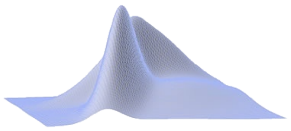
AMELIA

AMELIA - Data description v0.2.3.1

Jan Pablo Burgard, Florian Ertz, Hariolf Merkle, Ralf Münnich

13th January 2020



Economic and Social Statistics Department and Research Institute for Official and Survey Statistics – Prof. Dr. Ralf Münnich
Trier University

Contents

General contact
E-mail: amelia@uni-trier.de

Impressum/Imprint
Universität Trier - Campus I
Fachbereich IV - VWL
Professur für Wirtschafts- und Sozialstatistik
Universitätsring 15
D-54286 Trier
Germany

1 What is AMELIA?

AMELIA is an artificial dataset that enables and fosters comparative and reproducible research involving sampling design-based simulations. The purpose of such simulations is to compare and evaluate new sampling designs, different estimators and their variance estimators. The performance of survey statistics methods is often directly connected to the sampling design. Real sample data cannot be used for such simulations because of small sample sizes or other characteristics, which makes it difficult or impossible to implement specific sampling designs. Additionally, the use of household-level or person-level data is usually restricted due to statistical disclosure which complicates or circumvents comparative and reproducible research since data must not be disseminated to other researchers. Therefore, design-based simulations require a synthetic population that is available and from which samples can be drawn. Both, the population and the samples already drawn using different sampling designs are provided via the AMELIA platform (`http://www.amelia.uni-trier.de/`) to foster open and reproducible research.

Users are invited to further develop the dataset, for example by drawing new samples using other sampling designs. These new extensions can be listed in the respective section of the AMELIA platform.

AMELIA was developed within the scope of the project *Advanced Methodology for European Laeken Indicators* (*AMELI*[1]) for simulations concerning poverty measurement. The dataset was developed further for the project *Inclusive Growth Research Infrastructure Diffusion* (*InGRID*[2]). Within this project, a dissemination strategy for the dataset and its samples based on different sampling designs was developed and is now available via the AMELIA platform. This platform is maintained by the *Economic and Social Statistics Department at Trier University*[3].

The AMELIA dataset was created by drawing from conditional distributions. A detailed description of the generation process of AMELIA is given in Alfons et al. (2011) and Kolb (2013). The data generation approach used for the creation of AMELIA is described in Münnich and Schürle (2003). The platform and some properties of the dataset which are also presented in this paper are described in Merkle and Münnich (2016).

[1] `www.ameli.surveystatistics.net`

[2] `www.inclusivegrowth.be`

[3] `https://www.uni-trier.de/index.php?id=35894&L=2`

The AMELIA dataset and its samples are provided for the open source software R in the form of CSV files. The authors suggest the use of R and the RData files, since RData files are more efficiently compressed.

2 Version Control

This data description (v0.2.3.1) refers to the AMELIA dataset v0.2.3. The respective AMELIA release and its corresponding data description will always have the first three digits of the version number in common. Since all users of AMELIA are invited to develop the dataset further, there will be new releases in the future. Therefore, there might be an issue concerning reproducibility. Future releases might affect the results of former studies. Hence, older versions of the AMELIA dataset and its samples will still be available on the website. For a better understanding of new releases, the systematics of the version numbers are shown below.

- First digit: since this is a community-driven development, the dataset will probably never be finished in a symbolic manner. Thus, it is expected that this digit will always be zero. Results of former studies might not be fully reproducible with the new release.
- Second digit: a change in this digit indicates a change in variables or samples that have been released already in the past. Results of former studies might not be fully reproducible with the new release.
- Third digit: a change in this digit indicates that new variables or samples have been added. Results of former studies are still fully reproducible with the new release.
- Fourth digit: a change in this digit indicates a change of the data description. The data and samples are not affected.

Table 1 lists the versions of AMELIA released already. This table will be updated if there is a new version of the dataset available.

Release	**Date**	**Major developments**
v0.2.3	13th January 2019	Additional variables
v0.2.2	30th September 2017	Additional sampling designs
v0.2.1	26th January 2017	*First public release*

Table 1: AMELIA releases

3 Licence

The AMELIA dataset and its samples may be used for research and teaching free of charge upon the condition that this data description paper and the forthcoming paper by Burgard et al. (2017), once it is available, are properly cited in the data users publication(s) where AMELIA is used. The data user accepts to inform the authors via amelia@uni-trier.de in case she or he intends to provide additions like new variables or sampling designs as well as in case she or he made use of AMELIA in her or his publications. This notification should be sent no later than one month after release or publication.

4 Main Properties of AMELIA

The dataset is strongly oriented on EU-SILC and consists of approx. 10.0 mio. observations of 33 variables on personal level and approx. 3.7 mio. observations of 27 variables on household level. The main properties of AMELIA are shown in the following list.

- Large population size
- Household structure available
- Regional structure available
- Map available
- Samples using different sampling designs already drawn

AMELIA comprises the following four regional levels which are listed below in descending order of area size.

1. REG (Region)
2. PROV (Province)
3. DIS (District)
4. CIT (City/Community)

Table 2 describes the structure of the areas.

For further understanding of the regional structure of the dataset, see Figure 1 which depicts the first two regional levels, i.e. regions and provinces, using a

REG	PROV	DIS
1	1	1-4
	2	5-7
	3	8-10
2	4	11-15
	5	16-20
3	6	21-23
	7	24-27
	8	28-30
4	9	31-34
	10	35-38
	11	39-40

Table 2: Regional structure

synthetic map serving as an example. Currently, the provision of a map file is in preparation.

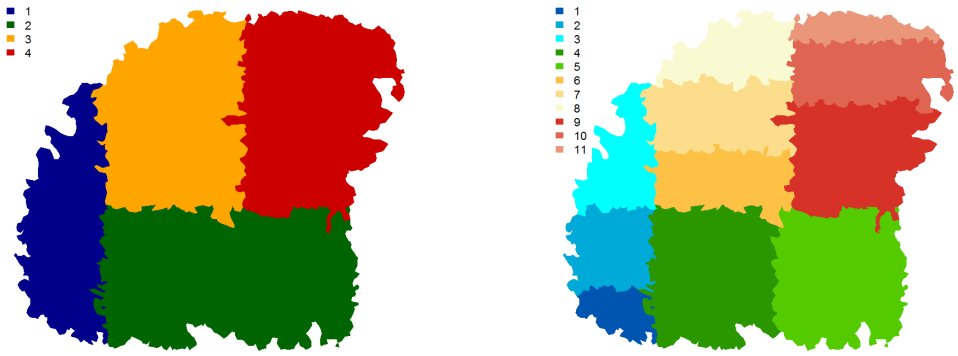


Figure 1: Regions (left) and provinces (right)

Figure 2 shows the districts and cities.

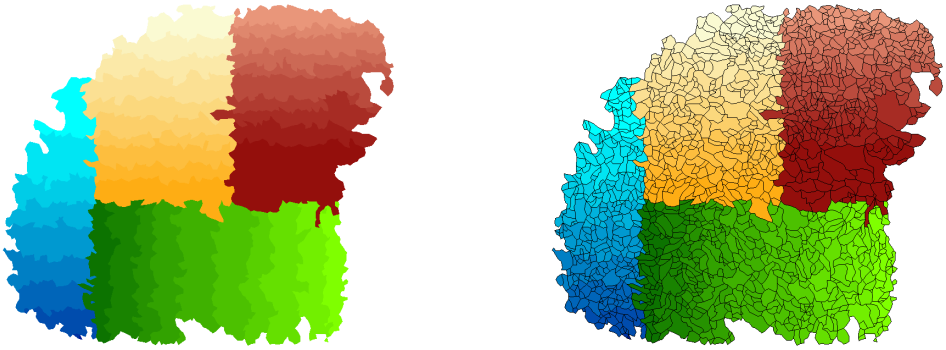


Figure 2: Districts (left) and cities (right)

5 Variables

A synthetic dataset contains the main structures of the original variables. An overview of the variables of the dataset is given in Tables 3 and 4.

Person-level variables

The following Table 3 shows the variables on person-level.

Variable	Variable (EU-SILC)	Name	Notes
AGE	PB010-PB140	Age	Censored at 80
BAS	RB210	Basic activity status	1: at work 2: unemployed 3: in retirement or early retirement or has given up business 4: other inactive person
CIT	-	City/Community	Regional identifier
COB	PB210	Country of birth	1: LOC 2: EU 3: OTH
DIS	-	District	Regional identifier
DOU	DB100	Degree of urbanisation	Degree of urbanisation of CIT 1: densely-populated area 2: intermediate area 3: thinly-populated area
EDI	HX090	Equivalised disposable (household) income	
EDU	PE010	Current education activity	1: in education 2: not in education
HHS	HX040	Household size	
HID	DB030	Household ID	
INC	-	Personal income	Sum of all personal income variables
ISCED	PE040	Highest ISCED level attained	0: ISCED0 1: ISCED1 2: ISCED2 3: ISCED3 4: ISCED4 5: ISCED5 or ISCED6
MST	PB190	Marital status	1: never married 2: married 3: separated 4: widowed 5: divorced
PID	RB030	Personal ID	Household ID + 2 digit serial number
PL070	PL070	Number of months spent at full-time work	0–12
PL072	PL072	Number of months spent at part-time work	0–12
PL080	PL080	Number of months spent in unemployment	0–12
PL085	PL085	Number of months spent in retirement	0–12
PL087	PL087	Number of months spent studying	0–12
PL090	PL090	Number of months spent in inactivity	0–12
PROV	-	Province	Regional identifier
PWHI	-	Person with highest income in household	1: person with highest income in household 2: not person with highest income in household
PY010	PY010	Employee Cash or near-cash income	
PY020	PY020	Non-Cash Employee income	

Variable	Variable (EU-SILC)	Name	Notes
PY050	PY050	Cash benefits or losses from self-employment	
PY070	PY070	Value of goods produced for own-consumption	
PY090	PY090	Unemployment benefits	
PY100	PY100	Old-age benefits	
PY110	PY110	Survivor benefits	
PY120	PY120	Sickness benefits	
PY130	PY130	Disability benefits	
PY140	PY140	Education-related allowances	
REG	-	Regional identifier	
RES	RB200	Residential status	1: currently living in the household 2: temporarily absent
SEM	-	Self-employment Dummy	1: self-employed 2: not self-employed
SEX	RB090	Sex	1: male 2: female
SOC	-	SOC	Social income
SUP	PL150	Managerial position	1: supervisory responsible 2: non-supervisory responsible
UEP	-	Unemployment profile	Unemployment of CIT 1: unemployment rate <3% 2: unemployment rate 3-5% 3: unemployment rate 5-7% 4: unemployment rate ≥7%

Table 3: Variables on person-level

Household-level variables

Some studies only require household-level data. As a consequence, the website provides an already aggregated household-level version of AMELIA. Note that the household-level dataset provides additional variables but not all person-level variables are aggregated in the household dataset. Table 4 lists the variables on household-level.

Variable	Variable (EU-SILC)	Name	Notes
CIT	-	City/Community	Regional identifier
DIS	-	District	Regional identifier
DOU	DB100	Degree of urbanisation	Degree of urbanisation of CIT 1: densely-populated area 2: intermediate area 3: thinly-populated area
EDI	HX090	Equivalised disposable (household) income	
EHHS	HX050	Equivalised household size	
HH050	HH050	Ability to keep home adequately warm	
HHS	HX040	Household size	
HID	DB030	Household ID	
HS010	HS010	Arrears on mortgage or rent payments	1: yes 2: no
HS020	HS020	Arrears on utility bills	1: yes 2: no
HS030	HS030	Arrears on hire purchase instalments or other loan payments	1: yes 2: no

Variable	Variable (EU-SILC)	Name	Notes
HS040	HS040	Capacity to afford paying for one week annual holiday away from home	1: yes 2: no
HS050	HS050	Capacity to afford a meal with meat (or veg. equiv.) every second day	1: yes 2: no
HS060	HS060	Capacity to face unexpected financial expenses	1: yes 2: no
HS070	HS070	Do you have a telephone?	1: yes 2: no – cannot afford 3: no – other reason
HS080	HS080	Do you have a colour TV?	1: yes 2: no – cannot afford 3: no – other reason
HS100	HS100	Do you have a washing machine?	1: yes 2: no – cannot afford 3: no – other reason
HS110	HS110	Do you have a car?	1: yes 2: no – cannot afford 3: no – other reason
HY010	HY010	Total household gross income	
HY020	HY020	Total disposable household income	
HY025	HY025	Within-household non-response inflation factor	
HY030	HY030	Imputed rent	
HY040	HY040	Income from rental of a property or land	
HY050	HY050	Family/children-related allowances	
HY060	HY060	Social exclusion not elsewhere classified	
HY070	HY070	Housing allowances	
HY080	HY080	Regular inter-household cash-transfer received	
HY090	HY090	Interest, dividends, profit from capital investments in unincorporated business	
HY100	HY100	Interest repayments on mortgages	
HY110	HY110	Income received by people aged under 16	
HY120	HY120	Regular taxes on wealth	
HY130	HY130	Regular inter-household cash transfer paid	
HY140	HY140	Tax on income & social insurance contributions	
INC	-	Personal income	Sum of all personal income variables
PROV	-	Province	Regional identifier
REG	-	Region	Regional identifier
SOC	-	Social income	
UEP	-	Unemployment profile	Unemployment of CIT 1: unemployment rate $<3\%$ 2: unemployment rate 3-5% 3: unemployment rate 5-7% 4: unemployment rate $\geq 7\%$

Table 4: Variables on household-level

The AMELIA dataset is available on person-level as well as on household-level. Every single variable of AMELIA is stored in a separate file. Thus, it is only necessary to load the variables that are actually needed. The samples currently available can be used for both datasets because all sampling designs consider entire households on the last stage.

6 Sampling Designs

There are several different sampling designs available which encompass simple random sampling without replacement (1.2), stratified sampling (1.4a, 1.4b, 1.4c) and two-stage stratified sampling (2.7). These sampling designs and their properties are shown in Table 5 which is based on Table 2.1 in Hulliger et al. (2011)

	$p_1(\cdot)$				$p_2(\cdot)$			
ID	PSU	Strata	π_{i1}	Alloc.	SSU	Strata	π_{i2}	Alloc.
1.2	HID	–	srs	–	–	–	–	–
1.4a	HID	PROV	srs	prop	–	–	–	–
1.4b	HID	DIS	srs	prop	–	–	–	–
1.4c	HID	DIS	srs	opt (EDI)	–	–	–	–
2.7	CIT	PROV*DOU	srs	prop	HID	–	srs	–

Notes: Alloc: allocation; prop.: proportional; srs: simple random sampling without replacement; $p_k(\cdot)$: sampling design at the kth stage; and π_I and π_{II}: sample inclusion probability at the first and second stage; *Variables*: HID: household identifier; CIT: municipality identifier; DIS: district; DOU: degree of urbanization; EDI: equivalised disposable income.

Table 5: Sampling Designs

Additionally, Table 6 displays the sampling fractions of each sampling design which are mostly similar to the sampling fractions of Table 2.2 in Hulliger et al. (2011).

ID(s)	$\mathbf{f_1}$	$\mathbf{f_2}$	**f**
1.2 / 1.4a / 1.4b / 1.4c	0.16%	–	0.16%
1.2 / 1.4a / 1.4b	1%	–	1%
1.2 / 1.4a / 1.4b	5%	–	5%
2.7	5%	20%	1%
2.7	16%	1%	1%
2.7	25%	20%	5%
2.7	6.25%	80%	5%

Notes: f: overall sampling fraction; f_1 and f_2 sampling fraction on first and second stage.

Table 6: Sampling fractions

The respective samples are stored in matrices with five columns. Every matrix comprises 100 samples. Table 7 shows the scheme of these matrices where HID represents the household ID, pi_I and pi_II are the inclusion probabilities on the first and second stage, pi is the overall inclusion probability and SPNO denotes the sample number.

HID	pi_I	pi_II	pi	SPNO
135	0.01	1	0.01	1
219	0.01	1	0.01	1
281	0.01	1	0.01	1
324	0.01	1	0.01	1
544	0.01	1	0.01	1
646	0.01	1	0.01	1
⋮	⋮	⋮	⋮	⋮
3781300	0.01	1	0.01	100

Table 7: Sample matrix scheme

For each sampling design and each sampling fraction according to Table 6, 10,000 samples are drawn and provided by 100 matrices comprising 100 samples.

7 Next Steps

The AMELIA dataset is currently expanded for the follow-up project *InGRID2*[4] that is part of the European Comissions *Horzion 2020*[5] (grant agreement No 730998). At this, the AMELIA dataset is extended by a longitudinal dimension making use of EU-SILC data. The AMELIA platform serves as a central data resource within InGRID2. Additionally, we plan to add wealth variables based on the European Central Bank's *Household Finance and Consumption Survey*. Moreover, the authors aim to register a Digital Object Identifier (DOI) to facilitate findability and assist version control.

List of Contributors

- Ralf Münnich (Senior local project manager and scientific coordinator)
- Jan Pablo Burgard (Junior local project manager)
- Florian Ertz (Junior local project manager)
- Jan-Philipp Kolb (Data generation)
- Hariolf Merkle (Maintenance and data generation)

[4] `www.inclusivegrowth.eu`
[5] `https://ec.europa.eu/programmes/horizon2020/`

- Simon Lenau (Implementation of sampling designs)
- Florian Volk (Editing)

Acknowledgements

This research was and is thankfully supported and funded by the European Commission through the projects *AMELI*, *InGRID* and *InGRID2*. We are thankful to our colleagues within and outside those projects who provided feedback and discussion helping to develop the dataset and the corresponding platform.

References

Alfons, A., Filzmoser, P., Hulliger, B., Kolb, J.-P., Kraft, S., Münnich, R., and Templ, M. (2011). Synthetic Data Generation of SILC Data. *AMELI Research Project Report WP6 - D6.2.*

Burgard, J. P., Kolb, J.-P., Merkle, H., and Münnich, R. (2017). Synthetic data for open and reproducible methodological research in social sciences and official statistics. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 11(3):233–244.

Hulliger, B., Alfons, A., Bruch, C., Filzmoser, P., Graf, M., Kolb, J.-P., Lehtonen, R., Lussmann, D., Meraner, A., Münnich, R., Myrskylä, M., Nedyalkova, D., Seger, J., Schoch, T., Templ, M., Valaste, M., Veijanen, A., and Zins, S. (2011). Report on the Simulation Results. *AMELI Research Project Report WP7 - D7.1.*

Kolb, J.-P. (2013). *Methoden zur Erzeugung synthetischer Simulationsgesamtheiten.* PhD thesis, Universität Trier, Universitätsring 15, 54296 Trier.

Merkle, H. and Münnich, R. (2016). The Amelia dataset - a synthetic universe for reproducible research. In Berger, Y. G., Burgard, J. P., Byrne, A., Cernat, A., Giusti, C., Koksel, P., Lenau, S., Marchetti, S., Merkle, H., Münnich, R., Permanyer, I., Pratesi, M., Salvati, N., Shlomo, N., Smith, D., and Tzavidis, N., editors, *InGRID Deliverable 23.1: Case studies, WP23 – D23.1.*

Münnich, R. and Schürle, J. (2003). On the simulation of complex universes in the case of applying the german microcensus. *DACSEIS research paper series*, No.4.