Discussion Paper

Improving the sampling strategy for the Community Innovation Survey using machine learning algorithms

Jonas Klingwort
Kees van Berkel
Jan van den Brakel

April 2026

Abstract

National statistical institutes (NSI's) are, besides register data, increasingly interested in using non-probability data to produce official statistics. Examples include information on the internet (e.g., social media messages or web-scraped data) and sensor data. Relying on these kinds of data sources implies that an NSI has no control over the availability of the data or its comparability over time. To minimize the consequences of such risks, this paper proposes utilizing information extracted from these types of data sources to improve the sampling strategy of a probability sample. This concept is illustrated with an application to the Community Innovation Survey (CIS). We consider three different data sources that provide information about the true distributions in the target population and show how this can be used in the sampling strategy. These three data sources are: (1) web-scraped data indicating the likelihood of a company being innovative, (2) administrative records of businesses receiving research and development subsidies, and (3) administrative data on the number of patents associated with each business. Using data from the CIS 2016, the paper studies the extent to which survey estimate accuracy can be improved by weighting to a population distribution informed by these auxiliary sources. The weighted estimates obtained through the generalized regression estimator are compared to those derived from the currently used Horvitz-Thompson estimator. This analysis assesses whether the existing weighting approach can be refined and identifies which auxiliary source yields the most accurate estimates. In this way, the paper contributes to the ongoing discussion on using traditional and novel data sources to enhance the quality of official statistics.

Keywords: weighting, GREG estimator, HT-estimator, web-scraping, data quality, accuracy, LLM, tf-idf, bag of words, logistic regression

Contents

1 Introduction

Traditionally, official statistics are based on probability samples compiled using design-based inference methods. Under this approach, administrative and register data are used for survey weighting, providing reliable auxiliary information such as demographic, geographic, or economic variables to enhance the precision of survey estimates and correct for nonresponse bias. The advantage of this traditional approach is that a national statistical institute (NSI) is in control over the availability of survey data, the comparability of sample estimates over time, the minimum required precision of sample estimates, and the operationalization of concepts of interest, such as in questionnaires (van den Brakel 2022). However, the data landscape has evolved in recent years with the emergence of new data sources, including web-scraped data, social media platforms, and sensor-generated data. These alternative data sources offer rich, real-time, and detailed auxiliary information, enabling the supplementation of traditional data and the more effective capture of dynamic population behaviors. These data sources are generally referred to as non-probability data, meaning that they are not obtained by drawing a probability sample from a finite population.

The question of how these new data sources can be used to compile official statistics. One approach is to use these new data sources as the primary data source for compiling official statistics. The literature on the use of non-probability samples for finite target population inference is rapidly growing. Valliant and Dever (2011) or Sakshaug et al. (2019) propose inference methods to combine probability samples with non-probability samples. For a review of inference methods for non-probability samples see Rao (2021), Beaumont (2020), and Wu (2022), and the references therein.

As pointed out by van den Brakel (2022), the use of these data sources as primary data for compiling official statistics implies an increased risk, since a NSI is no longer in control over the availability of the data, its comparability over time, the way variables of interest are operationalized, and the potential for selection bias. In this paper, it is proposed to use these data as auxiliary information in the sampling strategy of a sample survey. In this way, new data sources are combined with sample survey data using the standard design-based inference approach, which NSIs widely apply in the production of official statistics.

In this paper, we illustrate this approach with an application to the Community Innovation Survey (CIS). The CIS, which collects data on innovation dynamics across businesses (in this paper, the terms ‘businesses’ and ‘companies’ are used interchangeably), relies on an efficient sampling and weighting strategy to ensure its survey estimates accurately reflect the underlying population. The current weighting approach is based on the Horvitz-Thompson estimator (HT) and might have limitations in capturing the complexity and variability of the Dutch business landscape. With the available auxiliary data, there is an opportunity to revisit the weighting strategy for the CIS.

This paper analyzes how incorporating auxiliary information, based on web-scraped data and registers, can improve the CIS’s weighting strategy. Web scraping enables the automated extraction of website text from businesses, providing information about their products, services, and strategic focus. By analyzing this textual data, one can identify keywords, phrases, or themes indicative of innovation, such as references to ‘research and development’, ‘patents’, ‘new technologies’, or ‘innovations’. Natural language processing techniques can further classify companies as innovative or not by detecting sentiment or industry-specific markers of innovation. The approaches taken and methods developed by Statistics Netherlands to detect innovative companies or other industry-specific businesses using text from their websites are

described in several papers by Daas and van der Doef (2020), Daas and van der Doef (2021), Daas et al. (2024b), and Daas et al. (2024a). Of course, there are also other approaches in official statistics using web scraping to produce official statistics, for example Hackl (2016), Oancea and Necula (2019), Uriarte et al. (2019), and Kühnemann (2021). The two registers provide information on subsidies granted to businesses engaged in research and development or conducting technical-scientific research, and on the number of patents that relate to a business, potentially indicating innovation activity. Accordingly, three data sources that provide information about the distribution of related population totals will be used as auxiliary datasets in the analysis: (1) information on the likelihood of a business being innovative derived from web-scraped data, (2) administrative records on businesses receiving research and development (R&D) subsidies, and (3) patent data of businesses. We will analyze how this auxiliary information might improve the accuracy of survey estimates by applying the generalized regression estimator (GREG, see Särndal et al. (1992)) and comparing its accuracy with that of the currently applied HT estimator. The paper contributes to the broader debate on integrating new and traditional data sources in official statistics to improve data quality in official statistics.

The paper is organized as follows: the motivation of this paper is given in Section 2. The data sources and the required processing are described in Section 3. The weighting methods and applied methods are explained in Section 4. The results are presented in Section 5. In Section 6, the paper is discussed and concluded in Section 7.

2 Research background

Official statistics published by NSIs are predominantly based on data obtained through registrations or collected through surveys, usually based on a probability sample. Design-based and model-assisted inference methods have been the preferred methods for NSIs to produce official statistics for decades. There is, however, an increasing interest among NSIs in using non-probability samples or other large datasets — so-called big data — that are generated as a by-product of processes not directly related to statistical production purposes for official statistics. Examples include time and location of network activity available from mobile phone companies, social media messages from X (formerly Twitter) and Facebook, internet search behavior from Google Trends, information found on the internet, scanner data, and sensor data, like, e.g., satellite images, aerial images, and road sensor data. The advantage of such data sources is that they are generally less expensive compared to survey data collection, come in large volumes, and are sometimes collected at high frequencies. However, there are several drawbacks to using register data or non-probability data. In many cases, these data sources are made available by other institutes. As a result, an NSI has no control over the quality, availability, and stability of these data sources, which increases risks in the production of official statistics that data are not available or incomparable over time. In the following sections, we will describe in detail how the web-scraped data was derived. In the data section, we will also provide information on the two registers.

Daas and van der Doef (2020) developed a machine-learning algorithm to predict the number of innovative businesses in the Netherlands. This approach is based on a classification algorithm trained on text scraped from companies' websites. To this end, the URLs of about 3,600 companies from the CIS sample could be annotated based on the information provided in the sample. These annotated web-scraped texts are used to train a logistic regression model to classify companies and determine whether they are innovative. The model utilizes a

bag-of-words approach to predict the propensity that a company is innovative. This algorithm is used to classify the companies in the sampling frame for which a website URL is available. The purpose of the ML algorithm was to provide additional information on small innovative companies which are not included in the target population of the CIS. It is therefore important to note that the model was trained on survey data from medium- and large-sized companies and then applied to small companies.

One way to utilize these propensities is to classify companies as either conducting innovative activities or not. In this way, the prediction of the number of innovative companies is entirely based on the classification of web-scraped text using a machine learning algorithm. This approach has several drawbacks. Training a machine learning algorithm on a non-probability data source, such as web-scraped data in this case, that is annotated with sample data from the CIS, and subsequently relying on this algorithm to predict how many companies have innovative activities using the web-scraped data only, is usually suboptimal. Because the methodology relies entirely on an algorithm and ignores the available survey data, there is an increased risk of underestimating the variation of the target variable in the target population (MAKSWELL 2020, Section 2.1). Furthermore, the uncertainty of the predictions increases if the algorithm is applied to web-scraped data that follows a different distribution of innovative businesses than the sample of annotated data. This may happen, for example, if the algorithm is applied to web-scraped data obtained in other time domains or unobserved strata. In general the population observed with the webscraped data will not cover the intended target population. Accordingly there is an undercoverage which limit the use of these data as a single primary data source. Generalizing results to larger intended target populations generally requires strong assumptions that are usually not met.

To avoid extrapolating an algorithm to data sources outside the time domain used to train the algorithm, it is necessary to collect survey data for future periods. Once it is recognized that additional editions of the survey, in this case the CIS, are required to retrain the machine learning algorithm for future reference periods, an alternative way to use these data sources can be considered. In this paper, it is proposed to utilize information available from new data sources with sample data, for example, by using this information to make the sampling design of a survey more efficient or by using it as auxiliary information in a design-based inference approach. One approach is to create a stratum of companies in the sampling frame for which URLs for companies are available. Companies within this stratum can be sampled proportional to the propensity of being innovative by means of probability proportional to size sampling (PPS). Alternatively, substrata can be created based on these propensities, and sample fractions within these strata can be adjusted to reflect the average propensity of conducting innovative activities. The classification of companies into propensities of being innovative in several subclasses can also be used as auxiliary information in a design-based estimation procedure. Suppose the propensities are indeed good predictors of the phenomena of interest. In that case, using this additional auxiliary information in the sampling design or the estimator will reduce both the variance and nonresponse bias of the sample estimates. This is an alternative approach to utilizing new data sources, which better aligns with the standard design-based inference procedures applied by NSIs to produce official statistics in which survey data remains the primary data source. As a result, the NSI still has control over the availability of these data and the minimum required precision of the output. At the same time, the sample estimates are still based on the widely accepted design-based or model-assisted inference methods from the sampling literature.

3 Data

In this section, the CIS data, the three auxiliary sources, and their preparation will be described first. Some data exploration will be presented in Section 3.5.

3.1 Community Innovation Survey

The Community Innovation Survey (CIS) focuses on product, process, organizational, and marketing innovations. Innovation involves making changes to improve the company and enhance its success (see the Appendix for the exact definitions of innovation). The survey is conducted biennially and relies on companies self-reporting whether they innovate in these areas. Large companies (200 or more employees) are included in the sample with a sampling fraction of 1, while a stratified sample is taken from companies with 10 or more employees. The sample is stratified by size class (SC) and standard industrial classification (SIC or NACE 2.0 – Nomenclature générale des Activités économiques dans les Communautés Européennes). The 2016 edition used has 435 different strata and 1,665 units with a sampling fraction of 1 (take-all-strata). NACE 01–03 and NACE from 83 onward (e.g., government, education, and healthcare) are excluded from the target population. In 2016, the population size of the entire target population was 51,805, the sample size was 9,068, and the response was 7,340.

The key target variables refer to whether a business has introduced an innovation, distinguishing many different types of innovation. In revisiting the weighting strategy, we focus on five target variables: whether a business has introduced an innovation (INNO), whether a business has introduced a product innovation (INPDT), whether a business has introduced a process innovation (INPCS), whether a business is collaborating with other businesses regarding an innovation (CO), and the expenses for innovation, abbreviated as RALLX. The variable RALLX is continuous, while the other four are binary indicators. The variable RALLX contained a small amount of missing data and was imputed using mean imputation conditioned on SC and SIC. The missing data of the variable CO was imputed using mode imputation conditioned on SC and SIC. There was no item-nonresponse for the other variables considered. Furthermore, the reporting period for the variable RALLX covers one year (2016), while the other variables have a reporting period covering the last three years (2014–2016).

3.2 Administrative subsidy data

Companies engaged in research and development (R&D) or conducting technical-scientific research can use subsidies. This data is referred to as WBSO (Dutch abbreviation for Research and Development Promotion Act) throughout this paper. The WBSO is a tax scheme that allows for the reduction of R&D costs. The WBSO data has been used since 2022 to stratify the R&D survey (Klingwort and Krieg 2023), which has been shown to improve the sampling strategy. Currently, it is also considered in the weighting strategy of the R&D survey (Klingwort and Krieg 2025) where the effectiveness is evaluated.

From the 2016 WBSO data, four continuous variables will be considered in the weighting. First, WBSO_U (‘VASTGESTELDE_SO_UREN’), which informs on the realized hours for performing R&D that qualifies for WBSO subsidies. Second, WBSO_NL (‘VASTGESTELDE_SO_NIET_LOONKOSTEN’), which informs about the calculated amount based on established R&D hours or realized costs

directly attributable to realized R&D projects. Third, WBSO_L ('VASTGESTELDE_SO_LOONKOSTEN') informs about the realized labor costs for performing research and development work. Fourth, WBSO_A ('VASTGESTELDE_SO_AFDRACHTVERMINDERING') informs on the amount by which a social security withholding agent may reduce the total payroll taxes to be remitted (Statistics Netherlands 2021). It is understood that the definitions of the different WBSO variables do no correspond exactly with the definitions of the target variables in the CIS. This, however, does not obstruct the use of WBSO data as an auxiliary variable in the weighting scheme of the CIS, since only a correlation between these variables is required.

Since a business unit (BE) may appear more than once per year in the WBSO database, the sums of the variables are calculated per business. The data is linked to the 2016 CIS using a unique business identifier (BEID). Since not all BEs are eligible for WBSO or have not applied for WBSO, the information cannot be linked to every business. There are several reasons for not having WBSO. First, the BE does not have to pay taxes, so these businesses are not eligible for WBSO (many Institutions most likely do not pay taxes); second, a BE has no R&D activities; third, a BE is entitled but does not use the WBSO subsidies. Thus, missing WBSO data is not due to a linkage error, but rather it is missing because the BE does not, or is not allowed to, use WBSO. All WBSO variables must be filled to classify a unit as a 'WBSO unit'. Thirty units had only one WBSO variable filled. Those were also classified as 'Non-WBSO units'. A potential drawback of this data is that it only registers the granted WBSO subsides but not the rejected requests.

3.3 Administrative patent data

This administrative data source provides information on the number of patents associated with a company group. The number includes patents registered in the Netherlands and in the European Union. However, the CIS sampling unit is at the individual business level and not at the level of group-related businesses. The individual businesses from the CIS were linked to the business-group database to identify which group covers which individual business and how many unique patents a group registered. Thus, different individual businesses can belong to the same group. A count variable has been created to indicate the number of patents per group, which is taken into account in the weighting. A potential drawback of this data is a potential time lag between the innovative activity and the registration of the related patent.

3.4 Web-scraped data

This data source provides information on the likelihood that a business is innovative based on the text on its website. In this context, innovation refers to all activities aimed at innovation within a company, as defined by Statistics Netherlands (2019). Innovation involves making changes to improve the company and enhance its success. The methodology developed on how the data was scraped on the web, linked to the CIS observations, and how the propensities were derived from the text on the website is described in detail by Daas and van der Doef (2020), Daas and Wolf (2021), and Daas and van der Doef (2021) and is only described briefly here. In the first step, known URLs were linked with the corresponding business in the register. In the second step, the website texts were scraped. In the third step, these text files were represented as a frequency-annotated bag-of-words, and a document-term matrix was created. Fourth, the term frequency-inverse document frequency (tf-idf) was used to identify words that characterize the topics in a text. The fifth step consisted of the classification task, i.e., classifying whether a business is innovative based on its website texts. These classifications are based on a logistic regression. Three continuous variables will be considered in the weighting. First, the propensity

of a business being innovative is considered (Scraped_INNO_LOG). However, some businesses had multiple websites; accordingly, multiple propensities were available. Further, some of the multiple websites were in Dutch, and others in English. It was found that if a website was in English, the average propensity of being innovative was larger. The following experimental strategies were applied to solve this problem and have only one propensity per business: if there is little variation in the propensities, i.e., the standard deviation is less than 0.1, then the mean of the propensities is taken. Otherwise, the maximum of the 10% of the largest propensities is taken. This boils down to taking the maximum of all propensities in most cases.

In the course of this project, it was discussed that the model used to derive the propensities may not properly deal with the complexity of the data. The large number of explanatory variables in the bag-words approach, combined with the small sample size available to fit a logistic regression model, easily leads to high dimensionality problems and models that overfit the data. Therefore, it was decided to classify the scraped texts again using a large language model (LLM). This analysis is mainly experimental to compare the above-described propensities and evaluate both propensities' performance in the weighting model. DeepSeek was used for this task (Guo et al. 2024). The prompt contained the scraped URL text, the official Eurostat definition of what defines an innovative company, and the task of deriving a classification and propensities from the text. In particular, the prompt asked for general, product, and process innovation. The prompt used is shown in Appendix A. Three variables were derived: the propensities of the company being innovative (Scraped_INNO_LLM), the propensities of the company doing product innovation (Scraped_PRODINNO_LLM), and the propensities of the company doing process innovation (Scraped_PROCINNO_LLM).

3.5 Data exploration

CIS target variables

This section provides information on the distribution of the target variables in the CIS response. Table 3.1 shows the distribution of the four binary target variables (introduced an innovation (INNO), introduced a product innovation (INPDT), introduced a process innovation (INPCS), collaborating with other businesses regarding an innovation (CO)). Most units introduced an innovation, with more product than process innovations. About a fifth of the businesses collaborated with other businesses in the context of innovation.

Table 3.1 Distribution of the binary target variables in the CIS response.

Variable	No	Yes
INNO	0.35	0.65
INPDT	0.60	0.40
INPCS	0.66	0.34
CO	0.79	0.21

Table 3.2 shows the distribution of the continuous target variable (expenses for innovation (RALLX)). The distribution is right-skewed, as can be seen especially by the Q_{25} and Q_{50} statistics. Most businesses have low expenses, while some have large expenses.

Table 3.2 Distribution of the continuous target variable RALLX (in mln. euro) in the CIS response.

Min.	Q_{25}	Q_{50}	Mean	Q_{75}	Max.
0	0	73	2,505	669	1,195,886

Administrative subsidy data

Table 3.3 informs on the distribution of the WBSO indicator in the population, sample, and response. The sample design oversamples WBSO units and as a result more WBSO units are selected relative to their presence in the population. The response process does not change the distribution over WBSO and non-WBSO units.

Table 3.3 Distribution of WBSO indicator in the population, sample, and response.

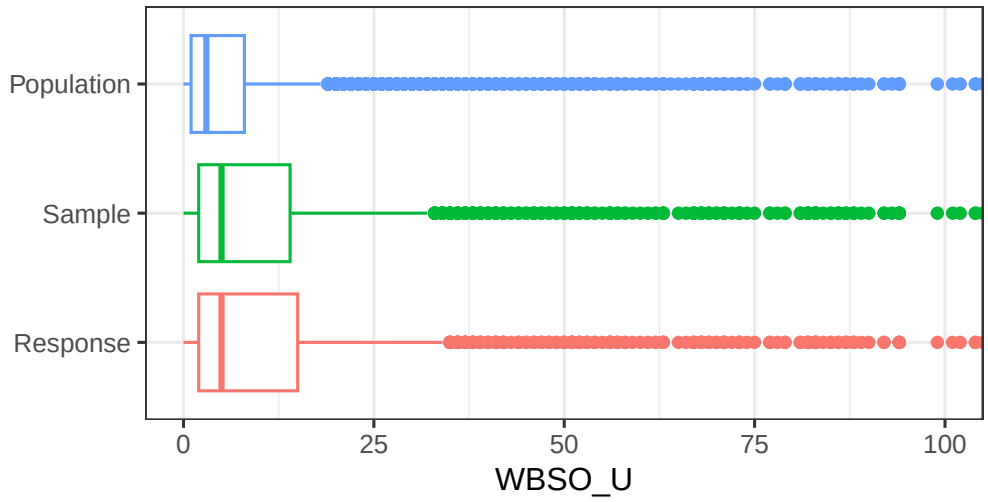
Source	Non-wbso unit	Wbso unit
Population	44,820 (0.87)	6,985 (0.13)
Sample	6,511 (0.72)	2,557 (0.28)
Response	5,178 (0.71)	2,162 (0.29)

Table 3.4 shows that for units with WBSO information, across all WBSO-related variables, the sample and response distributions consistently show higher means and medians than the population. Units with higher WBSO values are more likely to be included in the sample and response stages. Zero-valued quartiles in multiple variables suggest that many businesses do not have WBSO information. This may be evidence that the WBSO variable might not be a potential auxiliary variable to improve the weighting scheme. The distribution of the variable WBSO_U is shown in Figure 3.1. All other WBSO variables would result in a comparable plot. Only the scaling of the x-axis would differ.

Table 3.4 Distribution of the WBSO data (in thousand euro) in the population, sample, and response. The values have been scaled.

Source	Variable	Min.	Q_{25}	Q_{50}	Mean	Q_{75}	Max.
Population	WBSO_U	0.00	1.00	3.00	11.53	8.00	3871.00
Sample	WBSO_U	0.00	2.00	5.00	22.29	14.00	3871.00
Response	WBSO_U	0.00	2.00	5.00	24.08	15.00	3871.00
Population	WBSO_L	0.00	0.00	0.00	0.226	0.00	174.00
Sample	WBSO_L	0.00	0.00	0.00	0.579	0.00	174.00
Response	WBSO_L	0.00	0.00	0.00	0.638	0.00	174.00
Population	WBSO_NL	0.00	0.00	0.00	0.193	0.00	260.00
Population	WBSO_NL	0.00	0.00	0.00	0.512	0.00	260.00
Response	WBSO_NL	0.00	0.00	0.00	0.566	0.00	260.00
Population	WBSO_A	0.00	0.00	0.00	0.053	0.00	69.00
Sample	WBSO_A	0.00	0.00	0.00	0.142	0.00	69.00
Response	WBSO_A	0.00	0.00	0.00	0.156	0.00	69.00

Figure 3.1 Boxplots for WBSO_U split by population, sample, and response. For readability, values larger than 100 are not shown.



Administrative patent data

Table 3.5 shows that the population distribution differs from the sample and response distributions. The sample and response distributions are alike. The increase in patent units is most likely the result of oversampling WBSO units.

Table 3.5 Distribution of patent indicator in the population, sample, and response.

Source	Non-patent unit	Patent unit
Population	50,540 (0.98)	1,265 (0.02)
Sample	8,169 (0.90)	899 (0.10)
Response	6,549 (0.89)	791 (0.11)

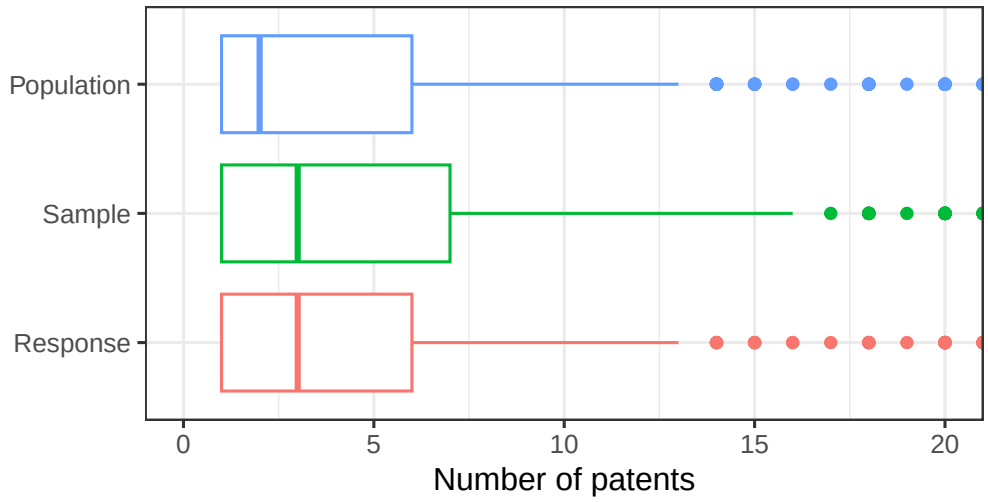
Table 3.6 shows that the sampling process selects units with higher patent values, as indicated by the increase in mean and median (for the sub-population with patents). The response stage slightly lowers this effect, but the distribution remains skewed toward higher patent values than the population.

Table 3.6 Distribution of the patent data in the population, sample, and response.

Source	Min.	Q_{25}	Q_{50}	Mean	Q_{75}	Max.
Population	1.00	1.00	2.00	16.35	6.00	1099.00
Sample	1.00	1.00	3.00	21.67	7.00	1099.00
Response	1.00	1.00	3.00	19.07	6.00	1099.00

The distribution of the patent variable is shown in Figure 3.2.

Figure 3.2 Boxplots for the number of patents split by population, sample, and response. For readability, values larger than 20 are not shown.



Web scraped data

Table 3.7 shows that while web-scraped units were less prevalent in the population, the sampling and response processes resulted in a more balanced distribution of scraped and non-scraped units. This is likely due to the fact that large companies are overrepresented in the sample and companies with less then 10 employees are not included. Further, large companies usually have websites while smaller companies don not. Scraped unit means that a business had at least one website. A non-scraped unit had no website and accordingly no derived propensity.

Table 3.7 Distribution of web-scrape indicator in the population, sample, and response.

Source	Non-scraped unit	Scraped-unit
Population	30,414 (0.59)	21,391 (0.41)
Sample	4,725 (0.52)	4,343 (0.48)
Response	3,716 (0.51)	3,624 (0.49)

Table 3.8 shows the population, sample, and response distributions for all web-scraped-based propensities. For 'INNO LOG', the distribution is relatively consistent across population, sample, and response data. The median increases slightly from population to sample and response. The mean follows a similar pattern. The upper quartiles and maximum values are identical across sample and response, suggesting a stable distribution after sampling. For 'INNO LLM', the distribution is tightly clustered around a median of 0.15 for all three groups. The mean is slightly higher in the sample and response compared to the population. The upper quartile is noticeably higher in the sample and response than in the population. Comparable patterns are found for 'INNO PROD LLM' and 'INNO PROC LLM'. 'INNO LOG' values are generally higher, with larger median, mean, and upper quartiles. The three LLM-based variables exhibit similar distributions, with nearly identical minima, medians, and maxima. The main differences appear in the means and upper quartiles, which tend to be slightly higher in the sample and response data. 'INNO LOG' has higher central values, indicating that propensities in this category are more widespread or intense than the LLM variants.

Table 3.8 Distribution of the web scraped data in the population, sample, and response.

Source	Variable	Min.	Q_{25}	Q_{50}	Mean	Q_{75}	Max.
Population	INNO LOG	0.000	0.051	0.380	0.470	0.940	1.000
Sample	INNO LOG	0.000	0.065	0.496	0.508	0.964	1.000
Response	INNO LOG	0.000	0.067	0.503	0.508	0.964	1.000
Population	INNO LLM	0.010	0.150	0.150	0.315	0.250	0.950
Sample	INNO LLM	0.010	0.150	0.150	0.342	0.750	0.950
Response	INNO LLM	0.010	0.150	0.150	0.342	0.750	0.950
Population	INNO PROD LLM	0.010	0.150	0.150	0.301	0.250	0.950
Sample	INNO PROD LLM	0.010	0.150	0.150	0.319	0.625	0.950
Response	INNO PROD LLM	0.010	0.150	0.150	0.318	0.600	0.950
Population	INNO PROC LLM	0.010	0.150	0.150	0.297	0.300	0.950
Sample	INNO PROC LLM	0.010	0.150	0.150	0.315	0.600	0.950
Response	INNO PROC LLM	0.010	0.150	0.150	0.316	0.600	0.950

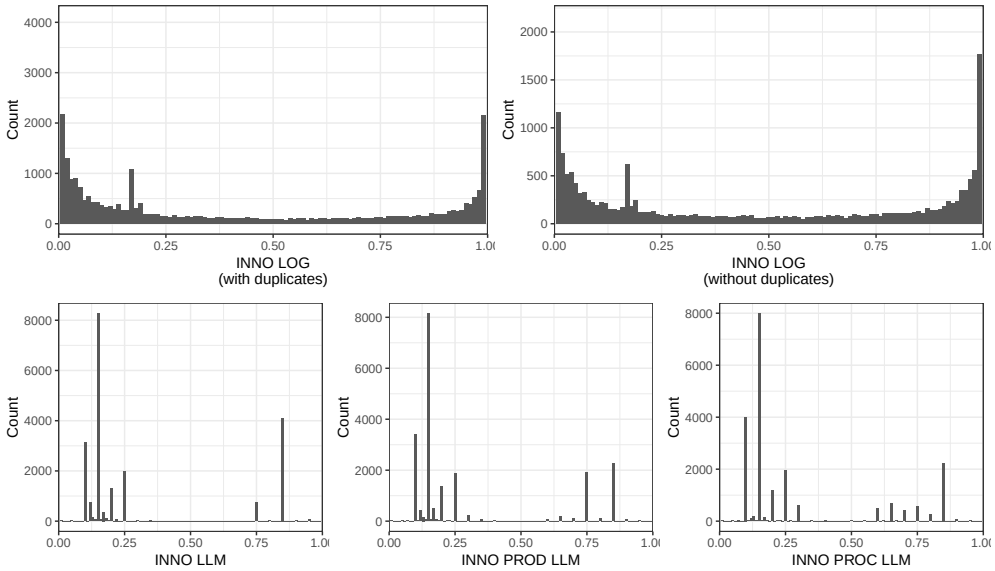
Table 3.9 shows the distribution by population, sample, response, propensity type, and language. English websites show higher innovation propensities for ‘INNO LOG.’ The LLM-based propensities do not show differences between language. The sample and response datasets consistently contain higher propensities, particularly in the upper quartiles. ‘INNO LOG’ captures a stronger, more dispersed signal than the LLM-based variables, which cluster tightly around 0.15.

Table 3.9 Distribution of the web-scraped-based propensities in the population, sample, and response split by language of website.

Source	Variable	Language	Min.	Q_{25}	Q_{50}	Mean	Q_{75}	Max.
Population	INNO LOG	NL	0.000	0.034	0.304	0.433	0.892	1.000
Population	INNO LOG	EN	0.000	0.171	0.793	0.598	0.987	1.000
Sample	INNO LOG	NL	0.000	0.044	0.421	0.474	0.939	1.000
Sample	INNO LOG	EN	0.000	0.171	0.868	0.630	0.987	1.000
Response	INNO LOG	NL	0.000	0.047	0.428	0.476	0.940	1.000
Response	INNO LOG	EN	0.000	0.171	0.833	0.625	0.987	1.000
Population	INNO LLM	NL	0.050	0.150	0.150	0.319	0.250	0.950
Population	INNO LLM	EN	0.010	0.100	0.150	0.302	0.250	0.950
Sample	INNO LLM	NL	0.100	0.150	0.150	0.348	0.750	0.950
Sample	INNO LLM	EN	0.010	0.100	0.150	0.317	0.750	0.950
Response	INNO LLM	NL	0.100	0.150	0.150	0.349	0.750	0.950
Response	INNO LLM	EN	0.010	0.100	0.150	0.317	0.750	0.950
Population	INNO PROD LLM	NL	0.050	0.150	0.150	0.302	0.250	0.950
Population	INNO PROD LLM	EN	0.010	0.100	0.150	0.296	0.250	0.950
Sample	INNO PROD LLM	NL	0.075	0.150	0.150	0.323	0.638	0.950
Sample	INNO PROD LLM	EN	0.010	0.100	0.150	0.308	0.400	0.950
Response	INNO PROD LLM	NL	0.075	0.150	0.150	0.321	0.600	0.950
Response	INNO PROD LLM	EN	0.010	0.100	0.150	0.307	0.363	0.950
Population	INNO PROC LLM	NL	0.050	0.150	0.150	0.298	0.300	0.950
Population	INNO PROC LLM	EN	0.010	0.100	0.150	0.291	0.250	0.950
Sample	INNO PROC LLM	NL	0.070	0.150	0.150	0.319	0.600	0.950
Sample	INNO PROC LLM	EN	0.010	0.100	0.150	0.302	0.550	0.950
Response	INNO PROC LLM	NL	0.070	0.150	0.150	0.320	0.600	0.950
Response	INNO PROC LLM	EN	0.010	0.100	0.150	0.303	0.600	0.950

Figure 3.3 shows the different propensity distributions. The first panel shows the propensities based on the bag of words, tf-idf, and the logistic regression. The second panel shows the same, but after the procedure to deal with duplicates (see Section 3.4). The third panel shows the propensities based on the LLM informing about general innovation. The fourth and fifth panels show the LLM-based propensities informing about product and process innovation. All panels show a U-curve, while the LLM-based propensities are more distributed towards 0 and 1.

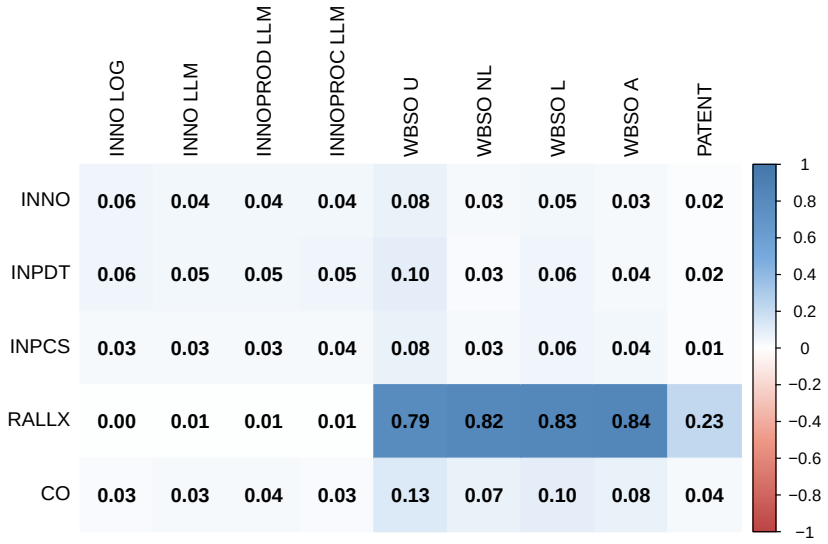
Figure 3.3 Distribution of the web-scraped-based propensities.



Correlations

Figure 3.4 shows the correlation coefficients (point biserial correlation) between the survey variables and auxiliary variables. The propensities of being innovative based on web-scraped data do not correlate with the survey variables. The WBSO and patent variables correlate highly with the survey variable RALLX, but do not correlate strongly with the other survey variables. The patent data correlates slightly with the survey variable RALLX. These findings indicate that the variable RALLX has the largest potential in combination with the GREG to improve the weighting scheme.

Figure 3.4 Correlation matrix of CIS target variables and auxiliary sources (web-scraped, WBSO, patent).



4 Methodology

4.1 Horvitz–Thompson estimator

The Horvitz-Thompson (HT) estimator is a general estimator that is completely based on the sampling design. The observations in the sample are weighted with the inverse of the inclusion probability of the respondents in the sample. Let y denote the variable of interest of the elements in the population U of size N. The parameters of interest are defined as population totals

$$Y = \sum_{i \in U} y_i. \tag{1}$$

For stratified sampling, the HT-estimator for a population total (1), (Horvitz and Thompson 1952) is defined as

$$\hat{Y}_{\mathsf{HT}} = \sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_i}{\pi_i}, \tag{2}$$

where y_i is the value of the variable of interest for unit i from stratum h for strata ($h = 1, ..., H$), S_h the sample for stratum h, and π_i the inclusion probability of unit i. An estimator for the variance of (2) is given by

$$\hat{\mathsf{Var}}(\hat{Y}_{\mathsf{HT}}) = \sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_h} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j}, \tag{3}$$

where π_{ij} is the joint inclusion probability of units i and j in stratum h. It is understood that for $j = i$ the second-order inclusion probabilities become $\pi_{ii} = \pi_i$. The standard error of the HT-estimator is obtained as

$$\hat{\mathsf{SE}}(\hat{Y}_{\mathsf{HT}}) = \sqrt{\mathsf{Var}(\hat{Y}_{\mathsf{HT}})}. \tag{4}$$

The coefficient of variation or relative standard error is

$$\hat{\mathsf{RSE}}(\hat{Y}_{\mathsf{HT}}) = \frac{\hat{\mathsf{SE}}(\hat{Y}_{\mathsf{HT}})}{\hat{Y}_{\mathsf{HT}}}. \tag{5}$$

4.2 Generalized regression estimator

The generalized regression (GREG) estimator attempts to improve upon the precision of the HT-estimator by incorporating auxiliary information in the estimator for which the distributions in the population are known from additional sources like registers. The GREG estimator is derived from a linear regression model that defines the relationship between the target variable and auxiliary variables:

$$y_i = \beta^\top \mathbf{x}_i + \epsilon_i \tag{6}$$

with $\mathbf{x}_i$ a q dimensional vector containing auxiliary variables for respondent i, β a q dimensional vector with regression coefficients and ϵ_i independently distributed residuals for which it holds that $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2 w_i$. It is assumed that for the variance of the residuals, the values for w_i are known up to a scaling factor σ^2. The generalized regression estimator for population totals is

$$\hat{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (\mathbf{X} - \hat{\mathbf{X}}_{\text{HT}})^\top \hat{\beta}, \tag{7}$$

with $\mathbf{X}$ a q vector with the known population totals of the auxiliary variables, $\hat{\mathbf{X}}_{\text{HT}}$ the HT-estimator for $\mathbf{X}$, similarly defined as $\hat{Y}_{\text{HT}}$ in (2) and $\hat{\beta}$ the HT-estimator for the regression coefficients:

$$\hat{\beta} = \left(\sum_{i \in S} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i w_i} \right)^{-1} \frac{\mathbf{x}_i y_i}{\pi_i w_i} \tag{8}$$

The weighting models to estimate the target variable $\hat{y}_i$ for respondent i, based on the vector of estimated regression coefficients, are specified in Section 4.3. A special case for stratified sampling is to define the weighting model strata-specifically. In this case, it follows that

$$\hat{Y}_{\text{GREG}} = \sum_{h=1}^{H} \left[\hat{Y}_{\text{HT}|h} + (\mathbf{X}_h - \hat{\mathbf{X}}_{\text{HT}|h})^\top \hat{\beta}_h \right], \tag{9}$$

with

$$\hat{\beta}_h = \left(\sum_{i \in S_h} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i w_i} \right)^{-1} \frac{\mathbf{x}_i y_i}{\pi_i w_i}, h = 1, \ldots, H. \tag{10}$$

Furthermore, $\hat{Y}_{\text{HT}|h}$ and $\hat{\mathbf{X}}_{\text{HT}|h}$ are the HT estimators for the stratum total of the target and auxiliary variables similarly defined as $\hat{Y}_{\text{HT}}$ in Equation (2). An estimator for the variance of $\hat{Y}_{\text{GREG}}$ under stratified sampling is given by

$$\hat{\text{Var}}(\hat{Y}_{\text{GREG}}) = \sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_h} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\hat{\epsilon}_i \hat{\epsilon}_j}{\pi_i \pi_j}, \tag{11}$$

with $\hat{\epsilon}_i = y_i - \hat{\beta}^\top \mathbf{x}_i$. The standard error of $\hat{Y}_{\text{GREG}}$ is defined as

$$\hat{\text{SE}}(\hat{Y}_{\text{GREG}}) = \sqrt{\text{Var}(\hat{Y}_{\text{GREG}})}. \tag{12}$$

and the coefficient of variation as

$$\hat{\text{RSE}}(\hat{Y}_{\text{GREG}}) = \frac{\hat{\text{SE}}(\hat{Y}_{\text{GREG}})}{\hat{Y}_{\text{GREG}}}. \tag{13}$$

See also Särndal et al. 1992, Ch. 6 for more details of the GREG estimator and its variance approximation.

4.3 Weighting models

The HT-estimator described in equation (2) is a design-based estimator and does not require the specification of a weighting model. The HT is solely based on the response y_i and the inclusion probability π_i. However, the GREG is a model-assisted estimator, and the weighting models must be explicitly specified. These will be described in the following.

In total, nine weighting models were specified next to the HT. Each model is used to estimate the five target variables. Four weighting models used the WBSO data (Section 3.2), four used the web-scraped data (Section 3.4), and one used the patent data (Section 3.3). All weighting models are specified in the same way: the stratification variable (x_1) of the CIS is used as a main effect to ensure that the population totals are calibrated per stratum. An auxiliary variable from either the WBSO, web-scraped, or patent data is used as a second main effect (x_2). As a third main effect (x_3), an indicator variable indicates whether the auxiliary information is available for unit i. A specific indicator variable was created for each of the three auxiliary sources (see Section 3). Finally, one interaction term is used: the interaction between the aforementioned indicator and the auxiliary variable. This interaction ensures that the GREG is applied for all units where domain-specific information is available, and the HT is applied for units without domain-specific information. The regression model used in the GREG estimator can be written as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_{23}(x_{2i} x_{3i}) + \epsilon_i \tag{14}$$

where β_0 is the intercept term, β_1, β_2, and β_3 are the coefficients for the main effects x_1, x_2, and x_3, β_{23} is the coefficient for the interaction between x_2 and x_3, and x_{1i}, x_{2i}, and x_{3i} are the values of x_1, x_2, x_3 for unit i. For the stratified case, the regression coefficients are taken strata-specific and are defined as

$$y_i = \beta_0^{(h)} + \beta_1^{(h)} x_{1i} + \beta_2^{(h)} x_{2i} + \beta_3^{(h)} x_{3i} + \beta_{23}^{(h)}(x_{2i} x_{3i}) + \epsilon_i, h = 1, \dots, H. \tag{15}$$

5 Results

Table 5.1 shows the distribution of the design weights for the HT and the regression weights for the GREG. The distribution of the weights is not strongly affected by the different weighting models under the GREG.

Table 5.1 Distribution of the design weights for HT and regression weights for GREG.

Model	Min.	Q_{25}	Q_{50}	Mean	Q_{75}	Max.
HT	1.138	1.138	2.091	7.058	7.107	278.091
INNO LOG	1.066	1.187	2.121	7.058	7.455	286.918
INNO LLM	1.052	1.183	2.127	7.058	7.465	288.321
INNO PROD LLM	1.034	1.183	2.178	7.058	7.464	288.971
INNO PROC LLM	1.069	1.185	2.126	7.058	7.460	287.586
WBSO U	0.228	1.089	2.072	7.058	7.024	278.091
WBSO NL	0.778	1.088	2.069	7.058	7.028	278.091
WBSO L	0.642	1.088	2.069	7.058	7.030	278.091
WBSO A	0.826	1.088	2.069	7.058	7.028	278.091
PATENT	1.095	1.147	2.091	7.058	7.112	278.091

In Figure 5.1 the relative standard error (RSE) is plotted on the y-axis for ten different weighting models on the x-axis. Results are presented in separate panels for five different variables. Generally, the differences between the RSEs are minor, with 'Expenses for innovation' being the exception. The RSEs obtained by the WBSO models are consistently the smallest. The web-based propensities produce the largest RSE most often. Also, they are usually larger than the RSE of the HT, i.e., using no auxiliary information in the weighting model produces a smaller error. The model using the patent data often has a comparable RSE as the HT, i.e., using no auxiliary information in the weighting model produces a similar error. The point and interval estimates (y-axis) per survey target variable (panel) and weighting model (x-axis) are shown in Figure 5.2. Generally, the point and interval estimates are comparable between the weighting models. 'Expenses for innovation' is the only exception.

Figure 5.1 Relative standard error by survey target variable and weighting model. The color indicates the source of the auxiliary information used for the weighting model.

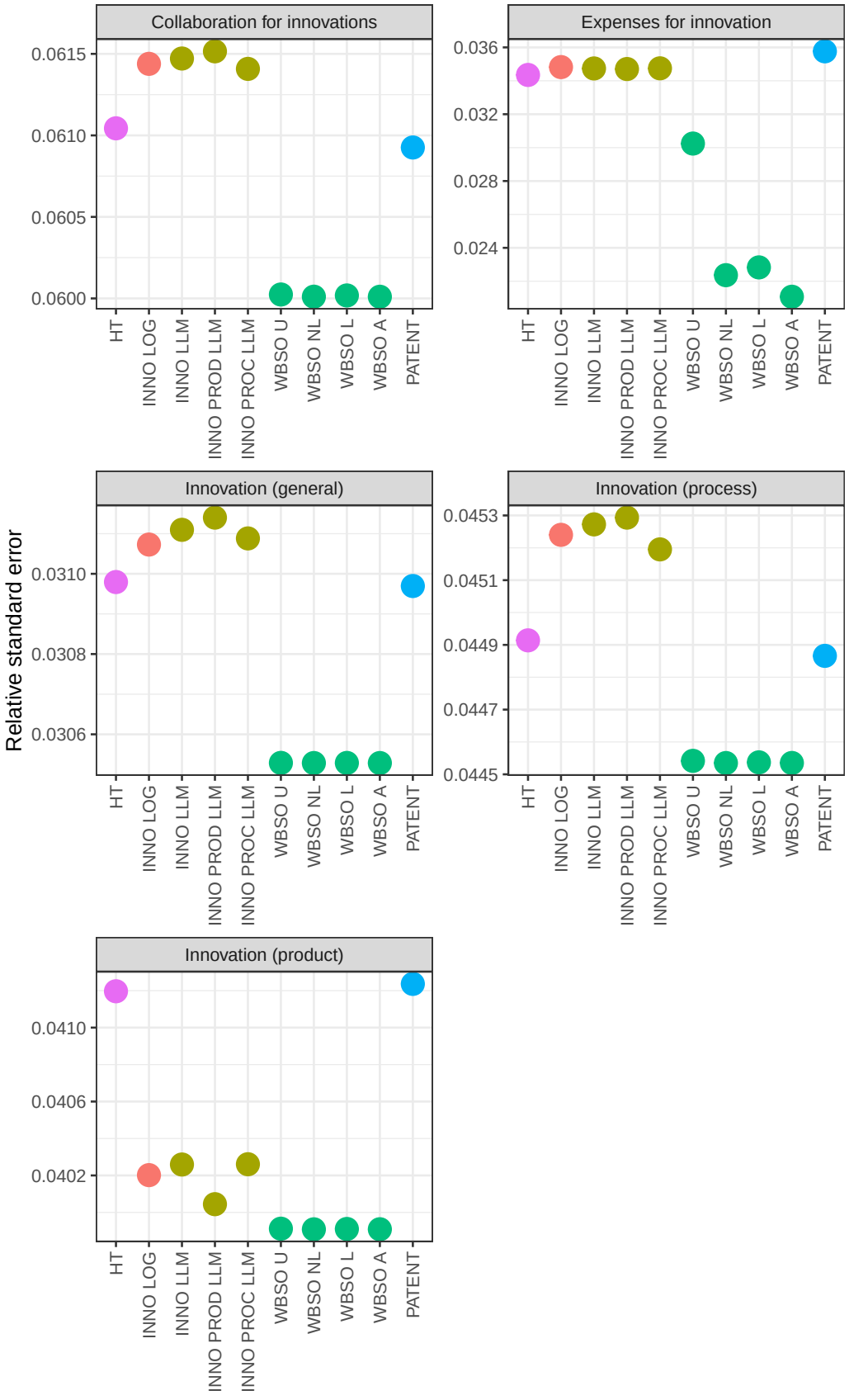
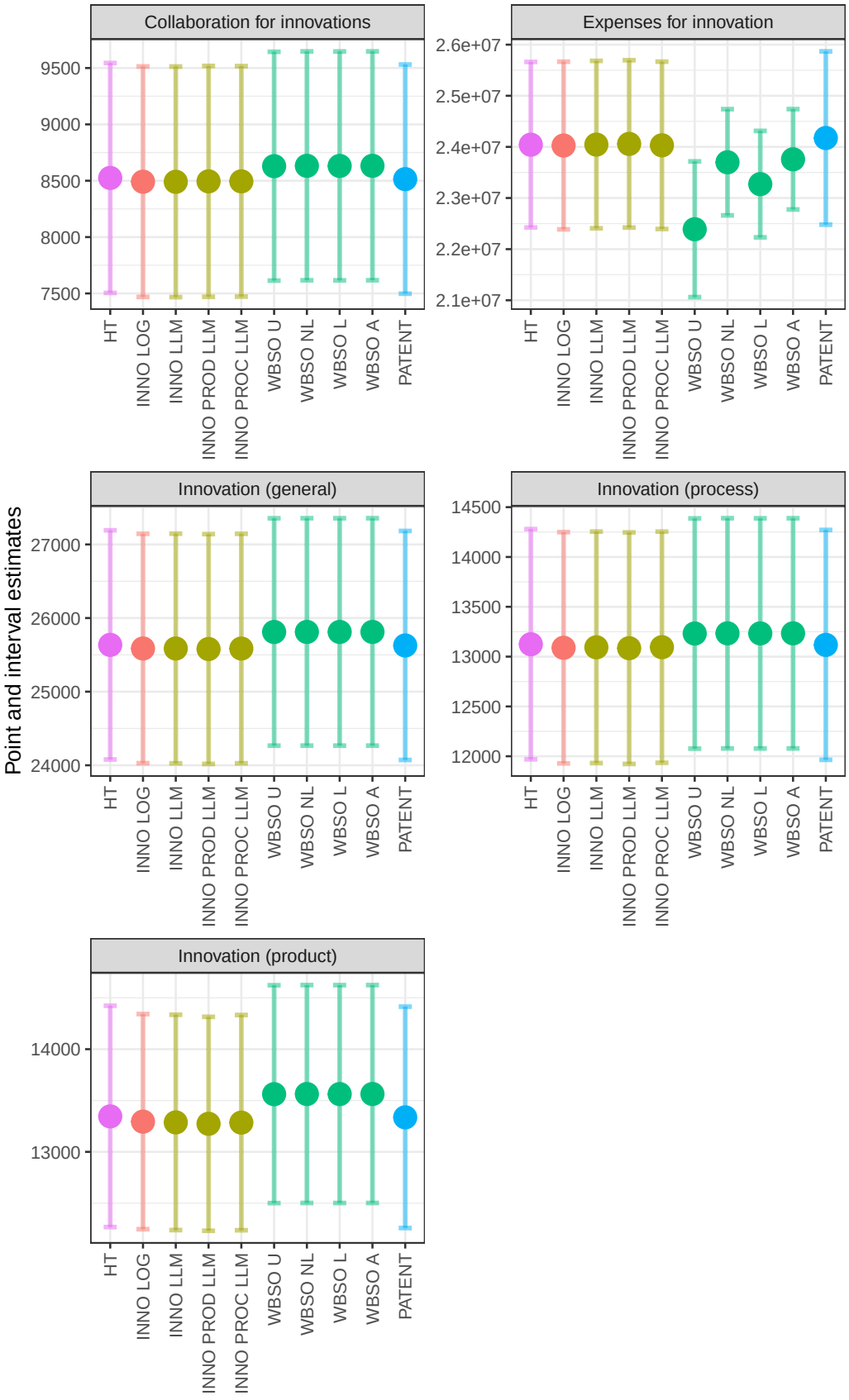


Figure 5.2 Point and interval estimates by survey target variable and weighting model. The color indicates the source of the auxiliary information used for the weighting model.



These results were also obtained on SIC-level. The highest level (letters A, B, C, D, E, F) was used for SIC. Six strata, five target variables, and ten weighting models result in 300 RSEs. These 300 RSEs are not shown here. Instead, Table 5.2 shows how often which weighting model achieved the smallest RSE. About 77% of the lowest RSEs were obtained by weighting models using the WBSO information. In about 23% of these cases, web-scraped-based propensities achieved the lowest RSE. The patent data only once had the lowest RSE. However, the HT never had the lowest RSE, i.e., on the SIC level, the weighting models using auxiliary information resulted in smaller RSEs.

Table 5.2 Count of weighting models with the lowest RSE on SIC niveau.

Model	Count
WBSO A	9
WBSO NL	7
WBSO U	3
WBSO L	3
INNO LOG	2
INNO LLM	2
INNO PROC LLM	2
INNO PROD LLM	1
PATENT	1
HT	0

6 Discussion

This study explored the potential of improving the sampling strategy for the Community Innovation Survey by evaluating three sources of auxiliary information: (1) web-scraped data on the likelihood of a business being innovative, (2) administrative records of businesses receiving research and development subsidies (WBSO), and (3) administrative data on the number of patents associated with each business. The aim was to assess whether these sources can contribute to a more targeted and efficient sampling process or inference procedure.

The results indicate that no auxiliary source consistently improved the sampling strategy across five studied target variables. The WBSO data is the most promising auxiliary variable, yielding measurable improvements for one of the five target variables (expenses for innovation). Using the patent data resulted not in any improvements. The innovation propensities – whether derived by a logistic model (Daas and van der Doef 2021) or from a large language model (Deepseek) – did not show a relation with the survey's innovation target variables, and thus, also could not improve the estimation procedure.

These results are, of course, disappointing. However, some limitations may have caused some of the findings, and show that more research is required.

With the WBSO information the best improvements were achieved. However, this data currently provides only information on whether or not businesses received subsidies. It does not provide information about businesses that applied for WBSO subsidies but were rejected since not every business with innovative activities according to the definitions of the CIS does qualify for the WBSO. Moreover, businesses qualifying for WBSO subsidies may not necessarily apply for it. As a results, businesses may be active in innovative fields but were dismissed for the WBSO subsidy and assigned to the group of units without WBSO in this study. Thus, the binary indicator does

not correctly reflect the complexity behind the WBSO subsidies. However, the information required to use an indicator with three groups (WBSO, no WBSO, rejected for WBSO) was unavailable in this study. Moreover, there may be mismatches in the data-linkage because WBSO from 2016 was used, but four target variables cover reporting periods from 2014–2016. Adding WBSO data from 2014–2015 may improve the correlation between the target variables and WBSO.

Second, the patent data is not available on the business unit level. This may cause errors in the linkage process, i.e., a business unit that is not innovative gets associated with several patents (which is assumed to be associated with innovation) because it belongs to a company group in which other businesses received patents. Such linkage errors will likely negatively impact the association between being innovative and the number of patents. Furthermore, the information in this data is time-delayed. The time a patent is applied for is well before the time at which this information is available in the database. For example, innovation can occur in the year t, and a patent is requested. However, the information is only available in the database at the time the patent is granted, say $t + x$. If no more innovation takes place in the years $t + x$, this has a negative effect on the association survey response and the auxiliary information.

Third, the large language model (LLM) likely failed to perform effectively because the text data had already been transformed using TF-IDF and bag-of-words representations before being fed into the model. These preprocessing techniques reduce text to numerical vectors based solely on word frequency, disregarding grammar, word order, and contextual meaning. While such methods can be useful for specific statistical models, they remove the semantic richness that LLMs rely on to understand relationships, intent, and nuance in language. As a result, the LLM was given a set of disconnected numerical features rather than coherent sentences, preventing it from applying its deep contextual understanding to the task.

Beyond these issues, there may also be a conceptual misalignment between the auxiliary sources and the survey's operational definition of innovation (see the Appendix for the exact definitions of innovation). For example, patent data, subsidies, and web-scraped indicators may capture primarily technological or formalized innovation, while the survey includes a broader range of innovative activities, such as process, organizational, or marketing innovations. This mismatch could reduce the strength of observed associations.

Nonetheless, the potential of web-derived and other non-traditional data sources remains important for future research. Further work could focus on refining how such content is collected and processed, as well as developing more effective methods for linking it to official business registers. For statistical agencies, these findings underscore that while auxiliary sources hold promise, they must be carefully validated and contextually matched to survey definitions before being integrated into operational sampling frameworks.

7 Conclusion

National statistical institutes (NSIs) are consistently interested in improving the quality of their statistical output. Traditionally, this is often achieved by including auxiliary information from registers or other administrative data sources. However, there is an increasing interest among NSIs in incorporating non-probability data sources, such as web content, sensor data, and social media, into the production of official statistics. While these sources can offer timely and

potentially rich information, their direct use has substantial risks, including selection bias, uncertain precision, limited comparability over time, no control over the availability of these data sources, and incomplete alignment with established statistical concepts. This study explored an alternative approach: using such data as auxiliary information to improve the sampling design or inference procedure of a probability survey. By doing so, these sources can be integrated within a design-based inference framework, which remains the foundation of most official statistical production, thereby mitigating many of the risks associated with their direct use. Although the present application did not yield consistent improvements, the approach remains promising. With continued methodological refinement, better alignment between auxiliary data and survey concepts, and advances in data linkage and processing, this strategy offers an option for NSIs to harness the value of emerging data sources while keeping the quality and reliability of official statistics.

Acknowledgments

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

Piet Daas was one of the authors of this paper, but unfortunately passed away during the writing of this paper. We do not put him as an author on the paper since we do not know if his views are similar to those of the other authors. We would like to acknowledge his important contribution to this paper and the line of work.

We would like to thank Linda Bruls, Rogier Goedhart, Yvonne Gootzen, Sabine Krieg, and Vera Toepoel for their effort to review this paper.

References

Beaumont, J.-F. (2020). “Are probability surveys bound to disappear for the production of official statistics?” In: *Survey Methodology* 46, pp. 1–28.

Daas, P., B. De-Miguel-Molina, and M. De-Miguel-Molina (2024a). “Identifying Drone Web Sites in Multiple Countries and Languages with a Single Model.” In: *Journal of Data Science*, pp. 1–14. DOI: 10.6339/23-JDS1087.

Daas, P., W. Hassink, and B. Klijs (2024b). “On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms.” In: *Journal of Official Statistics* 40.1, pp. 190–211. DOI: 10.1177/0282423X241235265.

Daas, P. and N. de Wolf (2021). “Identifying different types of companies via their website text.” In: *Research portal Eindhoven University of Technology*. URL: https://research.tue.nl/en/publications/identifying-different-types-of-companies-via-their-website-text.

Daas, P. and S. van der Doef (2020). “Detecting Innovative Companies via their Website.” In: *Statistical Journal of IAOS* 36.4, pp. 1239–1251. DOI: doi/10.3233/SJI-200627.

Daas, P. and S. van der Doef (2021). “Using Website texts to detect innovative companies.” In: *Center for Big Data Statistics (CBDS) Workingpaper* 01.21.

Guo, D. et al. (2024). *DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence*. arXiv: 2401.14196 [cs.SE]. URL: https://arxiv.org/abs/2401.14196.

Hackl, P. (2016). “Big Data: What can official statistics expect?” In: *Statistical Journal of the IAOS* 32.1, pp. 43–52. DOI: 10.3233/SJI-160965.

Horvitz, D. and D. Thompson (1952). “A generalization of sampling without replacement from a finite universe.” In: *Journal of the American Statistical Association* 47, pp. 663–685.

Klingwort, J. and S. Krieg (2023). *Steekproefontwerp Statistiek Research & Development*. Internal CBS report.

Klingwort, J. and S. Krieg (2025). *Statistiek Research & Development: Weging en Schatting*. Internal CBS report.

Kühnemann, H. (2021). “Anwendungen des Web Scraping in der amtlichen Statistik.” In: *AStA Wirtschafts- und Sozialstatistisches Archiv* 15.1, pp. 5–25. DOI: 10.1007/s11943-021-00280-5.

MAKSWELL (2020). *Methodological aspects of measuring SDG indicators with traditional and non-traditional data sources. Deliverable 2.2 Methodological aspects of using big data*. Tech. rep., pp. 1–66.

Oancea, B. and M. Necula (2019). “Web scraping techniques for price statistics – the Romanian experience.” In: *Statistical Journal of the IAOS* 35.4, pp. 657–667. DOI: 10.3233/SJI-190529.

Rao, J. (2021). “On Making Valid Inferences by Integrating Data from Surveys and Other Sources.” In: *Sankya B* 83, pp. 242–272.

Sakshaug, J. W., D. A. P. R. A. Wisniowski, and A. G. Blom (2019). “Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach.” In: *Journal of Official Statistics* 35, pp. 653–681.

Statistics Netherlands (2019). *Definition of the concept Innovation. Sept. 20, 2019*. URL: https://www.cbs.nl/en-gb/our-services/methods/definitions?tab=i#id=innovation.

Statistics Netherlands (2021). *Documentatierapport WBSO 2016–2020. Internal CBS report*.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Uriarte, J. I., G. R. Ramírez Muñoz de Toro, and J. M. C. Larrosa (2019). “Web scraping based online consumer price index: The “IPC Online” case.” In: *Journal of Economic and Social Measurement* 44.2-3, pp. 141–159. DOI: 10.3233/JEM-190464.

Valliant, R. and J. Dever (2011). “Estimating propensity adjustments for volunteer web surveys.” In: *Sociological Methods and Research* 40, pp. 105–137.
van den Brakel, J. (2022). “New data sources and inference methods for official statistics.” In: *Statistics in the public interest*. Ed. by A. Carriquiry, J. Tanur, and W. Eddy. Springer Verlag, pp. 411–432.
Wu, C. (2022). “Statistical inference with non-probability survey samples.” In: *Survey Methodology* 48, pp. 283–311.

Appendices

A LLM Prompt

Listing 1 Prompt for Deepseek

```
prompt <- paste0("First, Consider the following website text: ", data[i, urltext], ".

                 Second, consider this definition of innovation `Innovation is the use
                 of new ideas, products or methods where they have not been used
                 before.

                 For the Community Innovation Survey (CIS), an innovation is defined
                 as a new or significantly improved product (good or service)
                 introduced to the market, or the introduction within an enterprise
                 of a new or significantly improved process.

                 Innovations are based on the results of new technological
                 developments, new technology combinations, or the use of other
                 knowledge, acquired by the enterprise. The innovations may be
                 eveloped by the innovating enterprise or by another enterprise.
                 However, purely selling innovations wholly produced and developed
                 by other enterprises is not included as an innovation activity,
                 nor is introducing products with purely aesthetic changes.

                 Innovations should be new to the enterprise concerned: for product
                 innovations, they do not necessarily have to be new to the market
                 and, for process innovations, the enterprise does not necessarily
                 have to be the first one to have introduced the process.

                 Enterprises carrying out innovation activities cover all types of
                 innovators including product and process innovators, as well as
                 those enterprises with only ongoing and/or abandoned innovation
                 activities. The proportion of enterprises undertaking innovation
                 activities is also called the propensity (tendency) to innovate.

                 A product innovation is the market introduction of a new or a
                 significantly improved good or service.

                 A process innovation is the implementation of a new or significantly
                 improved production process, distribution method or support activity
                 for goods or services.'

                 Give me a classification (0 or 1) whether this website text belongs
                 to an innovative company considering the definition above.
                 Give me a probability whether this website text belongs to an
                 innovative company considering the definition above.

                 Give me a classification (0 or 1) whether this website text
                 belongs to a company with product innovations considering the
                 definition above.
                 Give me a probability whether this website text belongs to a
                 company with product innovations considering the definition above.

                 Give me a classification (0 or 1) whether this website text
                 belongs to a company with process innovations considering the
                 definition above.
                 Give me a probability whether this website text belongs to a
                 company with process innovations considering the definition above.

```

```
                Please answer in the form of a JSON-object containing the
                classification. The root needs to be a list. Your response
                must be only valid JSON so it can be read by a machine.

                Use always the following variable names in the same order:

                innovative_company_classification, innovative_company_probability,
                product_innovations_classification, product_innovations_probability,
                process_innovations_classification, process_innovations_probability")
```

The official definition of 'Innovation' from Eurostat is used, see: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Innovation#:~:text=Innovation%20is%20the%20use%20of,have%20not%20been%20used%20before.

Colophon

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

Prepress
Statistics Netherlands Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information