MCC Evaluation Microdata

Data Package

# Instructions

This template is informed by MCC’s Evaluation Microdata Documentation and De-Identification Guidelines. In addition to reviewing these Guidelines, MCC contractors responsible for preparation and documentation of evaluation-related microdata for public and/or restricted-access use should be familiar with the following US government guidelines for data de-identification and re-identification:

* [NIST 2015](http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf)
* [NIST 2016](http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf)

**MCC, the evaluator, and stakeholders should consider the following multi-stage process for data review and release**:

1. Evaluator and M&E PM should agree on expected DRB review date as early as possible to confirm. This should be scheduled at least one month before Evaluator’s contract expires.
2. Evaluator should submit full package to M&E PM. The package includes:

* One completed Section 1 of the DRB Data Package Worksheet for ALL data components (i.e. individual, household, and community data for one survey round are three data components with different risks)
* One completed Section 2 & 3 for EACH data component
* Datasets and code package(s)
* Informed consent(s)
* Questionnaire(s)
* Most recent Metadata file (for [Evaluation Catalog](https://data.mcc.gov/evaluations/index.php/catalog) entry)

1. M&E PM should review Metadata and [DRB](http://intranet.mcc.gov/department/DPE/Team/ME/Data%20Protection/2.%20Data%20Documentation%20and%20De-Identification/2_DRB%20Data%20Package%20-%20Cover%20and%20Worksheet.docx) Data Package Worksheet for clarity and completeness. This may require one round of revision based on the M&E PM requests for clarity and completeness.
2. Evaluator should submit full package to M&E PM. M&E PM and the M&E DRB members should establish a first-round review and feedback to the Evaluator on the proposed data de-identification process. This may require a second round of revision to the package.
3. Evaluator should submit full package to M&E PM for the confirmed MCC DRB review date at least 2 weeks prior to confirmed DRB review date.
4. If any feedback/revisions are required following MCC DRB review, Evaluator should revise and resubmit full package to M&E PM with documented responses to MCC DRB feedback to ensure timely virtual review and clearance of the full package. All final de-identification efforts and their impact on verification of analysis should be documented in the evaluator’s Transparency Statement available on the Evaluation Catalog.

All red font text are instructions in the Worksheet and must be replaced with standard black font with the contractor’s response.

**Unless otherwise agreed with MCC, the final document will be made public to complement/underlie the contractor’s Transparency Statement to document the data preparation and de-identification process required for the public and/or restricted-access microdata and any impact on the data for verifying evaluation analysis and broader data usability.**

# Section 1: Cover Sheet

**Overview of Data Package**

(*Instructions: Include a paragraph summarizing each data package component included in the package. For example, if the package includes household, individual, and community level data sets, please include a paragraph summarizing each of these three components, including information on the content and timing of the data collection*.)

This data package includes three datasets: **1) Household Roster 2) Extended Interview and 3)Time Use** as well as a merged dataset that contain all variables from the following two survey components.

**Household Roster –** This interview was administered with one person in the household (the head of household or the most knowledgeable member) and collected employment data on all members above the age of 15 belonging to the household. Data was collected for 8,533 households residing in 854 enumeration areas in all 7 regions of Kosovo. This dataset was reshaped to a long dataset such that each member of the household has one observation (a household with 5 members has 5 corresponding observations) for a total of 32,742 observations. The unique identifier for households is *hhid* and the unique identifier for individual members is *respondent\_id* (which is a combination of *hhid + hh\_member*).

**Extended Interview –** This interview was administered with one randomly selected male and one randomly selected female in the household, who were non-students aged 18-74 years. The Extended Interview captured detailed information on access to productive capital, household decision-making, time use, and gender attitudes. The time use survey was part of the extended interview, however that data was entered using a separate form, therefore it has been provided as a separate dataset (26 of the Extended Interview respondents refused to fill out the Time Use form). In total 8,604 extended interviews were completed, of which 3,710 were completed by females and 4,894 were completed by males. The unique identifier for individual members is *respondent\_id* (which is a combination of *hhid + f\_rid*), which was used to merge data from the extended interviews to the corresponding member in the household roster data set.

All datatsets can be cleanly merged using the *respondent\_id* variable.

Data collection was implemented through SI’s local partner, IDRA Research and Consulting, and occurred over a 17-week period between March 31,2017 and August 4, 2017.

**Complementary Data**

(*Instructions: Complementary data collection efforts are those efforts that complemented the data packages under review for de-identification, but do not necessarily require de-identification. The evaluator should list these data and provide a brief summary on how they connect to any data package components and affect the data package components’ de-identification. For example, if the geospatial data for the project infrastructure is collected and will be publicly released, it should be listed in the complementary data collection efforts*.)

This data package considers the following complementary data efforts:

The Kosovo Agency of Statistics provided the random sample of Enumeration Areas for the study, and the merged dataset also includes a variable (*ea\_ur*) on the urban/rural status of the EA. In addition, the dataset includes a variable on the total number of households (*ea\_hhs*) in the EA (from the listing exercise that was conducted prior to data collection in order to update the sampling frame and collect information on households).

**Data Package Folder Contents**

(*Instructions: Please list the Data Package Component File Name, and then include the File Names of each of the corresponding required documents [Metadata, Worksheet, Informed Consent, Questionnaire, Other docs]. Only one de-identification worksheet per survey is requested unless discussed.)*

Table 1: Data Package Components

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Package** | | | | |
| **Component** | **Worksheet** | **Informed Consent** | **Questionnaire** | **Other Documents** |
| * Final Roster EI TUS Merged Data Anonymized * Final Roster Data Anonymized * Final EI TUS Data Anonymized | MCC Kosovo LFTUS - Data Package Cover and Worksheet | MCC Kosovo LFTUS - Consent Statement | * FINAL Pre-Screen and Household Roster\_[Language] * FINAL Extended Interview\_[Language] | * FINAL Surveyor Manual\_[Language] * Anonymized Basic Codebook * Anonymized Distributions Codebook * Anonymized Label Book * RUN ME - Data Cleaning and Analysis |

# Section 2: Data Component Preparation Overview

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Response** | | | **Discussion/Explanation** |
| Data + Code Completeness | Complete | The MCC Kosovo LFTUS dataset includes results from the implementation of the study’s primary tools: the Household Roster and Extended Interview (including the Time Use module). The provided final, merged, de-identified dataset along with the analysis do-file allow new users to replicate researcher analysis. | | *To be considered Complete: The available data must allow new users to replicate evaluator analysis to the extent allowable by providing the full data set + analysis code. The constructed variables may also be included in a dataset, but if the dataset + code produces those variables, it is not necessary.*  *To be considered Incomplete: The available data only provides a sub-section of data as produced by the survey and/or the constructed variables only. Incomplete data files are limited in terms of full verification of analysis and/or broad usability of data and must be justified.* |
| Incomplete |
| Data Round(s): | Baseline only | Per the Period of Performance and scope outlined in the MCC Kosovo Task Order, the study just had a one-time data collection exercise and will have a one-stage data release. As such, there is no risk of inconsistency in data de-identification protocols. | | *MCC is willing to trade-off broad use of individual rounds for more consistent de-identification protocols across rounds of data. Therefore, unless there is specific demand for the baseline/interim only data, or contractual requirements, MCC prefers contractors to prepare all data rounds in one package.*  *If one stage only – please (i) confirm demand and/or contractual justification and (ii) discuss how preparation and release of this data as presented to the DRB may affect future data round releases.*  *If combination, please discuss if this file replaces any previously published datasets.* |
| Interim only |
| Endline only |
| Combination of rounds |
| Informed Consent and IRB | High restriction | The language outlined in the MCC Kosovo consent script is as follows: “Any information you provide that can identify you or other members of your household will be kept strictly confidential by the parties conducting this study, including MCC employees, employees of the survey firm, and the researchers, to the maximum extent permitted by the laws of the United States and the laws of Kosovo.” This statement denotes medium restriction requirements, and was used to inform the data de-identification process so that indirect identifiers are made available exclusively to Social Impact and MCC. If households were to be re-identified, it would constitute a breach of confidentiality and privacy promised by the informed consent and certain characteristics (e.g., poverty status) could potentially stigmatize respondents. | | *MCC assumes DIRECT identifiers are always removed from any public-use file. With this assumption: Please refer to the informed consent statement – does it require: High restriction: access to data that includes indirect identifiers is limited to the contractor only; Medium restriction: access to data that includes indirect identifiers is limited to the contractor and qualified researchers, including MCC; Low restriction: data with indirect identifiers may be made public.*  *Please discuss how the promises of confidentiality in the informed consent informed de-identification efforts. Please include any additional guidance provided by the IRB as applicable.* |
| Medium restriction |
| Low restriction |
| Geographic Identifiers | Highest (Region) | Listing average sample size (7 regions): 31,670  National population average:  N/A | No de-identification was performed on high-level/region-level geographic identifiers. | *Please provide justification on the identification/de-identification/complete removal of specific geographic regions. De-identifying at a higher geographic level may support privacy protection, but it may also reduce data usability. Please provide justification for recommendation.* |
| - Municipality | Listing average sample size (37 municipalities):  5,991    National population average:  46,842 | See description below. |
| - Settlement | Listing average sample size (423 settlements):  524  National population average:  N/A | Municipality and settlement (municipality\_hr, f\_munic, settlement\_\_hr, f\_sett) were de-identified in order to protect respondents at the equivalent of the village level. It was determined that municipalities are considered low enough levels of geographic identifiers to have limited effect on data usability since no de-identification was performed on high-level geographic identifiers. |
| Lowest ( Census Enumeration Areas) | Listing average sample size (854 EAs):  259  National population average:  N/A | Enumeration areas (variables: ea\_code, and others that contain EA codes: eaid hhid f\_hhid hhid\_hr ea\_code respondent\_id), which were informed by the 2011 Agricultural Census conducted in Kosovo, were de-identified by replacing them with random values, seeing as the risk to respondent privacy protection outweighed the risk of decrease in data usability. SI’s removal of the geographic unit names up through the EA level affords a substantial level of protection against this potential misuse of the data. |
| Knowledge of Treatment | High risk | Seeing as the MCC Kosovo LFTUS is a nationwide labor force survey and not an evaluation with designated treatment or control groups, knowledge of treatment does not apply in this research context. | | *In some cases, general knowledge of treatment areas and/or inclusion of a treatment variable can significantly increase re-identification risk depending on the population affected. Please provide assessment of this re-identification risk and recommendation if considered high/medium risk.* |
| Medium risk |
| Low risk |
| Publication Type | Public-use only | The publication type is public-use. In order to ensure protection of respondents and most applicable use of study findings, the data will be made available to the general public only after direct and indirect identifiers are recoded or removed from the dataset. Given that the data do not reveal particularly sensitive information, the release poses little risk to respondents, particularly after de-identification. | | *Please state for this data package: will there be public-use data only, restricted-use data only, or both and provide justification as this relates to enabling verification of evaluation results and/or broad usability of the data.* |
| Restricted-use only |
| Both |

# Section 3: Data Component Preparation Details

| **Specific Issues** | | **Risk Analysis** | | **Risk Mitigation** | |
| --- | --- | --- | --- | --- | --- |
| *Instructions* | *Response* | *Instructions* | *Response* |
| 1. | Who has significant financial, legal, cultural, or other incentives to re-identify survey respondents? | *List all potential threats[[1]](#footnote-2)* | Overall, the datasets do not contain very sensitive information. It is not expected that individuals or organizations beyond those with research motives will have serious incentives to attempt to de-anonymize these data.  Potential users:   * Kosovar National/Regional Government * Researchers |  |  |
| 2. | What is the potential value to these intruders? | *List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent) | Kosovar National/Regional Government: Local government entities may be incentivized to identify survey respondents because it may empower them to capture individuals whose actual income may not equal their reported income, and thus could disqualify them from receiving welfare or other benefits. In addition, they might want to identify businesses that have not been formally registered to impose fines.  Researchers: It is also possible that researchers may have incentives to attempt to re-identify respondents with certain characteristics (e.g., own their own businesses, engaged in agriculture) for the purposes of recruiting for other research. |  |  |
| 3. | What is the expected cost to these intruders to re-identify the data? | *Describe degree of difficulty for re-identification* | Costs would be high for those attempting to de-anonymize the data.  The removal of key geographic unit names up through the settlement level means a household would have to be identified within a region (there are seven regions total in Kosovo).  All unique identifiers, whether at the household or individual level have been generated by the data collection firm and/or SI, and therefore have no plausible link to any other dataset. Therefore, the datasets that will be made public do not contain any variable that can be used for linkage with another existing dataset outside of this study.  Participating Households: For households who participated in the survey, there could be financial cost of re-identification if they were to lose government benefits. Re-identification could also pose a psychological cost if they were to feel stigmatized or targeted by other researchers, or if they felt their trust in the confidentiality and privacy they were promised were to be violated. Given the anonymization measures, the probability of respondents incurring these costs is also negligible. |  |  |
| 4. | Assess availability of ‘linkage’ data that can be used to re-identify respondents. This includes other datasets or archives with information that can be used to re-identify individuals in the dataset. | *List all potential existing data* | There were several linkage data points in the data:  ea\_code = EA codes that can be linked to the KAS classification of EAs  ea\_hhs = number of households in the EA, which could be linked to KAS data | *Describe how to mitigate link to existing data that enables re-identification* | No link exists to an outside dataset after removal of ***municipality (municipality and f\_munic) and settlement (settlement and f\_sett) names***. The variables that remain in the dataset that uniquely identify households and individuals have no connection to any other data source, as the municipality and settlement codes were generated by IDRA and SI. |
| 5. | **Identity Disclosures:** What are the DIRECT identifiers in the raw data? | *List the DIRECT identifiers* (names, addresses, geographic information, government-issued ID numbers, etc.) | See Appendix A for a list of identifying variables that include names, locations, geographic information from the merged dataset. | *List all DIRECT identifiers removed from the dataset.* | Dropped direct identifying variables. **See Appendix A for full list.** |
| 6. | **Attribute Disclosures:** For GIS/GPS data, this distance data can be a direct identifier that is VERY useful analytically. Therefore, please describe how GIS/GPS data VALUE/USABILITY can be retained. | *List all GPS and/or GIS data.* | GPS coordinates of household locations (longitude, latitude) were collected and stored in the following variables, which have been dropped from the dataset: listing\_geocode, geocodelatitude , geocodelongitude , geocodealtitude , geocodeaccuracy, f\_listing\_geocode , f\_geocodelatitude , f\_geocodelongitude , f\_geocodealtitude , f\_geocodeaccuracy | *Describe process for de-identification. For example: introduce random errors into geographic data (GPS, GIS, etc.).*  *Displace urban points 0-2 km, rural points 0-5 km, and additional 1% of rural points 0-10 km[[2]](#footnote-3).* | Dropped the following variables with GPS information: listing\_geocode, geocodelatitude , geocodelongitude , geocodealtitude , geocodeaccuracy, f\_listing\_geocode , f\_geocodelatitude , f\_geocodelongitude , f\_geocodealtitude , and f\_geocodeaccuracy. |
| 7. | **Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data? | *List the identifying items/variables* | A number of variables contain values that could, either on their own or in combination, potentially identify respondents. Risk is first mitigated substantially by dropping the names of geographic units up through the EA levels. Additional re-coding and collapsing for variables with potentially sensitive information or outliers has been conducted as detailed in **Appendix B and C.** | *Describe top/bottom coding: set upper & lower bounds to remove outliers for continuous. Specify: are values set to the median, or other?*  *For large categories/datasets, the OMB suggests top coding at least the highest .5%; for smaller categories/datasets, top code the highest 3-5%. The same principles apply to bottom coding.[[3]](#footnote-4)* | Variables with potentially sensitive high outliers, for example expenditure and asset data, were top-coded in the following manner: All values at the 99th percentile and above were replaced with the average of those values. **See Appendix B and C for detail.** |
| *Describe any variables that require collapse and describe construction of new variables* | Any variables containing potentially private or sensitive data that were not dropped were considered for re-coding. After removal of geographic unit names, SI is comfortable leaving most variables in their original form. **See Appendix B and C for detail of any variables re-coded.** |
| *Describe any global re-coding to group observations into categories (e.g., age 0-5, 5-10, 65+, etc.). Ensure that the categories are neither too broad nor too narrow.* | **See above & Appendix B and C for detail.** |
| 8. | **Attribute Disclosures:** What variable combinations produce UNIQUE observations that create INDIRECT IDENTIFIERS (for example: individuals with high incomes, ages, or unique combinations, such as 17-year old widowers or contextually unusual racial/ethnic backgrounds) | *List the identifying items/variables:* | Re-coding and collapsing for variables that could indicate potentially sensitive information in combination with other variables has been conducted as detailed in **Appendix B and C.** | *For each identified rare data, describe the local suppression techniques employed to remove unique and rare data. Specify: are values set to missing, the median, or other?* | **See above & Appendix B and C for detail.** |

**Appendix A – Variables Dropped**

Variables removed include names, phone numbers, contact information, locations, and GPS coordinates, along with any other variables that contain such information. In addition, any variables with open-ended responses were coded into categorical variables and the original variables were dropped. All dropped variables dropped are listed below with variable labels.

|  |  |
| --- | --- |
| Variable Name | Variable Label |
| deviceid |  |
| subscriberid |  |
| simid |  |
| devicephonenum |  |
| supervisor | Select your Supervisor: |
| teamid | Select your team number: |
| enumerator | Select your name: |
| municipality\_hr | Please record the municipality where the survey takes place: |
| settlement\_hr | Please record the settlement where the survey takes place: |
| ea\_hr | Please record the enumeration area where the survey takes place: |
| strucuture\_id\_hr | Please record the structure ID where the survey takes place: |
| hh\_num\_hr | Please record the Household Number where the survey takes place |
| listing\_hhh\_name |  |
| hhh\_name | Please verify the name of the head of the household. If missing or incorrect, pl |
| listing\_geocode |  |
| geocodelatitude | Please record the geocoordinates for this household (latitude) |
| geocodelongitude | Please record the geocoordinates for this household (longitude) |
| geocodealtitude | Please record the geocoordinates for this household (altitude) |
| geocodeaccuracy | Please record the geocoordinates for this household (accuracy) |
| listing\_address |  |
| address | Please verify the address of the household. If missing or incorrect, please re-e |
| listing\_hhcounttotal |  |
| ps\_first | Please enter the household member’s first name (household member ): |
| text\_audit |  |
| phone\_r | Please record the respondent’s phone number: |
| first | Please enter the household member’s first name (household member ): |
| relationship\_other | Specify other. |
| absent\_r\_other | Specify other. |
| marital\_other | Specify other. |
| ethnicity\_other | Specify other. |
| lab\_ue\_reas\_other | Specify other. |
| job\_loc | Prevalent place (municipality) of work |
| job\_temp\_other | Specify other. |
| job\_temp\_dw | D1.d) Why does not want a permanent job? |
| job\_pt\_r\_other | Specify other. |
| job\_pt\_dw | D1.g) Why does not want a full-time job? |
| job\_emp\_other | Specify other. |
| job\_act\_o | D4) What is the main sector of ’s employer (or if self-employed, ${fir |
| job\_act\_other | Specify other. |
| job\_func\_o | D5) What is the main type of work that does for his/her job (name of o |
| job\_func\_f\_other | Specify other. |
| job\_loc\_other | Specify other. |
| ue\_reas\_other | Specify other. |
| ue\_seas\_d | E4.a) Please describe in more detail the type of work that usually und |
| ue\_act\_o | E6) What is the main sector of ’s most recent employer (including famil |
| ue\_act\_other | Specify other. |
| ue\_func\_o | E7) What is the main type of work that did for his/her last main job ( |
| ue\_func\_f\_other | Specify other. |
| es\_ia\_r\_other | Specify other. |
| es\_ia\_dwr | F2.a) Why does not want to work? |
| es\_act\_o | F7) What is the main sector of ’s target employer (including family and |
| es\_act\_other | Specify other. |
| es\_func\_o | F8) What is the main type of work that is pursuing in his/her search ( |
| es\_func\_f\_other | Specify other. |
| es\_field\_o | G3) What is the primary field of 's highest level of education complete |
| ed\_field\_w\_other | Specify other. |
| sdls\_other | Specify other. |
| consent\_n\_other | Specify other. |
| household\_image | With the permission, please take a photo of the façade of the building. |
| household\_notes | Please enter any additional important notes for this household. |
| ethnicity\_other\_coded | A6) What is ’s ethnicity? (select one): Specify Other |
| f\_munic | Please record the municipality where the survey takes place: |
| f\_sett | Please record the settlement where the survey takes place: |
| f\_ea | Please record the enumeration area where the survey takes place: |
| f\_strucuture\_id | Please record the structure ID where the survey takes place: |
| f\_hh\_num | Please record the Household Number where the survey takes place: |
| f\_listing\_geocode |  |
| f\_geocodelatitude | Please record the geocoordinates for this household (latitude) |
| f\_geocodelongitude | Please record the geocoordinates for this household (longitude) |
| f\_geocodealtitude | Please record the geocoordinates for this household (altitude) |
| f\_geocodeaccuracy | Please record the geocoordinates for this household (accuracy) |
| f\_name | Enter the respondent’s name: |
| energytype\_other | Specify other. |
| form\_mp\_other | Specify other. |
| form\_sp\_other | Specify other. |
| form\_more\_y\_other | Specify other. |
| form\_more\_n\_other | Specify other. |
| mfi\_mp\_other | Specify other. |
| mfi\_sp\_other | Specify other. |
| mfi\_more\_y\_other | Specify other. |
| mfi\_more\_n\_other | Specify other. |
| inforl\_more\_y\_other | Specify other. |
| inforl\_more\_n\_other | Specify other. |
| friend\_mp\_other | Specify other. |
| friend\_sp\_other | Specify other. |
| friend\_more\_y\_other | Specify other. |
| friend\_more\_n\_other | Specify other. |
| borrowednone | Why didn’t you borrow from any of these credit sources? |
| ent\_reg\_r\_other | Specify other. |
| ent\_choice\_other | Specify other. |
| ent\_act\_text | What is the main economic activity that this business is undertaking or will und |
| ent\_act\_other | Specify other. |
| ent\_sup\_other | Specify other. |
| ent\_ment\_other | Specify other. |
| ent\_const | What is the main obstacle in growing your business? |
| ent\_des\_w | Why is starting a new business not considered a desirable career choice in Kosov |
| rem\_purp\_other | Specify other. |
| rem2\_purp\_other | Specify other. |
| f\_discr\_h | In what ways were you not treated equally? |
| f\_ccare\_y | In what ways has child care access influenced your career decisions? |
| f\_transp\_y | In what ways has lack of access to safe, reliable, and/or affordable transport i |
| f\_equal\_h | In what ways are they not treated equally? |
| f\_discr\_fh | In what ways has concern about discrimination influenced your career decisions? |
| tusimage | Please take a photo of the front side of the completed paper time use diary. |
| tusimage2 | Please take a photo of the back side of the completed paper time use diary. |
| f\_phone | Please record respondent’s phone number: |
| others\_d | Please record who was present during this interview and at what points during th |
| respondent\_notes | Please enter any additional important notes for this respondent |
| f\_name\_tus | Enter respondent’s first name: |

**Appendix B – Variables Re-coded or Collapsed**

Variables were re-coded or collapsed on the basis of their potential to indicate unique characteristics (on their own or in combination) among households or individuals, and therefore potentially serve to identify these respondents. General approaches taken include: top/bottom coding of continuous data (99th percentile), and collapsed categories for categorical variables (when there were too few responses belonging to certain categories).

The following data was top-coded, with the top 99th percentile of data replaced with the mean of those values (rounded to the nearest integer), in order to retain the variable’s arithmetic average. Variables were only recoded this way when the high outlier values were deemed to present a plausible risk of re-identification or if they are generally considered private and sensitive information. Note that while overall arithmetic averages are maintained using this re-coding technique, equality of averages between the data used for analysis and the public data would no longer be maintained if sampling weights are applied or if sub-group analysis is performed.

|  |  |
| --- | --- |
| Variable Name | Variable Label |
| hhcounttotal | Including yourself, how many people are members of this household? |
| ps\_age | What is ’s age in years? |
| birth\_year | A3) What is ’s year of birth? |
| lab\_ed\_abs | C8) What is the total anticipated duration of ’s absence from work (in |
| job\_year | D2) How many years ago did he/she start his/her current job? |
| job\_hours | D3) How many hours per week does usually work in his/her main job? |
| job\_income | D7) What is the net monthly pay (€) gets from his/her main job? |
| pc\_land\_hm | How many Aris of agricultural land does your household currently own? |
| pc\_lstock\_l\_hm | How many large livestock (oxen, cattle) does your household currently own? |
| pc\_lstock\_s\_hm | How many small livestock (goats, pigs, sheep) does your household currently own? |
| pc\_poultry\_hm | How many Chickens, Ducks, Turkeys, Pigeons does your household currently own? |
| pc\_fish\_hm | How many fish ponds or fishing equipments does your household currently own? |
| pc\_farm\_nm\_hm | How many farm equipments (non-mechanized) does your household currently own? |
| pc\_farm\_m\_hm | How many farm equipments (mechanized) does your household currently own? |
| pc\_bus\_eq\_hm | How many nonfarm business equipments does your household currently own? |
| pc\_house\_hm | How many housing structures (residential) does your household currently own? |
| pc\_comm\_re\_hm | How many non-residential structures (commercial and other) does your household c |
| pc\_comp\_hm | How many computers or laptops does your household currently own? |
| pc\_dur\_l\_hm | How many large consumer durables (fridge, TV, sofa) does your household currentl |
| pc\_land\_ot\_hm | How many aris of other land not used for agriculture (residential or commercial) |
| pc\_transp\_b\_hm | How many means of transportation – non-motorized (bicycle) does your household c |
| pc\_transp\_m\_hm | How many means of transportation – motorized (motorcycle, car) does your househo |
| form | Bank/financial institution (non-MFI) |
| mfi | MFI (including NGO and group-based lending) |
| friend | Friends or relatives |
| form1\_tot | How much of the € was used for the main purpose you just specified {€}? |
| mfi1\_tot | How much of the € was used for the main purpose you just specified {€}? |
| friend1\_tot | How much of the € was used for the main purpose you just specified (€} |
| entre\_wgs | In what year did your business first pay salaries, wages, or payments in kind, i |
| entre\_ows | How many people, including yourself, own this business? |
| entre\_emp | Not counting the owners, how many people are currently working for this business |
| entre\_emp\_l | Of the person(s) working for this business, for how many do you pa |
| income\_hh | What is the total ANNUAL income in Euros (€) of all the members of your househol |
| income\_p | How much of your total ANNUAL household income (€) reported just now do you pers |
| rem\_tot | Over the past 12 months, what is the total amount of foreign remittances that yo |
| rem1\_tot | How much of the € you received was used for the main purpose you just |
| F\_consent\_n\_other | Specify other. |

**Appendix C – Recoded Categorical Variables**

The categorical variables outlined in the table below were recoded whenever a low number of observations meant that individuals could potentially be identified depending on their characteristics. For instance, there are relatively few individuals who hold PhDs across Kosovo’s seven regions, and respondents could be identified by their education status within their municipality, settlement, or EA. The same logic applies for ethnicity or business type, given the small number of non-Albanians and businesses operating in niche sectors within the LFTUS sample.

|  |  |
| --- | --- |
| Variable Name | Variable Label |
| ed\_highest | G2) What is the highest level of education completed by ? (select one) |
| ethnicity | A6) What is ’s ethnicity? (select one) |
| nationality\_c | A7.a) What is ‘s primary country of citizenship? |
| sdls\_other\_coded | H1) During the last week in which of the following categories would classify hi |
| ent\_act | What is the main economic activity that this business is undertaking or will undertake? |
| lab\_ue\_reas\_other\_coded | C4) Why did not work last week in his/her main job?: Specify Other |
| es\_ia\_r\_other\_coded | F2.a) Why does not want to work? |
| ent\_act\_text\_coded | What is the main economic activity that this business is undertaking or will und |
| ue\_seas\_c | E4.c) To which country does \_\_\_ most regularly go abroad for work? |

1. As stated in NIST 2016, de-identification practitioners should assume that de-identified US government datasets will be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-identification requirements accordingly. Although a specific dataset may not be seen as sensitive, de-identifying that dataset may be an important step in de-identifying another dataset that is sensitive. Alternatively, the adversary may merely wish to embarrass the US government agency or its partners. Thus, adversaries may have a strong incentive to re-identify datasets that are seemingly innocuous. [↑](#footnote-ref-2)
2. ICF International, Demographic & Health Surveys [↑](#footnote-ref-3)
3. Office of Management and Budget, Checklist on Disclosure Potential of Proposed Data Releases ([current link](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CCsQFjAA&url=http%3A%2F%2Fwww.fcsm.gov%2Fcommittees%2Fcdac%2Fcdac.html&ei=UN9vUpvxDZWt4APZyYD4Bg&usg=AFQjCNFwhGwSnNTfTDllVwmYgpJ2rdKEsg&bvm=bv.55123115,d.dmg)) [↑](#footnote-ref-4)