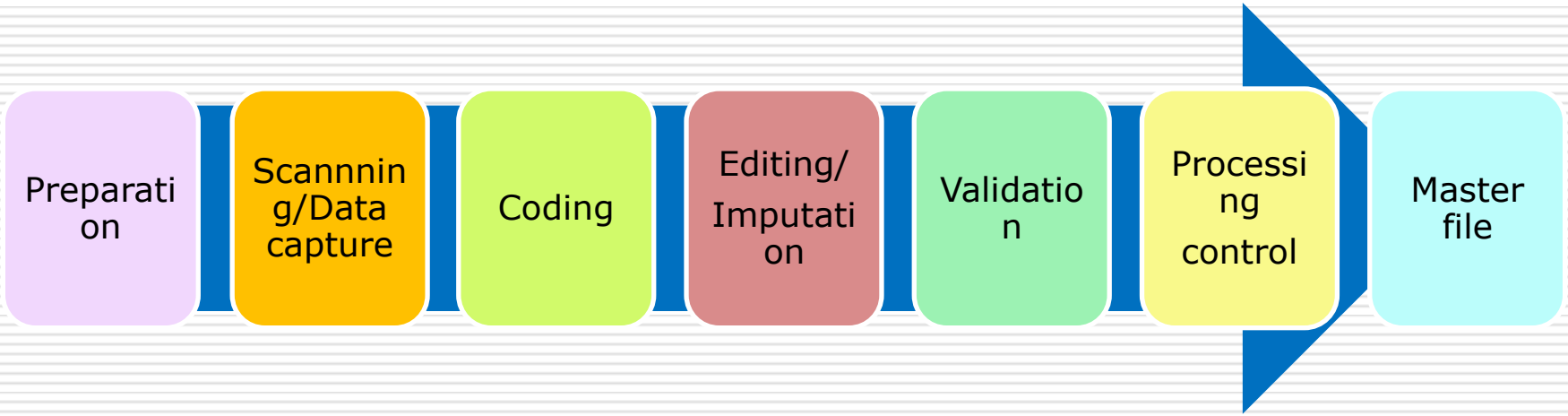
DATA VALIDATION-I
Evaluation of editing and imputation

United Nations Workshop on Evaluation and Analysis of Census Data,
1-12 December 2014, Nay Pyi Taw , Myanmar

Census processing overview

- Steps of data processing depend on the technology used in general, the process covers the following steps:



Validation

- It is a process of checking consistency in data after editing/imputation phase of the census:
  - Editing rules may be insufficient to identify all types of errors
  - Editing/imputation may introduce new errors in data because of incorrect application
  - Some unexpected patterns may not be identified with editing/consistency rules

Validation

- In general, two methods for data validation

  - Evaluation of performance of editing/imputation to ensure correct application or imputation
  - Analysing key aggregated data to check consistency among variables and with expected values/distribution to identify the unusual values/pattern

Basic definitions

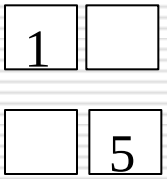
- **Editing:** List of rules to determine invalid and inconsistent data

- **Imputation :** The process of resolving problems concerning invalid or inconsistent data – and missing values- identified during editing

  - ***All records must respect a set of editing rules formulated to correct errors and finally disseminate reliable data***

Some examples for invalid data-Myanmar pilot census questionnaire

- **Age**
  - Equal to 99
    - Instruction – if it is greater or equal to 98, write 98
  - If age is written in one digit, such as



How to correct?

- **Place of birth, place of usual residence and place of previous residence**
  - If code given by enumerators is not consistent with the code list or code written in one or two digits

How to correct?

Some examples for inconsistent data-Myanmar pilot census questionnaire

- **Age and marital status**
  - If age of married person is below the minimum age at first marriage
- **Children ever born alive, living and dead children**
  - If number of children ever-born is not equal to the sum of number of living children and number of dead children
- **Last live birth and household deaths**
  - There is an infant birth who is not alive, but no infant death registered in the household deaths

What will be decision?

Some examples for inconsistent data-Myanmar pilot census questionnaire

- **Sex, age and relationship to the head of household**
  - If sex of the head of household and spouse is same
  - If age difference between the head of household and son/daughter is less than 13 or 14

- **Age, the highest completed level of education and occupation**
  - Age is 9, completed level is primary school and the person is secondary school teacher

What will be decision?

Assessing the performance of imputation

- After implementation of editing/imputation:
  - Data should be classified as follows :
    - Observed (consistent) data: the values which meet with all editing rules
    - Non-response or unknown : no value
    - Inconsistent data : the values which failed at least one editing rule
    - Imputed data for inconsistency –and non-response

    - For this analysis, all procedures performed in the database should be identifiable

Assessing the performance of imputation

1. Compare the distribution of the observed values with the distribution of the imputed values
   - if non-response and inconsistent data are distributed randomly,
     - no difference is expected between the distribution of the observed and the imputed values
   - If there are differences between the people who responded and those who did not or not give accurate data
     - The imputed data should not follow the same distribution as the observed data

Assessing the performance of the imputation

2. Compare the distribution of the observed values with the distribution of all values including the imputed values

- In general, imputed values should have a minimal effect on the distribution of the complete data
  - Unless the non-response rate is particularly high or the bias for certain characteristics

Table 2: Distribution of bedrooms — Thousands

Number of bedrooms	Observed responses N (1)	% (2)	Imputed responses N (3)	% (4)	Difference (Imputed-Observed) % (5)=(4)-(2)	Total Including imputed N (6)=(1)+(3)	% (7)	Change (total-observed) % (8)=(7)-(2)
0	62	0.3	5	0.8	0.5			
1	2,378	10.7	124	19.2	8.5			
2	6,097	27.4	192	29.8	2.3			
3	9,375	42.2	228	35.3	-6.8			
4	3,279	14.7	70	10.9	-3.9			
5	809	3.6	19	2.9	-0.7			
6	166	0.7	5	0.8	0.0			
7	39	0.2	1	0.2	0.0			
8 or more	27	0.1	1	0.2	0.0			
Total	22,232	100	645	100	0.0			

Source: England and Wales, Office for National Statistics, *2011 Census:Item Edit and Imputation: Evaluation Report*, June 2012

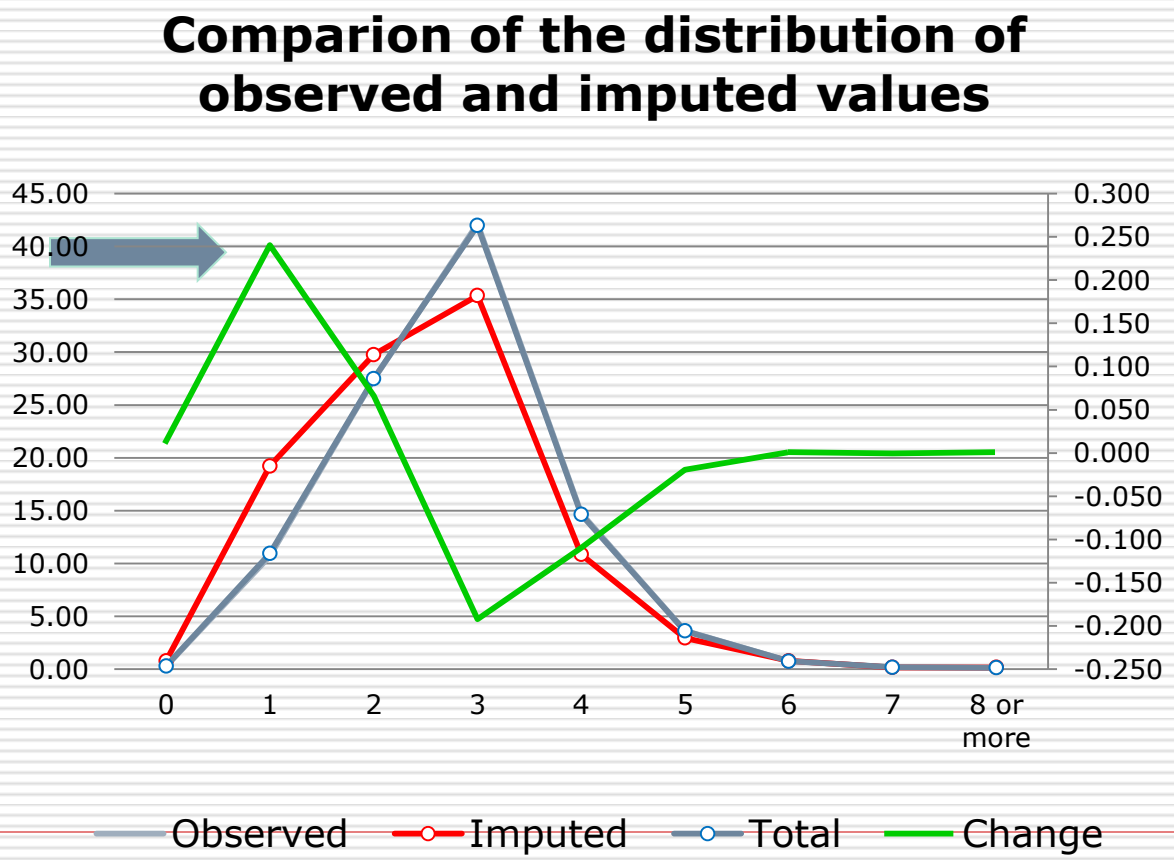
Table 2: Distribution of bedrooms

Number of bedrooms	Observed responses		Imputed responses		Difference (Imputed-Observed)	Total Including imputed		Change (total-observed)
	N	%	N	%	%	N	%	%
	(1)	(2)	(3)	(4)	(5)=(4)-(2)	(6)=(1)+(3)	(7)	(8)=(7)-(2)
0	62	0.3	5	0.8	0.5	67	0.3	0.014
1	2,378	10.7	124	19.2	8.5	2,502	10.9	0.240
2	6,097	27.4	192	29.8	2.3	6,289	27.5	0.066
3	9,375	42.2	228	35.3	-6.8	9,603	42.0	-0.192
4	3,279	14.7	70	10.9	-3.9	3,349	14.6	-0.110
5	809	3.6	19	2.9	-0.7	828	3.6	-0.020
6	166	0.7	5	0.8	0.0	171	0.7	0.001
7	39	0.2	1	0.2	0.0	40	0.2	-0.001
8 or more	27	0.1	1	0.2	0.0	28	0.1	0.001
Total	22,232	100	645	100	0.0	22,877	100	0.000

Source: England and Wales, Office for National Statistics, *2011 Census:Item Edit and Imputation: Evaluation Report*, June 2012

Assessing the performance of imputation

Maximum change



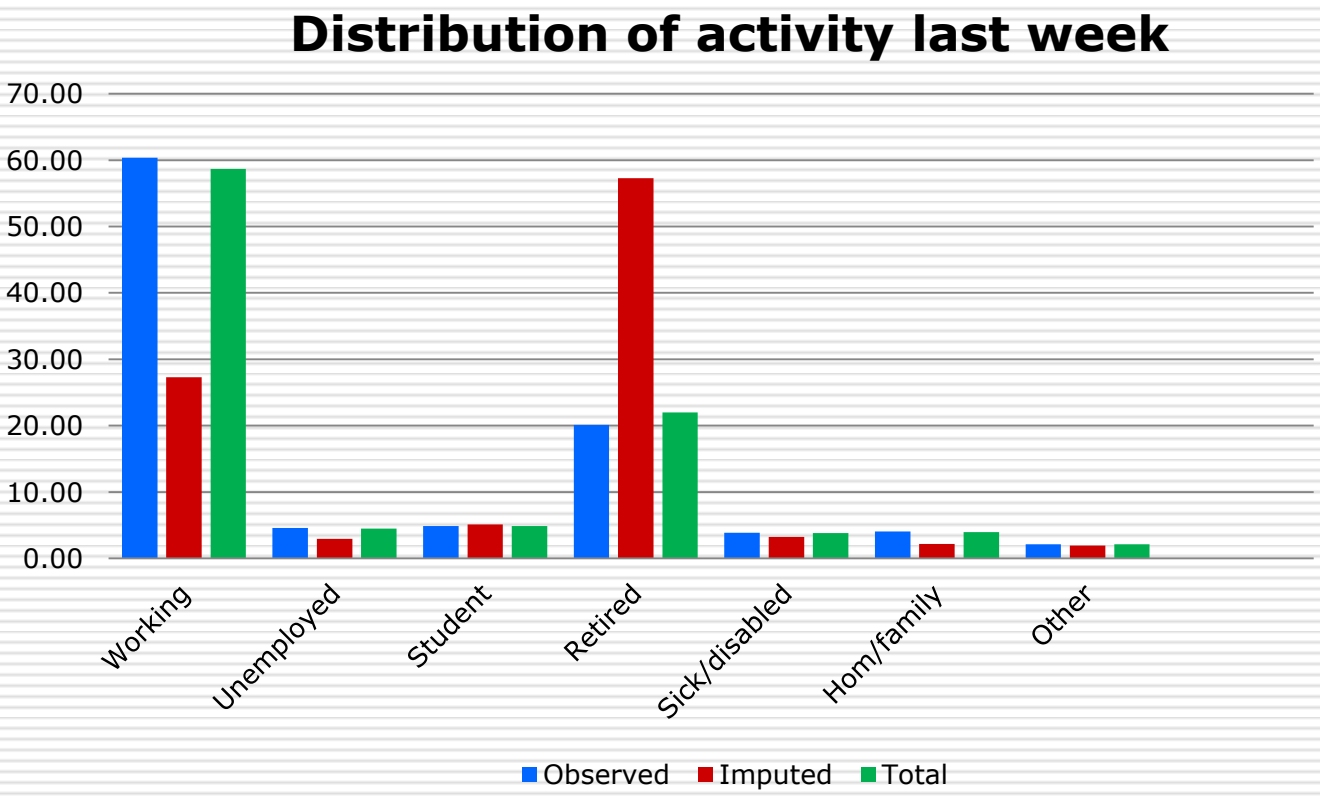
Assessing the performance of imputation

Distribution of economic activity last week, 2011 — Thousands

	Observed responses		Imputed responses		Difference (imputed - observed)	Total including imputed		Change (Total-observed)
	N	%	N	%	%	N	%	%
Working	24,653	60.4	602	27.3	-33.1	25,255	58.7	-1.7
Unemployed	1,880	4.6	65	2.9	-1.7	1,945	4.5	-0.1
Student	1,987	4.9	113	5.1	0.3	2,100	4.9	0.0
Retired	8,208	20.1	1,264	57.3	37.2	9,472	22.0	1.9
Sick/disabled	1,580	3.9	72	3.3	-0.6	1,652	3.8	0.0
Hom/family	1,653	4.0	48	2.2	-1.9	1,701	4.0	-0.1
Other	875	2.1	43	1.9	-0.2	918	2.1	0.0
Total	40,836	100	2,207	100	0.0	43,043	100	0.0

Source: England and Wales, Office for National Statistics, *2011 Census:Item Edit and Imputation: Evaluation Report*, June 2012

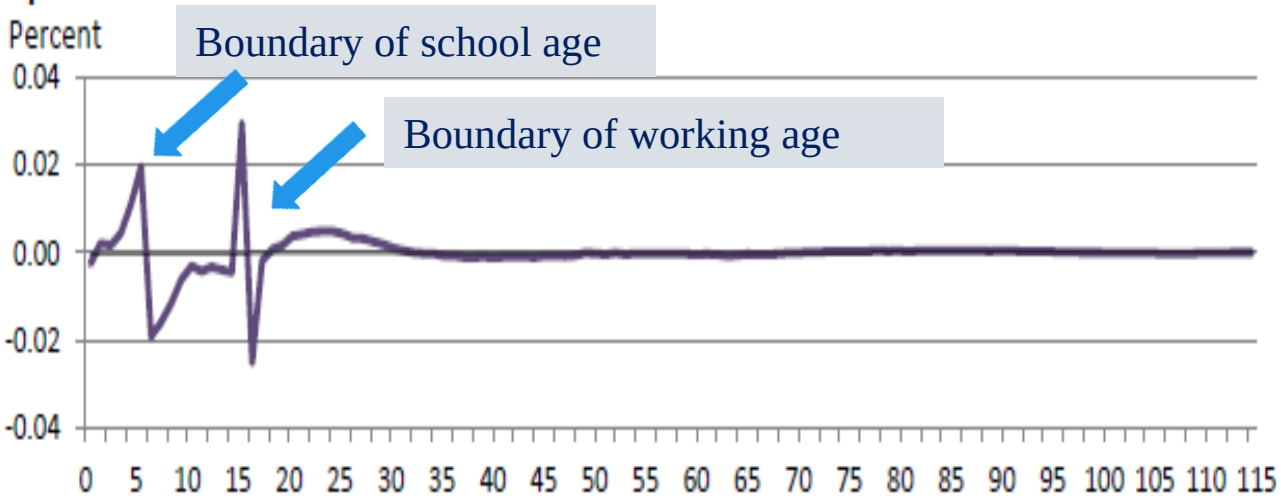
Assessing the performance of imputation



Source: England and Wales, Office for National Statistics, *2011 Census:Item Edit and Imputation: Evaluation Report*, June 2012

Understanding data editing and potential errors

Figure 6: Difference in the proportional distributions of single year of age before and after imputation



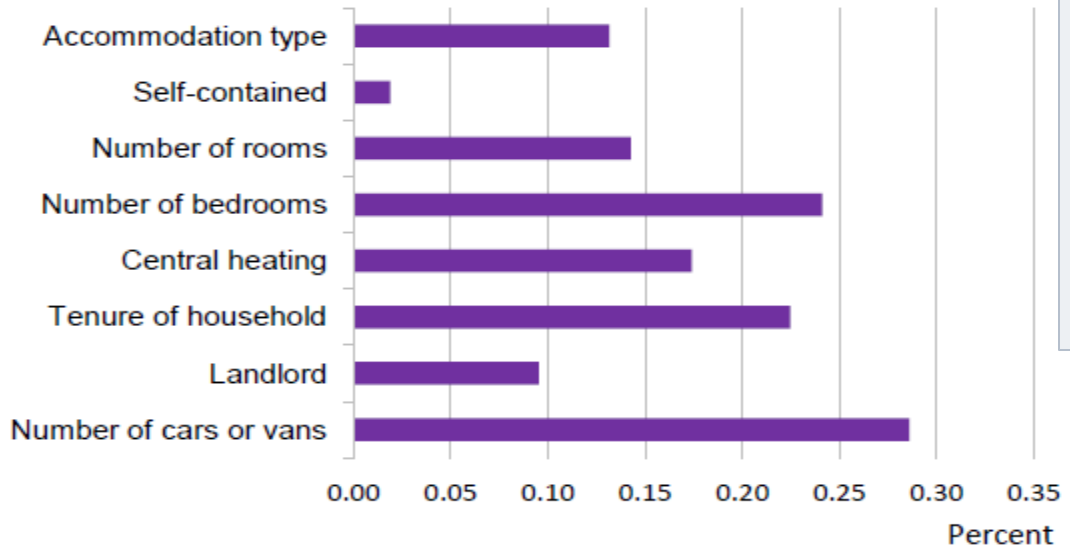
Source: England and Wales, Office for National Statistics, *2011 Census:Item Edit and Imputation: Evaluation Report*, June 2012

Assessing the performance of imputation

- Summary indexes at the variable level
  - Maximum absolute percent change
    - Maximum absolute percent change across all categories for each variable
  - Dissimilarity Index
    - Degree of change of two distributions (observed and total including imputed values) at the variable level
  - Imputation rate
    - Share of the imputed records in the total records

Assessing the performance of imputation

Figure 4: Maximum absolute percent change for any category after imputation - household questions

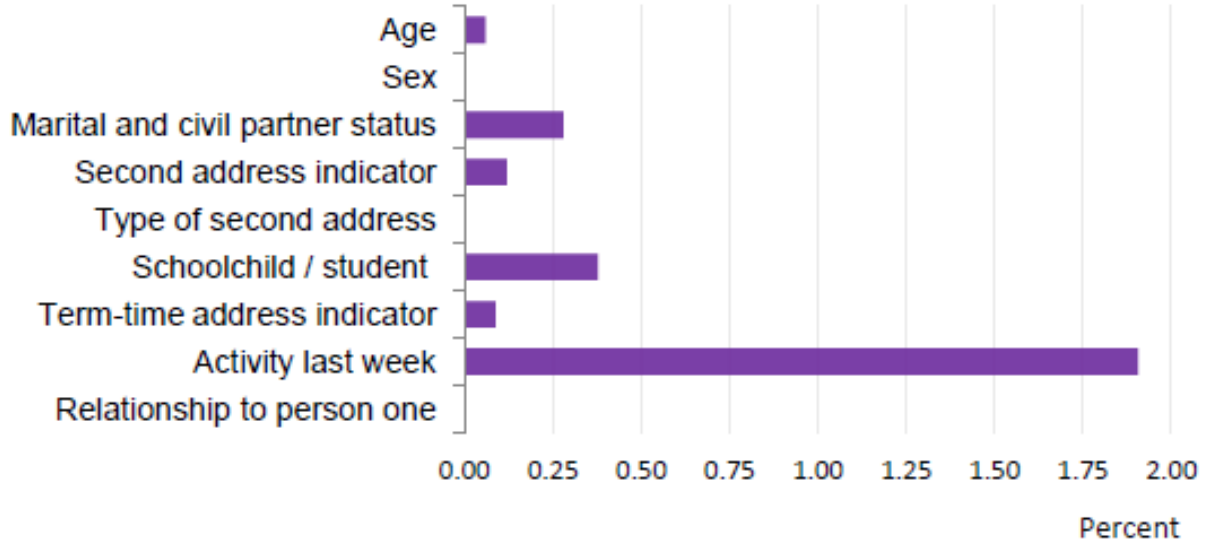


Maximum absolute percent change between the observed and final (imputed) distributions across all categories within each of the questions

Source: England and Wales, Office for National Statistics, *2011 Census:Item Edit and Imputation: Evaluation Report*, June 2012

Assessing the performance of imputation

Figure 5: Maximum absolute percent change in any category post imputation – demographic questions



Maximum absolute percent change between the observed and final (imputed) distributions across all categories within each of the questions

Source: England and Wales, Office for National Statistics, *2011 Census:Item Edit and Imputation: Evaluation Report*, June 2012

Index of dissimilarity

- To assess the degree of change induced by imputation on the initial distribution of variables

$$ID = \frac{1}{2}\sum_{k=1}^{K}\left| f_{y_k} - f^{*}_{y_k} \right|$$

Where;
k : categories of the variable
f : percentage distribution of the variable before imputation
f * : percentage distribution of the variable after imputation

Index of dissimilarity

$$ID = \frac{1}{2}\sum_{k=1}^{K}\left| f_{y_k} - f^{*}_{y_k} \right|$$

$0 \leq ID \leq 100$

- It assumes a 0 value when the two distributions before and after imputation are equal
- It is greater than 0 when they are different and reaches its maximum value of 100 when there is maximum dissimilarity between the two distributions
  - *when both are concentrated in one category which is different from each other*

Index of dissimilarity

Economic Activity Last Week, 2011 Thousands

	Observed responses		Imputed responses		Total including imputation		Absolute (observed-total)
	Number	%	Number	%	Number	%	|f-f*|
Working	24,653	60.4	602	27.3	25,255	58.7	1.7
Unemployed	1,880	4.6	65	2.9	1,945	4.5	0.1
Student	1,987	4.9	113	5.1	2,100	4.9	0.0
Retired	8,208	20.1	1,264	57.3	9,472	22.0	1.9
Sick/disabled	1,580	3.9	72	3.3	1,652	3.8	0.0
Hom/family	1,653	4.0	48	2.2	1,701	4.0	0.1
Other	875	2.1	43	1.9	918	2.1	0.0
Total	40,836	100.0	2,207	100.000	43,043	100.0	3.8
						DI	**1.9**

Source: England and Wales, Office for National Statistics, *2011 Census: Item Edit and Imputation: Evaluation Report*, June 2012

Assessing the performance of imputation

Population and Housing Census of Albania, 2011 Census

	Imputation Rate	Dissimilarity Index
Citizenship (Albania)	10.98	11.0
Had live-born children	9.81	1.3
Family nucleus	9.49	5.8
Highest completed level of education	7.55	2.2
Place of residence, 2001	5.72	2.6
Had a job last week of September	5.32	2.7
Ownership of the dwelling	7.3	6.0
Use of agricultural land	3.1	3.1

Individual dataset

Number of records = 2,800, 138

Number of variables = 66

Household dataset

Number of records = 722,262

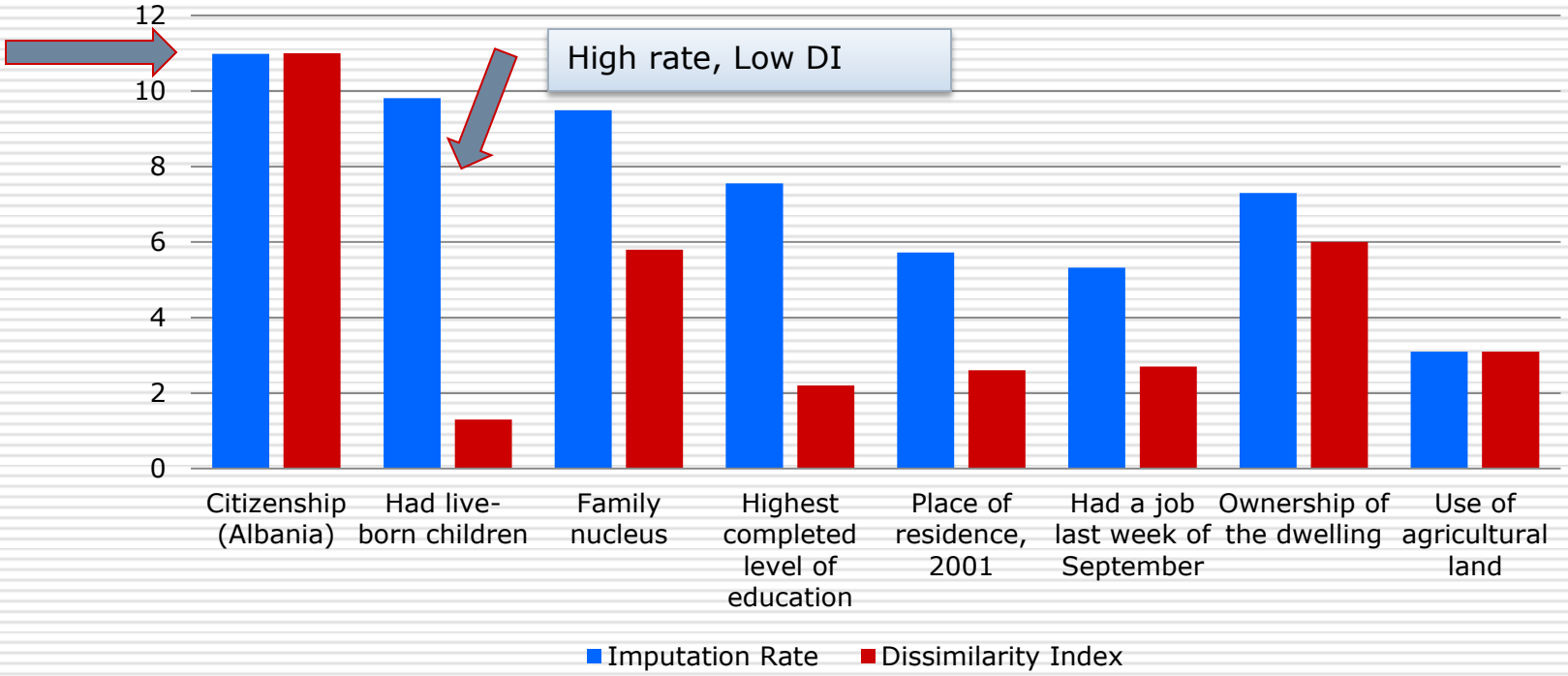
Number of variables = 30

Imputation Rate: Number of imputed records/ Total nunber of records*100

Source: Albania, Quality Dimensions of 2011 Population and Housing Census, May 2014

Assessing the performance of imputation

Comparison of imputation rate and dissimilarity index



Source: Albania, Quality Dimensions of 2011 Population and Housing Census, May 2014

Hands-on exercises

- England and Wales – 2011 Census

A. Marital and civil partnership

B. Distribution of highest level attended