university of groningen

Netherlands Integrated Data Infrastructure of Inequality in Organizations (NIDIO)

Version 1.1

Christoph Janietz & Zoltán Lippényi

May 9, 2025

Abstract

NIDIO is an open-source code infrastructure assisting with the use of Dutch administrative register data. It is built around administrative data provided by Statistics Netherlands (CBS) in the Microdata Services Remote Access Environment (RA). NIDIO facilitates the study of inequality within and between Dutch organizations by integrating various administrative source data into a harmonized three-level data structure (organizations, individuals, and jobs). NIDIO helps to reconstruct workers' demographic profiles and employment outcomes (e.g., wages) and links them to organizational characteristics.

Contact:
Christoph Janietz (c.janietz@rug.nl)
Zoltán Lippényi (z.lippenyi@rug.nl)

Project repository:
OSF
Code repository:
GitHub

Please cite as:
Janietz Christoph & Zoltán Lippényi (2024). NIDIO - Netherlands Integrated Data Infrastructure of Inequality in Organizations. DOI: 10.17605/OSF.IO/9B2XH

Funding acknowledgment:
The NIDIO code was developed by Christoph Janietz and Zoltán Lippényi as part of the research project "Beyond Boardrooms". This research was funded by an NWO Talent Scheme VIDI grant (project number: VI.Vidi.211.231).

Contents

1 Introduction

Microdata based on administrative registers are becoming increasingly available for social science research. While this development opens many exciting avenues for research, including the possibility of studying inequality within organizations, working with administrative microdata also presents new challenges. Perhaps the most fundamental and necessary step is the processing and linking of separate administrative source data to obtain an analyzable research dataset. Currently, practical open-access code that helps with this essential yet complex task is lacking. Statistical bureaus generally provide documentation that accompanies administrative datasets, but these documents are not intended to lay out the necessary steps towards obtaining a final dataset for analysis. The task of data processing is further aggravated by the often intricate concepts [1] and the very large number of observations that are commonly encountered in administrative data.

This lack of available resources motivated us to create the Netherlands Integrated Data Infrastructure for Inequality in Organizations (NIDIO) NIDIO is a code infrastructure that assists researchers with processing and linking administrative microdata provided by Statistics Netherlands (CBS) in the Microdata Services Remote Access Environment (RA). NIDIO is meant to ease the use of administrative microdata by creating data processing standards that are replicable and transferable across different research projects. In particular, the focus of NIDIO is the creation of linked employer-employee datasets that enable the incorporation of organizations and workplaces in the study of labor markets and social inequality.

In this manual, we explain the NIDIO code infrastructure in detail and provide suggestions for its use. In the remainder of this introduction, we provide a brief overview of the benefits and challenges of using administrative data for social science research and clarify how NIDIO helps researchers address these challenges.

[1] For example, different wage components which are based on complex administrative rules and procedures.

1.1 Benefits of Administrative Microdata

The main motivation for using administrative data for research on labor inequalities is that it provides a more reliable measurement of employment characteristics and wage earnings than traditional surveys, while also not suffering from survey non-response. Administrative wage data contain considerably less information on individuals than traditional surveys do, but this shortcoming can be mitigated by linking information from other administrative sources (e.g., demographic and birth registers). Combining different administrative data sources also creates new possibilities for research, such as identifying the social contexts in which individuals are embedded (e.g, neighborhoods, schools, and organizations) and analyzing their effect on employment and wage outcomes. For labor inequalities, the availability of data on full employee populations and the possibility of linking each employee to their employing organization (LEED data) have been particularly relevant. LEED data facilitate research on the role of organizations and workplaces in the generation of inequality. Social scientists have long emphasized that work organizations are relevant social contexts in which inequality is produced, but in-depth empirical scrutiny has long been hindered by a lack of appropriate data. Here, administrative data fill an important gap by allowing researchers to fully reconstruct the wage structure of individual organizations and to follow workers across different organizations over their careers. Consequently, the use of administrative data has proliferated in inequality research in the recent years.

1.2 Challenges of Administrative Microdata

However, administrative data also pose new challenges to social scientists. First, these data were originally generated for administrative purposes. As a result, underlying concepts and definitions often stem from legal or administrative frameworks, and are not always equivalent to or aligned with concepts and definitions that are common in academic research. Even when concepts roughly match, administrative categorizations may differ from the standard operationalizations in survey-based research. Second, the administrative population may not always be the same as the research population of interest. While full population coverage is a common major strength of administrative data, users must be precise in specifying their target population and adjust

their analysis accordingly. Moreover, researchers may occasionally encounter data omissions for administrative reasons that cannot be justified on the grounds of academic research. In certain situations, researchers may need to supplement administrative register data with survey data, which leads to questions regarding the representativeness of the realized samples. This also applies when combining different register datasets, in which systematic exclusions, differing coverage, and multiple levels of data may increase the complexity of data handling. Third, the large size of administrative datasets, in some cases hundreds of millions of observations, creates computational issues and necessitates efficient statistical coding, a problem that is not encountered when working with smaller-sized survey datasets.

1.3 How does NIDIO help?

While there are many challenges associated with the use of administrative register data, the growing number of insightful empirical studies that analyze such data shows that taking the effort can lead to fruitful results. However, the growing interest in administrative data has not yet resulted in systematic guidelines and code that support researchers in addressing these challenges. Although thorough documentation of individual administrative datasets by statistical offices such as Statistics Netherlands (CBS) exists, this documentation generally does not provide guidelines on how to process and link administrative data during research. Consequently, problems are often discovered and solved within the confines of individual research projects. This impedes cumulative learning in the user community and reduces the transparency and reproducibility of research in the absence of wider communication and code sharing.

NIDIO is an effort to fill this gap. It is an integrated code system that helps users overcome the challenges of working with administrative data in three key areas. First, NIDIO assists with pre-processing, reduction, and cleaning of the large and complex source data available in the Microdata Services Remote Access Environment (RA) of Statistics Netherlands (CBS). Second, NIDIO facilitates the linkage between different datasets by determining relevant unique identifiers that identify each observation. Third, NIDIO guides researchers in selecting and deriving a number of key variables for labor inequality research (such as wages, job tenure, or education) based on ad-

ministrative constructs in the source data. We describe these benefits of NIDIO in further detail in the subsequent chapters of this manual.

An important disclaimer is warranted at this point. NIDIO is a project by end users of CBS microdata. While constructing NIDIO, our aim was to develop standardized data processing procedures that only interfere as much as necessary in the decision-making space of prospective users. In other words, we restricted ourselves to what we considered essential data processing steps to retain a maximum of flexibility for using the NIDIO datasets. We have documented our underlying choices extensively in the code and this manual to ensure transparency. In our role as end users, we have no access to the original administrative records and limited capacity to correct mistakes that arise during data generation (i.e., errors during the creation and compilation of administrative records). Such errors can have different causes and may range from incorrect identifiers that lead to incomplete linkages to improbable values for certain variables. For a large part, such mistakes are filtered by Statistics Netherlands through their raw source data screening process.

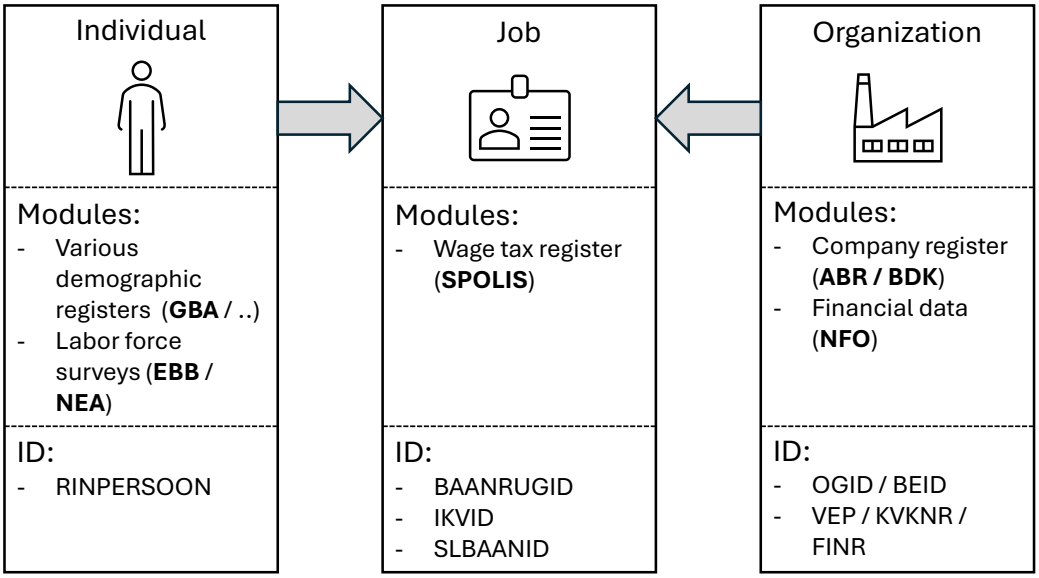

Figure 1: Structure of NIDIO

2 Structure of NIDIO

2.1 What is NIDIO?

NIDIO is an open-source code infrastructure that assists with the use of Dutch administrative register data. It is built around administrative data provided in the Microdata Services Remote Access Environment (RA) of Statistics Netherlands (CBS). NIDIO creates user-friendly datasets to study labor market inequality within and between Dutch organizations. The NIDIO code integrates various administrative source data into a harmonized longitudinal data structure that incorporates three analytical levels (organizations, individuals, and jobs). Using the NIDIO code, researchers can reconstruct workers' demographic profiles and employment outcomes (e.g., wages) and link them to organizational characteristics. The overall aim of NIDIO is two-fold:

1. NIDIO aims to simplify the use of administrative register data for social science research and
2. NIDIO aims to contribute to open and replicable science by improving the transparency of working with linked employer-employee data based on administrative registers.

NIDIO achieves these aims by addressing three challenges that are commonly

encountered while working with linked employer-employee register data:

- Preparing and linking administrative data sources involves several decision-making steps that are often undisclosed in published research. NIDIO provides transparent data processing routines that improve the reproducibility of completed analyses.
- Users with limited prior experience working with linked employer-employee register data face high startup costs during project setup. NIDIO eases time- and labor-intensive data processing by providing customizable installation tools.
- Translating administrative data into social science concepts and measures is a non-trivial task. NIDIO provides guidelines and best practices to bridge the gap between administrative and scientific data.

2.2 Levels - Modules - Datasets

NIDIO is structured around the concepts of levels, modules, and datasets (Figure 1). **Levels** are distinct units of analysis, of which we distinguish between organizations, individuals, and jobs. *Organizations* are described both as statistical units (OG - 'ondernemingengroep'; BE - 'bedrijfseenheid') and fiscal units (CBS Persoon; kvknr; finr) in the Dutch administrative data system. NIDIO uses a definition of organizations as entities who are "factual actors in the production process"[2] and that pay one or more individuals to work for them (i.e., employers). This definition includes private companies, governmental agencies, and non-profit organizations. Organizations who do not employ individuals are excluded from the NIDIO files during data processing.[3]

Individuals are defined as natural persons.[4] Individuals are recognized using the 'RINPERSOONS' prefix and the 'RINPERSOON' number in the Dutch administrative data system. Together, these two components form a unique pseudonymous persistent identifier. Individuals who are registered in the

[2] See CBS for further information.

[3] In technical terms: NIDIO restricts the population of organizations to those with linked job IDs of employees in the SPOLISBUS between 2006 and 2023.

[4] NIDIO includes all individuals registered in the Netherlands, thus also those who are not employed by an organization.

Personal Records Database (BRP)[5] can be consistently identified across different data topics within the Microdata Catalogue of Statistics Netherlands. NIDIO restricts its datasets to individuals registered in the BRP to facilitate data linkage.[6]

Jobs are defined as an existing employment relationship between an individual and an organization that generates wage earnings.[7] In other words, a given job ID establishes a relationship between an individual and an organization. An individual may hold multiple parallel jobs in different organizations at the same time. Jobs are identified using the variables 'baanrugid' (-2009) and 'ikvid' (2010-). The switch between these different job IDs marks a coding break in 2010. Therefore, jobs are also identified longitudinally using an additional variable 'slbaanid' constructed by CBS.[8] NIDIO additionally distinguishes 'main jobs' to identify the most economically relevant job among all existing jobs of an individual during a given time period. A main job is defined as the employment relationship of an individual with the highest absolute earnings during the reference period.[9]

Modules are distinct data components and are equivalent to the microdata topics listed in the CBS Microdata Catalogue (see Table 1).[10] Modules specify the source data in the CBS RA environment for processing by NIDIO. Researchers who intend to install a certain NIDIO module must have access to the corresponding original CBS microdata topic in their RA project.

Modules correspond to one of the three levels (organizations, individuals,

[5] In Dutch, its is referred to as the Basisregistratie Personen. See RVIG for further information. Registration is mandatory for individuals who intend to stay longer than four months in The Netherlands.

[6] In technical terms: NIDIO limits the population of individuals to those with the prefix 'R' in the variable RINPERSOONS.

[7] Note that NIDIO does not cover non-employment labor relationships, such as freelancing contracts, since it is currently not possible to link such employment forms to the organization for which the work is carried out. However, temporary agency work is included, but it is linked to the organization employing the worker and not to the client organization where contracted work is performed.

[8] See Spolislongbaantab (CBS) for further information.

[9] NIDIO users may identify these main jobs using the variable 'mainjob' in the module SPOLIS. See Spolishoofdbaanbus (CBS) for further information.

[10] It is recommended to consult the documentation of CBS for further technical details on each module. Links to CBS documentation material can be found in Table 1.

Module	Full name	CBS
ABR	Abr: Algemeen bedrijven register	Link
BDK	Bdk: Bedrijfsdemografisch kader	Link
NFO	Nfo: Financiën van niet-financiële ondernemingen	Link
GBA	Gbapersoontab: Persoonskenmerken van personen in de BRP	Link
OPL	Hoogsteopltab: Hoogst behaald/gevolgd opleidingsniveau	Link
KIND	Kindoudertab: personen en hun juridische ouders	Link
PARTNER	Partnerbus: Personen met partner met een adres	Link
EBB	Ebbnw: Enquête Beroepsbevolking nieuwe reeks 2003-2023	Link
NEA	Nationale Enquête Arbeidsomstandigheden (NEA)	Link
SPOLIS	Spolisbus: Banen en lonen volgens Polisadministratie	Link

Table 1: Overview of NIDIO modules

and jobs). ABR, BDK, and NFO hold *organization-level* data. ABR is the administrative register of the full population of organizations in the Netherlands that comprises organizational characteristics such as industry, location, and size. BDK contains additional demographic characteristics of organizations such as their founding date. NFO contains financial data from private non-financial companies. Companies report this information for taxation purposes. Note that NFO includes only a subset of the total population of employing organizations, as it does not include public companies and financial firms.

The GBA, OPL, KIND, PARTNER, EBB, and NEA modules contain *individual-level* data. GBA (short for GBAPERSOONTAB) contains demographic characteristics such as administrative sex categories and the country of birth of individuals registered in the Netherlands. OPL (short for HOOGSTEOPLTAB) contains information on the highest attained level of education of an individual. CBS derives educational codes from incomplete administrative records and supplements this administrative data with additional survey data.[11] KIND

[11] Education codes are not observed for all individuals in the population. For example, the education of individuals who have received their education abroad is not systematically represented in the administrative records. Most importantly, younger cohorts are overrepresented as their educational credentials are administered digitally, whereas those for the older generations predate digitization. CBS uses imputation and weighting methodology to derive population-level estimates for the whole Dutch population. See the HOOGSTEOPLTAB manual for detailed information. These weights are not always applicable when analyzing subpopulations and an adjusted weighting scheme may be required. This manual contains a short guide and example code on how to create custom weights using population marginals

(short for KINDOUDERTAB) is a register of children linked to their parents. The NIDIO code transforms this data into a 'parent file'. Each observation represents a parent-child link from the perspective of the parents. PARTNER (short for PARTNERBUS) identifies cohabiting partners and the period of their cohabitation. EBB (short for Enquête Beroepsbevolking) is the Dutch labor force survey administered by CBS. The NIDIO code extracts occupational codes of respondents (ISCO-08) from the EBB survey that can be linked to jobs in the administrative registers based on the survey date and main job status. NEA (short for Nationale Enquête Arbeidsomstandigheden) is the 'National Survey of Working Conditions' which is administered by CBS together with TNO. NEA survey data can be linked to administrative registers via the personal identifier 'RINPERSOON'. As in the case of the EBB module, NIDIO extracts the occupational codes of NEA survey respondents (ISCO-08) that can be linked to jobs in the administrative registers.

The SPOLIS module contains *job-level* data. SPOLIS (short for SPOLISBUS) contains information on paid wages and working hours, which employers report for taxation purposes. NIDIO aims to reduce the complexity of the SPOLIS source data and provides transparent data processing routines. NIDIO derives datasets containing one unique observation per job-year. This aggregation is achieved by summing earnings and working hours that are accumulated within the same job during the reference period.

Datasets are distinct files that NIDIO generates as output using CBS source data as input (Table 2). NIDIO generates one to four datasets for each module (i.e. data topic). Note that file sizes differ substantially. The smallest NIDIO dataset is the file **nidio_nea_occ_2006_2023.dta** with 16.8 MB, while the largest NIDIO dataset is **nidio_spolis_year_2006_2023.dta** with a file size of 50.8 GB.

Inside the RA environment, each NIDIO dataset can be linked to any other NIDIO dataset and many other original CBS datasets using *link variables*. The most important link variables are the calendar year (year), the organization IDs (ogid; beid), the individual ID (rinpersoon), and the job IDs (baanrugid, ikvid, slbaanid). Most NIDIO datasets are built around unique identifiers to facilitate linkage. For example, the combination of the variables year (calendar

and iterative proportional fitting (IPF) in section 5.2.

year) and beid (organization ID) corresponds to a single observation in the NIDIO dataset **nidio_abr_ogbe_register_2006_2023.dta**. This allows to link organizational-level data such as industry codes available in the ABR to the job-level data available in the SPOLIS by matching on the variables year and beid. An overview of unique identifiers in the NIDIO datasets is presented in Table 2.

Currently, NIDIO (v1.1) generates the following datasets (Table 2):

- **ABR: nidio_abr_og_register_2006_2023.dta**
  This dataset is a longitudinal register of the higher-level organizational units (OG) between 2006 and 2023. OG IDs are listed as yearly observations and include the start and end dates of their observation.
- **ABR: nidio_abr_og_size_2006_2023.dta**
  This dataset contains the number of employees per OG-year between 2006 and 2023. The number of employees is derived from SPOLIS and aggregated to the OG level using information from the BE level.
- **ABR: nidio_abr_ogbe_register_2006_2023.dta**
  This dataset is a longitudinal register of the lower-level organizational units (BE) between 2006 and 2023. BE IDs are listed as yearly observations and include the start and end dates of their observation. This file also contains relevant organizational characteristics observed at the BE level and links BE units to their overarching OG unit.
- **ABR: nidio_abr_ogkvk_register_2006_2023.dta**
  This dataset is a longitudinal register that links higher-level statistical organizational units (OG) to their corresponding identification as a legal entity (vepid) between 2006 and 2023.
- **ABR: nidio_abr_bekvk_register_2006_2023.dta**
  This dataset is a longitudinal register that links lower-level statistical organizational units (BE) to their corresponding identification as a legal entity (vepid) between 2006 and 2023.
- **BDK: nidio_bdk_be_2006_2023.dta**
  This dataset is a longitudinal register of additional demographic characteristics of the lower-level organizational units (BE) between 2007 and 2023. BE IDs are listed as yearly observations and include the start and end date of their observation

- **NFO: nidio_nfo_finances_2006_2023.dta**
  This dataset is a harmonized version of the financial data available in the NFO between 2006 and 2023.
- **NFO: nidio_nfo_og_2006_2023.dta**
  This dataset is a harmonized version of the financial data available in the NFO, which can be linked to the higher-level organizational units (OG). Each observation represents an unique OG-year.
- **SPOLIS: nidio_spolis_beid_register_2006_2023.dta**
  This dataset identifies lower-level organizational units (BE) that are employers between 2006 and 2023. BE IDs with at least one registered employee in the SPOLIS during a given calendar year are included in this dataset.
- **GBA: nidio_gba_rin_2023.dta**
  This dataset contains demographic characteristics of individuals registered in the GBR as of 2023. This dataset also includes additional country typologies merged via *landrefaktueel* stored in the RA utilities folder. A re-coded NIDIO version of the country typology is provided in addition to the original CBS classification.
- **OPL: nidio_opl_rin_2006_2023.dta**
  This dataset contains the highest attained level and field of education of individuals between 2006 and 2023. It includes education codes linked via the auxiliary education number (as ISCED 2011 codes) and education codes that are adjusted by CBS using additional imputation methodology (as SOI 2016/2021).
- **KIND: nidio_kindouder_parents_2023.dta**
  This dataset is a register of children linked to their parents. NIDIO transforms the original data to establish parent's unique identifiers as the *link variable*. The NIDIO dataset includes additional derived variables such as the total number of registered children of a parent until 2023, which are not included in the original source data.
- **PARTNER: nidio_partnerbus_rin_allyears.dta**
  This dataset contains spells of cohabiting partnerships until 2023.
- **EBB: nidio_ebb_occ_2006_2023.dta**
  This dataset contains the ISCO-08 occupation codes of respondents interviewed in the Dutch labor force survey (EBB) between 2006 and

2023. The occupation codes can be precisely merged to jobs in the administrative register data using a combination of the 'RINPERSOON' ID and the exact survey date.

- **NEA: nidio_nea_occ_2006_2023.dta**
  This dataset contains the ISCO-08 occupation codes of respondents interviewed in the Nationale Enquête Arbeidsomstandigheden (NEA) between 2006 and 2023. The occupation codes can be coarsely merged to jobs in the administrative register data using the rinpersoon ID.
- **SPOLIS: nidio_spolis_month_2006_2023.dta**
  This dataset includes all jobs of registered individuals employed in organizations included in the ABR in existence during September of a given year. Each observation represents a unique job during the month of September between 2006 and 2023. All compensation and working hours accumulated within the same job during this month are summed in one observation. Job IDs are made longitudinally consistent using SPOLISLONGBAANBUS (variable 'slbaanid'). The main jobs of individuals are identified using SPOLISHOOFDBAANBUS (variable 'mainjob').
- **SPOLIS: nidio_spolis_year_2006_2023.dta**
  This dataset includes all jobs of registered individuals employed in organizations included in the ABR. Each observation represents a unique job in a given year between 2006 and 2023. All compensation and working hours accumulated within the same job during the full calendar year are summed in one observation. However, categorical job characteristics, such as contract status (permanent or temporary employment), can change during the calendar year. These variables are assigned as the last observed status in a given year. Job IDs are made longitudinally consistent using SPOLISLONGBAANBUS (variable 'slbaanid'). The main jobs of individuals are identified using SPOLISHOOFDBAANBUS (variable 'mainjob').

Level	Module	Dataset	Unique Identifier	File size
Organization	ABR	nidio_abr_og_register_2006_2023.dta	year-ogid	203 MB
Organization	ABR	nidio_abr_og_size_2006_2023.dta	year-ogid	84.9 MB
Organization	ABR	nidio_abr_ogbe_register_2006_2023.dta	year-beid	674 MB
Organization	ABR	nidio_abr_ogkvk_register_2006_2023.dta	*none*	1.28 GB
Organization	ABR	nidio_abr_bekvk_register_2006_2023.dta	*none*	1.14 GB
Organization	BDK	nidio_bdk_be_2006_2023.dta	year-beid	1.66 GB
Organization	NFO	nidio_nfo_finances_2006_2023.dta	*none*	427 MB
Organization	NFO	nidio_nfo_og_2006_2023.dta	year-ogid	301 MB
Organization	SPOLIS	nidio_spolis_beid_register_2006_2023.dta	year-beid	66.6 MB
Person	GBA	nidio_gba_rin_2023.dta	rinpersoon	0.98 GB
Person	OPL	nidio_opl_rin_2006_2023.dta	year-rinpersoon	6.02 GB
Person	KIND	nidio_kindouder_parents_2023.dta	rinpersoon-rinchild	1.10 GB
Person	PARTNER	nidio_partnerbus_rin_allyears.dta	*none*	644 MB
Person	EBB	nidio_ebb_occ_2006_2023.dta	year-rinpersoon-svydate	144 MB
Person	NEA	nidio_nea_occ_2006_2023.dta	year-rinpersoon	16.8 MB
Job	SPOLIS	nidio_spolis_month_2006_2023.dta	year-rinpersoon-baanrugid/ikvid year-rinpersoon (if mainjob)	37.1 GB
Job	SPOLIS	nidio_spolis_year_2006_2023.dta	year-rinpersoon-baanrugid/ikvid year-rinpersoon (if mainjob)	50.8 GB

Table 2: Overview of NIDIO datasets

2.3 Important Notes

- **Source data & version numbers:** CBS regularly updates microdata files inside the RA environment and replaces older version numbers of microdata stored in the G:/ drive. NIDIO is programmed to automatically retrieve the current version number of source data for data processing. After the initial configuration of NIDIO, the file paths and version numbers of all source data currently in use can be checked with the command 'source_nidio'. For further details, see section 4.1.2.
- **Organizations as employers:** NIDIO reduces data to organizations with at least one employee during the reference year (see also section 2.2). All other organizations without employees (e.g., associations) are excluded from the NIDIO datasets.
- **Organizations registered in the ABR:** NIDIO excludes job observations in organizations that are not registered and tracked in the ABR. There are several BE IDs appearing in the SPOLIS source data with a leading letter "F" or "P" that cannot be linked to the general organization register.
- **Individuals registered in GBR:** NIDIO reduces data to individuals who are registered in the GBR (see also section 2.2). This is equivalent to individuals holding a constant rinpersoon prefix of 'R' in all NIDIO datasets. Therefore, NIDIO datasets do not contain the variable 'RINPERSOONS', but users may restore this variable using the command 'genrinpersoons' For further details, see section 4.4.3.
- **Jobs included in NIDIO:** Following these selection criteria, the NIDIO datasets **nidio_spolis_month_2006_2023.dta** and **nidio_spolis_year_2006_2023.dta** only include jobs of registered individuals (RINPERSOON==R) employed in registered organizations (BEID first sign $\neq$ F $\vee$ P). This leads to a loss of job IDs during data processing which is described in Tables 3 and 4.
- **SPOLISLONGBAANTAB & SPOLISHOOFDBAANBUS:** The installation of the NIDIO modules 'SPOLIS MONTH' and 'SPOLIS YEAR' requires access to the source data SPOLISLONGBAANTAB and SPOLISHOOFDBAANBUS in addition to the standard SPOLIS files. Access to these source data may need to be requested from CBS (microdata@cbs.nl). These data files are part of the SPOLIS component of the CBS microdata catalogue and do not incur additional costs.

Year	Source All jobs	RINPERSOONS==R Jobs of reg. ind.	NIDIO (SPOLIS MONTH) Jobs of reg. ind. in reg. org.
2006	7,821,419	7,708,906 (98.56%)	7,670,977 (98.08%)
2007	8,085,988	7,926,469 (98.03%)	7,884,070 (97.50%)
2008	8,180,730	7,998,700 (97.77%)	7,949,007 (97.17%)
2009	8,043,001	7,787,408 (97.95%)	7,826,799 (97.32%)
2010	8,161,109	7,994,872 (97.96%)	7,936,086 (97.24%)
2011	8,220,862	8,037,967 (97.78%)	7,969,983 (96.95%)
2012	8,130,651	7,952,213 (97.81%)	7,885,472 (96.98%)
2013	8,005,893	7,833,699 (97.85%)	7,746,377 (96.76%)
2014	8,005,359	7,836,607 (97.89%)	7,724,547 (96.49%)
2015	8,100,491	7,922,677 (97.80%)	7,818,659 (96.52%)
2016	8,238,899	8,049,986 (97.71%)	7,943,087 (96.53%)
2017	8,467,621	8,256,591 (97.51%)	8,153,979 (96.30%)
2018	8,709,505	8,708,407 (99.99%)	8,592,630 (98.66%)
2019	8,867,124	8,801,323 (99.26%)	8,706,502 (98.19%)
2020	8,738,864	8,651,547 (99.00%)	8,554,644 (97.89%)
2021	9,021,776	9,020,180 (99.98%)	8,915,374 (98.82%)
2022	9,236,047	9,234,554 (99.98%)	9,127,361 (98.82%)
2023	9,307,666	9,306,594 (99.99%)	9,189,718 (98.73%)

Table 3: Unique jobs in **nidio_spolis_month_2006_2023.dta**

Year	Source All jobs	RINPERSOONS==R Jobs of reg. ind.	NIDIO (SPOLIS YEAR) Jobs of reg. ind. in reg. org.
2006	10,971,506	10,744,081 (97.93%)	10,686,488 (97.40%)
2007	11,485,739	11,173,523 (97.28%)	11,113,590 (96.76%)
2008	11,560,579	11,190,810 (96.80%)	11,122,133 (96.21%)
2009	11,109,722	10,783,850 (97.07%)	10,712,334 (96.42%)
2010	11,496,623	11,137,507 (96.88%)	11,054,982 (96.16%)
2011	11,520,550	11,114,485 (96.48%)	11,038,235 (95.81%)
2012	11,127,211	10,749,095 (96.60%)	10,674,411 (95.93%)
2013	10,943,221	10,582,578 (96.70%)	10,463,523 (95.62%)
2014	11,056,526	10,690,840 (96.69%)	10,537,180 (95.30%)
2015	11,336,263	10,949,001 (96.58%)	10,807,645 (95.34%)
2016	11,631,841	11,206,292 (96.34%)	11,063,278 (95.11%)
2017	12,015,469	11,536,669 (96.02%)	11,394,904 (94.84%)
2018	12,395,976	12,385,701 (99.92%)	12,220,268 (98.58%)
2019	12,546,176	12,398,636 (98.90%)	12,265,526 (97.76%)
2020	12,096,355	11,870,386 (98.13%)	11,739,702 (97.05%)
2021	12,619,015	12,605,363 (99.89%)	12,460,261 (98.74%)
2022	13,211,030	13,196,191 (99.89%)	13,047,206 (98.76%)
2023	13,123,761	13,112,719 (99.92%)	12,951,611 (98.69%)

Table 4: Unique jobs in **nidio_spolis_year_2006_2023.dta**

3 Installation of NIDIO

3.1 Quick Installation Guide

NIDIO is installed by performing the following steps:

1. Place the NIDIO folder in your project drive (H:/) within the CBS RA environment.
2. Open the do-file install_NIDIO.do. You can find this file in the root directory ../NIDIO/ of the NIDIO file tree. Your working directory in STATA must be identical to the NIDIO root directory.
3. Execute the do-file install_NIDIO.do. Running the lines 10 and 11 of the do-file is mandatory. Executing these lines will locate and call the source data within the CBS RA environment.[12] You can customize the rest of your installation by using the command *install_nidio* together with one specified NIDIO module you wish to install.[13] For example, to install the module 'ABR', execute:

   ```
   install_nidio ABR
   ```

4. NIDIO will now be busy processing the CBS source data. This can take between 1 minute (module 'PARTNER') to 4 days (96 hours) (module 'SPOLIS_YEAR') depending on the selected module.
5. You will receive a notification in the STATA results window, when the installation of the module is complete. You can find the ready-to-use datasets in the corresponding folder under the file path ../NIDIO/Data/ [Module]/ after completion.

3.2 Module Dependency during Installation

The installation of several NIDIO modules is either dependent on access to or the prior installation of specific other modules. The following modules are affected:

[12] The respective code is stored in an additional do-file called config.do that is located in ../NIDIO/Code/.

[13] The lines 17 to 29 of the do-file install_NIDIO.do already list the available options. The installation of some modules is dependent on other modules. See next section for details.

- **ABR:** Installation of the NIDIO module 'ABR' requires *access to* SPOLIS in the CBS RA environment.
- **NFO:** Installation of the NIDIO module 'NFO' requires *prior installation* of the NIDIO module 'ABR'.
- **KIND:** Installation of the NIDIO module 'KIND' requires *prior installation* of the NIDIO module 'GBA'.

3.3 CBS Remote Access Environment

NIDIO is designed to operate within the Statistics Netherlands Microdata Services Remote Access Environment (RA). NIDIO processes sensitive source data protected by extensive privacy rules and regulations[14]. Disclosing information about individuals or organizations is strictly forbidden for any output using CBS microdata.

The CBS RA is a secure environment that limits data access to approved researchers who are employed by authorized research organizations. Researchers must submit a research project plan. Each individual research project is vetted and needs approval from Statistics Netherlands (CBS) before access to the CBS RA environment is granted. Only data requested for the project and approved for analysis can be accessed. The data must only be used for statistical analyses. Researchers are obliged to make the results of their data analyses available to interested parties free of charge.

The CBS RA environment is accessed by authorized researchers using a three-factor authentication system. This includes a personal password, SMS authentication, and use of a personal token. The CBS RA environment is only accessible using a VPN, which blocks all other parallel traffic. Once researchers have access to the RA and have requested the necessary data (Modules), they can make use of NIDIO within the secure environment during their analyses.

[14] Wet op het Centraal bureau voor de statistiek

3.4 Importing NIDIO into the CBS RA Environment

There are two ways to install NIDIO within the CBS RA Environment:

1. **Download NIDIO as a zip folder from the OSF repository.** The NIDIO code can be retrieved at https://osf.io/9b2xh/. This OSF repository is intended as the main gateway for the distribution of NIDIO. The OSF repository is synchronized with a GitHub repository in which the code is hosted and maintained. It is recommended to download the whole NIDIO folder as a zip-file via the OSF repository (see Figure 2).

   Users may then contact CBS (microdata@cbs.nl) to request an import of the NIDIO zip folder into their RA project environment.

2. **ODISSEI Storage Facility within the CBS Remote Environment** The very large NIDIO datasets **nidio_spolis_month_2006_2023.dta** and **nidio_spolis_year_2006_2023.dta** are also stored directly within the RA environment via the ODISSEI Storage Facility. Researchers at ODISSEI member organisations can request access to these files in their own RA projects (which is free of charge). Projects intending to use these NIDIO datasets need to have access to the original CBS microdata topic (SPOLIS) in their project. The usual fee for accessing the data applies. The stored files are the following:

   - **nidio_spolis_month_2006_2023.dta** (doi: 10.34894/YQIIQV)
   - **nidio_spolis_year_2006_2023.dta** (doi: 10.34894/63XWJB)

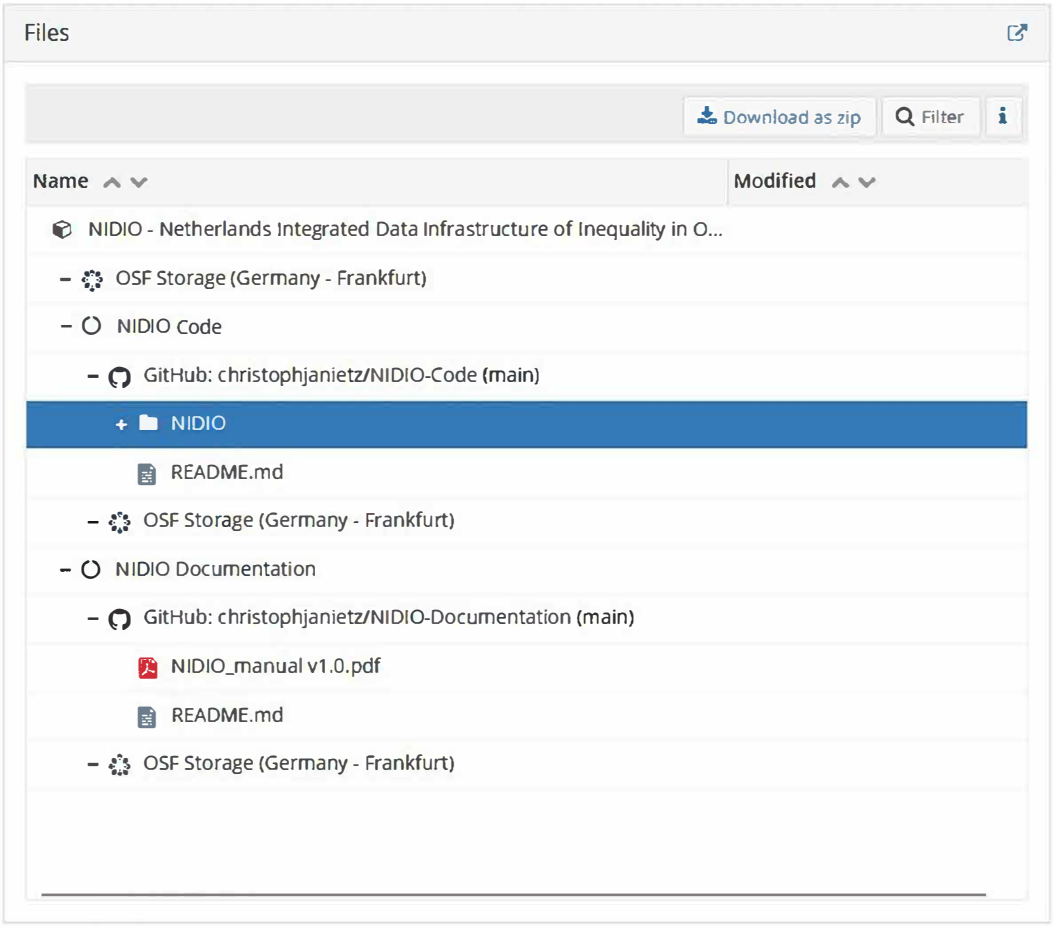

Figure 2: Downloading NIDIO from OSF repository (https://osf.io/9b2xh/)

4 Additional Functionalities

This section gives an overview of additional NIDIO functionalities that may be useful during the analysis of Dutch administrative register data in the CBS RA environment. These additional functionalities come in the form of user-written Stata commands.

The user-written commands provided with NIDIO have distinct functions. *Installation* commands simplify the initial setup of NIDIO. *Data filtering* commands subset very large NIDIO datasets (SPOLIS) using user-specified criteria. *Variable creation* commands generate useful additional variables using the NIDIO datasets.

Important: The commands must be activated by running the install_NIDIO .do do-file. All programming code is stored in the folder `../NIDIO/ Code/_PROGRAMS/`.

4.1 Installation

4.1.1 install_nidio

install_nidio - Install NIDIO modules with one line of code.

Syntax:

```
install_nidio module
```

Description:
install_nidio is a wrapper function that initiates the installation of a chosen NIDIO module. **install_nidio** will initiate the data processing after having first located the source data using the do-file install_NIDIO.do.
module stands for the code of the NIDIO module to be installed. The following modules can be installed using the respective code:

- ABR - Company register
- BDK - Company demographics
- NFO - Financial data of firms
- EBB - Occupational codes via Dutch Labor force survey (EBB)
- GBA - Demographic characteristics
- OPL - Highest education
- KIND - Registered children of parents
- NEA - Occupational codes via NEA
- PARTNER - Partnership register
- SPOLIS_MONTH - Jobs existing in September during reference year
- SPOLIS_YEAR - All jobs existing during reference year

Options:
None
Example:
Installation of NIDIO module SPOLIS (month format):

```
install_nidio SPOLIS_MONTH
```

4.1.2 source_nidio

source_nidio - Display file paths of CBS source data in the RA.

Syntax:

```
source_nidio
```

Description:
source_nidio lists the file paths of the CBS microdata used during data processing. This command creates a snapshot overview of the version number of each source data file used during data processing. CBS regularly updates files and replaces older version numbers. The snapshot overview of used version numbers may be stored for replication purposes.

Options:
None

Example:
List the file paths of the source data (CBS microdata):

```
source_nidio
```

4.2 Data Filtering

4.2.1 spolisselect

spolisselect - Loads a specified selection of the SPOLIS module.

Syntax:

```
spolisselect [namelist], data(string) [start(integer) end(integer)
jobtype(integer) mainjob(integer)]
```

Description:
spolisselect loads a user-specified subset of the SPOLIS data. The command is meant to speed up loading time while working with SPOLIS. Users may select a subset of variables via *namelist*. **Important:** SPOLIS module must be installed and dataset must be located in the folder `../NIDIO/Data/SPOLIS/`.

Options:
data: Specify data format of used SPOLIS file (required)

- data(month) for SPOLIS month format;
- data(year) for SPOLIS year format

start: Specify start year. Default is 2006.

end: Specify end year. Default is 2023.

jobtype: Select specific types of jobs.

- jobtype(1) for all job types (default);
- jobtype(2) for temporary agency, on-call, and standard jobs;
- jobtype(3) for only standard jobs

mainjob: Select only main jobs, if requested.

- mainjob(0) for all jobs (default);
- mainjob(1) for only main jobs

Example:
Select SPOLIS file in monthly format between 2011 and 2023 including all job types, but only main jobs:

```
spolisselect, data(month) start(2011) end(2023) mainjob(1)
```

4.2.2 spolisselectra

spolisselectra - Loads a specified selection of the NIDIO SPOLIS datasets stored in the CBS data storage.

Syntax:

```
spolisselectra [namelist], data(string) [start(integer) end(integer)
jobtype(integer) mainjob(integer)]
```

Description:
spolisselectra loads a user-specified subset of the SPOLIS data. The command is meant to speed up loading time while working with SPOLIS. Users may select a subset of variables via *namelist*. **Important:** Requires access to the NIDIO SPOLIS datasets in the CBS data storage facility within the RA environment.

Options:
data: Specify data format of used SPOLIS file (required)

- data(month) for SPOLIS month format;
- data(year) for SPOLIS year format

start: Specify start year. Default is 2006.

end: Specify end year. Default is 2023.

jobtype: Select specific types of jobs.

- jobtype(1) for all job types (default);
- jobtype(2) for temporary agency, on-call, and standard jobs;
- jobtype(3) for only standard jobs

mainjob: Select only main jobs, if requested.

- mainjob(0) for all jobs (default);
- mainjob(1) for only main jobs

Example:
Select SPOLIS file in monthly format between 2011 and 2023 including all job types, but only main jobs:

```
spolisselectra, data(month) start(2011) end(2023) mainjob(1)
```

4.2.3 orgsizeselect

orgsizeselect - Selects a subset of organizations based on size.

Syntax:

```
orgsizeselect, id(string) [min(integer) max(integer)
select(integer) n_org(integer)]
```

Description:
orgsizeselect identifies a subset of organizations based on size with size being defined as the number of employees. **orgsizeselect** reports the absolute number of organizations in the subset as well as the relative size of the subset. Users may restrict the dataset to the identified subset. Users may also save an additional variable that indicates the number of employees. **Important:** **orgsizeselect** is designed for the module 'SPOLIS'.

Options:
id: Select organization ID (required)

- id(beid) for BEID unit;
- id(ogid) for OGID unit

min: Specify minimum organization size Default is 1.

max: Specify maximum organization size. Default is 999999.

select: Reduce dataset in use to the selected set of organizations.

- select(0) for no (default);
- select(1) for yes

n_org: Create variable *N_org* holding organization size (i.e., number of employees) as integer.

- n_org(0) for no (default);
- n_org(1) for yes

Example:
Identify small organizations (maximum of 50 employees) at the BEID unit level and reduce the dataset to this subset.

```
orgsizeselect, id(beid) max(50) select(1)
```

4.3 Variable Creation

4.3.1 gentenure

gentenure - Create a job tenure measure using SPOLIS.

Syntax:

```
gentenure, data(string) gen(string) varformat(string)
```

Description:
gentenure creates a job tenure measure based on the variable 'job_tenure' in the NIDIO module SPOLIS. The variable 'job_tenure' holds the overall starting date of a job. Based on this information, **gentenure** computes the number of years (months) that have elapsed since the creation of the job. Jobs created before 2006 and ending before 2013 have a unknown starting date (i.e. are left-censored). **gentenure** computes a missing value for these cases.

Options:
data: Specify data format of used SPOLIS file (required).

- data(month) for SPOLIS month format;
- data(year) for SPOLIS year format

generate: Specify name of computed variable (required).

varformat: Unit of job tenure variable (required).

- varformat(month) for months as unit *(only for SPOLIS month format)*;
- varformat(year) for years as unit.

Examples:
Creation of a job tenure measure called 'jobten' using the SPOLIS year format.

```
gentenure, data(year) gen(jobten)
```

Creation of a job tenure measure called 'tenure_m' counted in months using the SPOLIS month format.

```
gentenure, data(month) gen(tenure_m) varformat(month)
```

4.3.2 genhwage

genhwage - Create an hourly wage measure using SPOLIS.

Syntax:

```
genhwage, data(string) generate(string) [concept(string)
denom(string) overwork(integer) real(integer)]
```

Description:
genhwage creates customized hourly wage measures. **genhwage** draws on different combinations of variables in the SPOLIS to define a numerator (option 'concept') and denominator (option 'denom'). The numerator is a combination of different forms of labor compensation, while the denominator is a working time measure. See the section 'Options for further details. Users may include overwork using the option overwork(1). In addition, **genhwage** optionally generates inflation-adjusted wages (i.e. real wages) using the official CPI calculated by CBS.

Options:
data: Specify data format of used SPOLIS file (required).

- data(month) for SPOLIS month format;
- data(year) for SPOLIS year format

generate: Specify name of computed variable (required).

concept: Numerator of the hourly wage measure. Choice is between various wage concepts.

- concept(basis) for basisloon (default);
- concept(extra) for basisloon + sbijzonderebeloning *(all non-regular remuneration which is part of gross pay including vacation pay, thirteenth month pay, and incidental salary (i.e. bonus)*;
- concept(bonus) for basisloon + sincidentsal *(amount paid in the pay period as incidental salary)*.

denom: Denominator of the hourly wage measure.

- denom(basis) for basisuuren *(net average number of hours worked excluding overtime hours)* (default);

- denom(regular) for regulierenuuren *(approximate number of hours actually worked excluding overtime hours)*

overwork: Includes overwork hours and compensation on request.

- overwork(0) for no (default);
- overwork(1) for yes *(Overwork is defined as additional work beyond stipulated hours in the employment contract that received a compensation surcharge)*

real: Calculates inflation adjusted wages using the formula $\frac{Wage}{CPI}$ on request.

- real(0) for no (default);
- real(1) for yes

Examples:
Creation of an hourly wage measure stored as the new variable 'real_hwage' using the SPOLIS month format. The new measure includes non-regular remuneration part of gross pay and is inflation-adjusted:

```
genhwage, data(month) gen(real_hwage) concept(extra) real(1)
```

Creation of an hourly wage measure stored as the new variable 'compensation' using the SPOLIS year format. The new measure includes incidental payments (i.e. bonus), actual working hours, and overwork compensation:

```
genhwage, data(year) gen(compensation) concept(bonus)
denom(regular) overwork(1)
```

4.3.3 genrinpersoons

genrinpersoons - Restores the variable 'rinpersoons'.

Syntax:

```
genrinpersoons
```

Description:
genrinpersoons restores the variable 'rinpersoons' that holds the prefix of the unique pseudonymous persistent person identifier. NIDIO datasets only include individuals who are registered in the GBR and have the constant prefix 'R' in the variable rinpersoons. NIDIO removes this prefix to speed up data processing. the command **genrinpersoons** restores this prefix. **genrinpersoons** can only be used in datasets that include individual IDs (variable 'rinpersoon').

Options:
None

Example:
Restore the variable 'rinpersoons':

```
genrinpersoons
```

5 Using NIDIO

In this section, we describe a specific research example using NIDIO (5.1). The aim of this example is to demonstrate the functionalities of NIDIO in an applied setting. Using NIDIO is not limited to this specific research example and NIDIO is applicable to a wide variety of research questions on inequality within and between work organizations. We hope users will draw inspiration from this example for how to use NIDIO in their own research.

In the outlined research example, we analyze total population data using variables that include observations for all organizations, individuals, and jobs. However, social scientists studying labor market inequality within and between organizations will often encounter situations in which certain relevant variables are only partially observed either for a subpopulation or for a sample. We discuss some additional recommendations for such situations (5.2). While analyzing administrative register data, the construction of a customized weighting scheme often represents a viable strategy to correct for selection bias when drawing inferences based on a subpopulation or sample.

5.1 Example: The Glass Ceiling in Larger Organizations

In this example, we focus on the glass ceiling phenomenon.[15] The glass ceiling is a metaphor used to portray the hidden, but systemic, barriers that prevent women from climbing career ladders within organizations. The glass ceiling can thus also be seen as a vertical form of job sex segregation. We use NIDIO to illustrate the extent of vertical job sex segregation in larger organizations in the Netherlands in 2022 and ask the following two descriptive research questions:

1. To what extent are women underrepresented in the highest-paying jobs within larger organizations in the Netherlands in 2022?
2. To what extent does this under-representation of women vary by industry?

To answer these questions, we require three NIDIO datasets: **nidio_spolis_month_2006_2023.dta**, **nidio_gba_rin_2023.dta**, and **nidio_abr_ogbe_register_2006_2023.dta**. We install these dataset by opening the do-file install_NIDIO.do in the NIDIO root directory and by executing the following commands in STATA:

[15] You can find an interactive data visualization using NIDIO here.

```
*Install the three required NIDIO modules
install_nidio ABR
install_nidio GBA
install_nidio SPOLIS_MONTH
```

The three required NIDIO datasets are now saved in their respective folders under `..NIDIO/Data/`. We continue the data preparation by opening the SPOLIS dataset that contains all existing jobs of registered individuals during September of a given year between 2006 and 2023. To use the SPOLIS dataset, we execute the following command:

```
* Load SPOLIS data and filter based on specific variables
spolisselect, data(month) start(2022) end(2022) jobtype(2)
mainjob(1)
```

By calling spolisselect this way, we open the SPOLIS dataset and also filter it by year, jobtype, and main job status. First, we restrict the data to the year 2022. Second, we exclude certain job types (internships, directors & large shareholders, and specific 'WSW' jobs, sheltered employment designated for workers with disabilities) from the analysis. Third, we limit the data to the main job of each individual worker (i.e., the most economically relevant job among all existing jobs of an individual). We then proceed with filtering the data by organization size. We execute the following command:

```
* Select larger organizations with at least 50 employees
orgsizeselect, id(beid) min(50) select(1)
```

In this execution of orgsizeselect, we restrict the job data to organizations (BE unit) with at least 50 main jobs. This is our definition of larger organizations for the remainder of this example. As a next step, we merge variables from the other two NIDIO datasets which we have installed before. Because we have reduced the SPOLIS dataset to main jobs, each observation now uniquely identifies one person in this dataset. Therefore, we use the variable 'rinpersoon' (individual ID) as a *linking variable* to merge individual-level information from the GBA dataset. The GBA dataset contains administrative sex categories (variable rin_sex) that we use to identify women and men in the SPOLIS data. We execute the following command:

```
*Merge administrative sex categories
merge 1:1 rinpersoon using "../nidio_gba_rin_2023.dta",
nogen keepusing(rin_sex) keep(3)
```

With a second merge, we obtain the variable be_industry from the ABR dataset. This variable identifies the industry affiliation of BE units based on the first digit of the Dutch industry classification SBI08. We need this information for our second research question. Because we merge an organizational-level variable, we now use the variable 'beid' (organization ID) as a *linking variable*. The combination of the variable 'year' and the variable 'beid' uniquely identifies observations in the ABR dataset. Therefore, we execute the following command:

```
*Merge industry codes
merge m:1 year beid using "../nidio_abr_ogbe_register_2006_2023.dta",
nogen keepusing(be_industry) keep(3)
```

As a next step, we create an hourly wage variable using the data on compensation and working hours available in the SPOLIS dataset. We construct this variable for the purpose of comparing compensation levels of women and men within organizations independent of working hours. We use the NIDIO command genhwage to construct this variable:

```
*Compute an hourly wage measure
genhwage, data(month) gen(hwage)
```

The resulting variable is called 'hwage' with the variable 'sbasisloon' in the numerator and the variable 'sbasisuuren' in the denominator. We exclude overwork hours and compensation. We do not adjust for inflation rates because of the cross-sectional nature of the present analysis (i.e., jobs in 2022). With the hourly wage measure in hand, we now can capture vertical job segregation within organizations. One possible strategy is to assign each individual worker to one out of five pay quintiles within their organization. The lowest pay quintile within an organization comprises the 20% lowest-paid jobs, while the highest pay quintile comprises the 20% highest-paid jobs. We employ this strategy and assign within-organization pay quintiles with the following command:

```
* Assigning individuals to within-organization pay quintiles
gquuantiles withinq = hwage, xtile nquantiles(5) by(beid)
```

After having assigned workers to their respective within-organization pay quintile, we now can investigate the extent of vertical job *sex* segregation in larger Dutch organizations. In particular, we estimate the share of women in the highest pay quintile within larger Dutch organizations and compare it

to the overall share of women working in these organizations. We carry out this estimation one time for the total population of large organizations and one time separately for each industry. To derive statistics, we execute the following commands:

```
* Dummy variable to obtain fractions using collapse
gen woman = 0
replace woman = 1 if rin_sex==2

* Total population of large organizations
collapse (mean) share_w = woman, by(withinq)
* Separately for each industry
collapse (mean) share_w = woman, by(be_indstry withinq)
```

We present these descriptive statistics in Figure 3 and 4.[16] Figure 3 displays the results for the total population of large organizations. We can see that women are under-represented in the highest pay quintile within organizations relative to their total employment share. While women hold a bit more than half of all jobs in larger Dutch organizations, they only occupy 42.89% of the jobs in the highest pay quintile within these organizations. The under-representation of women occurs predominantly in the highest pay quintile with representation almost reaching parity in the second-highest pay quintile (top 40% to top 20%).

Figure 4 additionally demonstrates that the glass ceiling phenomenon can be observed across a multitude of industries, even in those industries with a very large total share of women (e.g., healthcare and social work). However, there is also observable variation in the extent of vertical sex segregation across industries. For example, the under-representation of women at the top is much more pronounced in the financial industry (26.57% women in the top pay quintile) compared to the public administration sector (37.48%) despite a comparable total share of women who work in these industries.

This analysis is only an illustration and a starting point for further investigating the glass ceiling and gender inequality in Dutch organizations. The compiled NIDIO datasets provide ample opportunities for additional in-depth analysis.

[16] As an alternative to the outlined approach, users may want to weight worker observations by organization size. Without weighting, very large organizations have a greater impact on the estimated share of women within each pay quintile than other organizations.

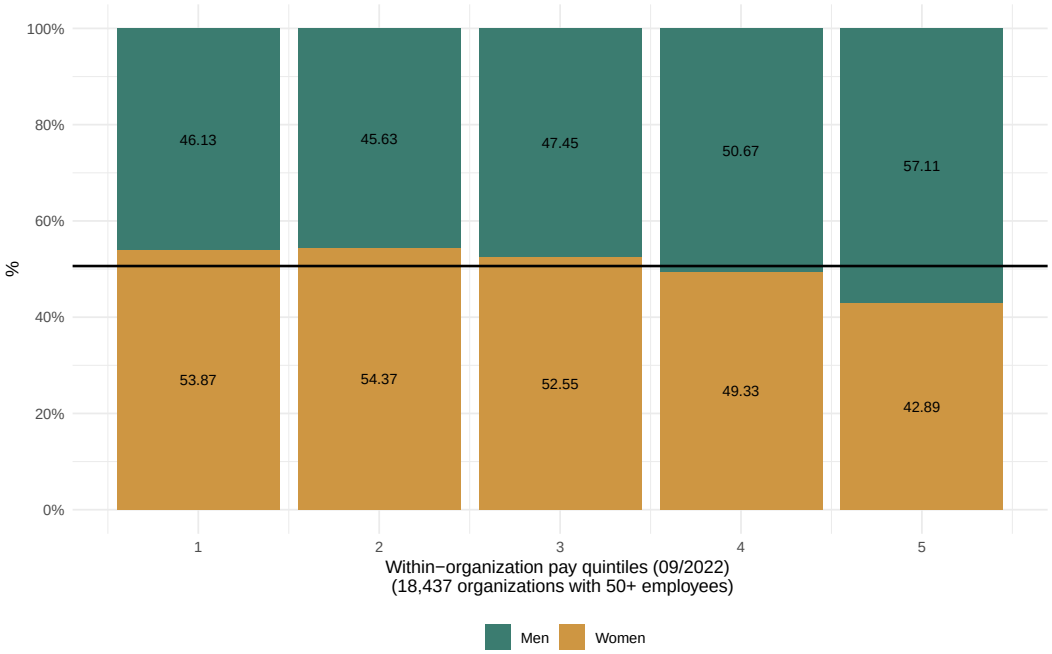

Figure 3: Share of women within pay quintiles of organizations

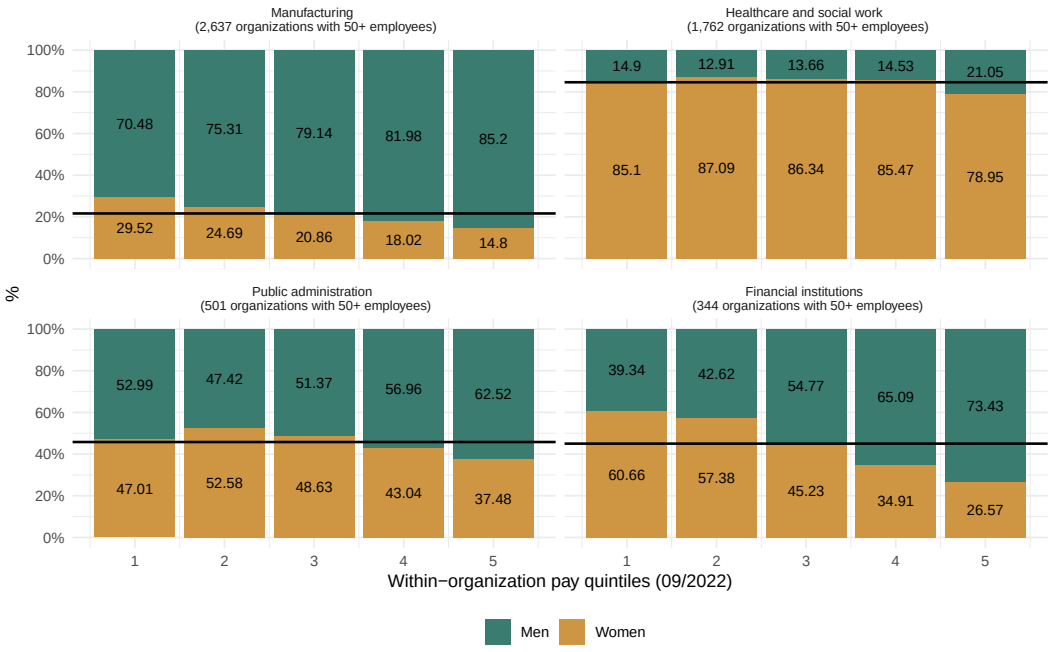

Figure 4: Share of women within pay quintiles of organizations by industry

5.2 Advice on Weighting using NIDIO

One of the main strengths of using administrative register data is its exhaustive population coverage. Many characteristics relevant to social science research are observable in the total population of organizations, individuals, or jobs. However, this is not always the case. In certain cases, important concepts and measures are observable only in a subpopulation or sample. In the context of NIDIO and the Dutch administrative register data system, two exemplary concepts observed for partial populations or samples are education and occupation:

1. *Education* is a central measure in labor market inequality research. In the Dutch administrative register data, education is only observed for a subpopulation (HOOGSTEOPLTAB). Underrepresented groups are older individuals who have graduated before the systematic digitization of educational data sources. Another underrepresented group are individuals who have completed their education abroad. CBS uses imputation and weighting methodologies to derive population-level estimates for the entire Dutch population. However, these weights are not always applicable, particularly when drawing inferences for the working population, which is a subpopulation of the total population. Therefore, an adjusted weighting scheme may be required.

2. *Occupation* is another key measure in labor market inequality research. Occupation is only observable via survey datasets in the Netherlands (most notably, the EBB and NEA), but can be linked to administrative data. However, unequal probabilities of selection and non-response may distort the representativeness of survey samples. The two surveys included in NIDIO (EBB and NEA) both provide survey weights to correct for sample selection bias, but these original survey weights might not always suffice. Users may want to combine occupational codes from both surveys to maximize the number of observations. But combining and using the original EBB and NEA survey weights is not a feasible strategy because of underlying methodological differences.

As the preceding examples illustrate, users of administrative register data who study inequality regularly encounter situations in which the data they are analyzing is only a subset of the population. In these situations, we recommend considering a customized weighting scheme, particularly when analyzing the data for descriptive purposes.

One viable strategy is to use variables for which the population margins are known as a benchmark. These population margins can be used to assess selectivity in the sample relative to the population in question and to construct custom weights if needed. These population margins can be computed for variables with exhaustive coverage. Examples of known population margins are key demographic characteristics of individuals, such as age, sex, and migration background. Known population margins are also available at the organizational level, such as size, industry, or sector.

Below, we present a routine based on the Stata package ipfraking written by Stan Kolenikov, which implements an iterative proportional fitting algorithm to construct and calibrate custom weights.[17][18] We outline three examples that consider an increasingly higher number of variables for which the population margins are known. Iterative proportional fitting is a particularly useful approach for constructing weights based on two or more population margins because this procedure avoids problems with sparse cells in high-dimensional tables.

```
* Applied examples of a customized weighting scheme with
* - poststratification weights (single margin)
* - iterative proportional fitting (2+ margins)
* using the ipfraking ado.

**********************
*Example 1
*Single margin: gender*
**********************

use "${wdir}/data/data_allyear", clear
keep if year==2022

* Setting up the totals
gen finalwgt=1
generate byte _one = 1
svyset _n
```

[17] Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal*, 14(1), 22-59. https://doi.org/10.1177/1536867X1401400104

[18] Kolenikov, S. (2019). Updates to the ipfraking ecosystem. *The Stata Journal*, 19(1), 143-184. https://doi.org/10.1177/1536867X19830912

```
version 14: svy: total _one, over(female, nolab)
matrix total_female = e(b)
matrix rownames total_female = female

* Obtaining the sample
gen sample = (rin_edu_isced2011!=.)
keep if sample==1
drop sample

* Calibrating the weights
ipfraking [pw=finalwgt], ctotal(total_female) replace

* Quality control
total _one [pw=finalwgt], over(female)
matrix list e(b), format(%12.0g)
matrix list total_female, format(%12.0g)

*****************************
*Example 2
*Two margins: age and gender*
*****************************

use "${wdir}/data/data_allyear", clear
keep if year==2022

*Age groups
egen agegr = cut(age), group(4)

* Setting up the totals
gen finalwgt=1
generate byte _one = 1
svyset _n
version 14: svy: total _one, over(female, nolab)
matrix total_female = e(b)
matrix rownames total_female = female
version 14: svy: total _one, over(agegr, nolab)
matrix total_agegr = e(b)
matrix rownames total_agegr = agegr

* Obtaining the sample
```

```
gen sample = (rin_edu_isced2011!=.)
keep if sample==1
drop sample

* Calibrating the weights
ipfraking [pw=finalwgt], ctotal(total_female total_agegr) replace

* Quality control
svyset, clear
svyset _n, weight(finalwgt)
svyset: tab agegr

************************************************
*Example 3
*Three margins: age, gender, migration background*
************************************************

use "${wdir}/data/data_allyear", clear
keep if year==2022

*Age groups
egen agegr = cut(age), group(4)

*Migration background
rename rin_miggen miggen

* Setting up the totals
gen finalwgt=1
generate byte _one = 1
svyset _n
version 14: svy: total _one, over(female, nolab)
matrix total_female = e(b)
matrix rownames total_female = female
version 14: svy: total _one, over(agegr, nolab)
matrix total_agegr = e(b)
matrix rownames total_agegr = agegr
version 14: svy: total _one, over(miggen, nolab)
matrix total_miggen = e(b)
matrix rownames total_miggen = miggen
```

```
* Obtaining the sample
gen sample = (rin_edu_isced2011!=.)
keep if sample==1
drop sample

* Calibrating the weights
ipfraking [pw=finalwgt], ctotal(total_female total_agegr total_miggen) replace


svyset, clear
svyset _n, weight(finalwgt)
```

6 Codebooks

Table 5 and Table 6 give an overview of all variables included in the NIDIO datasets. Users may consult these tables for an overview of the content of NIDIO and the clarification of specific variables covered by NIDIO.

- Table 5 lists all variables with a brief description of their content. It also lists the source data (i.e., the variables in the original CBS microdata files) that were used to derive a specific variable included in NIDIO.
- Table 6 lists the coding and value labels of categorical variables across NIDIO datasets. Some categorical variables are high-dimensional such as the ISCO-08 occupation classification, the SOI2021 field of study classification, or the official Dutch municipality codes. In these cases, links to detailed external code listings in original documentation material are provided in the table.

Table 5: Overview of variables by NIDIO dataset
(Note: Link variables in bold)

Variable	Description	Type	Original CBS variable
ABR: nidio_abr_og_register_2006_2023.dta			
year	**Calendar year**	**Numeric**	
ogid	**OG ID**	**Categorical**	**rog_identificatie; ogidentificatie**
og_sectorcode	Coordinated sector code of OG	String	rog_sectorcodegecoordineerd
og_sector	Harmonized and simplified sector code	Categorical	rog_sectorcodegecoordineerd
og_sector_alt	Longitudinally consistent sector code (2017-)	Categorical	rog_sectorcodegecoordineerd
og_ownership	OG ownership (non-financial firms & fin. institutions)	Categorical	rog_sectorcodegecoordineerd
og_start	Start of OG observation (since 01-07-2005)	Date	rog_datumontstaantoepassing
og_end	End of OG observation in calendar year (if applicable)	Date	rog_datumopheffingtoepassing
ABR: nidio_abr_og_size_2006_2023.dta			
year	**Calendar year**	**Numeric**	
ogid	**OG ID**	**Categorical**	**rog_identificatie; ogidentificatie**
og_employees	Number of employees in calendar year (from BE unit)	Numeric	
ABR: nidio_abr_ogbe_register_2006_2023.dta			
year	**Calendar year**	**Numeric**	
ogid	**OG ID**	**Categorical**	**rog_identificatie; ogidentificatie**
og_sectorcode	Coordinated sector code of OG	String	rog_sectorcodegecoordineerd
og_sector	Harmonized and simplified sector code	Categorical	rog_sectorcodegecoordineerd
og_sector_alt	Longitudinally consistent sector code (2017-)	Categorical	rog_sectorcodegecoordineerd
og_ownership	OG ownership (non-financial firms & fin. institutions)	Categorical	rog_sectorcodegecoordineerd

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
og_start	Start of OG observation (since 01-07-2005)	Date	rog_datumontstaantoepassing
og_end	End of OG observation in calendar year (if applicable)	Date	rog_datumopheffingtoepassing
beid	**BE ID**	**Categorical**	**rbe_identificatie; be_id**
be_start	Start of BE observation (since 01-07-2005)	Date	rbe_datumontstaantoepassing
be_end	End of BE observation in calendar year (if applicable)	Date	rbe_datumopheffingtoepassing
be_SBI93	Industry classification (SBI93) of BE	String	rbe_sbi93gecoordineerd
be_SBI08	Industry classification (SBI08) of BE	String	rbe_sbigecoordineerd
be_industry	Harmonized industry categories (1st digit of SBI08) of BE	Categorical	rbe_sbigecoordineerd
be_gksbs	Organization size of BE	Categorical	rbe_gksbsgecoordineerd
be_employees	Number of employees	Numeric	rbe_werkzamepersonen
be_lbe	Number of establishments within BE	Numeric	lbe
be_municipality_code	Municipality code (based on postcode)	String	gemcode
vepid	CBS Persoon ID (Kernpersoon BE)	Categorical	vep_identificatie_kernpersoon
vep_rechtvormcode	Legal form (via CBS kernpersoon) of BE	String	vep_rechtsvormcode
vep_legalform	Harmonized and simplified legal form	Categorical	vep_rechtsvormcode
vep_postcode_crypt	Postcode encrypted	String	postcode6_crypt
ogbe_start	Start of OG-BE link (since 01-07-2005)	Date	rop_datumontstaantoepassing
ogbe_end	End of OG-BE link in calendar year (if applicable)	Date	rop_datumopheffingtoepassing
ogbe_start_event	Event associated with start of OG-BE link	Categorical	vev_type; eventtype
ogbe_end_event	Event associated with end of OG-BE link	Categorical	vev_type; eventtype
ogbe_interruption	Within-year break of OG-BE link	Numeric	
ABR: nidio_abr_ogkvk_register_2006_2023.dta			
year	**Calendar year**	**Numeric**	

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
ogid	**OG ID**	**Categorical**	**rog_identificatie; ogidentificatie**
og_sectorcode	Coordinated sector code of OG	String	rog_sectorcodegecoordineerd
og_sector	Harmonized and simplified sector code	Categorical	rog_sectorcodegecoordineerd
og_sector_alt	Longitudinally consistent sector code (2017-)	Categorical	rog_sectorcodegecoordineerd
og_ownership	OG ownership (non-financial firms & fin. institutions)	Categorical	rog_sectorcodegecoordineerd
og_nrofvep	Number of CBS persoonen attached to OG	Numeric	
og_start	Start of OG observation (since 01-07-2005)	Date	rog_datumontstaantoepassing
og_end	End of OG observation in calendar year (if applicable)	Date	rog_datumopheffingtoepassing
vepid	CBS Persoon ID (Kernpersoon BE)	Categorical	vep_identificatie_kernpersoon
ogvep_start	Start of OG-VEP link (since 01-07-2005)	Date	rop_datumontstaantoepassing
ogvep_end	End of OG-VEP link in calendar year (if applicable)	Date	rop_datumopheffingtoepassing
ogvep_interruption	Within-year break of OG-VEP link	Numeric	
kvkid	Encrypted KVK dossier number	String	kvknr; ... (multiple)
finr	Encrypted fiscal ID	String	vep_finr; vep_finr_crypt
vepkvk_start	Start of VEP-KVK link (since 01-07-2005)	Date	vep_datumontstaantoepassing
vepkvk_end	End of VEP-KVK link in calendar year (if applicable)	Date	vep_datumopheffingtoepassing
ABR: nidio_abr_bekvk_register_2006_2023.dta			
year	**Calendar year**	**Numeric**	
beid	**BE ID**	**Categorical**	**rbe_identificatie; be_id**
be_nrofvep	Number of CBS persoonen attached to BE	Numeric	
be_start	Start of BE observation (since 01-07-2005)	Date	rbe_datumontstaantoepassing
be_end	End of BE observation in calendar year (if applicable)	Date	rbe_datumopheffingtoepassing
vepid	CBS Persoon ID (Kernpersoon BE)	Categorical	vep_identificatie_kernpersoon

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
bevep_start	Start of BE-VEP link (since 01-07-2005)	Date	rop_datumontstaantoepassing
bevep_end	End of OG-VEP link in calendar year (if applicable)	Date	rop_datumopheffingtoepassing
bevep_interruption	Within-year break of BE-VEP link	Numeric	
kvkid	Encrypted KVK dossier number	String	kvknr; ... (multiple)
finr	Encrypted fiscal ID	String	vep_finr; vep_finr_crypt
vepkvk_start	Start of VEP-KVK link (since 01-07-2005)	Date	vep_datumontstaantoepassing
vepkvk_end	End of VEP-KVK link in calendar year (if applicable)	Date	vep_datumopheffingtoepassing
BDK: nidio_bdk_be_2007_2023.dta			
year	**Calendar year**	**Numeric**	
beid	**BE ID**	**Categorical**	beid
be_start_bdk	Start of BE observation (BDK)	Date	datumonstaantoepassing_bdk
be_end_bdk	End of BE observation (BDK)	Date	datumopheffingtoepassing_bdk
be_start_abr	Start of BE observation (ABR)	Date	datumonstaantoepassing_abr
be_end_abr	End of BE observation (ABR)	Date	datumopheffingtoepassing_abr
be_founding	Founding date of organization	Date	oprichtingsdatumbedrijf
be_recht[...]code_bdk	Legal type of the organization (BDK)	Categorical	rechtsvorm
be_gksbs_bdk	Organization size of BE (BDK)	Categorical	grootteklasse
be_SBI08_bdk	Industry classification (SBI08) of BE (BDK)	Categorical	sbi
be_mkb	Middle-sized or small-sized company	Categorical	zelfstandigmkb
be_foreign	Foreign ownership	Categorical	buitenlands
be_uci	Country of Ultimate Controlling Institional Unit	String	uci
be_birth	Birth of organization during reference year	Categorical	oprichting
be_death	Death of organization during reference year	Categorical	opheffing

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
be_fastgrowth_bdk	Fast-growing comapny (BDK)	Categorical	snellegroeier
NFO: nidio_nfo_finances_2006_2023.dta			
year	**Calendar year**	**Numeric**	
finr	Encrypted fiscal ID	String	vep_finr_crypt
ogid	**OG ID**	**Categorical**	**ond_id**
source	Source of financial data	String	bron
oph	Reweighting factor at firm-level due to non-reporting	Numeric	oph
assets	Total balance (in 1000 Euros)	Numeric	b37
revenue	Net revenue (in 1000 Euros)	Numeric	r01
lcost	Labor costs (in 1000 Euros)	Numeric	r04
ccost	Capital costs (in 1000 Euros)	Numeric	r02
cdeprec	Capital depreciation (in 1000 Euros)	Numeric	r05_06
profit	Operating profit (in 1000 Euros)	Numeric	r07
result	Net result (in 1000 Euros)	Numeric	r20
NFO: nidio_nfo_og_2006_2023.dta			
year	**Calendar year**	**Numeric**	
ogid	**OG ID**	**Categorical**	**ond_id**
source	Source of financial data	String	bron
og_oph	Reweighting factor at firm-level due to non-reporting	Numeric	oph
og_assets	Total balance (in 1000 Euros) of OG	Numeric	b37
og_revenue	Net revenue (in 1000 Euros) of OG	Numeric	r01
og_lcost	Labor costs (in 1000 Euros) of OG	Numeric	r04
og_ccost	Capital costs (in 1000 Euros) of OG	Numeric	r02

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
og_cdeprec	Capital depreciation (in 1000 Euros) of OG	Numeric	r05_06
og_profit	Operating profit (in 1000 Euros) of OG	Numeric	r07
og_result	Net result (in 1000 Euros) of OG	Numeric	r20
SPOLIS: nidio_spolis_beid_register_2006_2023.dta			
year	**Calendar year**	**Numeric**	
beid	**BE ID**	**Categorical**	**beid; sbeid**
GBA: nidio_gba_rin_2023.dta			
rinpersoon	**Person ID**	**String**	**rinpersoon**
rin_cntbirth	Country of birth (respondent)	Categorical	gbageboorteland
rin_sex	Sex of respondent	Categorical	gbageslacht
rin_cntbirth_m	Country of birth (respondent's mother)	Categorical	gbageboortelandmoeder
rin_cntbirth_f	Country of birth (respondent's father)	Categorical	gbageboortelandvader
rin_nrprntsfrgnbrn	Number of parents born outside The Netherlands	Numeric	gbaaantaloudersbuitenland
rin_miggrp	Country of orgin (old CBS definition)	Categorical	gbaherkomstgroepering
rin_miggen	First or second generation migration background	Categorical	gbageneratie
rin_birthy	Year of birth (respondent)	Numeric	gbageboortejaar
rin_birthm	Month of birth (respondent)	Categorical	gbageboortemaand
rin_sex_m	Sex (respondent's mother)	Categorical	gbageslachtmoeder
rin_sex_f	Sex (respondent's father)	Categorical	gbageslachtvader
rin_birthy_m	Year of birth (respondent's mother)	Numeric	gbageboortejaarmoeder
rin_birthy_f	Year of birth (respondent's father)	Numeric	gbageboortejaarvader
rin_birthm_m	Month of birth (respondent's mother)	Categorical	gbageboortemaandmoeder
rin_birthm_f	Month of birth (respondent's father	Categorical	gbageboortemaandvader

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
rin_miggrp_imputed	Imputed country of origen	Categorical	gbaimputatiecode
rin_miggrp_cbs	Country of origin (new CBS definition)	Categorical	gbaherkomstland
rin_wrldrgn	World region (CBS typology)	Categorical	werelddeel2 (via Landaktueelref)
rin_wrldrgn_nidio	World region (NIDIO typology)	Categorical	
rin_wstrn	Western country (CBS typology)	Categorical	landtype (via Landaktueelref)
rin_nlbrn	Born in The Netherlands	Categorical	gbageboortelandnl
OPL: nidio_opl_rin_2006_2023.dta			
year	**Calendar year**	**Numeric**	
rinpersoon	**Person ID**	**String**	**rinpersoon**
edu_wgt	CBS education weight (if source=survey)	Numeric	gewichthoogsteopl
rin_edu_isced2011	Highest attained education ISCED2011 (via OPLNR)	Categorical	isced2011levelhb (via oplnrhb)
rin_edu_soi2016	Highest attained education SOI2016 (incl. CBS imputation)	Categorical	oplnivsoi2016agg4hbmetnirwo
rin_edufield_soi2016	Field of education SOI2016 (incl. CBS imputation)	Categorical	richtdetailiscedf2013hbmetnirwo
rin_edu_soi2021	Highest attained education SOI2021 (incl. CBS imputation)	Categorical	oplnivsoi2021agg4hbmetnirwo
rin_edufield_soi2021	Field of education SOI2021 (incl. CBS imputation)	Categorical	richtdetailiscedf2013hbmetnirwo
KIND: nidio_kindouder_parents_2023.dta			
rinpersoon	Person ID	String	rinpersoonma; rinpersoonpa
rin_nrchildren	Total number of registered children until current year	Numeric	
rin_mfchild	Legal role of parent	Categorical	
rinchild	Child ID	String	rinpersoon
child_birthy	Birth year of child	Numeric	gbageboortejaar
child_birthm	Birth month of child	Categorical	gbageboortemaand
childparent_link	CBS code describing means of established link	Categorical	xkoppelnummer

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
rinpalterp	Other Parent ID	String	rinpersoonma; rinpersoonpa
PARTNER: nidio_partnerbus_rin_allyears.dta			
rinpersoon	**Person ID**	**String**	**rinpersoon**
rinpersoonp	Partner ID	String	rinpersoonp
partnership_start	Start date partnership	Date	aanvangpartner
partnership_end	End date partnership	Date	aanvangpartner
EBB: nidio_ebb_occ_2006_2023.dta			
year	**Calendar year**	**Numeric**	
rinpersoon	**Person ID**	**String**	**rinpersoon**
rin_svydate_EBB	Survey date EBB (interview)	Date	sleutelebb
rin_svynr_EBB	Number of survey EBB (1-5)	Numeric	ebbstkpeilingnummer
rin_weight_EBB	Survey weight EBB (jaargewicht)	Numeric	ebbgewjaargewichta
rin_ISCO08	ISCO-08 occupation code	Categorical	ebbtw1isco2008v
NEA: nidio_nea_occ_2006_2023.dta			
year	**Calendar year**	**Numeric**	**jaar**
rinpersoon	**Person ID**	**String**	**rinpersoon**
rin_weight_NEA	Survey weight NEA (sample weights)	Numeric	weeg
rin_ISCO08	ISCO-08 occupation code	Categorical	isco08_unitgroup
rin_neaocc	Occupation codes NEA (-2010)	Categorical	beroep
SPOLIS: nidio_spolis_month_2006_2023.dta			
year	**Calendar year**	**Numeric**	
rinpersoon	**Person ID**	**String**	**(s)rinpersoon**

Continuation of Table 5

Variable	Description	Type	Original CBS variable
beid	**BE ID**	**Categorical**	**(s)beid**
baanrugid	Job ID POLIS (2006-2009)	String	baanrugid
ikvid	Job ID SPOLIS (2010-)	String	ikvid
slbaanid	Longitudinal Job ID	String	(via (S)POLISLONGBAANBUS)
scontractsoort	Type of contract	Categorical	(s)contractsoort
spolisdienstverband	Full-time / part-time employment	Categorical	(s)polisdienstverband
swekarbduurklasse	Working time category	Categorical	(s)wekarbduurklasse
scaosector	CAO sector	Categorical	(s)caosector
scao_crypt	Encrypted CAO code	String	scao_crypt
cao	CAO status	Categorical	scao_crypt
ssoortbaan	Job type	Categorical	(s)soortbaan
job_start_caly	Starting date of job within calendar year	Date	aanvbus; sdatumaanvangiko
job_end_caly	Ending date of job within calendar year	Date	eindbus; sdatumeindeiko
job_tenure	Overall starting date of job	Date	sdatumaanvangikvorg
sbaandagen_month	Job days (September)	Numeric	(s)baandagen
sbasisloon_month	Base pay (September)	Numeric	(s)basisloon
sbasisuren_month	Base working hours (September)	Numeric	(s)basisuren
sbijz[...]loning_month	Total extra compensation (September)	Numeric	(s)bizonderebeloning
sextrsal_month	End-of-year bonus (September)	Numeric	(s)extrasal
sincidentsal_month	Bonus pay (September)	Numeric	(s)incidentsal
slningld_month	Total pay (September)	Numeric	(s)lningld
slnowrk_month	Overtime compensation (September)	Numeric	(s)lnowrk
soverwerkuren_month	Overtime hours (September)	Numeric	(s)overwerkuren

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
sregul[...]uren_month	Regular working hours (September)	Numeric	(s)regulierenuren
svakbsl_month	Holiday allowance (September)	Numeric	(s)vakbsl
svoltijddagen_month	Full-time days (September)	Numeric	(s)voltijddagen
ft_factor	Full-time equivalent	Numeric	voltijddagen/baandagen
mainjob	Main job (most total pay)	Categorical	(via (S)POLISHOOFDBAANBUS)
SPOLIS: nidio_spolis_year_2006_2023.dta			
year	**Calendar year**	**Numeric**	
rinpersoon	**Person ID**	**String**	**(s)rinpersoon**
beid	**BE ID**	**Categorical**	**(s)beid**
baanrugid	Job ID POLIS (2006-2009)	String	baanrugid
ikvid	Job ID SPOLIS (2010-)	String	ikvid
slbaanid	Longitudinal Job ID	String	(via (S)POLISLONGBAANBUS)
scaosector	CAO sector	Categorical	(s)caosector
scao_crypt	Encrypted CAO code	String	scao_crypt
cao	CAO status	Categorical	scao_crypt
scontractsoort	Type of contract	Categorical	(s)contractsoort
ssoortbaan	Job type	Categorical	(s)soortbaan
job_start_caly	Starting date of job within calendar year	Date	aanvbus; sdatumaanvangiko
job_end_caly	Ending date of job within calendar year	Date	eindbus; sdatumeindeiko
job_tenure	Overall starting date of job	Date	sdatumaanvangikvorg
sbaandagen_year	Job days (Calendar year)	Numeric	(s)baandagen
sbasisloon_year	Base pay (Calendar year)	Numeric	(s)basisloon
sbasisuren_year	Base working hours (Calendar year)	Numeric	(s)basisuren

Continuation of Table 5			
Variable	**Description**	**Type**	**Original CBS variable**
sbijz[...]loning_year	Total extra compensation (Calendar year)	Numeric	(s)bizonderebeloning
sextrsal_year	End-of-year bonus (Calendar year)	Numeric	(s)extrasal
sincidentsal_year	Bonus pay (Calendar year)	Numeric	(s)incidentsal
slningld_year	Total pay (Calendar year)	Numeric	(s)lningld
slnowrk_year	Overtime compensation (Calendar year)	Numeric	(s)lnowrk
soverwerkuren_year	Overtime hours (Calendar year)	Numeric	(s)overwerkuren
sregul[...]uren_year	Regular working hours (Calendar year)	Numeric	(s)regulierenuren
svakbsl_year	Holiday allowance (Calendar year)	Numeric	(s)vakbsl
svoltijddagen_year	Full-time days (Calendar year)	Numeric	(s)voltijddagen
ft_factor	Full-time equivalent	Numeric	voltijddagen/baandagen
mainjob	Main job (most total pay)	Categorical	(via (S)POLISHOOFDBAANBUS)

Table 6: Codebook of categorical variables in NIDIO

Variable	Values
ABR: nidio_abr_og_register_2006_2023.dta	
og_sector	[11] Non-financial company
	[12] Financial organization
	[13] Governmental organization
	[15] Non-governmental non-profit organization
og_sector_alt	[11] Non-financial company
	[12] Financial organization
	[13] Governmental organization
	[15] Non-governmental non-profit organization
og_ownership	[0] n.a.
	[1] Public
	[2] Private
	[3] Foreign
ABR: nidio_abr_ogbe_register_2006_2023.dta	
og_sector	[11] Non-financial company
	[12] Financial organization
	[13] Governmental organization
	[15] Non-governmental non-profit organization
og_sector_alt	[11] Non-financial company
	[12] Financial organization
	[13] Governmental organization
	[15] Non-governmental non-profit organization
og_ownership	[0] n.a.
	[1] Public
	[2] Private
	[3] Foreign
be_SBI93	SBI93 documentation
be_SBI08	SBI08 documentation
be_industry	[1] Agriculture, forestry, and fishing
	[2] Mining and quarrying
	[3] Manufacturing
	[4] Electricity, gas, steam, and air conditioning supply
	[5] Water supply; sewerage, waste management and [...]
	[6] Construction

Continuation of Table 6	
Variable	**Values**
	[7] Wholesale and retail trade; [...]
	[8] Transportation and storage
	[9] Accomodation and food service activities
	[10] Information and communication
	[11] Financial institutions
	[12] Renting, buying, and selling of real estate
	[13] Consultancy, research and [..]
	[14] Renting and leasing of tangible goods and [...]
	[15] Public administration, public services, and [...]
	[16] Education
	[17] Human health and social work activities
	[18] Culture, sports, and recreation
	[19] Other service activities
	[20] Activities of households as employers
	[21] Extraterritorial organizations and bodies
be_gksbs	[0] 0 employees
	[10] 1 employee
	[21] 2 employees
	[22] 3-4 employees
	[30] 5-9 employees
	[40] 10-19 employees
	[50] 20-49 employees
	[60] 50-99 employees
	[71] 100-149 employees
	[72] 150-199 employees
	[81] 200-249 employees
	[82] 250-499 employees
	[91] 500-999 employees
	[92] 1000-1999 employees
	[93] 2000+ employees
be_municipality_code	Gemeentelijke Indelingen
vep_legalform	[1] Eeemanszaak
	[2] Eeemanszaak met meerdere eigenaren
	[5] Rederij
	[6] Maatschap
	[12] Vennootschap onder firma

Continuation of Table 6	
Variable	**Values**
	[25] Commanditaire vennootschap
	[35] Rechtspersoon in oprichting
	[43] Besloten vennootschap (bv)
	[57] Naamloze vennootschap (nv)
	[58] Europese naamloze vennootschap (se)
	[59] Europese cooperatieve vennootschap (sce)
	[67] Cooperatie
	[73] Kerkgenootschap
	[74] Stichting
	[77] Vereniging
	[87] Onderlinge waarborg maatschapij
	[90] Buitenlandse rechtsvorm
	[93] Europees economisch samenwerkingsverband
	[900] Verschillende publiekrechtelijke instellingen
	[998] Overige rrivaatrechtelijke Rechtspersoon
ogbe_start_event	[4] BE - Birth
	[6] BE - Merger
	[7] BE - Takeover
	[8] BE - Restructuring
	[9] BE - Demerger
	[10] BE - Breakup
	[12] BE - Combi Birth/Death
	[13] BE - Various
	[14] OG - Birth
	[16] OG - Merger
	[17] OG - Takeover
	[18] OG - Restructuring
	[19] OG - Demerger
	[20] OG - Breakup
	[21] OG - Combi Birth/Death
ogbe_end_event	[5] BE-Death
	[6] BE - Merger
	[7] BE - Takeover
	[8] BE - Restructuring
	[10] BE - Breakup
	[12] BE - Combi Birth/Death

Continuation of Table 6	
Variable	**Values**
	[13]BE - Various [15] OG - Death [16] OG - Merger [17] OG - Takeover [18] OG - Restructuring [19] OG - Demerger [20] OG - Breakup [21] OG - Combi Birth/Death
ABR: nidio_abr_ogkvk_register_2006_2023.dta	
og_sector	[11] Non-financial company [12] Financial organization [13] Governmental organization [15] Non-governmental non-profit organization
og_sector_alt	[11] Non-financial company [12] Financial organization [13] Governmental organization [15] Non-governmental non-profit organization
og_ownership	[0] n.a. [1] Public [2] Private [3] Foreign
BDK: nidio_bdk_be_2007_2023.dta	
be_rechtsvormcode_bdk	[1] Eenmanszaak [2] Maatschap, samenwerking [10] Vennootschap onder firma (VOF) [20] Commanditaire vennootschap (CV) [40] Besloten vennootschap (BV) [50] Naamloze vennootschap (NV) [60] Cooperative vereniging [70] Vereniging of stichting [900] Overheid [999] Overig of onbekend
be_gksbs_bdk	[0] 0 employees [10] 1 employee [21] 2 employees [22] 3-4 employees

Continuation of Table 6	
Variable	**Values**
	[30] 5-9 employees [40] 10-19 employees [50] 20-49 employees [60] 50-99 employees [71] 100-149 employees [72] 150-199 employees [81] 200-249 employees [82] 250-499 employees [91] 500-999 employees [92] 1000-1999 employees [93] 2000+ employees
	GBA: nidio_gba_rin_2023.dta
rin_cntbirth	Landaktueelref.pdf in RA
rin_sex	[1] Man [2] Woman
rin_cntbirth_m	Landaktueelref.pdf in RA
rin_cntbirth_f	Landaktueelref.pdf in RA
rin_miggrp	Landaktueelref.pdf in RA
rin_miggen	[0] Respondent & parents born in NL [1] First generation [2] Second generation
rin_sex_m	[1] Man [2] Woman
rin_sex_f	[1] Man [2] Woman
rin_miggrp_imputed	[0] Unchanged [0] Imputed [9] n.a.
rin_miggrp_cbs	Landaktueelref.pdf in RA
rin_wrldrgn	[1] Africa [2] America [3] Asia [4] Europe [5] Oceania
rin_wrldrgn_nidio	[1] West (Europe,US,Canada,AUS,NZ) [2] Asia and Oceania

Continuation of Table 6	
Variable	**Values**
	[3] Middle East +
	[4] Sub-Saharan Africa
	[5] Latin America and the Caribbean
rin_wstrn	[1] Western
	[2] Non-Western
rin_nlbrn	[0] Born outside NL
	[1] Born in NL
OPL: nidio_opl_rin_2006_2023.dta	
rin_edu_isced2011	[0] Early childhood education
	[1] Primary education
	[2] Lower secondary education
	[3] Upper secondary education
	[4] Post-secondary non-tertiary education
	[5] Short-cycle tertiary education
	[6] Bachelor's or equivalent level
	[7] Master's or equivalent level
	[8] Doctoral or equivalent level
rin_edu_soi2016	[1111] Basisonderwijs gr1-2
	[1112] Basisonderwijs gr3-8
	[1211] Praktijkonderwijs
	[1212] Vmbo-b/k
	[1213] Mbo1
	[1221] Vmbo-g/t
	[1222] Havo-, vwo-onderbouw
	[2111] Mbo2
	[2112] Mbo3
	[2121] Mbo4
	[2131] Havo-bovenbouw
	[2132] Vwo-bovenbouw
	[3111] Hbo-associate degree
	[3112] Hbo-bachelor
	[3113] Wo-bachelor
	[3211] Hbo-master
	[3212] Wo-master
	[3213] Doctor
rin_edufield_soi2016	SOI2016 documentation (p.20-33)

Continuation of Table 6	
Variable	**Values**
rin_edu_soi2021	[1111] Basisonderwijs gr1-2 [1112] Basisonderwijs gr3-8 [1211] Praktijkonderwijs [1212] Vmbo-b/k [1213] Mbo1 [1221] Vmbo-g/t [1222] Havo-, vwo-onderbouw [2111] Mbo2 [2112] Mbo3 [2121] Mbo4 [2131] Havo-bovenbouw [2132] Vwo-bovenbouw [3111] Hbo-associate degree [3112] Hbo-bachelor [3113] Wo-bachelor [3211] Hbo-master [3212] Wo-master [3213] Doctor
rin_edufield_soi2021	SOI2021 documentation (p.21-30)
KIND: nidio_kindouder_parents_2023.dta	
rin_mfchild	[1] Legal mother [2] Legal father
childparent_link	KINDOUDERTAB documentation (p.10-22).
EBB: nidio_ebb_occ_2006_2023.dta	
rin_ISCO08	ILO ISCO-08
NEA: nidio_nea_occ_2006_2023.dta	
rin_ISCO08	ILO ISCO-08
rin_neaocc	NEA 2009 documentation (p.93)
SPOLIS: nidio_spolis_month_2006_2023.dta	
scontractsoort	[0] Permanent Employment [1] Temporary Employment
spolisdienstverband	[1] Full-time [2] Part-time
swekarbduurklasse	[1] <12 hours [2] 12-<20 hours

Continuation of Table 6	
Variable	**Values**
	[3] 20-<25 hours
	[4] 25-<30 hours
	[5] 30-<35 hours
	[6] 35+ hours
scaosector	[1] Private
	[2] Non-profit
	[3] Government
cao	[0] No collective agreement
	[1] Sectoral agreement
	[2] Firm-level agreement
ssoortbaan	[1] Director / Large shareholder
	[2] Intern
	[3] WSW-er
	[4] Temporary agency worker
	[5] On-call worker
	[6] Standard
mainjob	[0] No
	[1] Yes
SPOLIS: nidio_spolis_year_2006_2023.dta	
scontractsoort	[0] Permanent Employment
	[1] Temporary Employment
scaosector	[1] Private
	[2] Non-profit
	[3] Government
cao	[0] No collective agreement
	[1] Sectoral agreement
	[2] Firm-level agreement
ssoortbaan	[1] Director / Large shareholder
	[2] Intern
	[3] WSW-er
	[4] Temporary agency worker
	[5] On-call worker
	[6] Standard
mainjob	[0] No
	[1] Yes