



# seriss

SYNERGIES FOR EUROPE'S  
RESEARCH INFRASTRUCTURES  
IN THE SOCIAL SCIENCES

Deliverable Number: 2.5

Deliverable Title: Report on auxiliary data in available country registers

Work Package: 2 – Representing the population

Deliverable type: Report

Dissemination status: Public

Submitted by: SHARE ERIC

Authors: Johanna Bristle (MEA), Sarah Butt (ESS ERIC HQ/CITY), Tom Emery (NIDI), Ruud Luijkx (TiU), Angelica M. Maineri (TiU), Senta-Melissa Pflüger (MEA), Annette Scherpenzeel (MEA), Stefan Zins (GESIS)

Date Submitted: November 2016

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 654221.





[www.seriss.eu](http://www.seriss.eu)  @SERISS\_EU

SERISS (Synergies for Europe's Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe's social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe's grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four year project (2015-19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey for Health Aging and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: Bristle, J., Butt, S., Emery, T., Luijck, R., Maineri, A.M., Pflüger, S-M., Scherpenzeel, A. and Zins, S. (2016) *Report on auxiliary data in available country registers*. Deliverable 2.5 of the SERISS project funded under the *European Union's Horizon 2020 research and innovation programme* GA No: 654221. Available at: [www.seriss.eu/resources/deliverables](http://www.seriss.eu/resources/deliverables)

# Auxiliary data in available country registers

## Content

Summary .....	4
1. Introduction.....	4
2. Auxiliary data in country registers.....	5
2.1 Survey descriptions.....	6
2.2 SERISS WP2 expert survey .....	10
2.3 Register documentation file .....	11
2.4 Findings .....	13
Availability.....	13
Access.....	15
Geographical identifiers.....	16
3. Auxiliary data sources at European level .....	18
3.1 European level data sources .....	18
3.2 Findings .....	20
Availability.....	20
Access.....	21
Geographical identifiers.....	21
4. Summary and Conclusion.....	22
References .....	24
Annex .....	24

# Summary

The potential for using auxiliary or contextual data for sample-based nonresponse adjustments recently gained more attention. However, identifying and accessing auxiliary data for nonresponse analysis, especially data which is of sufficiently high quality, presents a challenge. Data availability and access conditions can vary across countries and across organisations. This deliverable provides an inventory of auxiliary data that are available in registers used as sampling frames in the four major cross-European social surveys participating in SERISS: SHARE, ESS, GGP and EVS. Information is based primarily on findings from an expert survey among the researchers of these studies' country teams. Findings are augmented with information on auxiliary data sources on the European level. In addition to this summary report, an accompanying Excel file (SERISS\_WP2\_D2.5\_Annex2.xls) provides detailed information about the auxiliary data available in registers used by the four surveys across 24 countries. This resource provides an opportunity to compare and learn from the experiences of the four major cross-national surveys being conducted in Europe today and identify potential sources of auxiliary data for future use.

## 1. Introduction

This deliverable is part of SERISS work package 2 “Representing the population” which aims at developing ex ante and ex post strategies to obtain high-quality samples in social surveys. The second task of this work package explores the potential for exploiting the growing amount of pre-existing administrative data to better understand and overcome survey non-response and non-response bias. Understanding and reducing survey non-response is considered an ex ante solution to having a representative sample of the population, while using auxiliary data that is available for respondents as well as non-respondents can help to improve weighting strategies as a post-survey solution to nonresponse. Researchers from the four large cross-national studies united in SERISS work jointly on this aim: the European Social Survey (ESS), the European Values Study (EVS), the Gender and Generations Program (GGP), and the Survey of Health, Ageing, and Retirement in Europe (SHARE).

The potential for using auxiliary or contextual data for sample-based nonresponse adjustments recently gained more attention (Kreuter, 2013; Krueger & West, 2014; Olson, 2013). To be useful for sample-based nonresponse adjustments, auxiliary data have to fulfil two conditions: Firstly, auxiliary data need to be available for respondents and non-respondents (Groves, 2006). This is only the case if they are collected for all cases in the gross sample either directly through additional information from the sampling frames, through interviewer screening, the collection of process data (paradata) on e.g. call sequences, or through using contextual data from external data sources e.g. administrative records from other sources than the sampling frame or high-quality survey data, e.g. census information. Secondly, to be effective for nonresponse bias adjustments, auxiliary data need to be predictive for response propensities and correlate with substantive survey variables of interest (Groves, 2006).

Identifying and accessing auxiliary data for nonresponse analysis, especially data which is of sufficiently high quality, presents a challenge (Olson, 2013; Smith & Kim, 2013). Few studies

systematically assess the availability of auxiliary data for nonresponse analyses and those that do tend to be limited to a single country, often the USA. For example, Czajka (2013) explicitly assessed a compilation of administrative data sources (e.g. tax records) for nonresponse purposes and elaborated on the issues of coverage, unit of observation, data quality and completeness, timeliness, and legal restrictions. He stated that administrative records are of high accuracy for variables that directly serve the purpose of the register, e.g. payment amounts in tax records, but are of low accuracy if they are collected in secondary order, e.g. race (Czajka, 2013). Access to these data is often legally restricted. In the case of the US Census Bureau, the administrative records can only be used in-house and the use for nonresponse analyses is limited to surveys that are conducted by the Census Bureau (ibid.).

Poulain and Herm (2013) recently provided an overview about the availability of central population registers in European countries with a focus on the historical development, the link between census surveys and population registers, and opportunities of utilising register data for providing demographic statistics. Registers of Scandinavian countries can, under specific conditions, be accessed for research purposes on nonresponse and other areas for example (Olsen, 2011). The author declares the need for international, equal and free access possibilities for researchers.

In this report we add to the literature based on the experiences of the four major cross-European surveys (ESS, EVS, GGP and SHARE). The aim of this deliverable is to provide an inventory of auxiliary data that are available in registers used as sampling frames in the four major cross-European social surveys participating in SERISS. We will answer the following research questions:

- a) What auxiliary data are available from sampling frames used in the four major cross-European surveys of the moment?
- b) How are they accessible in practice?
- c) Which geographical identifiers are available in the registers?
- d) What are the overlaps and differences across surveys and across countries?
- e) What external data sources are available that provide harmonized contextual information at a European level and what do they offer?

For this purpose, we fielded an expert survey about the use of sampling frames and auxiliary data for surveys, to be answered by the researchers who are responsible for the sampling and data collection in the countries included in the four large surveys. All findings reported in the second chapter are based on experiences from sampling experts in the countries of the four studies. The third chapter comprises a compilation of auxiliary data available on the European level, collected and provided by Eurostat and Census HUB. These data sources provide small-area contextual data. In both chapters, we evaluate availability and accessibility of auxiliary variables chosen with regard to their potential for nonresponse analyses.

## 2. Auxiliary data in country registers

This chapter is based on results from the SERISS WP2 expert survey, which has been conducted among country teams of the European Social Survey (ESS), the European *Values* Study (EVS), the Gender and Generations Program (GGP), and the Survey of Health, Ageing,

and Retirement in Europe (SHARE). Details about the four studies, the expert survey as well as the main findings can be found in the following sections. Results from the survey are summarized along the dimensions auxiliary content of the registers, access opportunities, and geographical identifiers available for linkage.

## 2.1 Survey descriptions

### *European Social Survey (ESS)*

The European Social Survey (ESS) is an academically driven cross-national survey of public attitudes and behaviour that has been carried out in 36 European countries since 2001 (also see <http://www.europeansocialsurvey.org>).

The main aims of the ESS are to chart stability and change in social structure, conditions and attitudes in Europe and to interpret how Europe's social, political and moral fabric is changing; to achieve and spread higher standards of rigour in cross-national research in the social sciences, including for example, questionnaire design and pre-testing, sampling, data collection, reduction of bias and the reliability of questions; to introduce soundly-based indicators of national progress, based on citizens' perceptions and judgements of key aspects of their societies; to undertake and facilitate the training of European social researchers in comparative quantitative measurement and analysis; and to improve the visibility and outreach of data on social change among academics, policy makers and the wider public.

The ESS is a nationally representative cross-sectional general population survey. Every two years face-to-face interviews are conducted with a fresh representative sample of the resident population aged 15 and over and living in private households. Consisting of a core questionnaire that remains the same in every round alongside round-specific rotating modules, the survey covers a range of topics including: satisfaction with democracy, political engagement, health and wellbeing, social norms, attitudes to immigration, work and family, and attitudes to key areas of public policy including immigration, welfare state and energy and climate change. The ESS is carried out to the highest methodological standards and aims for optimal comparability or equivalence in all stages of data collection and processing. Detailed guidelines on implementation are provided in the [Specification for Participating Countries](#).

Seven rounds of the survey have been conducted to date with ESS Round 8 entering the field in September 2016. ESS data are available free of charge for non-commercial use and can be downloaded from the study website. ESS data are widely used by students, academic researchers, policymakers and journalists to inform debate on the key challenges facing European societies.

The leading principal of ESS sampling is that the selection has to be based on a probabilistic rule and that the sample sizes have to be planned in a way that makes estimates across countries comparable, by ensuring a similar level of precision, i.e. similar sampling errors of estimates. Thus convenience sampling and quota sampling are not permitted and the inclusion probabilities of the respondents should be known and calculable.

### *European Values Study (EVS)*

The European *Values Study* is a large-scale, cross-national, and longitudinal survey research program that investigates basic human values. The project originated in the 1970s, aiming at describing value differences, similarities, and changes within Europe. The study covers different value fields, such as life, family, work, religion, politics, and society (also see <http://www.europeanvaluesstudy.eu/>).

The study was fielded for the first time in 1981, and ever since it has been replicated every nine years (1990, 1999, 2008; the next wave will be fielded in 2017). In 1981, the study only covered 14 European countries, USA, and Canada; wave by wave, the survey research program has enlarged, reaching no less than 47 countries in wave 4 (2008), for a total of almost 70000 people interviewed. Thanks to its extensiveness in terms of time and country coverage, the study is therefore extremely valuable to researchers who want to investigate the stability and changes in values in Europe over time, as well as similarities and differences across countries.

The population to be covered by data collection is broadly defined as individuals aged 18 or older, with no upper age limits, which have address of residence in the country at the date of the beginning of fieldwork, regardless of nationality and citizenship or language. As for sampling, general common guidelines are set by the Methodology Group, which takes care of the quality of the project; national sampling procedures are then defined by each country team, under the general rule of probability sampling, and approved by the Methodology Group. In each country, between 1000 and 1500 respondents are interviewed at each wave.

Documentation on methodology, sampling procedures, questionnaires, as well as data are made available for download and free of charge on the GESIS Data Archive for the Social Sciences in Cologne, Germany. Between 2008 and 2015, data have been downloaded more than 30000 times. Moreover, a collection of articles and books using EVS data, the EVS Bibliography, currently contains more than 1300 publications, and can be accessed through the website.

### *Generation and Gender Program (GGP)*

In a nutshell, the GGP is best defined as a “harmonized, large-scale, longitudinal, cross-national panel study of individuals & contextual database” (also see <http://www.ggp-i.org/>). It is a longitudinal panel study covering the whole life course from 18 to 79 years of age. It collects both retrospective information on topics such as fertility, family formation and dissolution, as well as prospective information collected through subsequent waves of the survey, allowing us to see changes in people’s lives over time. It is a large-scale project involving data collection from about 10000 individuals per country (including both men and women). Such large sample sizes are necessary to study specific population subgroups such as migrants or people at the extreme ends of the income distribution, as well as to capture a sufficiently large number of life-events for statistical analyses.

The GGP is also a cross-national project currently covering 19 countries with data harmonized in a large database for cross-national comparisons. Moreover, 15 of these 19 countries have carried out subsequent waves of data collection (on the same individuals) allowing us to see changes over time in a variety of contexts.

The GGP covers a wide range of topics and collects data on: fertility and partnership histories, transition to adulthood, work-family balance, gender relations and gender division of housework, intergenerational exchange including informal and formal care, well-being and health, grandparenthood, and economic activity and retirement. Finally, the micro-level data

are also complemented by a contextual database providing information about policies and the economic environment at the regional and country-level that may affect individuals. Our Contextual Database contains information on more than 200 harmonized indicators, and tracks population trends and policy changes in 60 countries over the past 40 years. Both as a stand-alone tool for analysis and as a supplement to the individual-level database, this dataset is a powerful analytical component of the infrastructure which enables us to understand individuals' relationships and personal histories in the context of policy developments and social change.

As a research infrastructure, the GGP is built on the principle of open access. Micro-level data can be downloaded directly from the web after researchers have been granted access through a centralized registration and accreditation process. Meta-information and online analysis is possible for anyone through the NESSTAR system. The number of registered users for the GGP micro-level data has increased rapidly over the years and has now exceeded 3,200. As of 2016, the GGP is an emerging community of the ESFRI roadmap and aims to join the roadmap as a full member by 2020. In 2019, the GGP will be commencing with a new round of data collection and seeking to build on its current success through new innovative data collection methodologies and an expansion to countries that have thus far not participated in the program. The GGP is thus entering a new and exciting phase of its work at a time when the questions it seeks to answer have never been more relevant.

#### *Survey of Health, Ageing and Retirement in Europe (SHARE)*

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of approximately 123000 individuals aged 50 or older (more than 293000 interviews). SHARE's scientific power is based on its panel design that grasps the dynamic character of the ageing process. SHARE's multi-disciplinary approach delivers the full picture of the ageing process. Rigorous procedural guidelines and programs ensure an ex-ante harmonized cross-national design (also see [www.share-eric.eu](http://www.share-eric.eu)).

The first SHARE data collection started in 2004 and has been repeated bi-annually. With the most recent extension SHARE now covers 26 countries of the European Union as well as Switzerland and Israel. Jointly with harmonized data from the English Longitudinal Study of Ageing (ELSA) and the Irish Longitudinal Study on Ageing (TILDA), pan-European research on effects of our ageing societies and their implications can be extended to all EU countries. The data collection with all new members will be wave 7 of SHARE and start in 2017.

In wave 3, the SHARELIFE questionnaire collected retrospective life histories. This will be repeated in wave 7 for respondents who were not yet part of the panel in wave 3. SHARELIFE collects individual micro data over the respondents' entire life and can be linked with institutional macro data on the welfare state. It thereby allows assessing the full effect of welfare state interventions on the life of the individual. Changes in institutional settings that influence individual decisions are of specific interest to evaluate policies throughout Europe. Context variables were collected as a data set of institutional information on the welfare state in Europe. They span a period from 1960 to 2009, over the participating countries and were compiled through national efforts.

Sampling in SHARE is based on probability samples in all participating countries. SHARE represents the non-institutionalized population aged 50 and older. Spouses are also interviewed if they are younger than 50.

SHARE data are available for researchers free of charge and are very well used: about two publications based on SHARE data are published every week and more than 6000 researchers are registered as users. The publications cover economic, health and social science topics from retirement saving to active ageing and care at the end of life.

## 2.2 SERISS WP2 expert survey

To construct a comprehensive overview of the sampling frames which are used in the four studies and to create an inventory of the availability of auxiliary variables in these sampling frames, an expert survey was carried out. The expert survey asked about the name and type of register actually used for sampling, the responsible authority, the register's accessibility for different researchers and organisations, the amount of time it took to obtain a sample from it, the problems encountered, and the auxiliary variables obtainable from it. In addition, questions were included enquiring about other sources of auxiliary data that were used. The questionnaire was programmed as an electronic form and was sent by email to the country teams of each of the four studies, accompanied by an official invitation letter signed by the director of the respective study. The response rate of country teams and experts was 95% in SHARE and about 75% in the other three studies. Results on the use of registers as sampling frames and further methodological details of the survey can be found in Deliverable 2.1 of this work package, titled "The use of sampling frames in European studies" (Scherpenzeel et al. 2016). Annex 1 shows the questionnaire.

Out of the 81 country representatives that returned questionnaires, 50 indicated that a register had been used for sampling purposes in their study. Registers here include population or civil registers as well as electoral registers, health insurance registers or any other register covering the population. No use of telephone registers was reported in any of the countries' questionnaires. Thirty-one responses indicated that a geographical listing, a random route procedure or other sampling frames had been used.

With regard to the use of registers, there is little variation within countries about which register to use. In 21 out of the 24 countries where registers are used, all studies reported to use the same register. Table 1 summarizes the register names by country and studies using this register (based on the registers available in the SERISS expert survey).

In this report, results are reported for the subset of countries which used register data for sampling (50 responses across the four studies). Results are presented separately for the four studies. This provides a way to summarise the results while avoiding overlap from multiple uses of the same register across studies. It also recognises that there is some inconsistency across studies in the information provided on a given register. This inconsistency might partly be due to recall error, but might also point to variation in the amount of information provided to the studies. Each survey will only receive the auxiliary variables they explicitly ask for and might not be aware of the total amount of available variables in the registers. A further explanation for discrepancies in reporting might be the year of sampling. We collected information about sampling procedures for data collections that range from 2004 (wave 1 of GGP and SHARE) to 2017 (upcoming wave 7 of SHARE, wave 5 of EVS). Conditions might change over time. The drawback of presenting the results by study is the small number of cases per study, ranging from 7 to 18 cases. To find out more about registers used in specific countries we encourage the reader to make use of the detailed register overview in the Excel file "SERISS\_WP2\_D2.5\_Annex2.xls", which is the major contribution of this deliverable (available via [www.seriss.eu/resources/deliverables](http://www.seriss.eu/resources/deliverables)).

Table 1: Names of registers used in at least one of the four studies per country

	Register names	Studies using this register
<b>Austria (AT)</b>	(1) Data Door (2) Zentrales Melderegister (ZMR)	(1) SHARE (2) ESS, GGP
<b>Belgium (BE)</b>	Rijksregister/Registre national	ESS, EVS, SHARE
<b>Croatia (HR)</b>	Registar osiguranika Hrvatskog zavoda za zdravstveno osiguranje	SHARE
<b>Czech Republic (CZ)</b>	Census, Vital Statistics	GGP
<b>Denmark (DK)</b>	CPR	EVS, SHARE
<b>Estonia (EE)</b>	Population register	ESS, EVS, SHARE
<b>Finland (FI)</b>	Population Information System.	ESS
<b>France (FR)</b>	Echantillon-maître (master sample based on the rolling population Census)	SHARE
<b>Germany (DE)</b>	Einwohnermeldeamtregister	ESS, EVS, SHARE
<b>Hungary (HU)</b>	Népszámlálástartó (Population Register)	ESS, GGP
<b>Iceland (IS)</b>	Þjóðskrá / National Register	ESS, EVS
<b>Israel (IL)</b>	population registry	SHARE
<b>Italy (IT)</b>	(1) Register of electoral lists (2) Anagrafe	(1) EVS, SHARE (2) GGP
<b>Luxembourg (LU)</b>	Social security registry	SHARE
<b>Macedonia (MK)</b>	State Statistical Office of the Republic of Macedonia	EVS
<b>Malta (MT)</b>	Electoral register	EVS
<b>Netherlands (NL)</b>	BRP (former GBA):key register of persons	GGP, SHARE
<b>Norway (NO)</b>	Central Population Register	ESS, GGP
<b>Poland (PL)</b>	PESEL (Powszechny Elektroniczny System Ewidencji Ludności)	ESS, SHARE
<b>Portugal (PT)</b>	Register from the National Health Service	SHARE
<b>Slovenia (SI)</b>	The Slovenian Central Population Register	SHARE
<b>Spain (ES)</b>	Municipal Population Register	ESS, SHARE
<b>Sweden (SE)</b>	(1) Navet (2) Total Population Register (RTB)	(1) ESS, SHARE (2) GGP
<b>Switzerland (CH)</b>	SRPH - Stichprobenrahmen für Personen- und Haushaltserhebungen	ESS, EVS, SHARE

Source: SERISS expert survey on sampling frames

### 2.3 Register documentation file

A detailed documentation of all registers reported as being used on ESS, EVS, GGP or SHARE can be found in the Excel file “SERISS\_WP2\_D2.5\_Annex2.xls” (screenshot provided in Figure 1) which is available via [www.seriss.eu/resources/deliverables](http://www.seriss.eu/resources/deliverables). The documentation file lists information separately for each register used for the studies in countries that filled in the SERISS WP2 questionnaire. This file is the core outcome of this deliverable and has already been used in practice for establishing sampling procedures and choosing appropriate sampling sources in countries recently joining SHARE. Making the documentation file available to the public facilitates survey practitioners and researchers who need to decide on sampling strategies to compare current sampling practices and registers in use across the four major social science surveys in Europe and base their decisions on the manifold experiences of the countries participating in ESS, EVS, GGP and SHARE.

As the tables containing the full results of the expert survey are too large to include in the main body of this report, results are delivered as an Excel file supplementary to the report. We will however, summarize some of the information that can be found in detail in the following sections.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Sheet 1: Availability of auxiliary data in registers											
2												
3	Country	Study	Register	Year of data	Age	Gender	Ethnicity	Country o	Education	Income	Occupatic	Emp
4	AT	ESS	Zentrales Melderegiste	2014	Yes	Yes	No	Yes	No	No	No	No
5	AT	GGP	Zentrales Melderegiste	2012	Yes	Yes	No	Yes	No	No	No	No
6	AT	SHARE	Data Door	2011	No	No	No	No	No	No	No	No
7	BE	ESS	Rijksregister/Registre	2014	Yes	Yes	No	Yes	No	No	Yes	No
8	BE	EVS	Rijksregister	2017	Yes	Yes	No	Yes	No	No	No	No
9	BE	SHARE	Registre National / Rijl	2017	Yes	Yes	No	Yes	No	No	No	No
10	CH	ESS	SRPH - Stichprobenra	2016	Yes	Yes	No	Yes	No	No	No	No
11	CH	EVS	SRPH - Stichprobenra	2017	Yes	Yes	No	No	No	No	No	No
12	CH	SHARE	SRPH - Stichprobenra	2017	Yes	Yes	No	Yes	No	No	No	No
13	CZ	GGP	Census, Vital Statistic	2008	Yes	Yes	No	Yes	Yes	No	Yes	Yes
14	DE	ESS	Einwohnermelderegist	2016	Yes	Yes	No	No	No	No	No	No
15	DE	EVS	Einwohnermeldeamt	2017	Yes	Yes	No	No	No	No	No	No
16	DE	SHARE	Melderegister der Gen	2015	Yes	Yes	No	No	No	No	No	No
17	DK	EVS	CPR	2017	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
18	DK	SHARE	CPR	2013	Yes	Yes	Yes	Yes	No	No	No	No
19	EE	ESS	Population register	2014	Yes	Yes	Yes	Yes	No	No	No	No
20	EE	EVS	Rahvastikuregister / in	2017	Yes	Yes	No	No	No	No	No	Yes
21	EE	SHARE	Population register	2017	Yes	Yes	Yes	Yes	Yes	No	No	No
22	ES	ESS	Padrón Municipal de H	2016	Yes	Yes	No	Yes	Yes	No	No	No
23	ES	SHARE	Padrón Continuo	2011	Yes	Yes	No	Yes	Yes	No	No	No
24	ES-Girona	SHARE	Padró municipal d'hab	2013	Yes	Yes	No	Yes	No	No	No	No
25	FI	ESS	Population database w	2016	Yes	Yes	Yes	Yes	Yes	No	No	No
26	FR	SHARE	Echantillon-maitre (ma	2017	Yes	Yes	No	Yes	Yes	No	Yes	Yes
27	GE	EVS	Voter's list	2017	Yes	No	No	No	No	No	No	No
28	HR	SHARE	Registar osiguranika H	2015	Yes	Yes	No	No	No	No	No	No
29	HU	ESS	Népszégnnyilvántartó	2014	Yes	Yes	No	Yes	No	No	No	No
30	HU	GGP	Népszégnnyilvántartó	2004	Yes	Yes	No	Yes	No	No	No	No
31	IL	ESS		2016	Yes	Yes	No	Yes	No	No	No	No
32	IL	SHARE	population registry	2015	Yes	Yes	No	Yes	No	No	No	No
33	IS	ESS	Þjóðskrá. National Re	2012	Yes	Yes	No	Yes	No	No	No	No
34	IS	EVS	Þjóðskrá / National Re	2017	Yes	Yes	No	Yes	No	No	No	No
35	IT	EVS	Register of electoral lis	2008	Yes	Yes	No	No	No	No	No	No
36	IT	GGP	Anagrafe	2004	Yes	Yes	No	No	No	No	No	No
37	IT	SHARE	Registri elettorali sez	2015	Yes	Yes	No	No	No	No	No	No
38	LU	EVS	Social security registr	2017	No	No	No	No	No	No	No	No
39	LU	SHARE	Social Security Regist	2015	Yes	Yes	No	Yes	No	No	No	No
40	MK	EVS	Завод за статистика н	2017	Yes	Yes	No	No	No	No	No	No
41	MT	EVS	Electoral register	2017	Yes	No	No	No	No	No	No	No
42	NL	GGP	BRP (former GBA):key	2004	Yes	Yes	No	Yes	No	No	No	No
43	NL	SHARE	Formerly known as: G	2004	Yes	Yes	No	No	No	No	No	No
44	NO	ESS	Central Population Re	2016	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
45	NO	GGP	Central Population Re	2004	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
46	PL	ESS	PESEL (Powszechny	2016	Yes	Yes	No	Yes	No	No	No	No
47	PL	SHARE	Rejestr PESEL (Pows	2017	Yes	Yes	No	No	No	No	No	No
48	PT	SHARE	Register from the Nati	2015	Yes	Yes	No	No	No	No	No	No
49	SE	ESS	Navet	2016	Yes	Yes	No	Yes	No	No	No	No
50	SE	EVS		2017	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
51	SE	GGP	Total Population Regis	2004	Yes	Yes	No	Yes	Yes	Yes	No	Yes
52	SE	SHARE	Navet	2017	Yes	Yes	No	Yes	No	No	No	No
53	SI	SHARE	The Slovenian Central	2015	Yes	Yes	No	Yes	No	No	No	No
54												

Figure 1: Screenshot of register documentation file “SERISS\_WP2\_D2.5\_Annex2.xls”

## 2.4 Findings

### *Availability*

Three questions in the expert survey asked about the availability of auxiliary variables in the register data. Nonresponse literature has shown several indicators that might influence respondent's survey participation and nonresponse in a survey. Groves and Couper (1998) grouped them into four domains: social environment, household/respondent characteristics, survey design and interviewer characteristics. While survey design and interviewer characteristics can be influenced by the researcher, the environment and respondent characteristics are categorized as being "out of researchers' control" (ibid.). Therefore, it is even more important to find auxiliary variables in registers and administrative data that contribute to understanding the challenge of survey nonresponse. The characteristics asked about in the WP2 expert survey are based on previous literature (Bethlehem, Cobben, & Schouten, 2011; Groves & Couper, 1998), but also chosen with regard to the topics of the four studies under investigation (e.g. health, disability and death is especially relevant for a survey on ageing).

Figure 2 shows the availability of auxiliary variables in three domains: socio-demographic characteristics, household characteristics and neighbourhood characteristics. The bars show the percentage of the registers containing this information per study. Regarding socio-demographic characteristics, age and gender are typically retrieved with the sampling frame and available in all registers used for the ESS and GGP, all but one in SHARE and most in the EVS. Overall, the availability of further variables is limited, but many registers contain information on country of birth, marital status and death records. Not listed, but mentioned in the other category, many registers contain information about citizenship and some on migration. Household characteristics are shown in the second part of the figure. Half of the registers contain information on other persons living in the household like a partner or children. Other characteristics are available less frequently. In all studies some registers contain information on building type and household type. In addition, the register in Switzerland contains information on household size and has been used by three of the studies in our sample (ESS, EVS and SHARE). The availability of neighbourhood characteristics in registers is not so promising with many countries stating that there is no information of this kind in the register. For EVS and GGP information on population density can be retrieved from the register in 50% of cases (6 countries from the EVS and 3 from GGP), slightly less for degree of urbanization, unemployment rate, percentage of non-natives. Most of the country representatives using registers (80%) stated that this kind of information can be retrieved from other sources, predominantly from national or federal statistical offices, commercial data providers, other surveys, census data or interviewer observations. However, it varies on which geographical level the data are then available.

A detailed overview about indicators available for each register in the sample is listed in the Excel file in sheet "1\_Availability". An overview of individual characteristics in central population registers, based on a survey from 2010 in 20 European countries, can also be found in Poulain and Herm (2013). This article includes the availability of variables on birth, sex, citizenship, occupation, education, residence and migration.

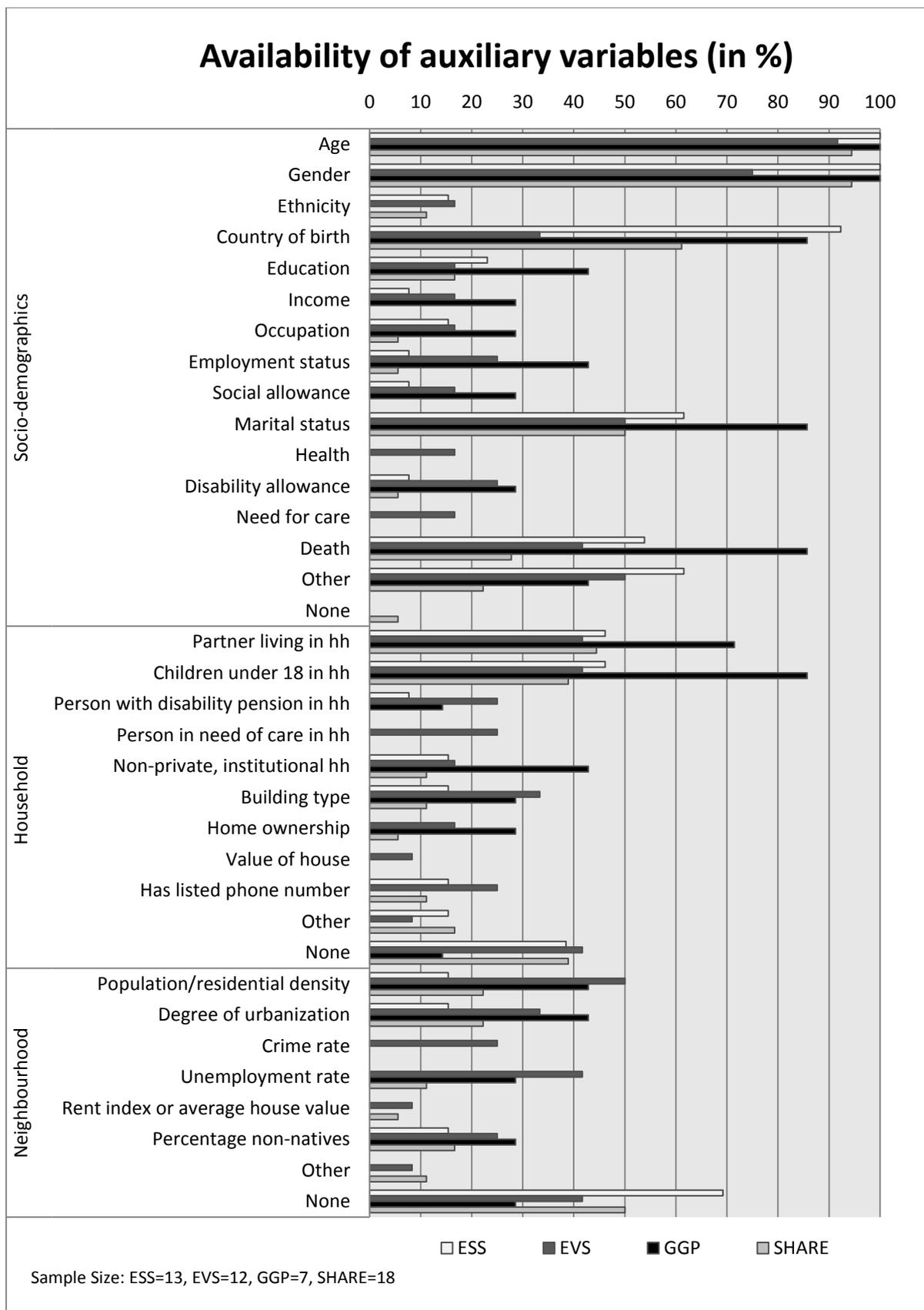


Figure 2: Availability of auxiliary variables

## Access

Access to registers is often restricted and sometimes they can only be accessed by statistical offices. Poulain and Herm (2013) emphasize the potential of administrative data for research in the fields of demography, sociology, epidemiology, but also state that “in practice, however, access is difficult and the obstacles encountered are a matter of concern” (ibid, p.205). The authors state in which areas of research register data are used, but they do not elaborate on how researchers gained access to these registers. In the SERISS questionnaire we therefore included questions on whether access is possible only nationally or internationally and further, who can obtain access to the auxiliary variables in the registers.

Overall, 40% of respondents to our questionnaire did not know about access conditions for using auxiliary data on the register. Out of the 30 informative answers we obtained, only in six countries international access is allowed. The registers providing both international and national access are from Austria, Czech Republic, Germany, Denmark, Estonia and Malta.

A further distinction was made according to the affiliation of researchers who are allowed to gain access to registers to use auxiliary data for research purposes (see Figure 3). Over all studies and responses, universities can gain access to these data in 21 different countries, non-commercial research institutes and affiliates in 20 different countries, 20 countries provide access for statistical offices. Fewer countries have registers where access is allowed for commercial survey organisations (10), commercial marketers (8), or others like private persons, governments, or anyone who is interested (4). In registers of 7 countries, no use is possible at all. The most generous registers can be found in Bulgaria, Germany, Denmark, Iceland, Montenegro, and Serbia, with providing access to all the user groups mentioned. The most restrictive registers with regard to access to user groups can be found in Croatia, Italy, and Portugal with no use at all. Details on the specific registers in all countries can be found in sheet “02\_Access” in the Excel file “SERISS\_WP2\_D2.5\_Annex2.xls”.

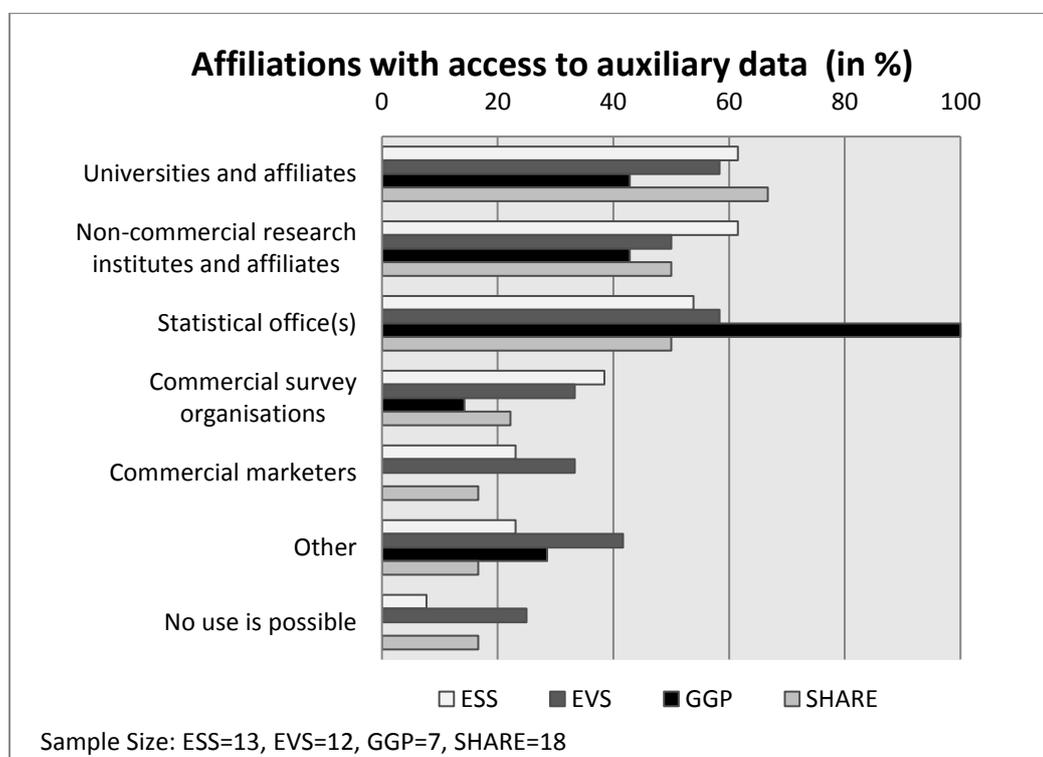


Figure 3: Affiliations with access to auxiliary data (multiple answers)

### *Geographical identifiers*

Auxiliary data can only be used for analysing survey nonresponse if sampling and other auxiliary data can be linked. If auxiliary data are obtained from the same register that is used for sampling, this is straightforward as long as there are not legal restrictions on use. However, based on answers from the WP2 expert survey, we know that most contextual information used by ESS, EVS, GGP and SHARE is not included in the register used for sampling, but can be obtained from other registers or other surveys (mostly) hosted by the national statistical offices. A prerequisite for the linkage to the sampling data then is the availability of geographical codes that then can be used for matching the two data sources.

In the questionnaire, we therefore queried the availability in the sampling registers of geographical codes ranging from postal code as a very small unit to NUTS 1 as a very broad regional code (e.g. in Germany, NUTS 1 are the federal states). The Nomenclature of Territorial Units for Statistics (NUTS) is a statistical system for dividing European regions into economic regions on different levels (Eurostat, 2015). We asked for the availability and usability of the three regional levels of the NUTS classification as well as Local Administration Units (LAU) which are smaller units than the NUTS and refer to municipalities or equivalent level. The codes reported in Figure 4 and Figure 5 are geographically nested and sorted by magnitude from the smallest unit asked about “postal code” to the broadest unit asked about “NUTS-1”. Other often refers to municipality codes or providing the whole address (from which all other geographical codes can be derived). “None” means that none of the categories asked about was indicated as being available.

The postal code is the geographical code most often included in the register data. 40 % of the ESS and GGP registers, almost 60 % of the EVS, and 80 % of the SHARE registers contain this information. Although it is available in registers from 18 different countries overall, the postal code can be used for matching survey data only in 11 countries. These generous conditions apply to registers from Austria, Belgium, Germany, Finland, Hungary, Iceland, Italy, Luxembourg, Poland, Sweden and Slovenia. The respondents of the expert survey stated that the geographical level at which auxiliary data from statistical offices is provided varies a lot across countries and across indicators. Assuming the external data is provided LAU-2 level, external contextual information can be linked to the sampling register for 42% of the ESS registers, 50% of the EVS registers, 72% of the GGP and 44% of the SHARE registers. Registers from nine countries do not provide any geographical code for matching their data to survey sampling data.

A detailed overview with all geographical codes available in all registers can be found in sheet “03\_Geocodes” in the Excel file “SERISS\_WP2\_D2.5\_Annex2.xls”.

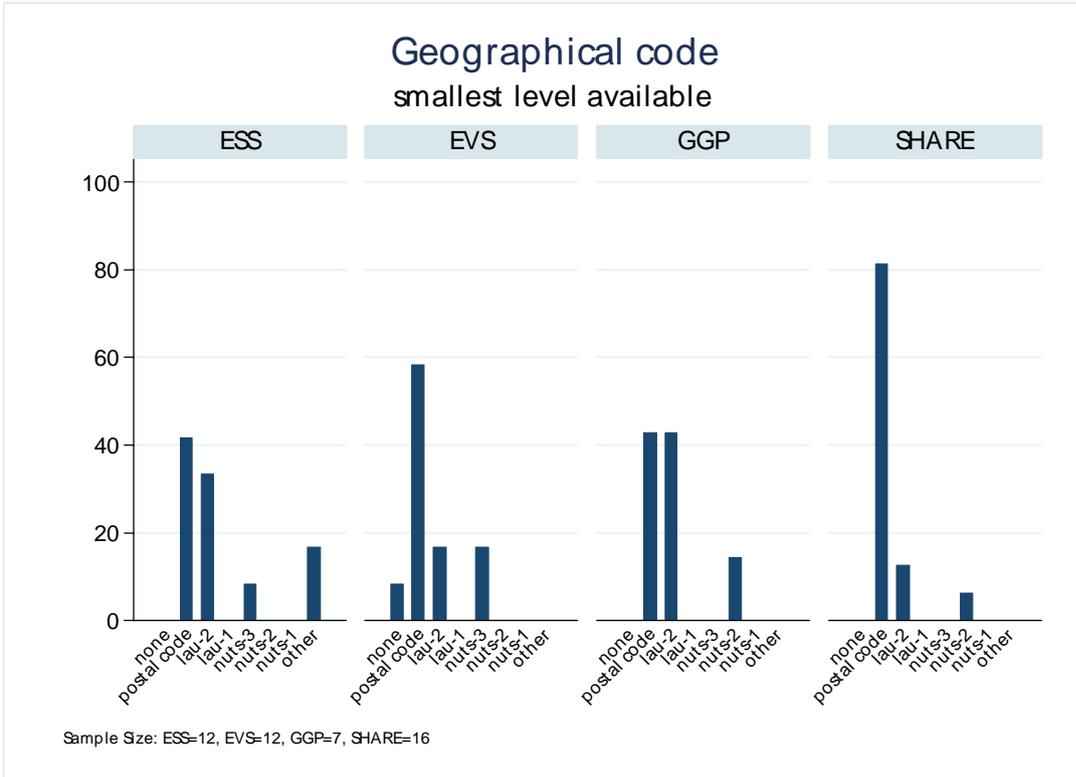


Figure 4: Available geographical codes in sampling registers

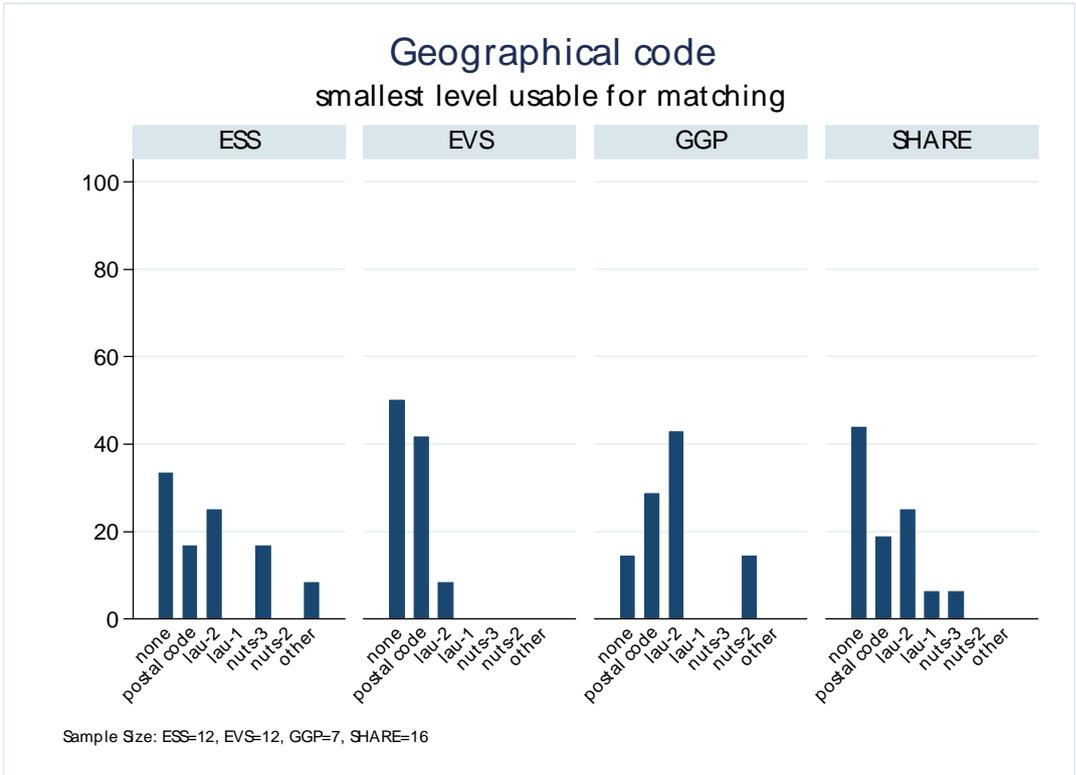


Figure 5: Usable geographical codes in sampling registers

### 3. Auxiliary data sources at European level

In this chapter we provide an overview of the two most important sources of contextual auxiliary data at European level, Eurostat and Census Hub, which may be relevant for nonresponse analysis. Based on the findings from our expert survey, we know that contextual data most often is not included in the registers but available externally. The main criteria for selecting the data sources for this chapter were that: data were available at, or could be aggregated to, small-area level, that data were of high quality and that data were available for many countries that also participate in the ESS, EVS, GGP or SHARE. In the following sections, the two sources identified are described shortly and then compared along the same dimensions as in chapter 2, namely the availability of indicators, the possibility of access, and geographical identifiers.

#### 3.1 European level data sources

Although termed as European-level data sources, the two sources discussed here contain data for different groups of countries. Table 2 provides an overview of which countries (participating in the ESS, EVS, GGP and SHARE) the European data sources provide information for in general.

##### *EUROSTAT*

Eurostat is one of the Directorates-General of the European Commission and is the leading provider of official statistics on Europe. It is the statistical office of the European Union (EU). One of the main tasks is the provision of comparable statistical information over member countries and different regions. While data collection is conducted by the countries themselves, Eurostat consolidates and post-harmonizes the different sources. To get a broad range of detailed data, Eurostat combines data from several different studies on specific topics. This includes the European Community Household Panel, the European Labor Force Survey, Community Innovation Survey, the European Union Statistics on Income and Living Conditions, the Structure of Earnings Survey, the Adult Education Survey, the European Road Freight Transport Survey, the European Health Interview Survey, Continuing Vocational Training, Community Statistics on information Society, and the Micro-Moments-Dataset. The resulting data can be used in various spheres of public life including politics, businesses, education sector and journalism. For detailed information see <http://ec.europa.eu/eurostat/web/microdata/european-community-household-panel>.

##### *Census Hub*

The Census Hub is a uniquely rich data source based on the national population and housing census exercises undertaken in the EU member states and EFTA countries during 2011. It was created by the European Statistical System to disseminate the results of the several censuses in Europe. Census data represent an important source of statistical information on different regional levels by collecting data about the population size and major demographic and economic characteristics of each country. While the national practices in terms of methods and data sources differ substantially, the web tool (Census Hub) enables users to easily access comparable data by ordering it similarly across countries. Another advantage of this

data source is that users are not restricted to accessing a small number of pre-defined tables but are able to specify and download their own cross-tabulations. The data are validated by the National Statistical Offices and are not re-validated by Eurostat. For more information about the European Census 2011 and access to the Census Hub tool see: <http://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/2011-census>.

**Table 2: Availability of European-level data in the 37 countries responding to the SERISS WP2 expert survey**

	<b>Eurostat</b>	<b>Census Hub</b>
<b>Austria (AT)</b>	X	X
<b>Belgium (BE)</b>	X	X
<b>Bulgaria (BG)</b>	X	X
<b>Croatia (HR)</b>	X	X
<b>Cyprus (CY)</b>	X	X
<b>Czech Republic (CZ)</b>	X	X
<b>Denmark (DK)</b>	X	X
<b>Estonia (EE)</b>	X	X
<b>Finland (FI)</b>	X	X
<b>France (FR)</b>	X	X
<b>Georgia (GE)</b>		
<b>Germany (DE)</b>	X	X
<b>Greece (GR)</b>		X
<b>Hungary (HU)</b>	X	X
<b>Iceland (IS)</b>	X	X
<b>Ireland (IE)</b>	X	X
<b>Israel (IL)</b>		
<b>Italy (IT)</b>	X	X
<b>Lithuania (LT)</b>	X	X
<b>Luxembourg (LU)</b>	X	X
<b>Macedonia (MK)</b>		
<b>Malta (MT)</b>	X	X
<b>Montenegro (ME)</b>		
<b>Netherlands (NL)</b>	X	X
<b>Norway (NO)</b>	X	X
<b>Poland (PL)</b>	X	X
<b>Portugal (PT)</b>	X	X
<b>Romania (RO)</b>	X	X
<b>Russia (RU)</b>		
<b>Serbia (RS)</b>		
<b>Slovakia (SK)</b>	X	X
<b>Slovenia (SI)</b>	X	X
<b>Spain (ES)</b>	X	X
<b>Sweden (SE)</b>	X	X
<b>Switzerland (CH)</b>		X
<b>Ukraine (UA)</b>		
<b>United Kingdom (GB)</b>	X	X

## 3.2 Findings

### Availability

There is a wide range of indicators across different topics accessible from the data sources made available via Eurostat and Census Hub. Some of these indicators might be more important for nonresponse analyses than others.

In spite of the extremely wide range of available indicators, Table 3 gives a general overview of the most important domains covered in the different studies and provides examples of variables contained in the data. For exact descriptions and partly also detailed variable lists of each study please see the specific homepages.

**Table 3: Overview of available indicators**

	<b>(AREAS OF ) INDICATORS</b>
<b>Eurostat</b>	<p>Different ones for every section/survey covered by EUROSTAT</p> <p><u>Areas:</u> demographics, employment, education, health...</p> <p><u>Exemplary variables:</u> sex, age, nationality, NACE, ISCO, employment rates, self-employment, working time, household statistics, distribution of income</p> <p>(for detailed lists of every survey see:  <a href="http://ec.europa.eu/eurostat/web/microdata/overview">http://ec.europa.eu/eurostat/web/microdata/overview</a>)</p>
<b>Census Hub</b>	<p><u>Areas:</u> demographics, family and household, employment and education, migration and mobility, housing arrangements, data on housing...</p> <p><u>Exemplary variables:</u></p> <p><u>Individual:</u> sex, age, marital status, family status, household status, current activity status, occupation, industry, status in employment, place of work, educational attainment, size of locality, place of birth, citizenship, housing arrangements</p> <p><u>Families/ households:</u> type of family nucleus, size of family nucleus, type of private household, size of private household, tenure status of household</p> <p><u>Dwellings:</u> living quarter, type of ownership, number of occupants, useful floor space, number of rooms, water supply system, toilet facilities, bathing facilities, type of heating, type of building, period of construction</p> <p><u>Detailed lists can be found here:</u>  <a href="http://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/2011-census">http://ec.europa.eu/eurostat/web/population-and-housing-census/census-data/2011-census</a>.</p>

### Access

One of the most important factors when working with external auxiliary data is the possibility of accessing them for research purposes. The level at which the data is provided plays an important role in determining access (aggregated or microdata). As the findings of the expert survey showed, access is often restricted or is only possible for certain institutions. Across the two European data sources considered here, aggregated data are almost always available free-of-charge to everyone online, access to confidential microdata is significantly restricted and granted only under certain conditions (see Table 4).

Table 4: Overview of accessibility

	<b>ACCESS OPTIONS</b>	
	<b>Microdata</b>	<b>Aggregated data</b>
<b>Eurostat</b>	Only for scientific purpose by application with research proposal	Free access
<b>Census Hub</b>	Access to Census microdata varies by country, but no only aggregated data in the Census Hub	Free access

### Geographical identifiers

In order to use auxiliary data from external sources for nonresponse analyses, it must be possible to link the contextual data to sample records at the small-area level. Eurostat and Census hub data provide opportunities for linkage at several levels from national to regional down to municipal level. However, when investigating these data sources it becomes clear that the options for linkage vary between data sources and also between variables within the studies). Most often small-area level codes are available for internal procedures (e.g. sampling) but not publicly available for research purposes (see Table 5). The most promising data comes from Census Hub since they provide some information on the municipal level.

Table 5: Overview of available geographical identifiers

	<b>GEOGRAPHICAL IDENTIFIERS</b>
<b>Eurostat</b>	Mostly national data (e.g. marital status, education, employment, net earnings, average household size), but also some variables available on a regional level (NUTS2) (e.g. population density, births, deaths, total fertility rate)
<b>Census Hub</b>	Choice of national, NUTS2, NUTS3 or municipal level (LAU2), but not all variables available for all countries or on all levels

## 4. Summary and Conclusion

The main objective of this report is to provide an overview of auxiliary data available in country registers used by the four large cross-national studies collaborating as part of SERISS: the European Social Survey (ESS), the European *Values* Study (EVS), the Gender and Generations Program (GGP), and the Survey of Health, Ageing, and Retirement in Europe (SHARE). The information is primarily based on findings from the SERISS WP2 expert survey among the researchers of these studies' country teams. These findings have been augmented with information from data sources on the European level. All registers and data sources were evaluated along the dimensions of content, access options, and geographical identifiers available for linking data.

The main outcome of this deliverable is a detailed record of the registers used across the four surveys in 37 different countries. Overall, 50 country teams (out of a total of 81 responding) reported using register data for sampling and provided information on the variables available, access conditions and opportunities for linkage. Full details of the responses to the SERISS WP2 expert survey can be found in the Excel file "SERISS\_WP2\_D2.5\_Annex2.xls" This document is a resource for researchers and survey practitioners needing to carry out sampling and/or nonresponse analysis for any survey being carried out in the participating countries. It provides an opportunity to compare and learn from the experiences of the four major cross-national surveys being conducted in Europe today and identify potential data sources for future uses.

Summary findings from the data gathering exercise are presented in this report. Overall, most registers contain auxiliary information, but the availability is often limited to general socio-demographic characteristics. Household characteristics are included less often. Information about the neighbourhood is very rarely included, but can be obtained from external sources for 80% of the listed registers. Who can access this information in the registers is less clear. The most generous registers in terms of international access come from Austria, Czech Republic, Germany, Denmark, Estonia and Malta; in all other countries access is restricted to national users. Universities, statistical offices and non-commercial research institutes are allowed to obtain access more often than commercial survey organisations or commercial marketers. Furthermore, the level of geographical codes for combining data sources varies. Postal code (the lowest level of data asked about) is available in registers from 18 different countries overall, but can be used for matching survey data only in 11 countries (Austria, Belgium, Germany, Finland, Hungary, Iceland, Italy, Luxembourg, Poland, Sweden and Slovenia). Sometimes other municipal codes or addresses are also included, but can rarely be used for linkage (Austria and Estonia are commendable exceptions).

The auxiliary information available within registers is limited and most country teams rely on external sources of auxiliary data to supplement (or instead of) register data. This report therefore also provides an overview of key data sources, available at a European level for deriving contextual information that might be of relevance for nonresponse analyses. Data from Eurostat and Census Hub have been evaluated. While the variety of possible indicators available is vast, it is important to note that not all information can be obtained for small-area contextual levels and although both evaluated sources contain geographical identifiers at very small levels, they are usually not available for external use by researchers. Census Hub provides the most promising compilation of data as census surveys are gold standard surveys

and the Hub provides an easy-to-use compilation of the national census surveys from 30 out of the 37 countries that responded to the expert survey.

Using administrative data from registers or other auxiliary sources e.g. European surveys or censuses for nonresponse analyses both have some pros and cons. The main advantage of using auxiliary data from registers is its availability for all gross sample units. Under the condition that auxiliary data is equally accessible for additional data as for sampling, the data linkage should be straightforward since both come from the same source. The main disadvantage might be a limitation of content and missing information on some observations. Registers are installed for a specific purpose and only variables that serve that purpose are included. For example, socio-economic information, which is of high interest for nonresponse analyses, is rarely included. . Furthermore, legal barriers limiting the use of administrative records need to be investigated and overcome to facilitate using administrative records for nonresponse and nonresponse bias analyses. The second option is to derive auxiliary data from other survey data at national or European level, preferably gold standard surveys like census data. The main advantage is the variety of informational content, but the main weakness is that components of total survey error (coverage, sampling, nonresponse, measurement, and processing error, see Groves et al., 2004) apply to these data sources in the same way as they apply to the survey data under investigation. For example, census data is seen as the gold-standard for survey data and often used for estimating response bias in items of other surveys, but has the drawback of being conducted only every ten years (Czajka 2013). The extent of the potential error needs to be considered when using other survey data.

This review of auxiliary data in sample registers has highlighted that there is significant variation in data availability and access conditions not only across countries but also across surveys and organisations within countries. Future efforts within and beyond SERISS should strive, therefore, to foster knowledge exchange enabling survey practitioners to make sampling decisions informed by others' experiences and to increase transparency of obtaining uniform access to information in registers for all interested parties. The next steps within this work package will be to use the SERISS Survey Network to gather experts and interested to facilitate exchange and collaboration. We will also organise a workshop to discuss available data sources and their potential use in nonresponse analysis.

## References

- Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. New York: Wiley.
- Czajka, J. L. (2013). Can Administrative Records Be Used to Reduce Nonresponse Bias? *The ANNALS of the American Academy of Political and Social Science*, 645(1), 171-184. doi: 10.1177/0002716212463313
- Eurostat. (2015). Nomenclature of territorial units for statistics. NUTS 2013/EU-28. European Union. Retrieved from <http://ec.europa.eu/eurostat/web/nuts/publications>. doi: 10.2785/53780. .
- Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5), 646-675.
- Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.
- Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Hoboken, New Jersey: John Wiley & Sons.
- Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information*. Hoboken, NJ: John Wiley & Sons.
- Krueger, B. S., & West, B. T. (2014). Assessing the Potential of Paradata and other Auxiliary Data for Nonresponse Adjustments. *Public Opinion Quarterly*. doi: 10.1093/poq/nfu040
- Olsen, J. (2011). Register-based research: Some methodological considerations. *Scandinavian Journal of Public Health*, 39(3), 225-229. doi: 10.1177/1403494811402719
- Olson, K. (2013). Paradata for Nonresponse Adjustment. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 142-170. doi: 10.1177/0002716212459475
- Poulain, M., & Herm, A. (2013). Central Population Registers as a Source of Demographic Statistics in Europe. *Population-E*, 68(2), 183-212. doi: 10.3917/pope.1302.0183
- Scherpenzeel, A., Maineri, A., Pflüger, S. et al. (2016). *Report on the use of sampling frames in European studies*, Deliverable 2.1 of Synergies for Europe's research infrastructures in the Social Sciences (SERISS) project.
- Smith, T. W., & Kim, J. (2013). An Assessment of the Multi-level Integrated Database Approach. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 185-221. doi: 10.1177/0002716212463340

## Annex

1. Questionnaire on National Registers - Generic version
2. Detailed register information "SERISS\_WP2\_D2.5\_Annex2.xls"