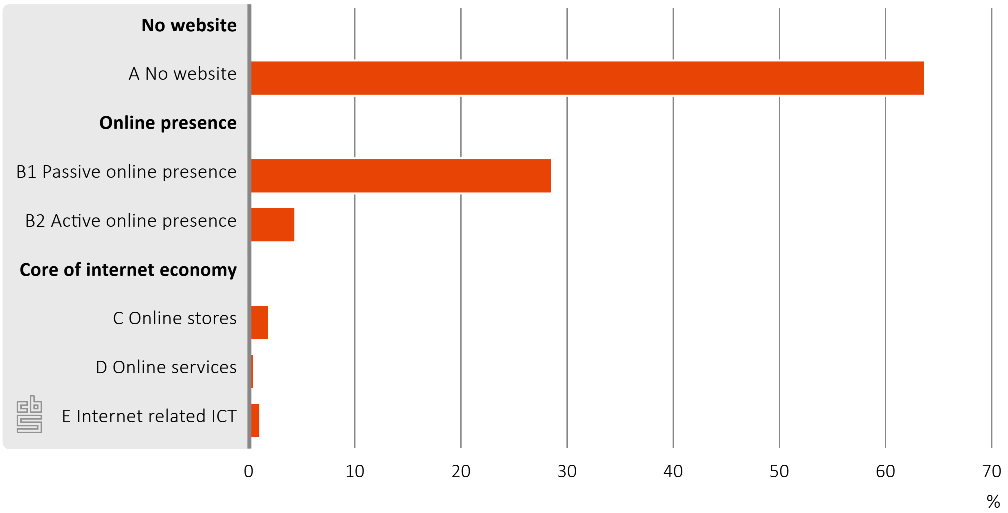
For each of these variables, the regression analyses produces coefficients. These coefficients show the direction and magnitude of the effect of the corresponding variables on e-commerce turnover. We can thus employ these coefficients and variables to estimate the turnover from e-commerce for businesses for which actual statistics are not available. However, we found that the model was not producing plausible results for small business. This is to be expected because the data upon which the regressions were run (the Retail

Survey) did not include businesses with fewer than 10 employees. We therefore visited the websites of a sample of small businesses with websites to search for evidence that these businesses were generating revenue through other means than e-commerce. We found that the majority of the majority of these businesses had only e-commerce as means of generating turnover. We therefore simply assumed that all businesses with e-commerce websites with fewer than 10 employees generate at least 50% of their turnover from e-commerce. We therefore classify these businesses as online stores.

6. Results

The application of the above methodology to our data facilitates the analysis of the internet economy. The first question of interest is, how many Dutch businesses have a website, and if they do, into which category does that website fall? We find that 36% of the Dutch companies are present on the internet, as shown in Figure 6.1. Of that, 3.3% belong to the core of the internet economy (categories C, D and E). The turnover of the core of the internet economy is € 104 billion which constitutes 8% of the Dutch business economy. The core provides 345,000 jobs. The composition of the core, and its size in respect to businesses with an online presence and businesses without a website is given in Figure 6.1.

6.1 Number of business by internet category, 2015



In this chapter, we discuss the most interesting results which give insights into the nature of these categories and thus into the importance of the internet economy. Section 6.1 focusses on businesses without a website and the differences between these businesses and the businesses which do have a website. The businesses with an internet presence (both passive and active) are discussed in more detail in Section 6.2. Section 6.3 concerns the core of the internet economy according to this study: the online stores, the other online services and internet related ICT. The complete set of tables with all the results can be found in Appendix C.

Demographics of Dutch businesses

It is important for a proper interpretation of the results to have a complete picture of the Dutch business population in 2015. There are large differences in the numbers of businesses across different sectors and sizes. For example there are almost 1,2 million businesses with one employee. That is 77% of all the 1,5 million businesses. There are also big differences between sectors, from several hundred companies in the sector Mining and quarrying to more than 300 thousand companies in the sector Consultancy, research and other specialised business services. These differences influence our outcomes to a great extent.
The table below present the demographic distribution of Dutch businesses across sectors and size and can be helpful when interpreting the results.

Demographics of Dutch businesses, 2015

		Size				
Sector		1 employee	2–49 employees	50–249 employees	250 or more employees	Grand Total
A	Agriculture, forestry and fishing	38,070	33,770	130	10	71,980
B	Mining and quarrying	250	110	30	10	390
C	Manufacturing	37,080	20,190	1,960	390	59,610
D	Electricity, gas, steam and air conditioning supply	670	280	20	20	990
E	Water supply; sewerage, waste management and remediation activities	890	530	80	30	1,520
F	Construction	124,750	25,490	660	90	150,990
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	140,760	80,370	1,920	330	223,380
H	Transportation and storage	22,600	12,560	730	140	36,030
I	Accommodation and food service activities	22,110	29,000	320	50	51,480
J	Information and communication	68,090	12,610	500	90	81,290
K	Financial institutions	74,070	9,300	160	60	83,590
L	Renting, buying and selling of real estate	17,820	6,480	160	30	24,480
M	Consultancy, research and other specialised business services	264,180	38,350	750	150	303,430
N	Renting and leasing of tangible goods and other business support services	45,490	15,690	1,210	320	62,710
O	Public administration, public services and compulsory social security	130	100	330	200	760
P	Education	58,510	6,270	710	330	65,820
Q	Human health and social work activities	101,970	23,760	760	570	127,050
R	Culture, sports and recreation	81,780	9,930	210	20	91,940
S	Other service activities	78,950	13,620	200	30	92,800
T	Activities of households as employers	10	10			20
U	Extraterritorial organisations and bodies	0	0			10
Grand Total		1,178,160	338,380	10,820	2,870	1,530,240

6.1 Businesses without a website

Most self-employed businesses do not have a website
The results of this study show that almost 64% (975,000 businesses) do not have a website (category A of the internet economy). Of all the categories, this category is by far the largest. However, if we also see that the majority of businesses without a website consist of only one

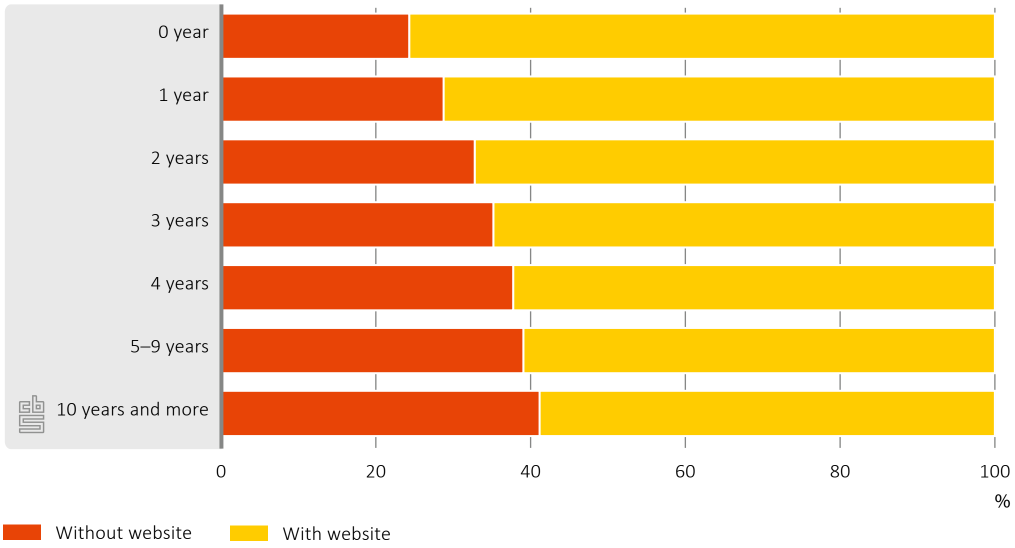
employee (more than 80%), it makes more sense. For example, self-employed contractors or financial advisors don't necessary need a website to do business, especially given the cost of setting up a website in comparison to using, for example, a Facebook page. Unfortunately we could not measure these professional Facebook pages in this study. If we consider businesses with 10 or more employed workers we see that almost 80% have a website. We also find evidence of a potential relationship between the age of a business and whether the business has a website. Almost half of the companies without a website, 45%, is founded less than five years ago. This suggests that businesses are focussing on their core business before considering an online presence, or that many small, young businesses opt to use social media sites such as Facebook to facilitate an online presence. Another possible explanation is that younger businesses also have younger websites, to which there may not yet be any links, and as such, these cannot be found by Dataprovider. Finally, our process of merging with the ABR, and in particular the manner in which we have dealt with overlap, is likely to have biased our results slightly towards older companies which are more established both economically, in terms of revenue, and in terms of how established they are on the internet.

Other than age and number of employees, we also find evidence than certain sectors generally make less use of the internet. The results show that sectors like 'agriculture, fisheries and forestry', 'construction industry', 'financial institutions' and 'transportation and storage' often do not have websites. This result is not surprising, but highlights the important result fact that the importance of the internet economy varies between sectors of the economy.

6.1.1 Businesses with website broken down by SIC and size, 2015

		Website/Size				
Sector		With website 1 employee	With website 2–49 employees	With website 50–249 employees	With website 250 or more employees	With website Total
		%				
A	Agriculture, forestry and fishing	8	13	72	90	10
B	Mining and quarrying	19	46	85	88	32
C	Manufacturing	34	66	87	98	47
D	Electricity, gas, steam and air conditioning supply	27	30	95	100	31
E	Water supply; sewerage, waste management and remediation activities	21	68	88	100	42
F	Construction	17	52	89	100	23
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	39	56	80	98	46
H	Transportation and storage	14	34	80	94	23
I	Accommodation and food service activities	32	48	69	86	42
J	Information and communication	50	72	84	92	54
K	Financial institutions	9	41	85	98	13
L	Renting, buying and selling of real estate	22	43	92	100	28
M	Consultancy, research and other specialised business services	36	64	84	94	40
N	Renting and leasing of tangible goods and other business support services	32	58	80	93	40
O	Public administration, public services and compulsory social security	15	67	91	96	76
P	Education	40	69	91	98	44
Q	Human health and social work activities	32	52	86	96	36
R	Culture, sports and recreation	40	61	86	100	42
S	Other service activities	31	53	78	93	34
T	Activities of households as employers	10	60			27
U	Extraterritorial organisations and bodies		67			33
Grand Total		31	51	84	96	36

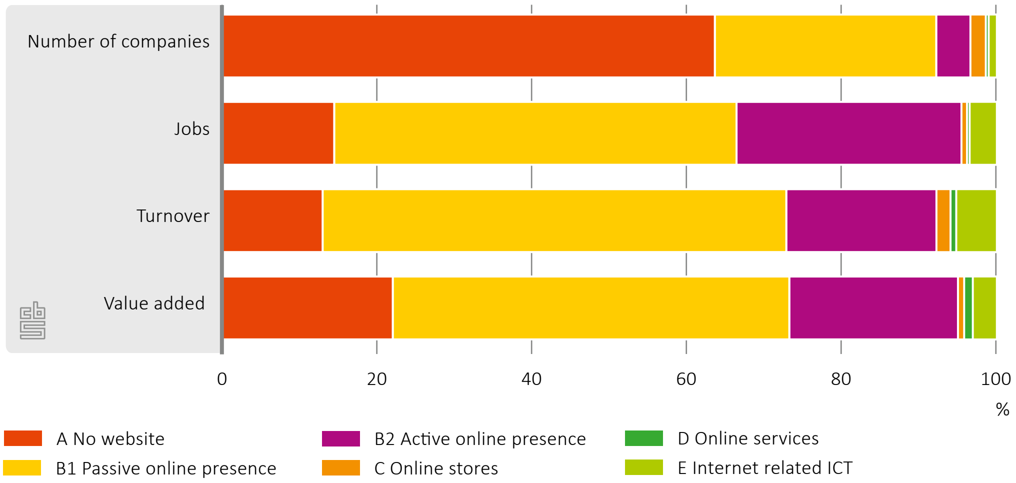
6.1.2 Businesses with or without a website broken down by age, 2015



Almost all medium and large sized businesses have a website
While we find evidence that small businesses are less involved in the internet economy, we find that 36% of all business do have a website. These businesses vary from only being present to fully being dependent on the internet for generating income. Almost all large businesses of more than 250 employees have a website. Also, more than 80% of the medium sized businesses (50–249 employees) are present on the internet. Sectors such as ‘Information and communication’, ‘Public administration’, ‘Education’, ‘Wholesale and retail trade’ and ‘Manufacturing’ are particularly prominent in terms of the share of businesses which have a website.

Relative low turnover and number employees in businesses without a website
This study shows that the majority (64%) of businesses do not have a website. However, in terms of turnover and employment this category is a relatively small part of the Dutch

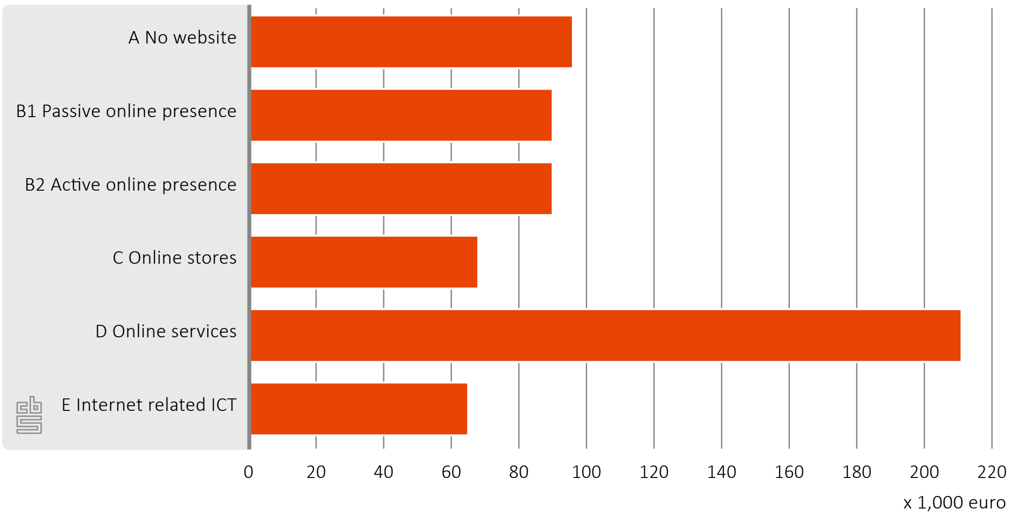
6.1.3 Relative distribution of number of companies, jobs, turnover and value added by Internet categories, 2015



economy. These 975,000 businesses represent only 13% of turnover and 14% of the employees. This makes sense because of the number of self-employed businesses which do not have a website. In turn, this means that the majority of the turnover and employment is related to businesses that do have an internet presence, although most of that is only a passive presence.

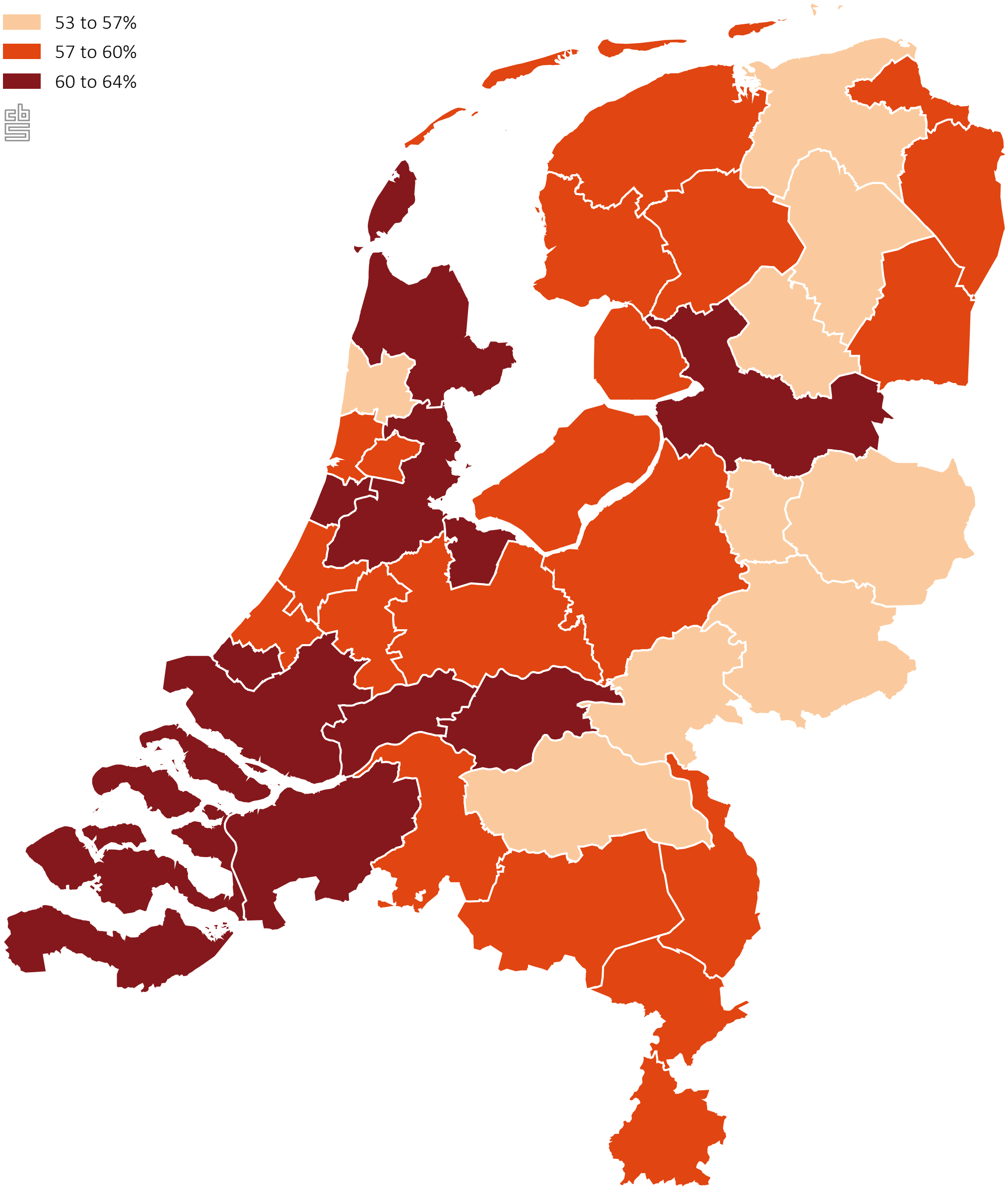
No clear pattern between businesses without a website and productivity
Labour productivity for businesses outside of the internet economy is higher than for those in the core of the internet economy. The exception is category D. For category D labour productivity is almost twice as high as the average for the business economy (€ 91,000). This is due to online booking websites which have a high value added and relatively few employees. The labour productivity in the other two categories of the core of the internet economy is below average. Business without a website and the online presence businesses (categories B1 and B2) have a labour productivity that is comparable to the average of the business economy. We can also analyse labour productivity by SIC group. The top 5 most productive SIC groups in decreasing order are 'mining and quarrying', 'real estate activities', 'energy provision', 'financial services' and 'water supply; sewerage; waste management and remediation activities'. These are not sectors which are generally associated with the internet economy. This goes some way to explaining the lower labour productivity in the internet economy compared to outside of the internet economy. It seems then that businesses which have little need of the internet to conduct economic activity are by no means less productive.

6.1.4 Labour productivity by internet category, 2015



Relatively many businesses without a website in Zeeland and Noord-Holland
Finally (for businesses without a website) we can analyse the regional distribution of the Local Business Units for the businesses without a website. In general, the results show small differences in the internet economy between regions. Considering businesses without a website, we see that especially in south-west part of the Netherlands and in the province of Noord-Holland there are relatively more businesses without a website. In the east of the country, we see the opposite.

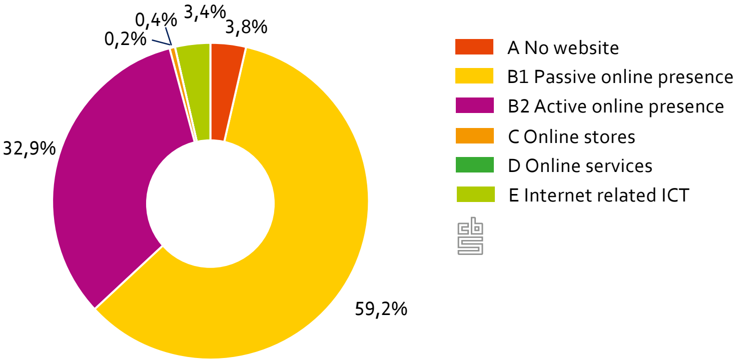
6.1.5 Relative regional distribution of branches of businesses without a website (category A), 2015



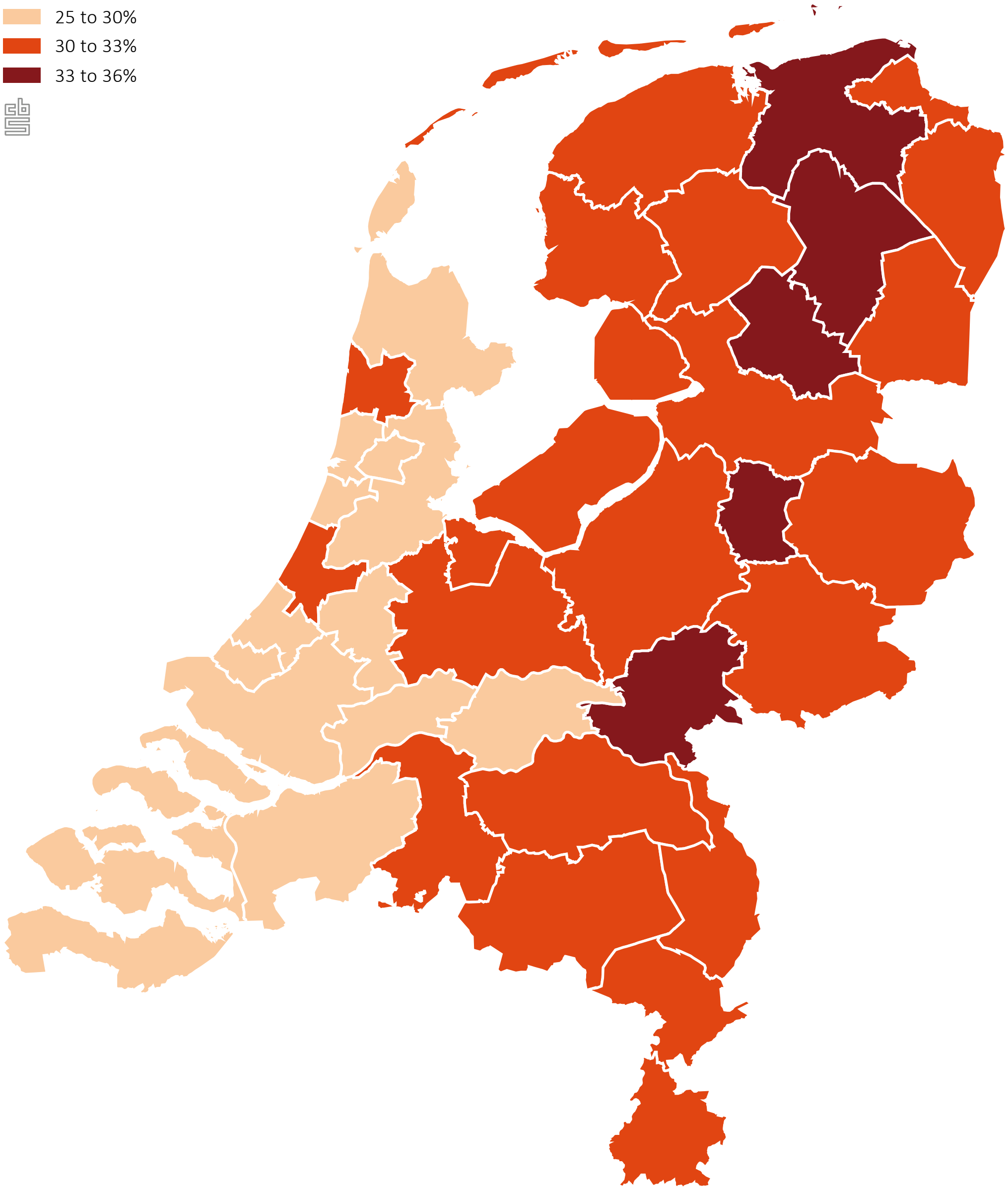
6.2 Internet presence

Businesses with an internet presence (category B1 and B2) do have a website but are not allocated to the core of the internet economy (categories C, D and E). Business in category B1 and B2 do not generate income directly through or with the internet, but their website may indirectly help in their business activities. For example, websites serve a marketing function as well as providing information which encourages consumption of their goods or services. Internet presence businesses are divided into two categories: category B1 termed passive online presence and category B2 termed active online presence. Passive online presence businesses provide only information on, or marketing for their business activities. Active online presence businesses provide some services on their website to support their core businesses activities (category B2: active online presence).

6.2.1 Relative distribution of companies with 250 or more employees over the internet categories, 2015



6.2.2 Relative regional distribution of branches of businesses with passive online presence (category B1), 2015



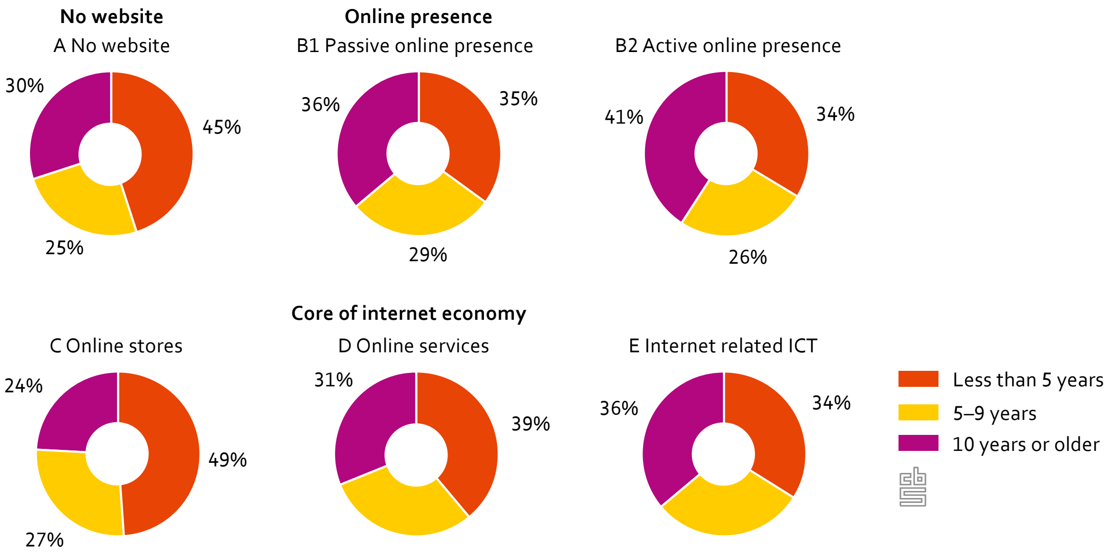
Many companies have a passive online presence
There are around 438,000 companies with a passive online presence, which constitutes 29% of all Dutch businesses in 2015. In terms of turnover and employment, this is the largest category of the internet economy. Passive online presence businesses have a turnover of € 816 billion and they represent more than 4 million jobs. This result is explained by the large number of large businesses within this category. Almost 60% of the companies with 250 or more employees (1,700 businesses in total) have a passive online presence.

Businesses with an internet presence most often fall within the sectors 'Consultancy, research and other specialised business services', 'Wholesale and retail trade; repair of motor vehicles and motorcycles' and 'Human health and social work activities'. The regional distribution of these businesses shows that they are more predominantly located in the north and east of country, as shown in Map 6.2.2.

4.4% of all businesses have a website with an active component
68,000 businesses (4.4% of all businesses) have a website with an active component (active online presence, category B2). This means that it is possible to book, order or reserve something on the website of these businesses, but these activities are not their core activity. This group covers for example a restaurant where you can book a table via the website or large business with a only a small online store. 25% of the businesses fall into the sector 'Wholesale and retail trade; repair of motor vehicles and motorcycles'. This is by far the largest sector for this category. The next largest category is 'Consultancy, research and other specialised business services' at 12%.

Over 40% of businesses with an active online presence are 10 years or older
Just like the passive online presence, the category active internet presence contains relatively many medium and large size companies. They account for a turnover of € 264 billion and 2.3 million jobs. The number of jobs in this category is somewhat striking as nearly 30% of all jobs in the Netherlands can be found in the businesses in this category . Finally, these are the oldest companies of all categories. More than 40% were founded in 2005 or earlier. Analysis of the regional distribution of these companies shows that they are relatively evenly spread across all regions. We therefore do not present a map for this case.

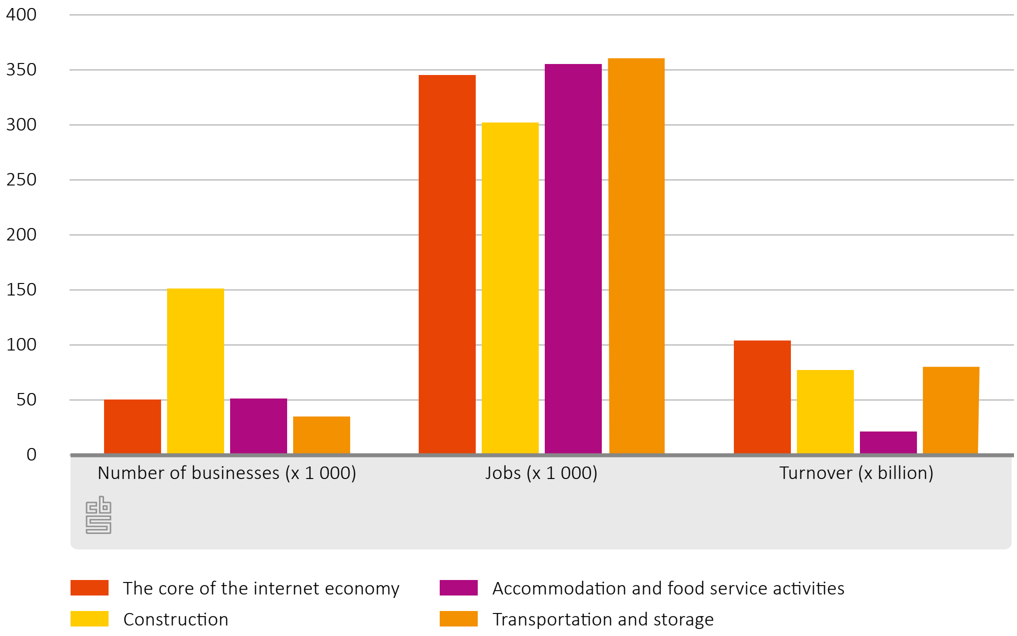
6.2.3 Relative age of businesses for the internet categories, 2015



6.3 The core of the internet economy

The online stores, the other online services and the internet related ICT (categories C, D and E) can be considered the core of the Dutch internet economy. Their income is directly generated through the internet (online stores and other online services) or with the internet (internet related ICT). In total, the core of the internet economy consists of more than 50,000 companies representing 3.3% of all Dutch businesses. Together they account for a turnover of € 104 billion, which is nearly 8% of the turnover of the Dutch business economy (See box 1). There are 345,000 jobs in businesses which exist in the core of the internet economy. In terms of magnitude , the core of the internet economy is roughly the same as the sectors 'construction' or 'accommodation and food service activities' or 'transportation and storage'. (see figure 6.3.1).

6.3.1 The core of the internet economy compared to other sectors



In this section we discuss the most interesting results separately for the three categories of the core of the internet economy. Each of the different categories in the core, it turns out, have quite different characteristics.

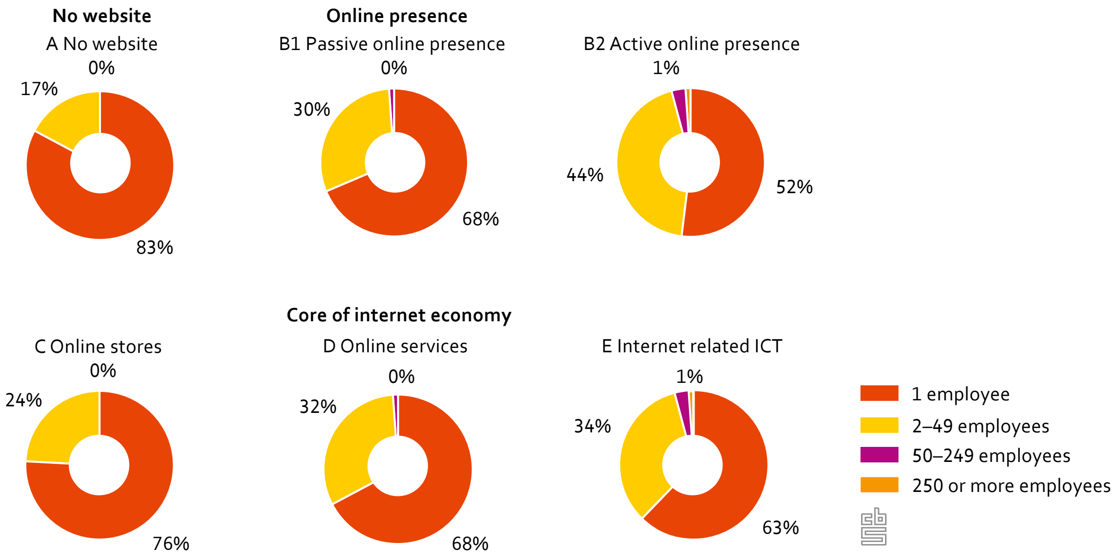
6.3.1 Category C: Online stores

2% of all businesses are categorised as online stores
This study shows that the Netherlands have in total 28,500 online stores in 2015, which constitutes 2% of all businesses. In this study, we have defined online stores as businesses that generate more than half of their income through their e-commerce website. This means that businesses with e-commerce websites are not necessarily considered online stores in this study when e-commerce is not their principle method for generating turnover. For

example, large clothing retailers often have an e-commerce website as a complement to their high-street stores, where the high-street stores generate the majority of their income. Most online stores originate in obvious sectors: approximately half of them belong to the retail sector and almost 15% to the wholesale. This also means that more than 30% of the online stores are found in a less obvious sector from the traditional SIC classification: 'Information and communication' and 'Manufacturing' for example. This suggests that the SIC classification is not successfully identifying all online store as such. This study shows that sectors other than retail are using e-commerce to sell their products directly to consumers.

An average of 2.5 websites per online store
It is common for online stores to have multiple e-commerce websites. The results show that there are nearly 70,000 websites that belong to only 28,500 thousand businesses: an average of 2,5 e-commerce websites per online store. Online stores account for a turnover of € 23 billion. There are 46,000 jobs in the online stores in 2015.

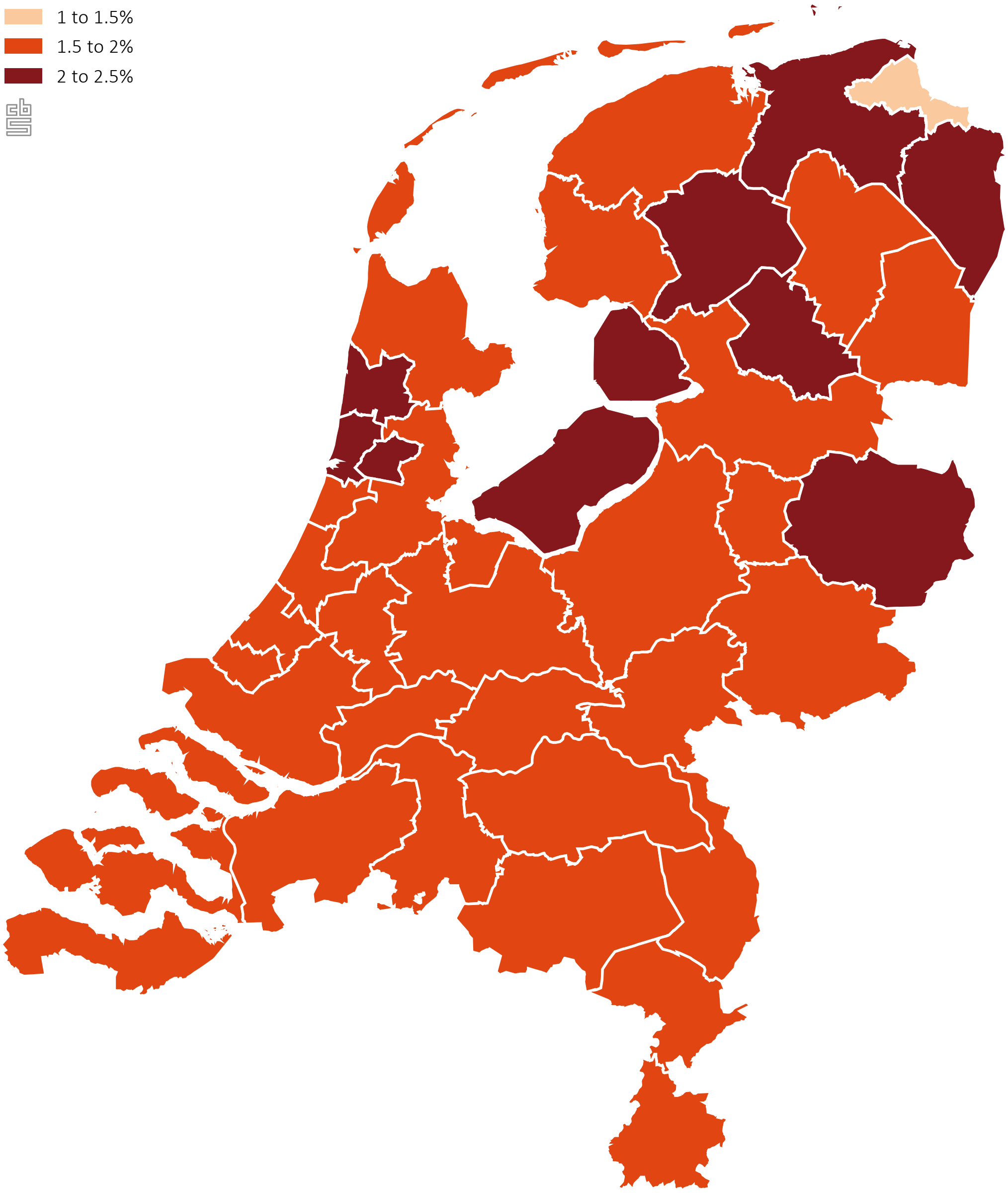
6.3.1.1 Relative size of businesses for the internet categories, 2015



75% of the online stores have only 1 employee
Three quarters of the online stores (22,000 businesses) have only 1 employee. Only a little over 100 online stores have more than 50 employees. This is related to the fact that online stores is the 'youngest' category of all. Almost half of the online stores are founded in the last 5 years. The results support the belief that it is relatively easy to set up an online store on your own.

The locations of the online stores are remarkably often in the northern half of the Netherlands, roughly above the latitude of Amsterdam.

6.3.1.2 Relative regional distribution of branches of online stores (category C), 2015



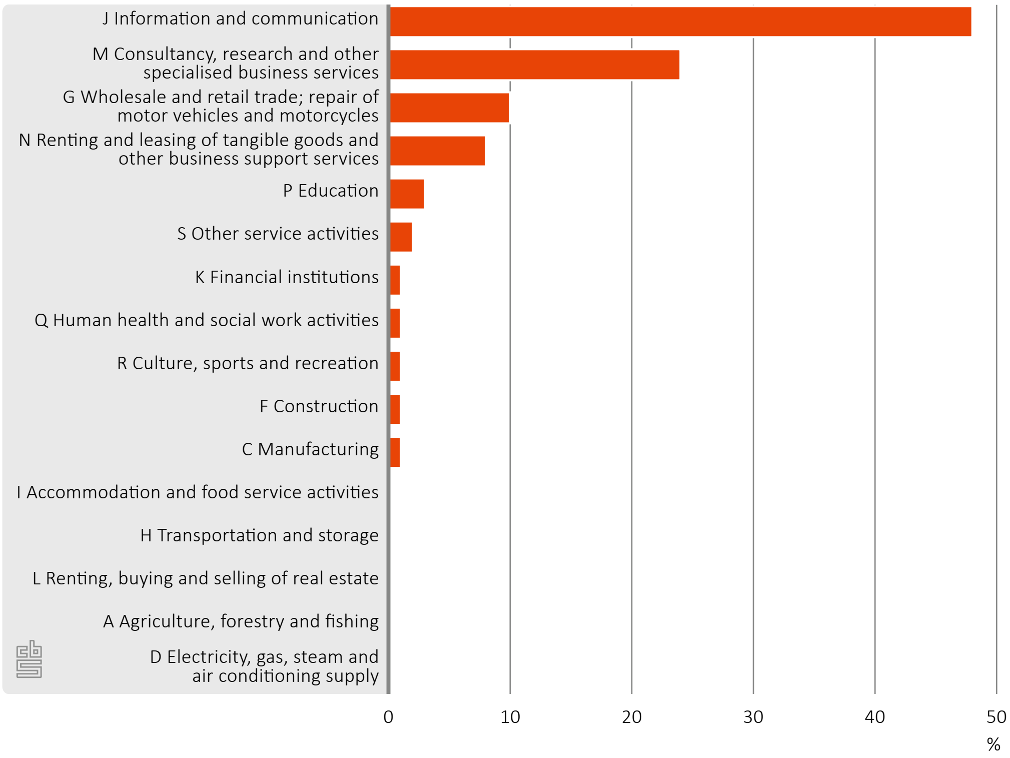
6.3.2 Category D: Online services

Online services only exist because the internet offers them a platform to provide services. Among the online services for example, are price comparison sites, dating sites, online games and auction sites. This is not (yet) a big category of businesses. In 2015, there were 5,700 business in the category other online services. These businesses had a turnover of € 10 billion, which is less than 1% of the total for the Dutch business economy. They account for a total of 26,000 jobs.

The online services are not well represented in the current SIC classification. They are spread across various sectors such as ‘Information and Communication’ (2,700 businesses), ‘Consultancy, research and other specialised business services’ (1,400 businesses), ‘Wholesale and retail trade; repair of motor vehicles and motorcycles’ (600 businesses) and

‘Renting and leasing of tangible goods and other business support services’ (400 businesses). This study is the first to present results of the online services as an separate category of businesses.

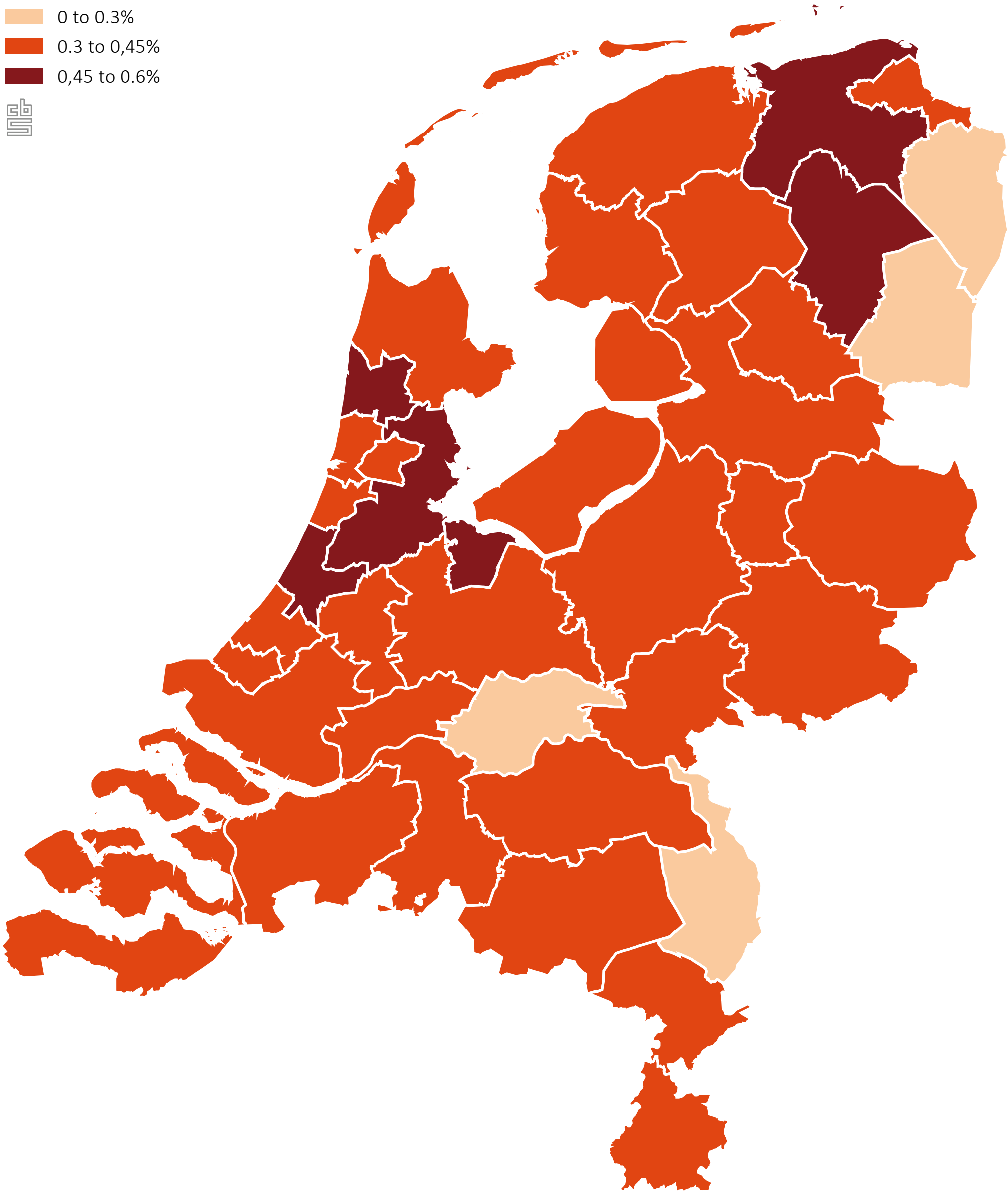
6.3.2.1 SIC categories for the online services, 2015



Similarly to online stores, the online services are relatively young. Nearly 40% of all online services businesses are under five years old. Our results show that there are only a few large businesses that are an online service: just over 50 businesses in this category have more than 50 employees. The majority (68%) consists of 1 employee.

Online services often in the regions around Amsterdam and Groningen
Although the number of businesses in this category is still relatively low, they are more likely than average based in the regions around Amsterdam and Groningen. Groningen has a reputation for innovative economic activity, particularly related to digital and internet based industries, and Amsterdam can be considered a hub for entrepreneurship. As such, these results concur with our expectations.

6.3.2.2 Relative regional distribution of branches of online services (category D), 2015



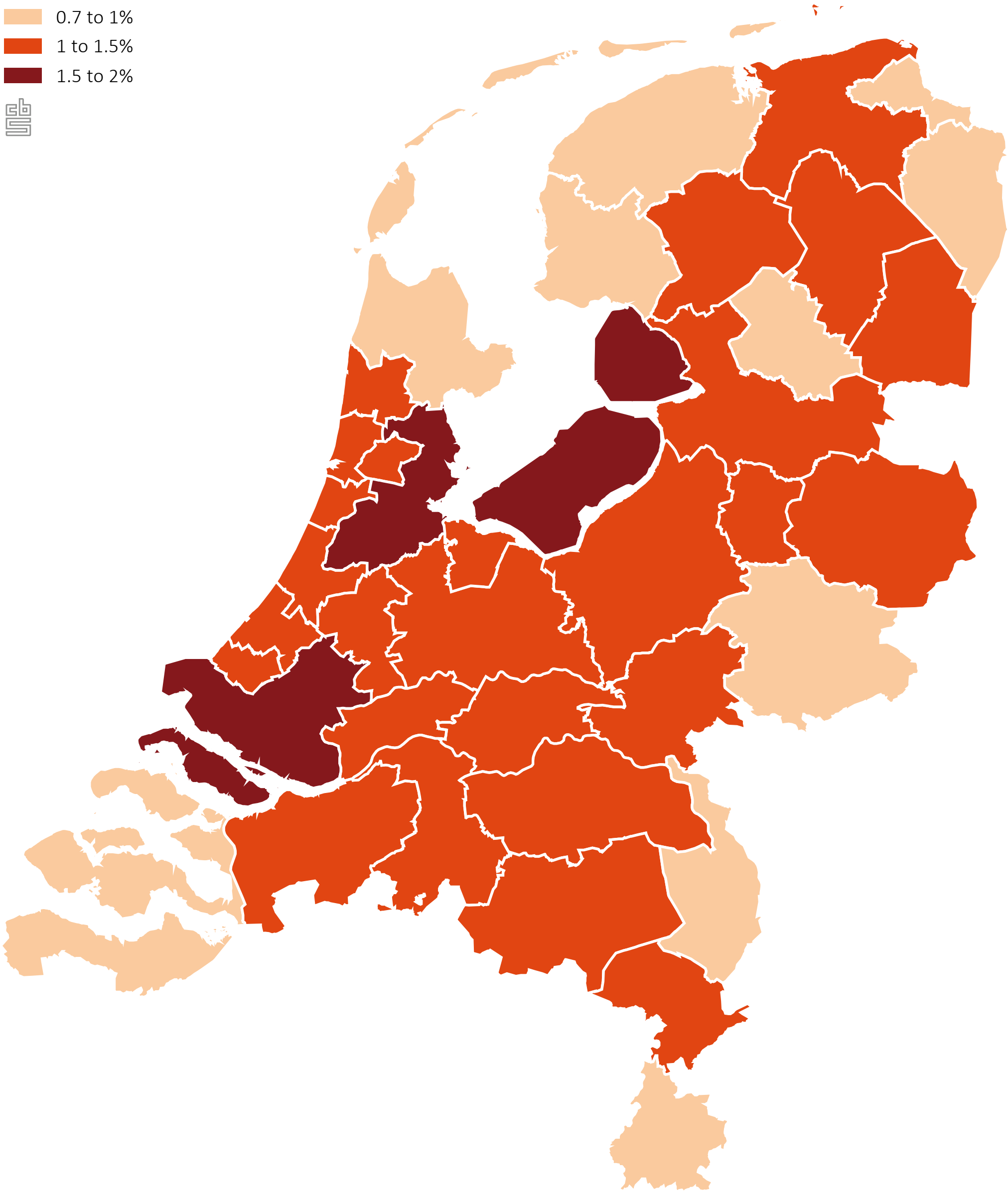
6.3.3 Category E: internet related ICT

16,000 businesses in Category E
There are 16,000 companies in category E in 2015. Internet related ICT businesses include web designers, software developers, hosting companies but also the internet consultancy. These are often companies in the sectors ‘Information and communication’ (7,900 businesses) and ‘Consultancy, research and other specialised business services’ (4,700 business), but also in ‘Wholesale and retail trade; repair of motor vehicles and motorcycles’ (1,300 businesses).

Statistics Netherlands publishes annually on the ICT sector . According to that study, there are around 70,000 businesses in the Dutch ICT sector in 2015, which is a lot more than the number of businesses in Category E. This has to do with the definition of the ICT-sector that includes the ICT-services, the ICT-wholesale and the ICT-industry. This is much more general than our definition because we only consider the internet-related ICT. For example, in our study, we only include the software developers that develop internet related software such as antivirus software but not software developers that create general software such as word processing software.

273,000 jobs in internet related ICT
Businesses in Category E generated a total turnover of € 71 billion and provided 273,000 jobs in 2015. The companies in this category are slightly older and larger than the online stores and online services businesses. 36% are founded more than 10 years ago. Approximately 4% have 50 or more employees. Especially the mid-sized companies (50–249 employees) are well represented. Businesses in Category E are more often based around Amsterdam and Rotterdam, and also in the province of Flevoland.

6.3.3.1 Relative regional distribution of branches of Internet related ICT (category E), 2015



7. Discussion

In this section, we begin by addressing the four research aims and the central research question outlined in the introduction. We will then proceed to discuss the strengths and limitations of this study. This study is experimental and exploratory in its nature and as such it is important that the strengths and weaknesses are explicitly discussed. Again, in line with this research as experimental and exploratory, we discuss the possibilities for future research, with a specific emphasis on how to build on the strengths of this study, and most importantly, to address the weaknesses.

7.1 Conclusion

To begin with, we will quickly summarise how the work in this study has addressed the research aims. We will then proceed to discuss the central research question. The first research aim is:

To construct a definition of the internet economy that: 1) reflects the beliefs on what the internet economy is and 2) is pragmatic considering the possibilities of big data analyses.

We have ensured that this aim has been met by the manner in which the definition was constructed. Firstly, we explored the big data provided by Dataprovider to come up with initial ideas for the definition. We ensured that the definition reflected the beliefs on what the internet economy is, firstly by basing our definition on existing definitions of similar concepts and secondly, by developing our definition in collaboration with the steering group consisting of representatives the Dutch government, business-world, academia, Google and Dataprovider. The result is a definition which is both pragmatic, big data based, and broadly accepted. The internet economy is defined as the set of businesses which are, according to the methods described in this study, identified as having an online presence, being online stores, providing online services, or providing internet related ICT. The second research aim is:

To show the importance and size of the internet economy as part of the Dutch economy.

While we have produced a definition of the internet economy which is pragmatic and broadly acceptable, there are many challenges in operationalizing the definition in a way which would facilitate answering this question fully and with full certainty. There are also conceptual and technical issues to consider which always complicate answering questions about the economic importance of any sector or technology. However, analysis of the businesses within our definition of the internet economy provides many useful insights into the nature of the economy. The results suggest that relatively few businesses (36%) have a website and thus a relation to the internet but these businesses account for a disproportionately large share of the turnover (87%) and jobs (86%). The core of the internet economy (online stores, online services and internet related ICT) is with 3.3% of the number of business, 4.4% of the jobs and 7.7% of the turnover a modest but significant sector in our economy. We have also shown how the importance and size of the internet economy varies spatially and explored the characteristics of the businesses which make up the internet economy, such as their size and age. The third research aim is:

To show, by proof of concept, the possibilities of new measurement methods for producing statistics.

The internet is one of the most promising avenues for big data collection and thus for increasing the availability of timely and relevant data. This study has made an important step in demonstrating how big data from the internet can be combined with existing standard data sources in order to further capitalise on the valuable insights provided by big data from the internet. We have developed a complex linking methodology (based on different linking key combinations) which facilitates an optimised link between big data from the internet and the other standard sources. While there are many challenges to overcome in combining big and standard data sources (which will be discussed later), this study has shown that meaningful and novel insights can indeed be generated in this way. The fourth research aim is:

To explain differences from standard statistical concepts/classifications and existing statistics.

As has already been mentioned, the standard method for studying particular sectors of the economy is to rely on the SIC codes. The methodology in this study begins not by classifying businesses, but by classifying websites. The classification of businesses is made on the basis of the classification of the websites which link to the business. This is a fundamentally different statistical approach to classifying businesses.

In section 3, we have discussed definitions which are related to the concept of the internet economy. In some cases, the definition used in this study (henceforth 'our definition') defines an internet economy which can be loosely thought of as 'subset' of the internet economy conforming to other definitions. In other cases, there is overlap between our definition and other definitions, but neither can be considered a subset of the other. Our definition defines an internet economy which is a subset of the internet economies derived by NIESR and Growth Intelligence (2013) and the OECD (2013). In the case of NIESR and Growth Intelligence (2013), our definition results in a subset because it exclude digital activities not related to the internet. In the case of OECD (2013), our definition results in a subset because it excludes the social and cultural values of the internet. The definitions employed by CBS (2016) of the ICT economy and of the Boston Consulting Group (2011), define economies which overlap with the economy resulting from our definition. In the cases of the CBS ICT economy definition, this is because only the internet related elements of the ICT economy are included in our definition. In the case of the Boston Consulting Group, it is more challenging to identify the source of the overlap because their definition is a macro-data approach instead of the micro-data approach adopted in this study. In any case, the definitions share some conceptual similarities. On this basis, we can expect overlap. Finally, the central research question is:

To explore the possibilities to deepen our understanding of the importance of the internet economy.

In this study, we have employed a method which demonstrates how it is possible to gain insights into the internet economy by combining existing CBS micro-data with big data available from the internet. To properly explore the possibilities, we need to analyse the strengths and weaknesses of this method.

7.2 Strengths

The key strength in this study is the combination of big data from the internet with Statistics Netherlands’ micro-data. We have demonstrated how these two sources of data can be complementary. The big data from the internet provided by Dataprovider is very contemporaneous due to monthly updates and provides a wealth of information on the businesses behind the website. CBS micro-data provides reliable and detailed statistics on businesses. Furthermore, the use of standard CBS micro-data facilitates comparison to other traditional statistics. The combination of big data from the internet and CBS micro-data is a key strength because it opens up new possibilities for analysis and insight.

The use of big data in this study both facilitates and requires innovative methodologies. One particularly strong aspect of our methodology is the level of detail and flexibility in the categorisation of websites. The allocation of websites to a given category is achieved by first allocating the website to a sub-category. Sub-categories can be easily added or removed to improve or modify the definition. This is particularly valuable given the speed at which the internet economy develops. The last revision of the SIC was in 2008. Since 2008 there have been a lot of new business opportunities related to the internet that have not yet been incorporated in the SIC, if indeed they will ever become significant enough to merit inclusion. For example, app builders do not have their own SIC code yet but we do have them as a subcategory of our category E in this study. As new facets of the internet economy develop, they can be easily included in the method in this study. Using the internet to do this is advantageous because businesses are more likely to change information on their website than to contact the CoC to change their registered economic activity.

Further, any new kind of activity, business or otherwise, can potentially be studied using the keywords in Dataprovider data. For example, social enterprises are becoming more common but there are only a few data sources which can provide insight into this phenomenon. An analysis of Dataprovider keywords could potentially identify social enterprises, or indeed any possible way to group businesses according to key words such as ‘biological’, ‘fair trade’ or ‘sustainable’. The Dataprovider database in combination with the GBR could potentially be used to provide insight into all of these types of business. Combining the GBR with Dataprovider has the potential to free research questions from the confines of the SIC to define any subset of the economy. This holds in as far as that subset of the business economy has an online representation which is sufficiently large and representative.

The flexibility of the method is demonstrated by considering other innovative aspects of this study. This study identifies online services as a distinct category of the internet economy. While there is quite some attention given to ICT and online stores, online services have been shown to be an important part of the internet economy too. It is also a part of the internet economy which will be interesting to study into the future as an increasing portion of services may come to be provided digitally. This study is also innovative in the way in which it allows for the peripheral category of online presence. This category respects that it is often difficult to say, in black and white terms, whether a business belongs to the internet economy or not. In general, the extent to which a given business is involved in the internet economy can be best considered as a continuum rather than as, black and white, inside or outside the internet economy. Having the peripheral category of internet presence is a useful step in the direction of recognising this reality.

In general then, the strengths relate to the different types of data used, the combination of many data sources and the flexibility in, and applicability of, the methodology. This is a relatively short list of strengths, but each of these strengths carry significant weight.

7.3 Limitations

In comparison to a short list of strengths each carrying much weight, there are many limitations to this study. Often, a given limitation is not particularly problematic, but they all need to be made explicit and carefully evaluated. We will consider the limitations of the data, the definition and the method. Finally, we discuss the extent to which we can be confident in the results.

Regarding the data, the principal limitation is that the Dataprovider data does not include all Dutch websites. While the coverage is estimated to be high at 95%, the consequence is that not all businesses can be linked to a website. For example, some websites from foreign business that have a branch in the Netherlands are more likely to be missed. This means that it is likely that the size of the internet economy is somewhat underestimated in this study. Another weakness in the data also leads us to suspect that the size of the internet economy is somewhat underestimated. This is because many business do not have an ‘individual’ website, but make use of Facebook and other social media outlets to facilitate an online presence. These companies cannot be identified as part of the internet economy within this study. It is likely to be the smaller businesses that make use of social media for an online presence because this is cheaper than constructing and maintaining one’s own website. This suggests a degree of bias in our results towards larger businesses. A possible way to address this weakness is to acquire data from Facebook on the number of pages which classify themselves as businesses. While it may not be possible to make any links to the ABR on the basis of these data, there may be some opportunities to better understand to what extent our results are biased.

Regarding the definition, it is important to note that the process of constructing the definition has been to a large extent determined by the availability of data. We have thus not started from a purely theoretical perspective, or a perspective determined in the first instance by the ‘beliefs’ about what the internet economy is. The process was to consider what was possible with the data, and to reconcile this with what was pragmatically possible. A good example is that we did not have information on the extent businesses use the internet for their services. Businesses like banks provide a lot of online services which have significant economic value (online banking). In this regard, banks could be categorised as an online service (category D). However, our definition categorises them as having an active online presence (category B2). In practise, banks belong to both category B1 (consider online information about mortgages), category B2 (active online presence, one can apply for a credit card online) and also category D (online banking). A pragmatic choice was made to allocate banks to category B2. It is not possible to split the turnover of banks between these three categories. Similar issues exist for many kinds of businesses. In all cases, decisions were made which as far as possible reconcile pragmatic considerations with the beliefs about what the internet economy is.

Also regarding the definition, some of the assumptions are more conservative than others. In two major respects, the definition is conservative. Firstly, businesses are only included in the core categories if it is reasonably certain that the majority of their business activities fall into one of the core categories. In this way, businesses which are defined as having internet presence may also perform some business activities associated with the core. Secondly, only businesses with a website were allocated to the core of the internet economy. Although it seems unlikely, it could be the case for example that a web-designer does not necessarily have a website and such that web-designer would not be included in the definition. The definition is not at all conservative, on the other hand, because all of the turnover of a business which falls under the definition (core and internet presence) is allocated to the internet economy.

Regarding the method, there are two key limitations. The first concerns the link to the GBR. It has not been possible to link every website to the correct business. There is nothing to stop a website from putting the CoC and telephone numbers of another business, for whatever reason, on their own website. It is also not possible to ascertain with certainty the extent to which incorrect links have been made. There is also no guarantee that the telephone numbers or CoC numbers on the website have been correctly (without typos) placed on the website, even if it was intended to use those of the business behind the website. Further, large, complex business with multiple BUs per EG are difficult to link with the GBR because an EG often only has one website for all of its BUs.

The second limitation regarding the method concerns the issues of multiple websites per CoC, multiple CoCs per EG and BUs belonging to multiple categories (overlap). We constructed decision rules to deal with these problems and we are confident that these decision rules are the best approach. As such, the categories are, in broad brush terms accurate, but we cannot guarantee that all businesses do not fall into the incorrect category.

All the above weaknesses were considered in depth throughout the research process in order to ensure that they were dealt with in the most appropriate way. However, it is not possible to always know the extent of the pejorative effect that these limitations have on the results. Nonetheless, some of the results do give an insight into the cumulative effect of the limitations. Of particular interest is that only 36% of Dutch businesses have a website according to the results. This will be in part due to the use of existing social media outlets to provide an internet presence and partly due to the conservative nature of many aspects of the method. This is particularly problematic because of the possibility that certain types of businesses, in particular businesses with only 1 employee are structurally under-represented in the results. The bias towards larger businesses is a cause for concern because it is conceivable that many smaller businesses rely on the internet to a greater extent than larger businesses.

7.4 Suggestions for further research

Our recommendations for future research relate firstly to improving the methodology and secondly to broadening the application of the method. The method needs to be improved in ways which address the weaknesses discussed in the previous section. In the first instance, this means including data from Facebook or other sources in order to complement the Dataprovider data. This will help to reduce any bias towards larger companies and better understand the true scope and nature of the internet economy. The method can also potentially be improved by adopting machine learning. In this case, machine learning would involve constructing an algorithm to allocate businesses to a given category. Within this project, we experimented with allocating websites to categories based on the keywords of those websites. Unfortunately, the keywords contain too much white noise for machine learning algorithms to function[12] and there was not enough time available in this project to come up with a method to overcome this. Nonetheless, the concept of machine learning offers many possibilities for improving the method. One option which was not explored within this study is to construct a machine learning algorithm which uses information from not only websites but also from GBR. The GBR, which is in principal devoid of all the white noise present on the internet, could potentially provide enough structure in the data to facilitate effective machine learning. Such an approach would be equally methodologically challenging and innovative. It would thus require much time and effort to firstly determine its feasibility and secondly to operationalize it. This approach could, however, be extremely valuable, because machine learning is a vital tool to fully capture the information within, and potential of big data.

The method as it is now can also be more broadly applied in the future. Within the Netherlands, it would be useful to repeat the study of multiple years in order to obtain longitudinal data. Having an insight into trends in this case would facilitate an insight into questions such as whether businesses in the internet economy grow faster than businesses outside of it, or whether the survival chances of businesses with a website are better or worse than those without. It could also be useful to repeat the study in other countries. Dataprovider has data on the internet of more than 40 countries. If the methodology can be repeated in a sufficiently similar way, then comparisons can be made between the internet economy in different countries. This would allow us to answer questions regarding the size and composition of the internet economy in different countries.

Finally, there is also the possibility to revisit the definition of the internet economy. As repeatedly mentioned, our definition has been the result of a pragmatic process of reconciling the beliefs about the internet economy with the definitional possibilities dictated by the data. Therefore, an open mind should be kept regarding the possibilities to better define the internet economy, especially if more data can be acquired to expand what is possible regarding the internet economy. An important area where more understanding would be beneficial is the role of the internet at the level of individual business. As mentioned, businesses can be involved in different activities which fall under different categories of the internet economy and each of these different activities can be to different extents dependent on the internet. A better understanding of these issues would allow the indicators to be modified so as to account for the share of turnover, for example, that can be considered as allocable to the internet economy.

[12] Dataprovider use machine learning in relation to e-commerce websites. This functions well, but our experience suggests that the diversity of activities and prevalence of white noise within the other categories precludes effective machine-learning.

References

Bean, C. H. (2016). Independent Review of UK Economic Statistics.

Boston Consulting groep (2011). Interned. Hoe het internet de Nederlandse economie verandert.

CBS (2015). ICT, kennis en economie 2015.

Dataprovider (2016) How the Dutch use TLDs. What top-level domains do Dutch registrants use and how well is the coverage of Dataprovider?

McKinsey Global Institute (2011). Internet matters: The Net's sweeping impact of growth, jobs, and prosperity.

National Institute of Economic and Social Research and Growth Intelligence (2013). Measuring the UK's digital economy with big data.

Organisation for Economic Co-operation and Development (2013), 'Measuring the Internet Economy: A Contribution to the Research Agenda', OECD Digital Economy Papers, No. 226, OECD Publishing. http://dx.doi.org/10.1787/5k43gjg6r8jf-en.

Appendix A: Dataprovider Dataset

Group	Field	Example data	Description
Geolocation	Country	The Netherlands	Where the business is actually operation
	Region	Noord-Holland	The state or province where the business operates
	Zip Code	9723 HS	The zip code of the business
	Zip Code Quality	75%, (from 0–100%)	How certain we are that this is the right zip code
	City	Amsterdam	The city where the business operates
	Address	Grote Markt 23	The address of the business
	lat / long	32.70179, –97.62637	Geographic coordinates of the business
Business	Company Name	R. Goulooze Holding B.V.	The name of the company
	Legal Entity	B.V.	The legal entity of this business
	Chamber of Commerce	30125826	The chamber of commerce number of the business
	Bank Account Number	325324603	The bank account number of the business
	IBAN Number	NL36RABO0325324603	The IBAN bank account number of the business
	BIC Number	RABONL2U	The BIC number of the bank that the company uses
	Bank	Rabobank	What bank the company uses
	Tax Number	NL115441359B01	The tax number of the company
	Phone Number	31(0)306579252	The most important phone number of the company
	Phone Number Quality	83%, (scale from 0–100%)	How certain we are if this is the right phone number
	Phone Numbers	31(0)325324603	The phone numbers we have found
Content	Hostname	www.shirtplanet.nl	The URL of the company
	Title	ShirtPlanet.nl De Leukste T-shirts Online	The title that is used on the website

Group	Field	Example data	Description
	Description	ShirtPlanet leuke grappige t-shirts kopen van baby rompertje tot extra grote maten.	The meta description that is used on the website
	Keywords	shirtplanet, shirt, grappige, rompertje, verjaardag, kado, leuke, humor, funny	The most important words that are used on the website
	Category	Fashion	Dataprovider has xx default categories and based on the content of the websites assigns it accordingly
	Authors	Frans de la Haije	The author that is found in the html code, often the designer of the website
	Copyright	MediaCT	The copyright notice that is found in the html code, often the designer of the website
	Multi Language	No	The website contains copy in one or more languages
	Language	NL, EN	What languages are used on the website
eCommerce	Online Store	Yes	The website contains an online store
	eCommerce Probability	37%, (scale from 0–100%)	How certain we are that the website contains an online store
	Shopping Cart Software	Magento	What kind of shopping cart software does the website use
	Trustmarks	Thuiswinkel waarborg,	The trustmarks the website uses
	Delivery Services	TNT, DPD	What delivery services are offered
	Payment Methods	Mastcard, VISA, PayPal	What payment methods are offered
	Payment Services Provider	PayPal, Docdata	What online payment services are used
	Currency	EUR	In what currency the priced are advertised
	Average Price	29	The average price of the products offered
	Products	3,924	Estimation of how many products are offered on the website
Marketing	Alexa Rank	17,601,874	How does the website rank on visitors (lower score is more visitors) according to Alexa
	Tweeds	15	How many tweets are sent by the Twitter account that is published on the website
	Facebook Shares	23	How many posts are shared on the Facebook account that is published on the website
	Incoming Links	115	How many incoming links the website has
	Refering websites	12	How many other websites are linking to the website
	Anchor texts	www.shirtplanet.nl, t-shirts kopen, tshirts, online t-shirtsbestellen	What Anchor text is used in the incoming links
	Site traffic	4,078	How many unique visitors the website has, this is an estimate by Dataprovider
	Analytics Id	UA-9625634-15	The Google Analytics ID that is found on the website
	Adsense Id	PUB-5709822177168445	The Adsense ID that is found on the website
	Analytics Id	Google Analytics	What kind of statistics software the website uses
	Ad Network	Google Adsense	Which Adnetworks the website works with
	Affiliates	Daisycon	Which affiliate networks the website works with
	Social	Twitter, LinkedIn	Does the website use Facebook, Twitter, LinkedIN, Pintrest, Google +
	Social Profiles	www.linkedin.com/company/feederlines, www.twitter.com/feederlines	The social profiles the website publishes
	Live Chat Software	CoBrowser	Visitors of the website can chat with the people behind the website
	Jobs	Yes	The website has jobs advertised
Technical	CMS	Plone	What Content Management System is used on the website
	Scripting Language	PHP	What scripting language is used to build the website
	Technical evaluation	6,7 (scale from 1–10)	How is the website coded, the W3C is the benchmark

Group	Field	Example data	Description
	SEO Score	84% (scale from 0–100%)	How well does the website use all the HTML elements to tell search engines what the website is about
	Flash	Yes	The website uses Flash software
	RSS	Yes	The website offers a RSS feed
	Login	Yes	The website has content that's protected by a login and password
	HTML version	XHTML 1.0 Strict	What version of HTML is used to code the website
	Generator	Plone – http://plone.org	What HTML generator is used to code the website
	Mobile version	No	Is there a mobile version of the website or is the website responsive
	Mobile App	App Store	Does the website refer to apps in the Apple App store
	Maps	Bing Maps	The website makes use of maps
	Libraries	MooTools, Slimbox, SWFObject	What scripting libraries are used
Hosting	Top Level Domain	Nl	The top level domain of the website
	Subdomain	www	Subdomain of the website
	Domain	shirtplanet.nl	The domain of the website
	Hosting Country	NL	In what country the website is hosted
	Domain age	120	The number of months ago the domain was registered
	IP Address	141.255.181.112	The IP address of the website
	AS number	51,686	ID of the owner of the IP block
	AS Company	Antagonist B.V.	Name of the company that belongs to the AS number
	Reverse DNS lookup	www.antagonist.nl	Domain associated with IP address
	Operating System	Ubuntu	The operating system of the server where the website is hosted
	Webserver	Apache/2.2	The software used to deliver the website
	Server Signature	Apache/2.2.14 (Ubuntu)	Information about the hosting server that is included in the header response
	SSL Certificate	No	The website uses Secure Sockets Layer for secure connections
	Status codes	200,404	The status codes of the pages found during indexation
	Average load time	109 Kb/s	The average loadtime of the website
	CDN	Akamai, Google API	Does the website use a content delivery network
	Video	Vimeo, Youtube	What third party streaming video suppliers the website uses
	Parking	GoDaddy	There's no website, it is parked with a default page from the webhoster
	Email provider	Gmail	Hosting provider of the email mentioned in the MX records of the DNS

Appendix B: Keywords Categories D and E

Internet category	Subcategories	Topic	Keywords: base	Keywords: combination
Category D: Online services	Leisure	Hotels	hotel, hotels, resort, resorts, hostel, bedenbreakfast, hostels, booking	hotels
		Flights	vliegtickets, vliegticket, vliegwinkel, tickets, ticket	vluchten, luchtvaart, airlines, luchtvaartmaatschappijen
		Holidays	reizen, rondreizen, reisspecialist, reis, vakantie, vakanties, travel, travels	online, website, site, vergelijk, vergelijken
		Food	thuisbezorgd, iens, eten, recepten, restaurant, restaurants, koken	eten, online, vergelijk, vergelijken
	News and entertainment	News	nieuws, weer, news, media, magazine	online
		Blogs	blog, blogger	online
		Vlogs	vlog, vlogger	online
		Games	spelletjes, spelletje, game, games, spellen, spelen, speel, spel	online, website, site, gratis
		Videos, music	tv, film, films, filmpjes, filmpje, movie, movies, serie, series, clip, clips, videos, video, muziek	online, streaming, stream, kijk, bekijk
		Books	boeken, books, boeken	online
		e-learning	learning, learn, cursus, cursussen, leer, leren	online
		Gambling	gokken, casino, casinos, gok, gokkasten	online
		Adult	escort, porno, porn, seks, seks, naakt, kinky, gay, erotic	
	Business	Advertising	adverteren, adverteer, reclame, reclamebureau, reclamebureaus	online, optimalisatie, adwords
		Finance	bankieren, bank, beleggen, onlinebeleggen, beleggers, belegger, beleggingen	online, digitaal, bitcoin, digitale
		Consultancy	advies, consultancy	online
		Jobs	vacaturesite, vacaturesites, vacatures, vacature, uitzendbureau, uitzendbureaus, uitzendwerk, uitzendkrachten, vakantiewerk, vakantiebaan, vakantiebanen, bijbaan, bijbanen, baan, banen, recruitment, recruiter, job, jobs, carriere, career	online, zoek, zoeken, vind, vinden
	Retail	Housing	woning, woningaanbod, huis, huizen, koopwoningen, makelaars, makelaar, huurwoningen, hypotheken, hypotheek, funda	online, zoek, zoeken, vind, vinden
		Price comparison	energie, verzekering, abonnementen, prijzen, aanbieders, goedkoopste, bespaar, autoverzekering, prijsvergelijk, prijsvergelijking, prijsvergelijker, abonnement, mobiel	vergelijker, vergelijker, vergelijk, vergelijken
		Tickets	tickets, ticketing, ticket	online
		Auctions	veiling, onlineveiling, veilingen, onlineveilingen, auction, auctions, veilingmeester, internetveiling, internetveilingen	online
		Online trade	marktplaats, tweedehands, speurders	online

Internet category	Subcategories	Topic	Keywords: base	Keywords: combination
	General services	Dating	datingsite, dating, date, sexdate, sexdating, daten	online
		Visualisations	visualisatie, visualisaties, visualiseer, animatie, animaties, ontwerp, ontwerpen	online, 3d
		Transport	car, meeliften, lift, liften, meerijden, 9292, reisplanner, routeplanner, parkeren, auto	online, app
		Online payment	diensten, payment, betalen, betaaloplossingen	online
Category E: Internet related ICT	Hosting and cloud	Webhosting	hosting, domeinnaan, webhosting, server, service, cloud, development, ontwikkeling,	websites, opslag, cloud, online
		Cloud services, Datacentres	clouddiensten, datacenter, datacentrum, cloudcomputing, massaopslag, dataopslag, databanken, datawarehouse, gegevensverwerking, ontdek, levert, gebruik, omzet, zakelijk, leverancier, aanbieders, handelaar, database	websites, opslag, cloud, online
		Website design, developing	webdesign, internetbureau, websitebouw, webportal, webportals, developer, websites, design, ontwerp, software, develop, mobiele, ontwikkelaar, developer, ontwikkelen, developer, ontwikkeling, development, grafisch, origineel, bekeken, bediening, ontwikkeling, beeld, beelden, geluid, marketing, interactief, privacy, interactieve, interactie, bewegende, verhaal, animaties, multimedia, bouwers, bouw, bouwen, gebouwd, producent, handelaar, ontwerp	software, pakketten, virtual, online, websites, web
		App design, developing	mobiele, iphone, android, tablet, mobile, opdracht, uitgeverij	app, apps, digital, digitale, applicaties
	Software	Software products and services	software, develop, ontwikkelen, produceren, uitgeven, adviseren, databanken	
	Marketing and consultancy	Internet marketing	ecommerce, emarketing, adwords, internetmarketing, zoekmachine, zoekmachines, b2b, b2c, verkoop, inkoop, online, marketing, seo, optimalisatie, socialemedia, sociale, analytics, adverteerders, gedreven, intelligence, privacy	online, websites, elektronisch, internet, netwerk, gegevens, search, business
		Internet consultancy	internet, online, ontwikkeling, consultancy, research, onderzoek, advies, analyse, analyseren, techbedrijf, technologie, boekhoudleverancier, communicatie, wiki, community, hangout, skype, berichtendienst, communicatie, youtube, tweets, twitter, snapchat, facebook,	internet, consultancy, time, online, digital, digitale, internet, webdiensten, netwerk, it, ict,

Internet category	Subcategories	Topic	Keywords: base	Keywords: combination
	Infrastructure and security	Firewalls	firewall, firewalls, cyber, vpn, spyware, antispam, netwerkbeveiliging, internetbedreigingen, hacking, hackers, security, beveiliging, cybercrime, spam, phishing, tracking, pharming, risk, risico, incident, virus, managed, oplossingen, oplossing, beschermt, specialist, instellen, inbraak, verminking, wachtwoorden, creditcardgegevens, beschermen, beveiligd, blokkade	applicaties, geavanceerde, cyber, web, internet, webdiensten, netwerk, gegevens, eigendom, publieke, veiligheid
		Datamining & Big Data	robots, automatiseren, crawler, dataverzameling, datamining, textmining), webdesign, data, leverancier, aanbieder, handelaar, tekst, specialist, toepassen, google, overweegt, ontdek, patronen, intelligentie, ontwikkel, ingeprogrameerd, platform, recognition, machine, verzamelen, gebied, bedrijfsgegevens, verzamelt, science, big	data, kunstmatige, telecommunicatie, software, pakketten, learning, virtual, online, websites, app, apps, digital, digitale, elektronisch, applicaties, geavanceerde, opslag, cloud, internet, consultancy, webdiensten, netwerk, gegevens, veiligheid, it, ict

Appendix C: Complete set of tables

1. Number of businesses in the interneteconomy broken down by size, 2015

		Size			
	Total	1 employed person	2–49 employed persons	50–249 employed persons	250 or more employed persons
Total	1,530,240	1,178,160	338,380	10,820	2,870
Internetcategory					
Category A: no website	974,490	808,380	164,230	1,770	110
Category B1: passive online presence	437,770	299,330	130,490	6,260	1,700
Category B2: active online presence	67,710	34,890	29,710	2,160	950
Category C: online stores	28,430	21,620	6,700	100	10
Category D: online services	5,650	3,820	1,790	40	10
Category E: internet related ICT	16,190	10,130	5,470	500	100

Source: CBS..

2. Number of businesses in the interneteconomy broken down by sector and size, 2015

			Sector																				
		Total	A Agriculture, forestry and fishing	B Mining and quarrying	C Manufacturing	D Electricity, gas, steam and air conditioning supply	E Water supply; sewerage, waste management and remediation activities	F Construction	G Wholesale and retail trade; repair of motor vehicles and motorcycles	H Transportation and storage	I Accommodation and food service activities	J Information and communication	K Financial institutions	L Renting, buying and selling of real estate	M Consultancy, research and other specialised business services	N Renting and leasing of tangible goods and other business support services	O Public administration, public services and compulsory social security	P Education	Q Human health and social work activities	R Culture, sports and recreation	S Other service activities	T Activities of households as employers	U Extraterritorial organisations and bodies
Total		1530240	71980	390	59610	990	1520	150990	223380	36030	51480	81290	83590	24480	303430	62710	760	65820	127050	91940	92800	20	10
Internetcategory	**Size**																						
Category A:	Total	974490	64430	270	31440	680	890	115580	121400	27810	30110	37780	73000	17540	183570	37760	180	37000	80900	53010	61140	10	0
no website	1 employed person	808380	35110	200	24320	490	710	103200	85800	19400	15000	34130	67500	13810	169570	30850	110	34970	69380	49160	54640	10	0
	2–49 employed persons	164230	29280	60	6860	190	170	12310	35200	8250	15000	3560	5470	3710	13870	6630	30	1960	11390	3830	6450	0	0
	50–249 employed persons	1770	40	0	260	0	10	70	390	140	100	80	20	10	120	250	30	60	110	30	40	0	0
	250 or more employed persons	110	0	0	10	0	0	0	10	10	10	10	0	0	10	20	10	10	20	0	0	0	0
Category B1:	Total	437770	6390	120	23610	250	570	33260	64160	7150	15100	28190	8990	5240	103100	20490	400	22700	38560	32600	26900	0	0
passive online presence	1 employed person	299330	2470	50	10580	150	170	20320	32330	2750	5540	23640	5630	3180	82800	12330	20	19380	28810	28180	21020	0	0
	2–49 employed persons	130490	3830	50	11270	70	320	12320	30690	3790	9420	4460	3200	1960	19800	7290	50	2890	8980	4330	5770	0	0
	50–249 employed persons	6260	80	20	1460	20	50	540	980	500	120	70	110	90	400	680	230	320	420	80	90	0	0
	250 or more employed persons	1700	10	10	310	10	20	80	170	110	20	20	50	20	100	190	110	110	350	10	20	0	0
Category B2:	Total	67710	930	10	2960	60	70	1380	17080	920	5880	3300	1150	1570	8410	3170	160	5200	6840	5120	3520	0	0
active online presence	1 employed person	34890	360	0	1090	30	10	670	7390	350	1390	2330	620	730	5970	1520	0	3440	3190	3470	2340	0	0
	2–49 employed persons	29710	560	0	1580	10	40	640	9090	470	4370	910	490	770	2330	1300	10	1240	3240	1560	1110	0	0
	50–249 employed persons	2160	10	0	230	0	10	60	460	80	100	40	20	60	80	250	70	320	230	90	60	0	0
	250 or more employed persons	950	0	0	70	10	10	10	140	30	20	20	20	10	30	100	70	200	190	10	10	0	0
Category C:	Total	28430	180	0	1330	0	0	500	18880	90	230	1430	230	80	2310	460	0	500	460	840	930	0	0
online stores	1 employed person	21620	100	0	950	0	0	380	14170	70	120	1180	170	50	1940	340	0	410	370	670	710	0	0
	2–49 employed persons	6700	80	0	370	0	0	120	4630	20	110	250	60	20	370	120	0	90	80	160	210	0	0
	50–249 employed persons	100	0	0	10	0	0	0	60	0	0	0	0	0	0	0	0	10	0	10	10	0	0
	250 or more employed persons	10	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Category D:	Total	5650	10	0	40	0	0	40	570	10	20	2720	60	10	1360	440	0	150	60	50	100	0	0
online services	1 employed person	3820	0	0	20	0	0	20	320	0	10	1960	30	10	990	210	0	110	40	40	70	0	0
	2–49 employed persons	1790	10	0	20	0	0	20	240	10	20	740	40	0	370	210	0	40	20	10	30	0	0
	50–249 employed persons	40	0	0	0	0	0	0	0	0	0	20	0	0	0	20	0	0	0	0	0	0	0
	250 or more employed persons	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Category E:	Total	16190	40	0	240	0	0	230	1300	40	140	7870	170	50	4680	390	10	270	240	310	220	0	0
internet related ICT	1 employed person	10130	20	0	120	0	0	160	750	20	60	4850	130	40	2920	230	0	210	180	270	170	0	0
	2–49 employed persons	5470	20	0	100	0	0	80	510	20	70	2690	40	10	1600	140	0	50	40	40	50	0	0
	50–249 employed persons	500	0	0	10	0	0	0	30	0	0	290	0	0	140	20	0	0	0	0	0	0	0
	250 or more employed persons	100	0	0	10	0	0	0	10	0	0	50	0	0	20	10	10	10	10	0	0	0	0

Source: CBS.

3. Number of businesses in the interneteconomy broken down by age and size, 2015

			Age						
		Total	0 year	1 year	2 year	3 year	4 year	5–9 year	10 or more years
Total		1,530,240	166,910	142,880	115,110	113,370	99,290	403,170	489,520
Internetcategory	**Size**								
Category A: no website	Total	974,490	126,330	101,790	77,380	73,480	61,810	245,670	288,030
	1 employed person	808,380	116,870	92,350	69,210	65,570	54,630	212,590	197,170
	2–49 employed persons	164,230	9,400	9,360	8,110	7,840	7,110	32,790	89,630
	50–249 employed persons	1,770	60	80	50	70	70	270	1,170
	250 or more employed persons	110	10	10	10	0	10	20	60
Category B1: passive online presence	Total	437,770	31,160	31,320	29,090	31,300	29,570	125,920	159,410
	1 employed person	299,330	27,230	26,420	23,800	25,480	23,890	96,790	75,720
	2–49 employed persons	130,490	3,800	4,760	5,120	5,680	5,500	28,050	77,570
	50–249 employed persons	6,260	110	120	130	110	140	800	4,860
	250 or more employed persons	1,700	30	30	30	40	40	290	1,250
Category B2: active online presence	Total	67,710	4,780	4,950	4,450	4,390	4,150	17,360	27,620
	1 employed person	34,890	3,860	3,720	3,140	2,990	2,710	10,350	8,130
	2–49 employed persons	29,710	880	1,170	1,250	1,340	1,360	6,500	17,210
	50–249 employed persons	2,160	30	40	50	50	60	320	1,610
	250 or more employed persons	950	10	20	20	10	30	190	680
Category C: online stores	Total	28,430	3,070	3,290	2,720	2,580	2,260	7,650	6,860
	1 employed person	21,620	2,850	2,970	2,400	2,130	1,830	5,850	3,590
	2–49 employed persons	6,700	220	310	320	450	430	1,780	3,190
	50–249 employed persons	100	0	0	0	10	0	10	70
	250 or more employed persons	10	0	0	0	0	0	0	10
Category D: online services	Total	5,650	440	460	440	460	410	1,690	1,770
	1 employed person	3,820	380	370	320	350	310	1,200	890
	2–49 employed persons	1,790	70	80	120	110	90	480	840
	50–249 employed persons	40	0	0	0	0	0	10	30
	250 or more employed persons	10	0	0	0	0	0	0	10
Category E: internet related ICT	Total	16,190	1,130	1,060	1,040	1,160	1,080	4,880	5,840
	1 employed person	10,130	1,000	880	780	880	810	3,360	2,410
	2–49 employed persons	5,470	110	180	230	270	260	1,410	3,020
	50–249 employed persons	500	10	10	20	20	20	80	340
	250 or more employed persons	100	0	0	0	0	0	20	70

Source: CBS.

4. Turnover in the interneteconomy broken down by size, 2015

		Size			
	Total [1]	1 employed person	2–49 employed persons	50–249 employed persons	250 or more employed persons
	billion euro				
Total [1]	1,361	82	382	381	517
Internetcategory					
Category A: no website	177	48	96	31	2
Category B1: passive online presence	816	27	216	278	293
Category B2: active online presence	264	3	50	52	160
Category C: online stores	23	2	6	6	9
Category D: online services	10	0	2	2	6
Category E: internet related ICT	71	1	12	12	47

Source: CBS.

[1] Total of the business economy. This does not include NACE categories A, K, O-U.

5. Employees in the interneteconomy broken down by size, 2015

		Size			
	Total [1]	1 employed person	2–49 employed persons	50–249 employed persons	250 or more employed persons
	x 1 000				
Total [1]	7,418	254	2,001	1,333	3,752
Internetcategory					
Category A: no website	1,049	171	635	201	42
Category B1: passive online presence	3,827	62	958	769	2,038
Category B2: active online presence	2,131	13	303	296	1,519
Category C: online stores	44	5	20	11	7
Category D: online services	25	1	12	4	8
Category E: internet related ICT	264	2	74	51	137

Source: CBS.

[1] Categories do not add up to the grand total because 1 % of the employees could not be matched to a company.

6. Jobs of employees in the interneteconomy broken down by size, 2015

		Size			
	Total [1]	1 employed person	2–49 employed persons	50–249 employed persons	250 or more employed persons
	x 1 000				
Total [1]	7,875	281	2,134	1,395	3,954
Internetcategory					
Category A: no website	1,125	187	682	212	45
Category B1: passive online presence	4,037	71	1,017	802	2,147
Category B2: active online presence	2,257	15	325	312	1,605
Category C: online stores	46	6	21	12	7
Category D: online services	26	1	12	5	8
Category E: internet related ICT	273	2	76	53	142

Source: CBS.

[1] Categories do not add up to the grand total because 1 % of the jobs of employees could not be matched to a company.

7. Production value, value added and employment in the interneteconomy, 2015

	Production value (basic prices)	Value added (basic prices)	Employed persons	Employed persons (fte)
	billion euro		x 1 000	
Total [1]	951	404	5,611	4,456
Internetcategory				
Category A: no website	176	89	1,195	930
Category B1: passive online presence	537	208	2,812	2,290
Category B2: active online presence	170	88	1,304	979
Category C: online stores	8	3	53	40
Category D: online services	9	4	25	21
Category E: internet related ICT	51	13	219	194

Source: CBS.

[1] Total of the business economy. This does not include NACE categories A, K, O-U.

8. Local branches in the interneteconomy broken down by province, 2015

		Internetcategory					
	Total	Category A: no website	Category B1: passive online presence	Category B2: active online presence	Category C: online stores	Category D: online services	Category E: internet related ICT
Total	1,701,830	1,010,910	514,860	117,420	30,690	6,820	21,140
Province							
Groningen	50,210	28,060	16,550	3,670	1,020	240	660
Friesland	62,590	36,220	19,920	4,470	1,170	250	570
Drenthe	44,020	24,670	14,730	3,180	840	170	440
Overijssel	102,930	57,500	33,480	8,210	2,060	420	1,260
Flevoland	38,420	22,220	11,780	2,740	920	150	620
Gelderland	200,510	114,190	64,880	14,730	3,710	750	2,260
Utrecht	142,380	82,760	44,360	10,230	2,320	620	2,090
Noord-Holland	334,880	205,880	95,300	22,120	5,650	1,540	4,390
Zuid-Holland	342,690	210,700	96,870	23,120	5,920	1,260	4,820
Zeeland	36,030	21,870	10,460	2,730	590	120	270
Noord-Brabant	253,220	151,460	77,190	15,910	4,840	980	2,840
Limburg	93,950	55,370	29,340	6,320	1,650	340	930

Source: CBS.

9. Local branches in the interneteconomy broken down by COROP-area, 2015

		Internetcategory					
	Total	Category A: no website	Category B1: passive online presence	Category B2: active online presence	Category C: online stores	Category D: online services	Category E: internet related ICT
Total	1,701,830	1,010,910	514,860	117,420	30,690	6,820	21,140
COROP-area							
Oost-Groningen (CR)	11,300	6,550	3,540	820	260	30	100
Delfzijl en omgeving (CR)	3,490	2,090	1,090	240	40	10	30
Overig Groningen (CR)	35,420	19,430	11,930	2,610	720	200	530
Noord-Friesland (CR)	29,610	17,100	9,450	2,110	560	120	270
Zuidwest-Friesland (CR)	15,050	8,860	4,770	1,010	240	60	110
Zuidoost-Friesland (CR)	17,930	10,260	5,710	1,340	370	60	190
Noord-Drenthe (CR)	17,070	9,210	5,990	1,280	320	80	190
Zuidoost-Drenthe (CR)	14,390	8,470	4,550	940	250	30	150
Zuidwest-Drenthe (CR)	12,560	6,990	4,180	970	270	50	100
Noord-Overijssel (CR)	34,200	19,450	10,990	2,570	650	150	390
Zuidwest-Overijssel (CR)	13,530	7,230	4,670	1,190	220	50	170
Twente (CR)	55,200	30,830	17,820	4,450	1,190	220	690
Veluwe (CR)	66,460	37,890	21,160	5,150	1,220	240	800
Achterhoek (CR)	39,090	22,070	12,930	2,830	720	140	390
Arnhem/Nijmegen (CR)	67,720	37,240	22,990	5,100	1,310	300	790
Zuidwest-Gelderland (CR)	27,250	16,990	7,800	1,650	460	60	280
Utrecht (CR)	142,380	82,760	44,360	10,230	2,320	620	2,090
Kop van Noord-Holland (CR)	36,550	22,300	10,570	2,580	680	130	290
Alkmaar en omgeving (CR)	24,090	13,590	7,610	1,920	560	110	290
IJmond (CR)	17,470	10,440	5,200	1,240	360	60	180
Agglomeratie Haarlem (CR)	27,550	16,190	8,470	1,900	520	150	320
Zaanstreek (CR)	14,820	8,850	4,310	1,120	300	60	180
Groot-Amsterdam (CR)	180,690	114,100	49,030	11,310	2,640	870	2,750
Het Gooi en Vechtstreek (CR)	33,700	20,400	10,110	2,050	580	170	400
Agglomeratie Leiden en Bollenstreek (CR)	39,310	23,300	11,720	3,040	650	150	450
Agglomeratie 's-Gravenhage (CR)	83,060	52,950	21,960	5,570	1,230	290	1,060
Delft en Westland (CR)	23,500	14,120	7,040	1,550	400	80	310
Oost-Zuid-Holland (CR)	32,020	19,170	9,560	2,160	620	140	370
Groot-Rijnmond (CR)	128,960	79,350	36,090	8,490	2,350	480	2,210
Zuidoost-Zuid-Holland (CR)	35,840	21,810	10,510	2,310	670	120	430
Zeeuwsch-Vlaanderen (CR)	9,420	5,880	2,610	660	170	30	70
Overig Zeeland (CR)	26,620	16,000	7,850	2,070	420	90	200
West-Noord-Brabant (CR)	61,870	37,630	18,290	3,920	1,180	230	610
Midden-Noord-Brabant (CR)	45,800	27,260	14,220	2,800	890	170	460
Noordoost-Noord-Brabant (CR)	69,250	41,690	20,880	4,330	1,310	240	800
Zuidoost-Noord-Brabant (CR)	76,300	44,880	23,800	4,860	1,460	340	960
Noord-Limburg (CR)	24,250	14,130	7,720	1,660	450	70	220
Midden-Limburg (CR)	21,940	13,010	6,750	1,450	400	80	250
Zuid-Limburg (CR)	47,750	28,220	14,880	3,210	800	190	460
Flevoland (CR)	38,420	22,220	11,780	2,740	920	150	620

Source: CBS.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2015–2016	2015 to 2016 inclusive
2015/2016	Average for 2015 to 2016 inclusive
2015/’16	Crop year, financial year, school year, etc., beginning in 2015 and ending in 2016
2013/’14–2015/’16	Crop year, financial year, etc., 2013/’14 to 2015/’16 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
CCN Creatie, The Hague

Design
Edenspiekermann

Information
Telephone +31 88 570 7070
Via contact form: www.cbs.nl/information