ZA6294 / ZA6295

MESARAS 2013:
Mobility, Expectations, Self-Assessment and Risk Attitude of Students

- Read Me -
MESARAS 2013: READ ME, version: 05/2016

This document provides a concise description of relevant study features or indicates where to find further possibly relevant information in the documentation material and highlights potential limitations to be considered.

The available documentation and auxiliary material comprises (aside from this READ ME):

- Project and methodological report
- Codebook
- Variable list
- Questionnaire (German original and translated English version)
- Data set for geo-referencing, based on key variables to link postal codes with centroids’ coordinates and districts (e.g. INKAR data).

The following key elements deserve some initial attention and are addressed subsequently in alphabetical order:

**bound variables, centroids, codebook, data collection, data format, data quality, derived variables, distance concept, geo-referencing, postal code, project and methodological report, Public-use-file (PUF), questionnaire, questionnaire variables, representativeness, sample, Scientific-use-file (SUF), Stata commands, system variables, target group, topics, variable list, variable names**

<table>
<thead>
<tr>
<th>Key element</th>
<th>Remarks</th>
<th>Further information</th>
</tr>
</thead>
<tbody>
<tr>
<td>bound variables</td>
<td>For some variables, two different versions are available, labelled ‘lower bound’ variables with suffix [_lb] and ‘upper bound’ variables with suffix [_ub]. Usually, upper bound variables follow the overall distribution of the lower bound variables very closely. Deviations occur for most variables with a single digit frequency in the full sample of 2589 observations. They originate mostly from either multiple or unclear ticks in case of a scale item. Larger deviations occur for questions referring to monetary sums or percentages, since respondents provided more often an interval (e.g. ‘200-250 €’, here the lower bound variable [X_lb] would contain ‘200’ whilst ‘250’ would be stored in the upper bound variable [X_ub]). Derived variables naturally inherit this property from their parental variable(s). Both types are included in the SUF for transparency reasons and to enable any user to check whether results are sensitive to the choice of either [_lb] or [_ub] variables. Alternatively, it is possible to condition on those observations with unambiguous response behaviour.</td>
<td>further information</td>
</tr>
<tr>
<td>centroids</td>
<td>Postal code areas’ centroids are the pivotal element to evaluate individuals’ mobility. They are a respective area’s arithmetic mean of all coordinate points (given by latitude and longitude) encompassed by this area (basically, a centroid is the point where one could balance a two-dimensional representation of such an area on the tip of a needle). Therefore, centroids represent the most central point within such an area and may thus serve as reference points to calculate distances between two areas.</td>
<td>see ‘distance concept’, PMR (p.7)</td>
</tr>
<tr>
<td>codebook</td>
<td>The codebook provides detailed information on all variables included in the SUF (and those fully available in the PUF). For string-type variables some examples are shown (e.g. postal codes). Generic numeric variables are described based on means, standard deviation and percentiles. Numerically coded variables are reported with their value labels and frequency tabulation. In all cases, the occurrence of missing values is reported as well.</td>
<td>document: codebook</td>
</tr>
<tr>
<td>data collection</td>
<td>The survey was conducted as a self-administered paper-based questionnaire with mainly standardised items. It was implemented into either so-called ‘orientation weeks’ for beginning students or one of the first introductory lectures in the first semester, taking place between 02.10.2013 and 23.10.2013. At each university, permission to conduct this survey has been granted by the economics department and the person(s) who is (are) responsible for the session into which it was implemented.</td>
<td>document: data collection</td>
</tr>
<tr>
<td>data format</td>
<td>The final versions of the Scientific and Public use Files have been prepared in Stata</td>
<td></td>
</tr>
<tr>
<td><strong>Section</strong></td>
<td><strong>Content</strong></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>14.1.</td>
<td>Using Stata 13 (or older), you might run into compatibility issues. The size of both files (SUF and PUF) is around 8 MB.</td>
<td></td>
</tr>
<tr>
<td>data quality</td>
<td>After a pre-test phase (including a small Alpha-test and a Beta-test in a group similar to the target group), the survey has been supervised in the field by the principal investigator. Since the Beta-test demonstrated a too high error rate in case of automatic data entry, all questionnaires have been manually entered. This raw data has undergone several consistency and plausibility checks in a first analytical step. Identified errors or unclear cases have been corrected on a case-by-case basis. In a subsequent step, all observations have been fully cross-checked with all responses in the corresponding questionnaires once more.</td>
<td></td>
</tr>
<tr>
<td>derived variables</td>
<td>Derived variables in the data files provide additional information and are not directly extracted from the questionnaire. This includes for instance calculated distances or aggregating variables. They are easily recognised by their prefix, starting with [d]. They also indicate directly their ‘parental’-variables they are based on and come usually in the format [dABB_ABB], where [A] represents the Block (1-6), [BB] the question number within the block (01-18). Additional calculations are reported in the ‘remarks’-column in the variable list.</td>
<td></td>
</tr>
<tr>
<td>distance concept</td>
<td>The standard distance concept in the MESARAS project is geographic distance (as the bird flies), measured between postal code areas' centroids. All these distances (in kilometre) have been calculated in Stata, using the geodist-command, relying on the standard ellipsoid (WGS 1984). The project and methodological report additionally refers to road distance and travel time, both related to the reachability model ('Erreichbarkeitsmodell') by the BBSR. Since this is proprietary BBSR-data, it cannot be included in the current version of the MESARAS-data files.</td>
<td></td>
</tr>
<tr>
<td>geo-referencing</td>
<td>Geo-referencing in the MESARAS-context allows linking individuals to a specific spatial unit (usually postal code areas) at a certain point in their life. This in turn enables to evaluate the interconnectedness of individuals' decisions and local conditions (e.g. indicators on the district level taken from INKAR online). All operations can be performed using the key variable PLZ, reported in GEO-variables, both in the MESARAS-data sets and the link data set (MESARAS_plz_georef.dta). The latter provides centroid coordinates and corresponding district codes for a given postal code area. Links of PLZ and geographic coordinates are based on a data set (so-called Shape-file), provided by Christl (2013). At the time the MESARAS project has been initiated, this was the most suitable (and complete) data set on German post code areas available. Information on district codes (based on the so-called ‘Regionalschlüssel’) are taken from a list of municipalities (‘Gemeindeverzeichnis, Gebietsstand: 31.12.2012’, released by the German Federal Statistical Office). It is imperative to note that the assignment of PLZ and districts is not necessarily static across years. The links have been verified for the period of 2012 (the last full year before the survey has been conducted) and for PLZ stored in the variables [unplace], [unzip], and a list of alternative study destinations’ PLZ [q306xxa_, q306xxb_]. There is roughly a dozen PLZ which cannot be mapped into a single district. These cases contain no district code in [district_code_incomplete], but a warning. However, using [district_code_force] allows merging these cases to the district with the largest overlap as well. A careful application of this method is suggested.</td>
<td></td>
</tr>
<tr>
<td>postal code</td>
<td>Postal codes (PLZ, sometimes labelled ZIP) respectively associated areas are the pivotal elements in the MESARAS-context, offering the opportunity to perform detailed geo-referenced analysis of rather precisely measured mobility. Most GEO-variables contain PLZ values. Participants’ self-reported PLZ have been verified. If respondents stated instead of a PLZ a city (and a state) and this city could unambiguously identified using the official post code search (provided by ‘Deutsche Post’), the PLZ of this city's central station (a ‘natural’ reference point) has been inserted. The same holds for all potential university locations (stated in Block 3 of the questionnaire) since economic departments and campuses are often not in the same area as the university's official main building. If no unambiguous identification was possible (e.g. if ‘Freising, BAY’ or just ‘Frankfurt’ is stated) the original entry was preserved. This implies that only unambiguously identified spatial references are used in subsequent geo-referenced analyses.</td>
<td></td>
</tr>
<tr>
<td>project and methodological report (PMR)</td>
<td>This report provides an overview of the overall project and methodological considerations. Covered methodological aspects are: survey design, sample statistics, representativeness, introduction to centroid-concepts, and a discussion of distance concepts. Furthermore, it depicts the overall survey background, introduces the main survey topics (Block 1-6). Additionally, a variety of general descriptive statistics is presented and some basic hypotheses in the context of mobility and personality are explored.</td>
<td></td>
</tr>
</tbody>
</table>
| Public-use-file (PUF) | Publicly available data, comprising all variables (without upper bound variables) from the SUF. To ensure full respondents' full anonymity, the content of mostly all GEO-
variables and those with rare case responses have been replaced by a reference to the SUF or re-categorised. Information on missing values is preserved.

questionnaire
The questionnaire has been implemented as paper-based self-administered questionnaire, comprising eight DIN A4 pages. The original version has been handed out in German (‘MESARAS_Fragebogen’). The English version (‘MESARAS_questionnaire’) in the documentation material is a (rather literal) translation. Arrows guided respondents in case of filter questions.

questionnaire variables
These variables (prefix [q]) can directly be linked to a specific question or item in the questionnaire. Their format is [qABBCCD], where [A] represents the Block (1-6), [BB] the question number within the block (01-18), and [CC] is the sub-item number (00-10, strictly following the order of appearance in the questionnaire). In rare cases, [D] represents an additional classification (a-e), e.g. in cases of multiple stated other reasons or destinations.

representativeness
Drawing on matriculation data, the universities’ sample is evaluated based on gender composition and age (as well as share of students from abroad). For the LUH-sub-sample, access to administrative data allowed also evaluating representativeness related to a level grades and performing more exact distributional tests.

sample
The sample consists of 2589 student participants. Related to the target group definition, the majority are beginning students in their first semester enrolled in an economics (Bachelor) programme at one of seven universities in Northern and Middle Germany (Bielefeld, Clausthal-Zellerfeld, Dortmund, Halle, Hannover, Magdeburg, Muenster). The ‘accessibility rate’ (number of participants in the target group / number of newly matriculated) amounts in the full sample to 68.3%.

Scientific-use-file (SUF)
Full data set, containing all variables (including upper bound variables). It can be used (free of charge) via on-site access at the GESIS Secure Data Center (Cologne, Germany) under the conditions of ‘Access category C’ (specification of user and analysis intention).

Stata commands
Within the MESARAS project, the following user-written Stata-commands have been found to be very helpful:
1. shp2dta: importing GIS data (shp-files) into Stata
2. spmap: creating maps in Stata (a wonderful time sink)
3. geodist: calculating geographic distance between two coordinate points
4. geonear: identifying closest neighbour(s) based on their coordinates
All these commands can be installed via ssc install.

system variables
System variable (prefix [s]) provide general or background information (not directly extracted from the questionnaire), e.g. with respect to item completion rates.

target group
For reasons of comparability and to ensure the existence of geographic alternatives, students enrolled in economics programmes have been targeted (programmes are basically offered at every university). Focus on beginning students was required to sample individuals who just recently made an important mobility-related decision. The inclusion of the seven universities in Northern and Middle Germany resulted from granted permission, their ‘adjacent’ location and proximity (distance matters) to the principal investigator’s institution.

topics
The six main thematic topics (mentioned subtopics are not comprehensive) covered in the survey are:
1. Socio-demographic background
2. Mobility: past, current and future mobility; intra-national and cross-border perspective
3. Study: alternative study locations; funding; decision criteria
4. Risk attitude and time preferences
5. Expectations: focusing on future mobility behaviour; scenario-specific decisions
6. Self-Assessment: performance; Big-5; language; social preferences

variable list
Contains the most concise overview of available variables in the data set. Beyond, design types of underlying questionnaire items and variable content classifications provide additional guidance. The ‘remarks’-column informs about calculation methods of derived variables and sources in case of borrowed (and subsequently modified) items.

variable names
All variable names consist of a prefix (indicating their origin) and a short name (indicating their content). Sometimes a suffix may be attached (see bound variables). For convenience reasons, restricting variable names to the short component might be helpful. This Stata-loop does the trick:

```
foreach x of varlist _all {
    local var_split = strpos("x","")
    local var_new = substr("x", var_split+2,)
    rename `x' `var_new'
}
```