Notater
Documents
2014/27

Statistics Norway
Statistisk sentralbyrå

Susie Jentoft

Imputation of missing data among immigrants in the Register of the Population's Level of Education (BU)

Documents 2014/27

Susie Jentoft

Imputation of missing data among immigrants in the Register of the Population's Level of Education (BU)

Statistisk sentralbyrå • Statistics Norway
Oslo–Kongsvinger

Documents In this series, documentation, method descriptions, model descriptions and standards are published.

© Statistics Norway
When using material from this publication, Statistics Norway shall be quoted as the source.

Published June 2014

ISBN 978-82-537-8956-9 (electronic)
ISSN 1891-5906
Subject: Education

Symbols in tables	**Symbol**
Category not applicable	.
Data not available	..
Data not yet available	…
Not for publication	:
Nil	-
Less than 0.5 of unit employed	0
Less than 0.05 of unit employed	0.0
Provisional or preliminary figure	*
Break in the homogeneity of a vertical series	—
Break in the homogeneity of a horizontal series	|
Decimal punctuation mark	.

Preface

This documentation describes a proposed method for imputing missing data among immigrants in the Register of the Population's Level of Education (BU) in Norway.

This report has been written by Susie Jentoft with much input from Aslaug Hurlen Foss from the Division for Statistical Methods who started this work back in 2005. Additionally Alice Steinkellner and Anne Marie Rustad Holseter from the Division for Education Statistics have made valuable contributions to this work. A number of others have provided information on the various imputation methods used at Statistics Norway including Johan Fosen, Li-Chun Zhang, Ole Villund and Paul Inge Severeide.

Statistisk sentralbyrå, 10 March 2014

Hans Henrik Scheel

Abstract

In Norway, the Register for the Population’s Level of Education (BU) contains information on all residents, 16 years of age and older. While in general, missing data is minimal in this register (around 3 percent), increases in immigration to Norway in recent years is creating knowledge gaps about the level of education of its new residents. Census-surveys have filled some of the missing data, however, around 20 percent of immigrants still have an unknown level of education. With rising non-response in these census-surveys, the problem is unlikely to disappear.

An imputation method is proposed to address the non-response bias from these surveys and to create a “complete” dataset. A missing at random (MAR) assumption is made, with imputation being based on a nearest neighbour technique called predictive mean matching. The auxiliary variables used to find a matching donor include gender, age, occupation, income, length of time living in Norway, citizenship and country of origin.

Results from the imputed dataset show some small overall changes. In general, imputation reduces the percentage of both the highest and lower education levels. Comparison of the imputation results among Swedish immigrants with data from Statistics Sweden shows the imputation proposed is adjusting the data in the right direction.

Contents

1. Introduction

Statistics Norway maintains a register of the population's level of education for all residents 16 years of age and over (Befolkningens utdanningsnivå: BU). The data comes from a number of sources including educational institutes and the governmental student loan agency. Immigrants who have studied abroad are generally missing in this register. With an increase in the number of people immigrating to Norway in recent years, this is increasingly becoming a problem with the register.

To address this issue a number of census-surveys have been completed. This involved sending questionnaires to immigrants who were missing level of education data in the register. The first was in 1991 followed by one in 1999 (Fosen, Johnsen et al. 2000) and then two recent surveys in 2011 and 2012 (Steinkellner and Holt 2013). The recent census-surveys were only sent to those who were 20 years of age or older. While these surveys have been compulsory to respond to, non-response was still a problem at around 36 percent for 2011/12 including postal returns (Steinkellner and Holt 2013). As of 1st October 2012, a total of 122 950 or 3 percent were missing (total register population of 4 061 984) information on level of education in BU. Among immigrants, a total of 108 654 or 20 percent are missing from the register (of a total of 533 050 immigrants).

Missing data is primarily a problem at this level when those missing data are in some way different to those that we have information on. This can introduce a bias to the statistics we produce and may not accurately reflect the true values in the population.

This document investigates those with missing data in the BU register and suggests an imputation method which could be used to fill-in missing values. The work for this started back in 2005 with an investigation on how the 1999 survey could be used to impute missing data (Foss 2006). Here we extend this work using a nearest-neighbour imputation method to estimate the distribution of level of education among immigrants.

1.1. Project goal

To investigate missing patterns among immigrants in the BU register and to create a complete dataset for level of education for immigrants using imputation methods.

1.2. Current imputation practices in other registers at Statistics Norway

A number of registers at Statistics Norway already use imputation methods for missing data. Three of these are briefly mentioned here as examples. Further imputation methods are used and have been suggested by the Division for Statistical Methods for use in survey data.

1.1.1. Occupation

Occupation is currently imputed in the Norwegian register for people who are employed and are missing data. This is a categorical, hierarchical (for the most part) code system. The imputation is strongly based on the level of education and industry the person works in (if available). Individuals are stratified into groups which are as homogenous as possible, with reference to occupation. For groups that are not homogeneous, occupation codes are stochastically drawn from the distribution seen in the observed members of the group. Cross-referencing is done with results from the labour force survey.

1.1.2. Register based employment statistics

For the calculation of register based statistics on employment status, a complex micro-integration approach is taken. In this case, many registers (more than 10) are used and harmonised to create the statistics. Individuals are classified into groups based on the support for employment found in the register information. For example, if an individual is missing from the employer/employee register but is registered in the wage sum register they are classified into an uncertain group. If an individual is in the employer/employee register and is registered in the wage sum register they are placed in a certain-of-employment group. We may consider the uncertain groups as missing data (but with good auxiliary information). Employment status for individuals in the uncertain groups is determined using an income cut-off value which is based on the estimated number of employed from the quarterly Labour Force Survey (Fosen 2011). Only those with registered incomes over the cut-off value are assumed to be employed. This approach may be seen as a kind of restricted logistic regression within groups using income as the explanatory variable.

1.1.3. Household statistics

In order to calculate statistics at a household level, individuals are joined into a household unit. This is done primarily using the address information provided from the Central Population Register (CPR) and the register on ground properties and addresses called Matrikkelen. If information is missing or partially missing from these registers, it is not always possible to deduce household units. In this case a nearest neighbour approach is taken using a two-step process. This is described further in Zhang and Hendriks (Zhang and Hendriks 2012).

2. Comparison of immigrant and general populations

Level of education will be specified using broader levels than is found in the register today. This reflects the reporting level that will be used for the data, as well as the uncertainty around the estimates. Level of education will be recorded in 5 levels, as shown in table 1.

Table 1. Level of education explanation

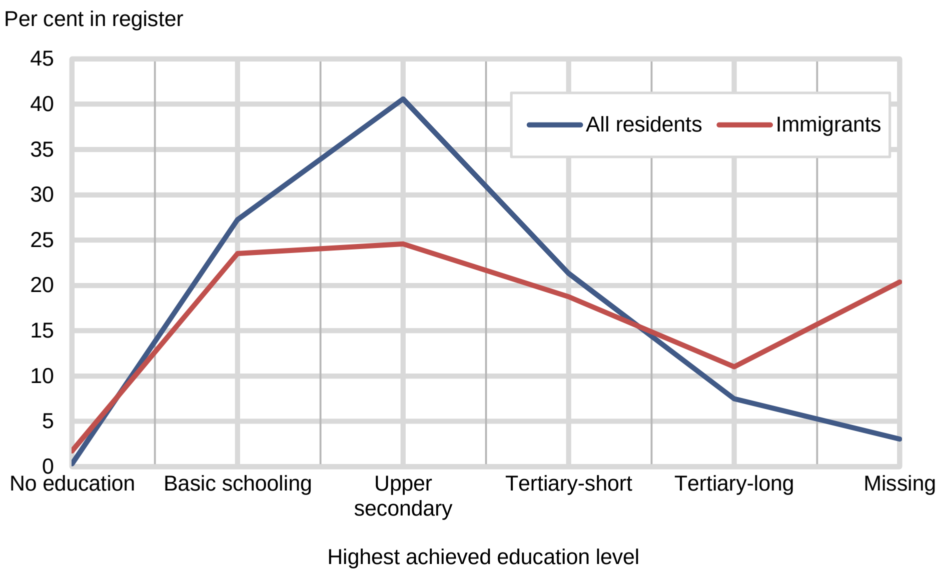
New education level	New education level explanation	Previous level (`nivaa`)
1	No formal education completed	0
2	Basic school education	1,2
3	Upper secondary education	3,4,5
4	Tertiary education - short (up to 4 years)	6
5	Tertiary education -long (over 4 years)	7,8

We have defined immigrants as those who were not born in Norway and have parents and grand-parents who were not born in Norway (Dzamarija, Andreassen et al. 2013). This definition includes those whom have attained Norwegian citizenship. Table 2 and figure 1, show that the distribution of level of education varies between the full register and that of Norway’s immigrant population. In general, the immigrant population has a higher proportion of both the lowest and highest education levels compared to the general population. The proportion of missing data is much greater among the immigrant population.

Table 2. Level of education in register including immigrants (as of 01.10.12)

		0	1	2	3	4	5	6	7	8	Missing	Total
All residents	Previous grouping	12 520	14 505	1 093 393	526 844	1 014 636	106 727	866 348	277 324	26 737	122 950	4 061 984
	New education level	12 520	1 107 898		1 648 207			866 348	304 061		122 950	4 061 984
Immi-grants	Previous grouping	9277	14 378	110 996	19 939	107 344	3662	99 991	49 894	8915	108 654	533 050
	New education level	9277	125 374		130 945			99 991	58 809		108 654	533 050

Figure 1. Level of education in full register and among immigrants



2.1. Coverage of the register

An important point to consider in the register is also whether the missing group is actually in the target population: ie do they still reside in Norway? This may be of particular concern among immigrant populations whom may be more likely to emigrate out of the population. If individuals do not register that they have emigrated from Norway and do not return to Norway, there will be an element of over-coverage in the register. One reflection of this may be in the high postal return rate in the surveys (around 10 per cent during the last survey). Individuals that have emigrated and not registered their move will most likely be identified eventually and updated, but there may be a long time-lag.

Under-coverage in the register may also be a problem. This is when individuals have moved to Norway but are not registered in the population register. If under- and over-coverage are similar in size and structure (characteristics), then the register will be a good representation of the population. However, under-coverage is likely to be less of a problem than over-coverage due to the incentives for registering in Norway (health cover, welfare etc.). Vassenden (2001) has previously stated that the register coverage of immigrants in Norway is probably fairly good, particularly if seen over a period of a few years but a formal analysis of this hasn't been performed recently.

3. Using auxiliary variables to investigate missing data

Statistics Norway holds a wide range of auxiliary information which is available for all (or a large proportion of) immigrants. This information is important to investigate for those that are missing in the register in order to build up a method for imputing level of education. Generally in missing data treatments, we want to use auxiliary information that is correlated with both the interest variable (level of education) and the response propensity/missingness. Table 3 lists the variables investigated.

Table 3. List of auxiliary variables used to look at missing data patterns.

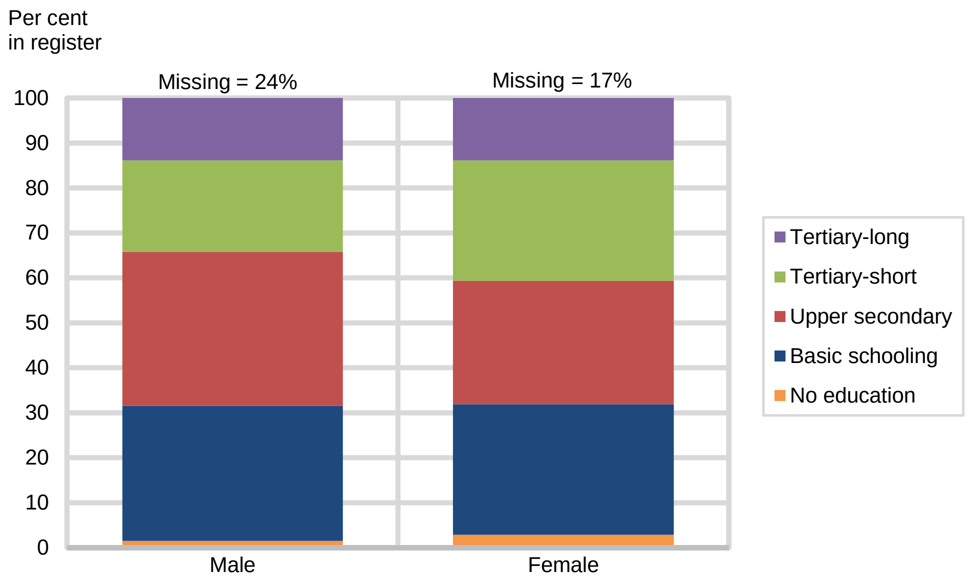
Auxiliary variables	Explanation	Code name
Country of origin	This is generally the registered nationality of the individual prior to coming to Norway	`landbak`
Gender	The gender of the individual	`kjoen`
Age	Age of the individual	`alder`
Parents education	Parents level of education.	`sosbak`
Source	Indicates where the data has comes from. Levels 07 & 08 are education surveys (2011 and 2012)	`kilde`
Citizenship	Country of citizenship.	`statborg`
Region	Region that the individual lives in	`fylke`
Occupation	Registered occupation using 10 standard groupings from the AA-register. There are 287 920 individual immigrants with registered occupational codes.	`yrk_kode`
Residence time	The length of time since the individual registered as living in Norway. This is 0 if they registered in 2012, 1 if register in 2011 etc.	`botid`
Income	The registered income from wages and self-employment (net income) during the calendar year for 2011 for the individual.	`yrkesinntekt_2011`

While the parents' level of education is likely to be highly correlated with our variable of interest, missing data is a problem with this variable. Only around 5 percent of those missing level of education had values for this variable. As a result, it was not seen as a viable option in the imputation model. The region variable did not show great variations in distributions for level of education so was not used in modelling and is not shown further here. The source of the data is obviously not available for those missing in the register, so is not used in the modelling, however, it is addressed in a later paragraph.

3.1. Gender

Level of education appears to vary little among genders. In general, females show a slightly higher percentage of tertiary education compared to males. A higher percentage of males are missing in the register. Gender differences may interact with other variables, particular country of origin.

Figure 2. Level of education, by gender for immigrants



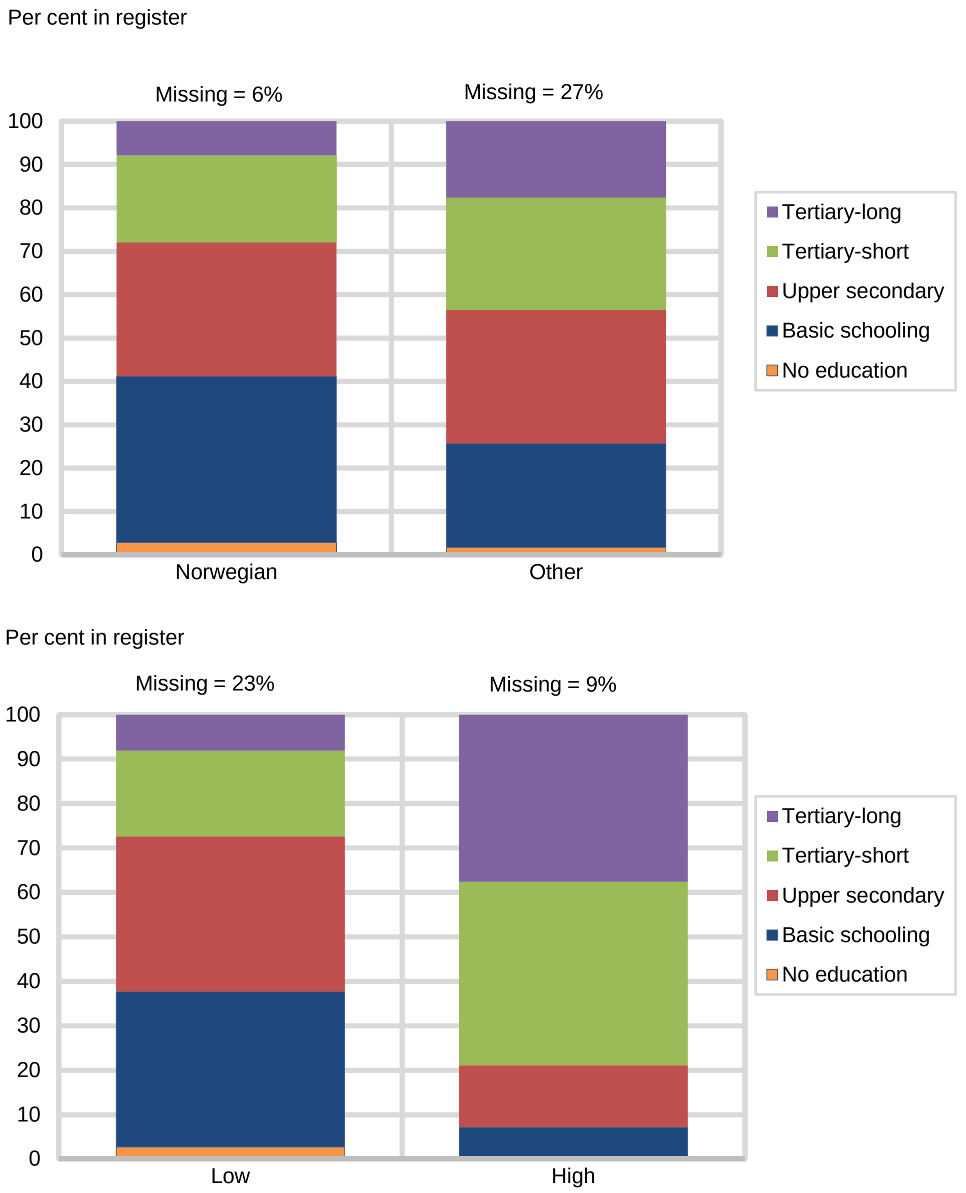
3.2. Citizenship

A comparison between immigrants that are now Norwegian citizens and those that are not, showed differences in level of education. Those that are not Norwegian citizens, generally had a higher level of education and greater levels of missing data in the register.

3.3. Occupation

Breaking occupation into high and low educational requirement groups shows strong differences in level of education. The high educational requirement group includes managers, professionals, technicians and associate professionals, military and government workers. The low educational requirement group includes those working in clerical support, sales and service, agricultural, forestry and fisheries, craft and trades, plant and machine operators, other fields and those missing registered occupational codes. The low educational requirement group had in general higher levels of missing data in the register.

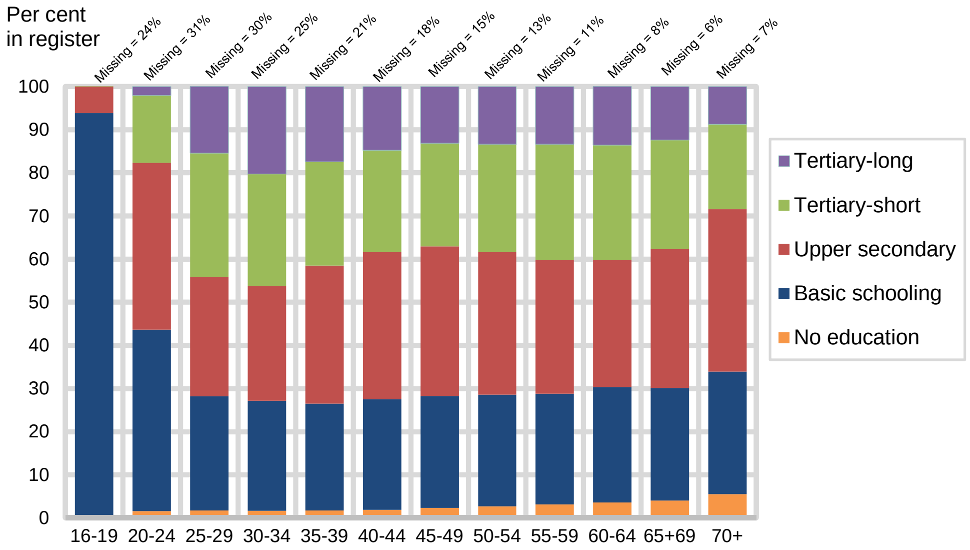
Figure 3 and 4. Level of education by citizenship (left) and by occupation high/low educational requirements (right) for immigrants



3.4. Age

Age is presented in 5-year groups with the youngest group being a 4-year group: 16-19 year olds. 15-year olds are included in the BU register but have a very high percentage of missing data. Of the 4714 immigrant 15 year olds, only 37 contained data on level of education. We have therefore excluded them in this report and from imputation. The oldest group contains those 70 years and over. The two groups of those under 25 years of age show quite a different level of education pattern to those above 25 years of age. The proportion missing in the register decreases with age. It is important to note that the recent surveys were only sent to those 20 years of age or older. This explains the high level of missing data among the younger groups. In general for those 25 years and older, level of education decreases with age.

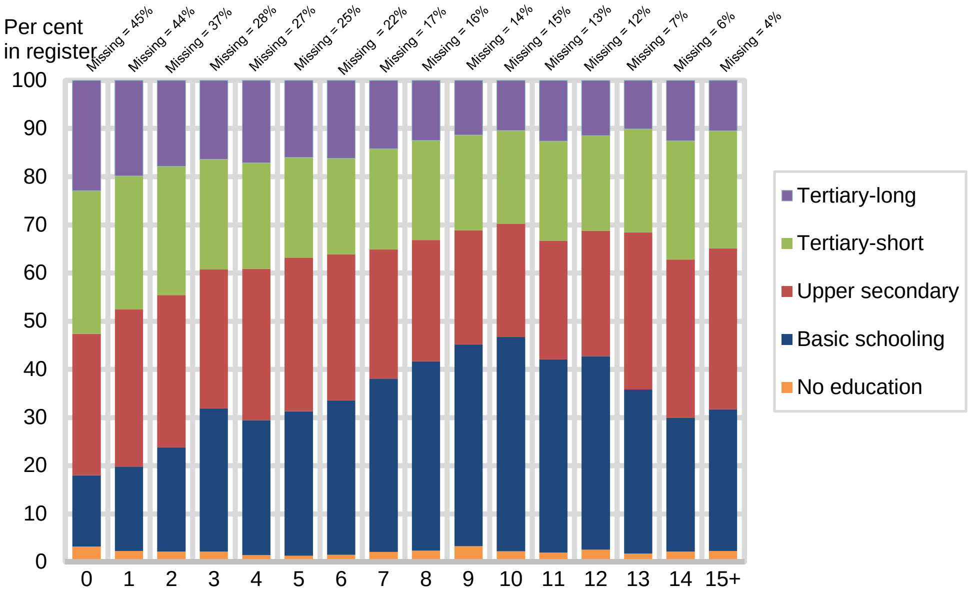
Figure 5. Level of education, by age groups for immigrants



3.5. Length of residence

It appears that level of educational and the proportion missing in the register decreases with length of residence. At around 10 years of residence there is perhaps a small reverse trend where level of education increases. Those who have immigrated here most recently have the most missing data in the register.

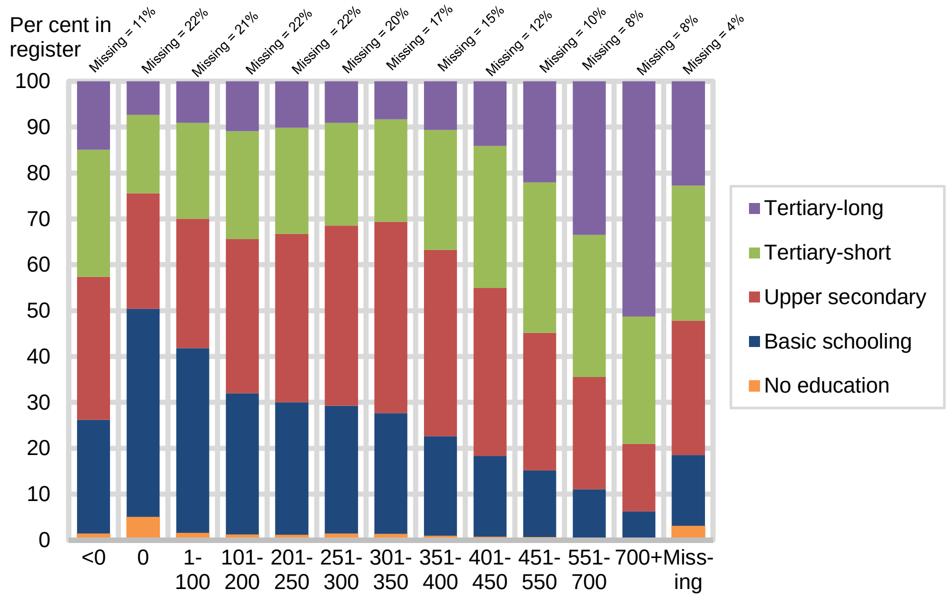
Figure 6. Length of residence time in Norway in years for immigrants. Zero indicates they registered in 2012 and 15 includes 15 or more years



3.6. Income

The income variable shows a strong correlation with level of education. Here income is the amount received from wages and self-employed income (net). People with higher incomes generally had higher education levels. Those with negative incomes were more similar to those with higher incomes than zero income, therefore the absolute income value was used in the imputation model.

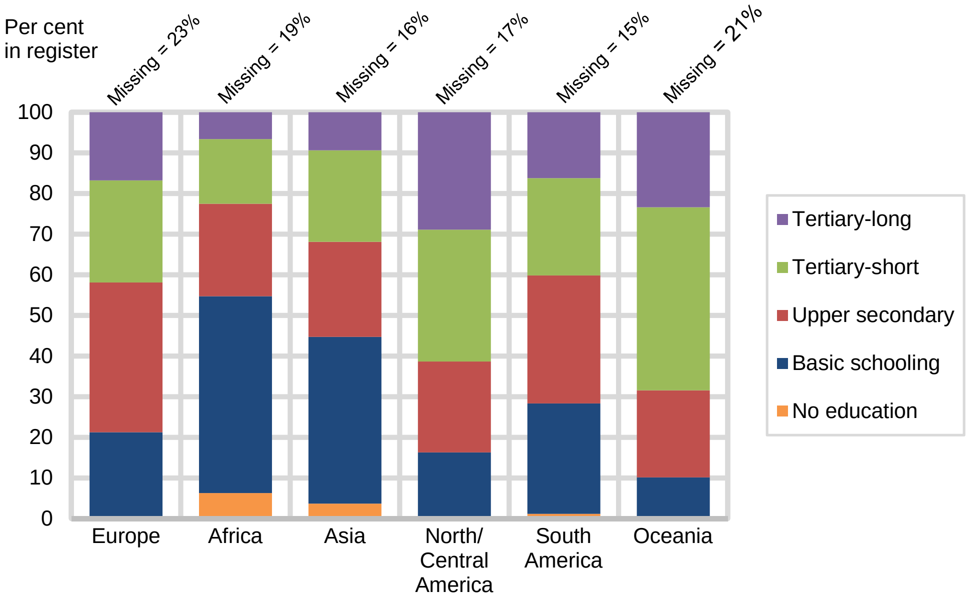
Figure 7. Level of education, by income group for immigrants expressed in NOK 1 000



3.7. Country of origin

The country of origin is known from previous studies to correlate with level of education. It is also likely to interact with other variables. The figure below shows large variations in level of education between world regions.

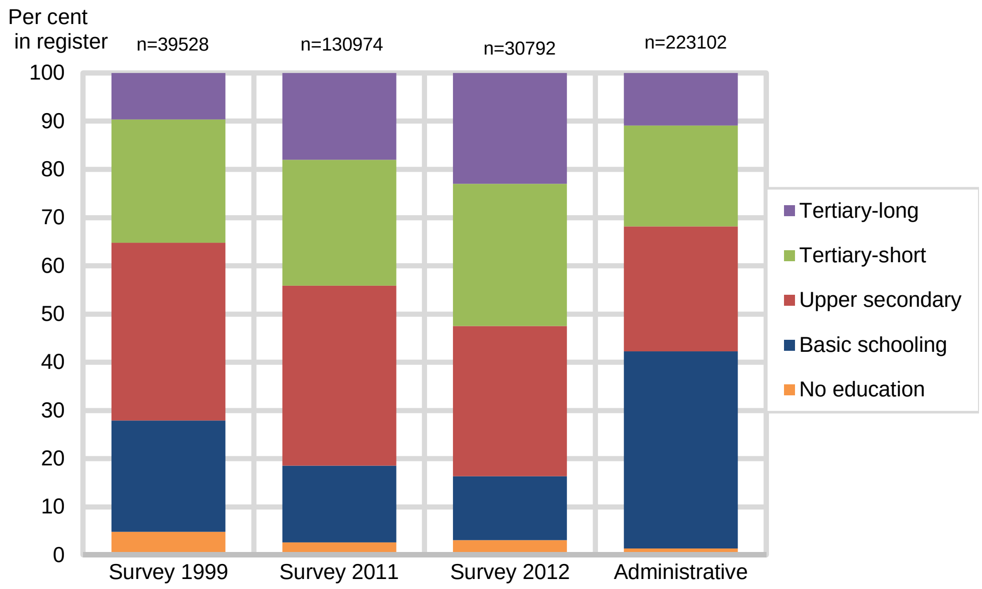
Figure 8. Level of education, by region of origin for immigrants



3.8. Sources of data

The BU register data comes from a wide range of sources (see Appendix A) but falls into 2 categories, survey and administrative sources. The following graph shows a comparison of the distribution of level of education from 3 of the surveys and from all other administrative sources (and the 1991 survey). The n-values represent how many people in the BU register have information from that particular source. We can see that those who have responded to the most recent surveys show an increasingly higher level of education than that seen in the previous surveys and from administrative sources. There are perhaps two main reasons that could explain this. Firstly, more recent immigrants may in fact have a higher level of education than previously seen. Alternatively, it may be that non-response in the surveys, which has increased in recent years, is giving a selective bias towards those with higher levels of education.

Figure 9. Level of education of immigrants, by data source



4. Imputation

Work prior to this project (Foss 2006) and interest from the Division of Education Statistics has led to this investigation on whether imputation could be used for handling missing data for level of education among immigrants. Imputation is when missing data points are estimated in some way and filled-in, producing a "complete" dataset. This allows easier construction of tables and aggregation. It can also improve the quality of the data if the approach adjusts for selective non-response (Waal, Pannekoek et al. 2011).

4.1. Imputation method

Previously, three cold deck imputation methods using different variable groupings have been investigated (Foss 2006). One difficulty with standard cold or hot deck imputation is the restriction of how many auxiliary variables can be used because imputation cells/groups become too small. Nearest Neighbour Imputation (NNI) is a special type of cold or hot deck imputation which uses a distance function to find the donor closest to the missing value with respect to auxiliary information. It is more flexible in allowing more auxiliary information to be used and was therefore the preferred method in this case. NNI is a commonly used imputation technique for missing data at other statistical bureaus (Fay 1999; Rancourt 1999) and is already used for some registers and survey variables at Statistics Norway.

NNI has advantages over other imputation techniques (for example regression imputation) in that it will only impute occurring values (ie. Level of education will never be imputed as a negative value or higher than those specified) and can make use of all types of auxiliary information. It is said to be semi-parametric and while it may make use of an imputation model (as the distance function), it does not fully rely on it, making it less sensitive to model misspecifications (Durrant 2005). Additionally, Chen and Shao show that NNI can provide asymptotically valid distribution and quantile estimators (Chen and Shao 2000).

For imputation of level of education we suggest using predictive mean matching to create the distance function for NNI. This method was first described by Little (Little 1988) and has the advantage of creating distances based on the predictive power of the auxiliary information. In general, this method involves fitting a regression model to the data available and predicting values for both missing and non-missing individuals. The closest predicted value to that of the missing individuals is then selected as the donor. The observed value of the donor is used to impute the missing individual.

In detail, the following steps were taken where y_i is the observed level of education for non-missing individual *i*. A linear regression of the target variable was fitted for non-missing y values:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \varepsilon_i$$

Where x_{1i} is a binary variable for the gender of individual *i*, x_{2i} is the age of individual *i*, x_{3i} *is the* income from wages and self-employment earnings (net) expressed as the log of the absolute value of occupational income for individual *i*, x_{4i} is a binary variable for high and low educational requirement occupations for individual *i*, x_{5i} is a binary variable for Norwegian and non-Norwegian citizens for individual *i*, and x_{6i} is the length of time that individual *i* has lived in Norway (up to 15 years).

The estimated β coefficients are then used to find the predicted values $\hat{y}_i$ for both missing and observed values of *y*. For a given missing value y_k, let y_k^* denote the imputed value. Then $y_k^* = y_d$ where *d* is the donor in the response sample that minimizes, for all units *i* in the response sample,

$$D(k,i) = | \hat{y}_i - \hat{y}_k |$$

If there are several donors that minimize *D*, one is selected randomly. This means the results are partially stochastic.

In the approach described here, we assume that the missing data follows a missing at random (MAR) pattern. This means that whether or not data are missing may depend on the values of the auxiliary variables but not on our interest variable, level of education, itself. If this assumption does not hold, it is said to be not missing at random (NMAR) or nonignorable (Durrant 2005) and is discussed later on in this report.

Nearest neighbour imputation was performed within groups. Figure 5 shows that the distribution among those under 25 years of age is very different to the rest of the population and so these are also imputed separately. Additionally, income is missing in nearly 10 percent of the immigrant population. As this is a continuous variable we can not group them together in a separate group. Therefore we decided to impute this group separately using a model without income. This resulted in four main groups. The regression model shown previously was fitted to the groups. AIC backward selection showed that residence and nationality were not important in the model for those under 25 years of age and were dropped. The following table shows the variables used in each of the four main groups.

Table 4. Models used for imputation within the four main groups.

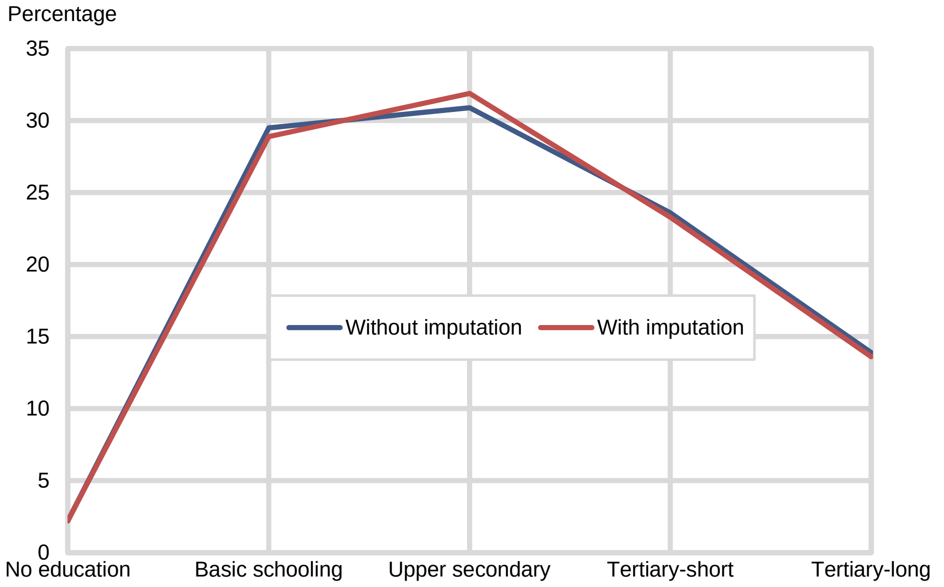
Group	Model
Under 25 years without income variable	Education = gender + occupation + age
Under 25 years with income variable	Education = gender + occupation + age + income
25 years and over without income variable	Education = gender + occupation + age + citizenship + length of residence
25 years and over with income variable	Education = gender + occupation + age + citizenship + length of residence + income

The country origin of an individual has consistently been seen as a vital player in their level of education (see figure 8) (Fosen, Johnsen et al. 2000; Steinkellner and Holt 2013). Therefore, imputation is done within country of orign in addition to within the four main groups shown in table 4 if there were enough observations. If there were fewer than 30 observations at a country level, the country was allocated to a general group for each of the world regions. The donors used were from both survey and administrative sources. Analysis for this work was done in R version 2.15.2.

4.2. Imputation Results

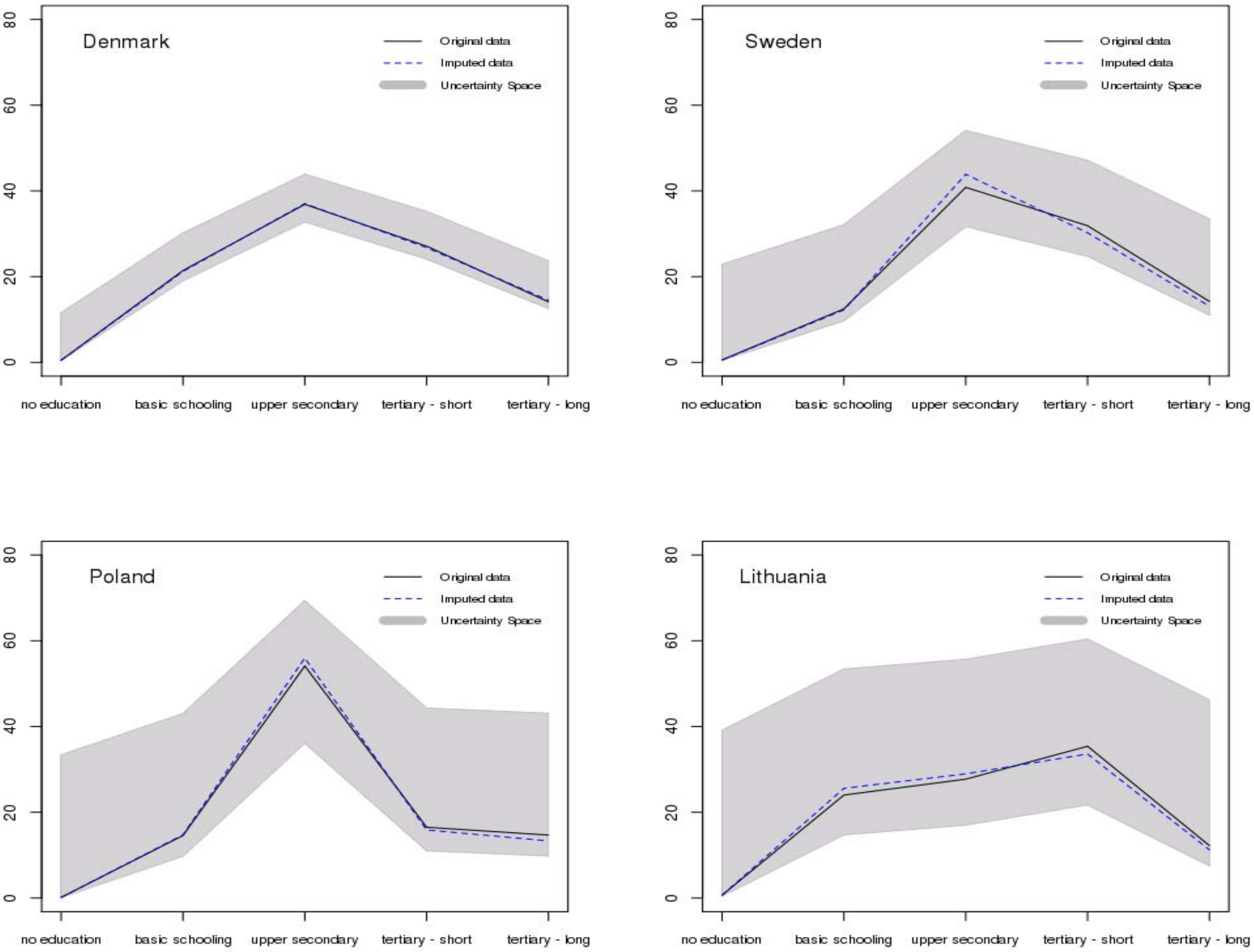
Figure 10 shows the overall effect of the imputation on level of education among immigrants. There is no large change but a small increase in the level of upper secondary education (level 3) and a decrease in the tertiary education levels (4 and 5). This likely reflects an element of selective non-response in the recent surveys.

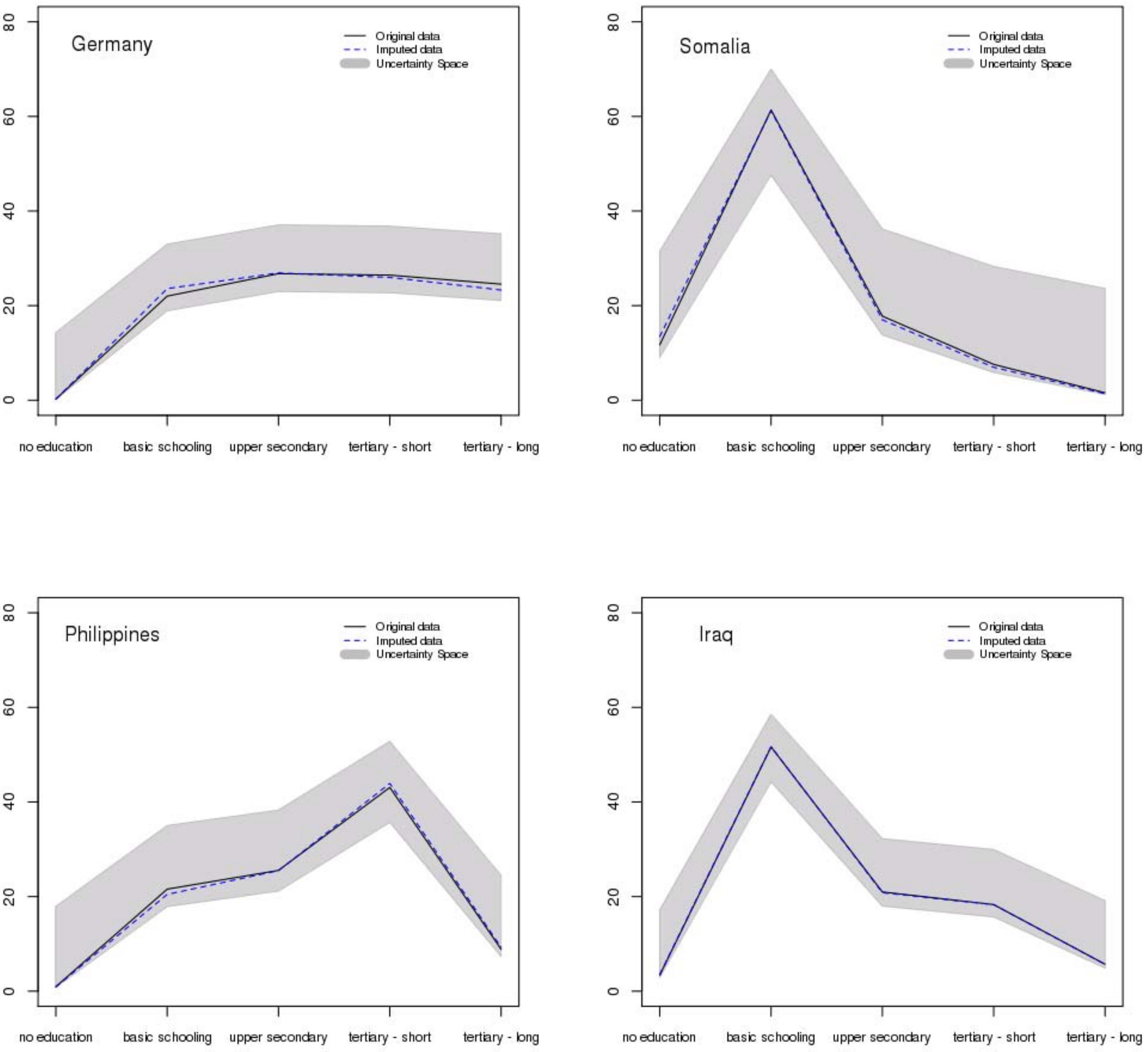
Figure 10. Level of education among immigrants. Solid line indicates values excluding missing data and dotted line gives the level including imputed values

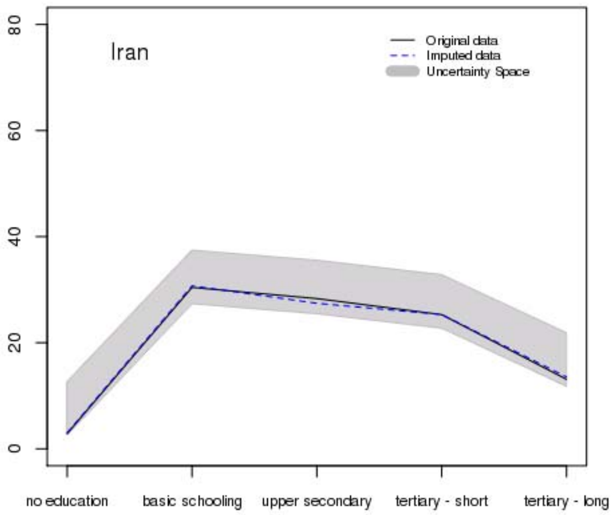


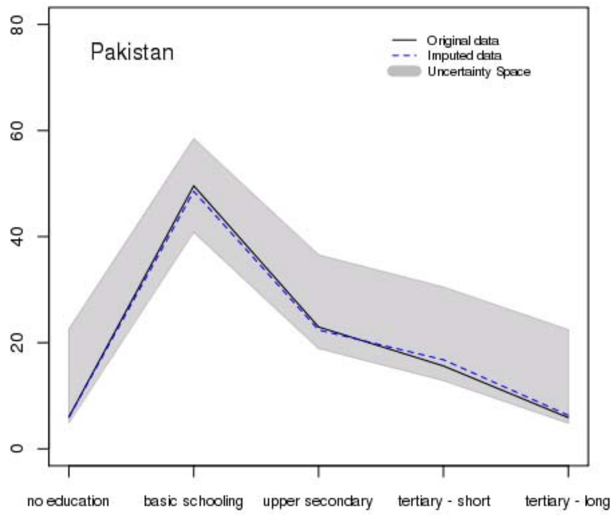
At a country of origin level, some changes are seen but they are not dramatic in most cases. Those countries with very high percentages of missing data and low absolute numbers had the greatest differences between values with and without imputation. The following are graphs of the top 10 countries in terms of their: absolute number of immigrants, absolute number of individuals missing from the register, greatest percentage of missing data (where there was at least 30 individuals from that country). The grey areas represent the uncertainty space defined as follows: for each education level, if all individuals with missing data are imputed with that education level a maximum level is found. Similarly, if no individuals with missing data are imputed at that level, a minimum percentage is created. This is a reflection of how much missing data there is for a country of origin group.

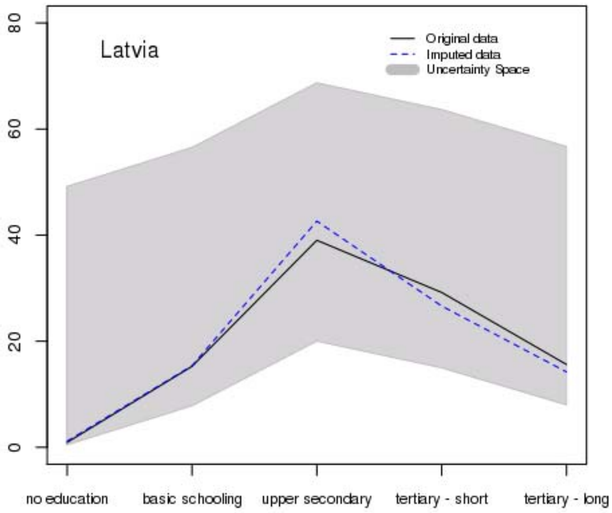
Figure 11. Example countries with high absolute numbers, high missing data and/or high percentage missing. Figures show original percentages of level of education among immigrants in the register excluding missing values, percentages including imputed values and the uncertainty relating to the percentage missing.

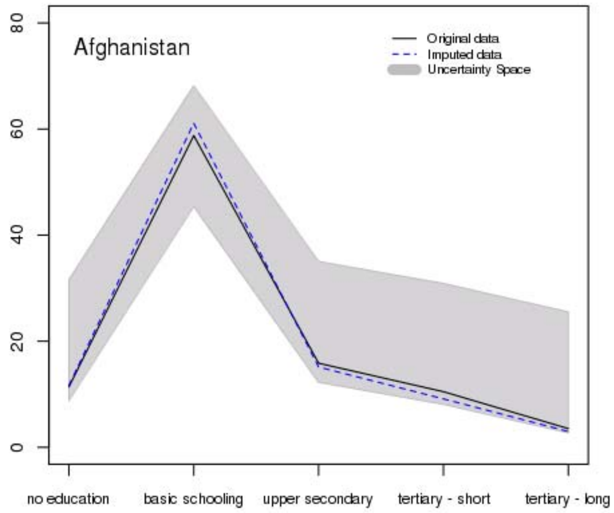


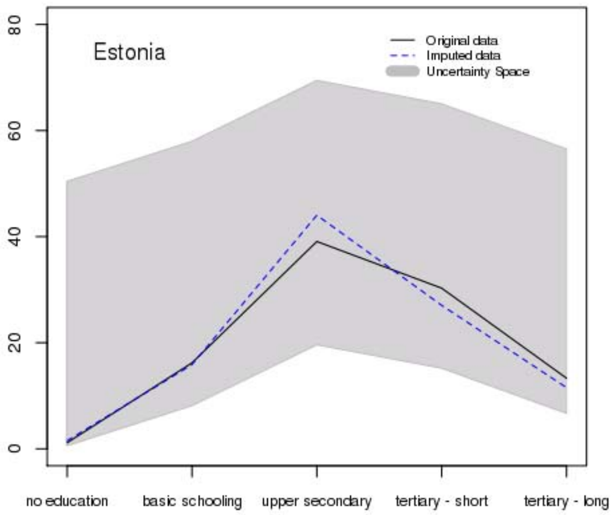

Germany
Somalia
Philippines
Iraq
Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long

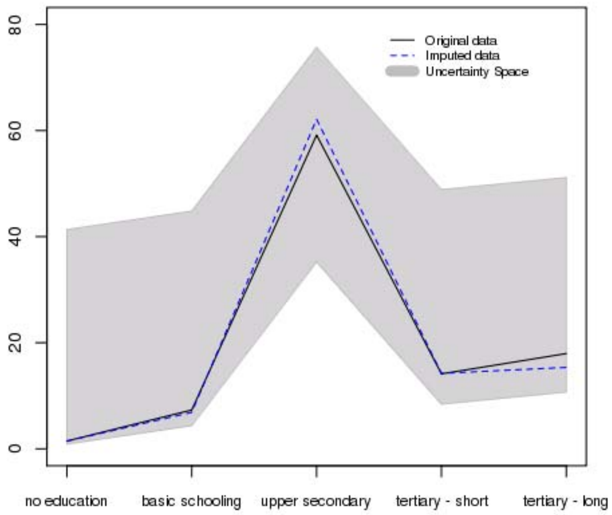

Iran
Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long

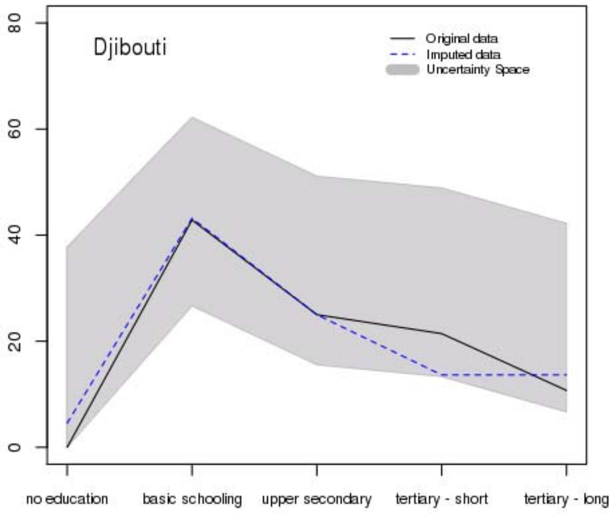

Pakistan
Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long

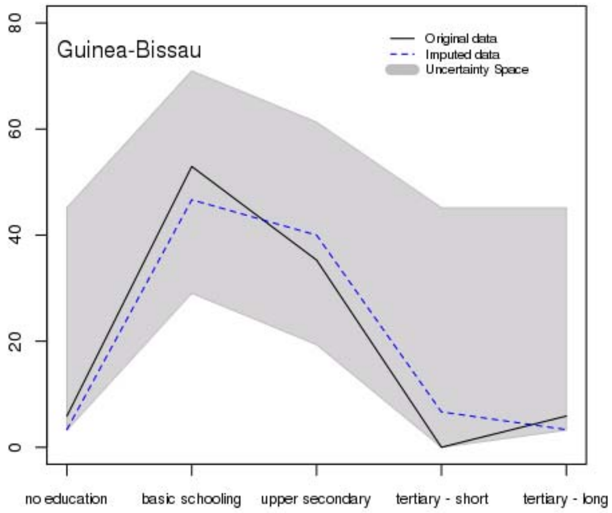

Latvia
Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long

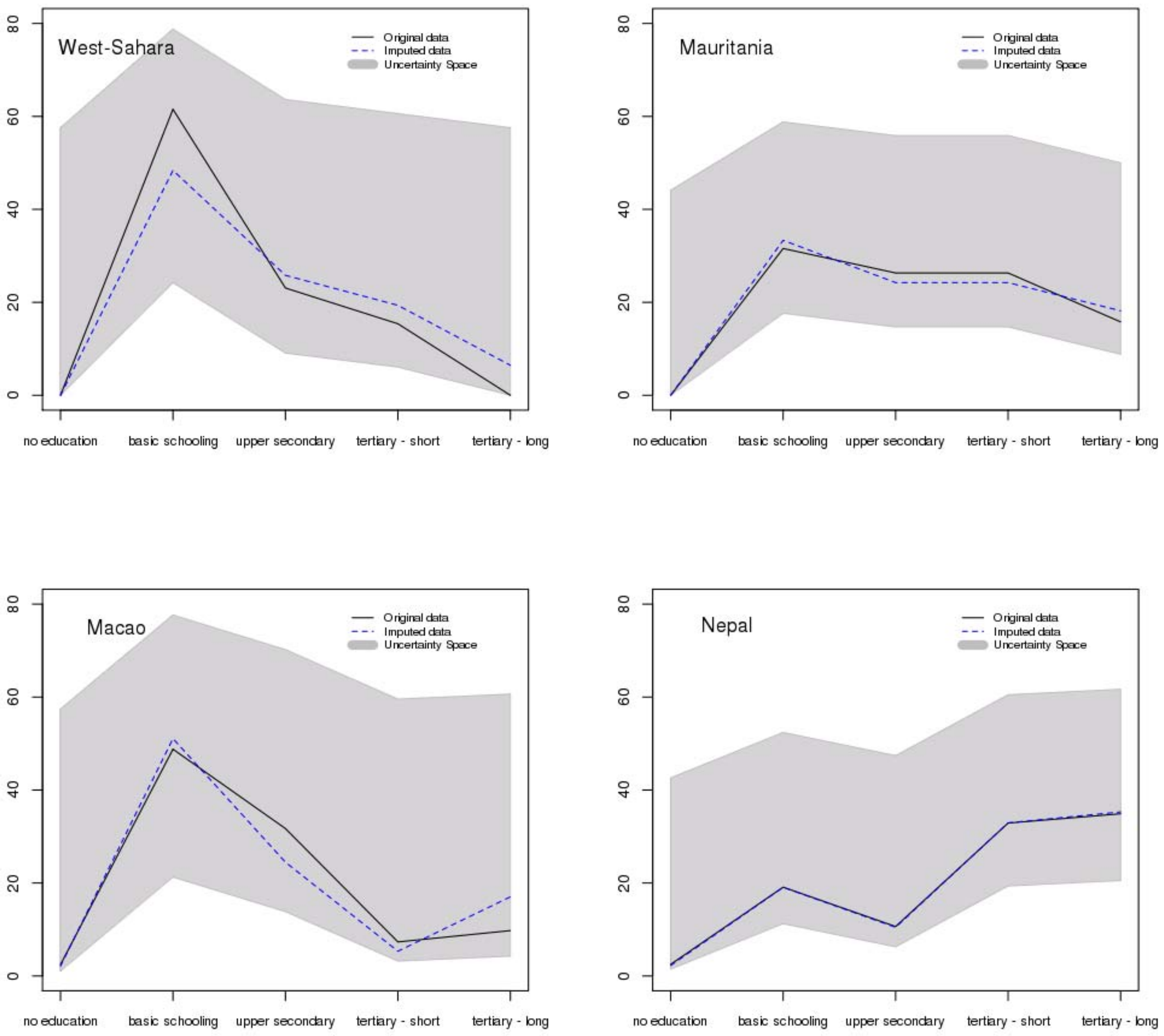

Afghanistan
Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long


Estonia
Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long


Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long


Djibouti
Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long


Guinea-Bissau
Original data
Imputed data
Uncertainty Space
0
20
40
60
80
no education
basic schooling
upper secondary
tertiary - short
tertiary - long



4.3. Quality testing using Swedish immigrants

Statistics Sweden (SCB) is the National Statistical Institute for Sweden. They have extensive registers including information on who is living in the country, who emigrates out of the country and to where. From their databank, we were able to collect aggregate data on the level of education of those that emigrated from Sweden to Norway in each of the years 2008-2012. The figures from SCB only include those born in Sweden and those that have permanently moved to Norway. Short and long tertiary education were aggregated as the data from SCB discriminated only between higher education up to 2 years duration and over 2 years. In the Norwegian register, the distinguishing is made between 4 years duration or less and over 4 years. To make this comparable we have aggregated. While there are uncertainties in this data, it provides an opportunity to compare our imputation results with an independent data source.

Table 5 shows the absolute numbers within education levels recorded for Swedish immigrants to Norway in 2010-2012 according to SCB and Statistic Norway. There is a large discrepancy in the absolute numbers with SCB recording over 1.5 times of those registered in the Register for the Population's Level of Education (BU) in

Norway for some years. As we do not have access to micro-data, we are not able to deduce who is included in the data from SCB and not in our records.

Table 5. Level of education for Swedish immigrants 2010 to 2012. SCB is data from Statistics Sweden, BU is register data from Statistics Norway and BU Imp shows register and imputed values from Statistics Norway.

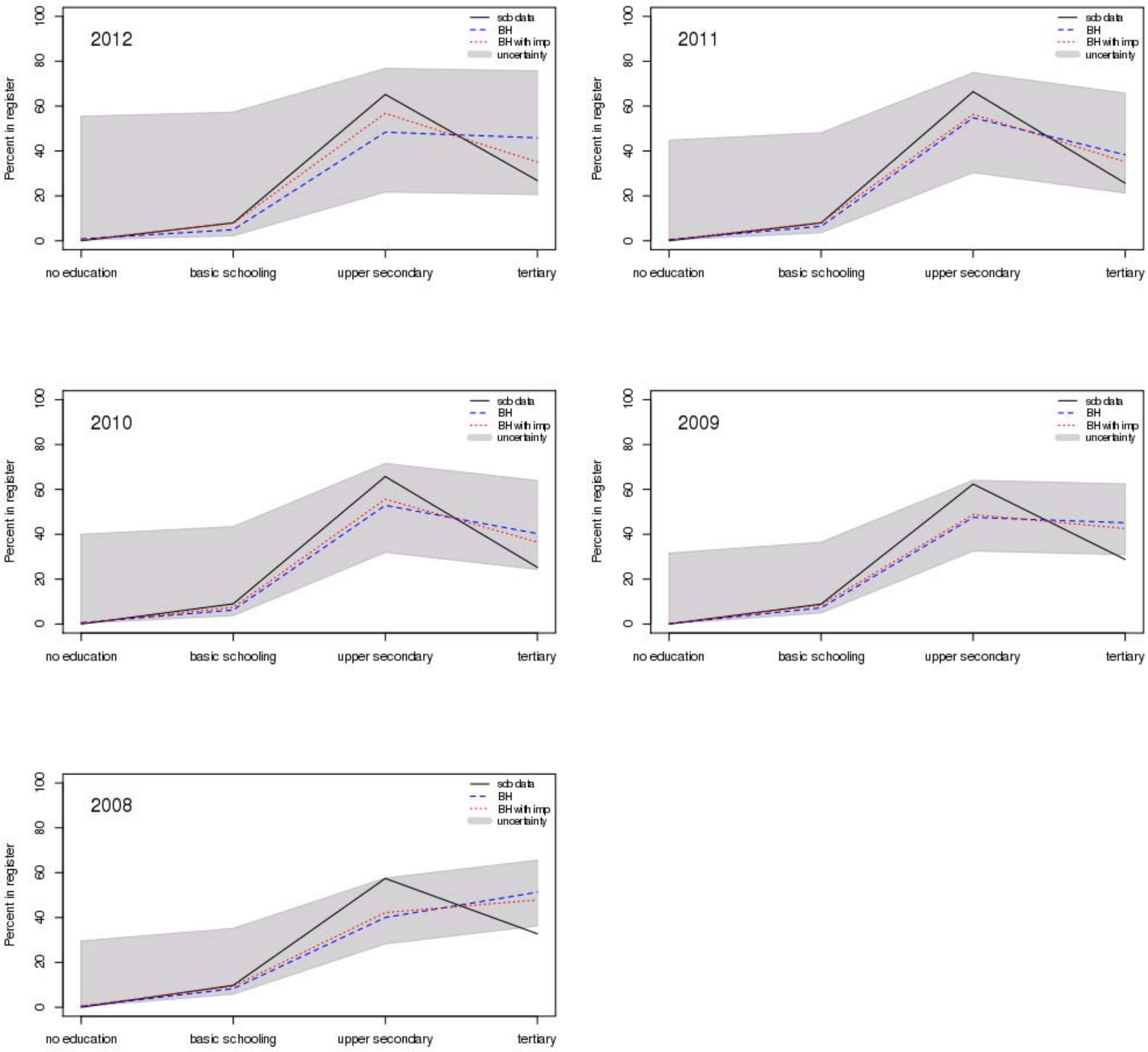
Education level	2010			2011			2012		
	SCB	BU	BU Imp	SCB	BU	BU Imp	SCB	BU	BU Imp
No formal education …	-	10	12	-	11	19	-	11	21
Basic schooling ………	438	108	235	427	142	375	309	65	274
Upper secondary education ………………	3205	916	1 549	3 580	1 213	2 214	2 528	639	1 633
Tertiary education ……	1233	698	1 081	1 382	848	1 390	1 040	606	1 017
Missing …………………	22	1 145	0	22	1 784	0	26	1 624	0
Total …………………….	4898	2 877	2 877	5 411	3 998	3 998	3 903	2 945	2 945

Figure 12, shows the proportions of the Swedish immigrant population in 2009-2012 within the 4 education levels. We did not receive information on how many did not have any formal education so are assuming it is approximately zero. Both registry data and registry data with imputations are shown in the figure. We see that the data including imputations is more similar to the data we received from SCB than when missing data is ignored. This indicates that the imputation is correcting the data in the right direction. However, results including imputations are still very different to the data from SCB. This perhaps indicates a non-ignorable missing data pattern that the imputation method is not correcting for.

The grey areas in figure 12 indicate the absolute uncertainty area. They are calculated for each education level separately by imputing all missing values with a single education level. This then gives the maximum proportion that the education level can have. Likewise, the proportion for each level is calculated when none of the imputed values are that level to give the minimum proportion. This relates to the level of missingness in the register.

Interestingly, in 2008 and 2009 the data from SCB drops below the minimum boundary for tertiary education. This means that even if we impute all the missing values in the register (BU) to something other than this level, it will not be enough to reduce that proportion to the levels seen in the SCB data.

Figure 12. Registered level of education using SCB data, SSB register data (BU) with and without imputation. The grey area represents the absolute uncertainty space.



4.4. A note on non-ignorable missing data (NMAR)

Up until this point we have assumed that those who are missing level of education are different with respect to the variables investigated but not according to the level of education itself. However, it is possible that this assumption does not hold. Level of education is regularly used as a weighting variable at Statistics Norway and often correlates with non-response, whereby higher education generally leads to higher response probabilities (Bjørnstad 2013).

All the surveys on level of education have involved reminders sent to initial non-respondents. In 2011 and 2012, two waves of reminders were sent. There is a general idea that those who initially do not respond but are contacted through reminders may be more similar to non-respondents than those that respond initially. Using this assumption, we could apply a non-ignorable missing data approach through using only those that respond after reminders as donors. However, in the level of education surveys, Pedersen and Falnes-Dalheim observed that the representativity of the respondents decreased with additional reminders (Pedersen and Falnes-Dalheim 2012). In essence, those that responded after initial reminders were *less* similar with respect to auxiliary information to non-respondents than those that responded first. It is therefore very speculative to apply this approach in this setting. Other more complex non-ignorable missing data approaches are outside the scope of this investigation. The proposed imputation method considers many, high-quality register variables and is an appropriate method for adjusting for missing data in the given setting.

4.5. Implementing this imputation method

4.5.1. Frequency

Statistics on level of education among immigrants are published yearly. Therefore, the imputation method described may be repeated yearly and is proposed to be implemented in 2014 publications. It is advisable to also repeat the imputation for all those that are missing and have previously received imputed values. This may result in changes to level of education at a micro-data level, however the imputation is done with the focus of adjusting the data at an aggregate level. Given the number of variables included in the models, some changes in the number and the structure of new immigrants should not be a problem for the imputation method.

4.5.2. Input data

The imputation of a given individual should not restrict further questions on level of education if they arise. For example, the labour force survey has questions on level of education if an individual is selected for the survey and has a missing value in the register. This should still be the case in the future.

Imputed values should not be used as donors in subsequent imputation procedures nor should they be used to impute variables that are included in the imputation model.

4.5.3. Limitations

The imputation method described here is for creating a complete dataset which can be used to build aggregate tables and figures representative of the immigrant population. The goal of the imputation method is to “fill-in” missing data to create statistics and not necessarily to impute the “correct” value at an individual level. As the imputation method considers country of origin, length of time living in Norway, gender and age, tables at these levels using imputed values is appropriate. It should not be used for researching correlations with variables that are not considered here. Researchers wanting to use level of education to investigate correlations should develop their own imputation methods that are appropriate to their research question.

5. Conclusion

It is important for Norway as a country to understand the level of education of its residents. Immigrants are an increasing proportion of the population and have high levels of missing data in the BU register. Surveys should continue to be used for collecting data from new residents, however there is likely to always be non-response and gaps in this register.

The imputation method proposed in this document provides a way to adjust for the bias that occurs when not all individuals respond to surveys. The imputation method recommended is a nearest-neighbour variant based on predictive mean matching.

Appendix A. Code for data sources in level of education register (BU).

```
’01’ = ’Fra Folke- og boligtellingen 1970’
’02’ = ’Fra Avsluttafilen’
’03’ = ’Fra lånekassa’
’04’ = ’Fra undersøkelsen om skolegang 1999’
’05’ = ’Fra Folke- og boligtellingen 1980’
’06’ = ’Tilganger fra FS-skoler på BHU-2000’
’07’ = ’Fra undersøkelsen om utdanning 2011’
’08’ = ’Fra undersøkelsen om utdanning 2012’
‘10’ = ‘Avsluttet grunnskole’
’11’ = ’DUF,Datasystemet for utlendings-/flyktningsaker-Utlendingsdir.’
’20’ = ’LINDA-elev/LINDA-avsluttet’
’21’ = ’LINDA-fagopplæring’
’22’ = ’Arbeidsdirektoratet’
’23’ = ’Diskett/skjema videregående’
’24’ = ’Militære videregående skoler’
’26’ = ’Folkehøgskoler’
’27’ = ’Nettskoler’
’30’ = ’Nasjonal vitnemålsdatabase (NVB)’
’31’ = ’Autorisasjonsregisteret for helsepersonell (HPR)’
’40’ = ’FS universitet’
’41’ = ’FS høgskoler’
’42’ = ’M-STAS universitet’
’43’ = ’M-STAS høgskoler’
’44’ = ’Diskett/skjema universitet’
’45’ = ’Diskett/skjema høgskoler’
’46’ = ’Militære høgskoler’
’47’ = ’Doktorgradsregister NIFU’
’48’ = ’Statens lånekasse for utdanning’
’49’ = ’DBH Database for høyere utdanning’
’50’ = ’Etterrapporterte grader’
’99’ = ’Uoppgitt(99)’
```

References

Bjørnstad, J. F. (2013). Standardisert frafallshåndtering for person- og husholdningsundersøkelser. Delprosjekt 1: Oversikt over status om frafall og frafallsbehandling. Interne dokumenter. Oslo, Norway, SSB. **6/2013**.

Chen, J. and J. Shao (2000). "Nearest Neighbour Imputation for Survey Data." Journal of Official Statistics **16**(2): 113-131.

Durrant, G. B. (2005). Imputation Methods for handling Item Nonresoponse in Social Sciences: A Methodological Review. NCRM Methods Review Papers, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampton. **NCRM/002**.

Dzamarija, M. T., K. K. Andreassen, et al. (2013). Dokumentason av registerbasert statistikk over innvandrere os norskfødte med innvandrerforeldre. Interne dokumenter. Oslo.

Fay, R. E. (1999). Theory and application of nearest neighbour imputation in the census 2000. Second ASC., AMSTAT.

Fosen, J. (2011). Register-based employment statistics. micro-integration and quality-perspective life-cycle. A case study. ESSnet on Data Integration: Report on WP4 Case Studies.

Fosen, J., A. K. Johnsen, et al. (2000). Frafall blant innvandrere: En undersøkelse av frafall i Utdanningsundersøkelsen 1999 og i valgundersøkelser blant innvandrere. Notater. Oslo, Norway, SSB.

Foss, A. H. (2006). Utkast til imputering av uoppgitt i registeret over Befolkningens høyeste utdanning. Unpublished document

Little, R. J. A. (1988). "Missing-Data Adjustments in Large Surveys." Journal of Business & Economic Statistics **6**(3): 287-296.

Pedersen, H. E. and E. Falnes-Dalheim (2012). Non-response and representativity in a survey on education completed abroad. Q2012 Conference. Athens, Statistics Norway.

Rancourt, E. (1999). Estimation with nearest neighbour imputation at Statistics canada. Second ASC., AMSTAT.

Steinkellner, A. and A. K. J. Holt (2013). Undersøkelse om utdanning fullført i utlandet 2011/2012: Dokumentasjonsrapport. Interne dokumenter.

Vassenden, K. (2001). Hvor stor er innvandringen til Norge? Samfunnsspeilet. **2**.

Waal, T. d., J. Pannekoek, et al. (2011). Handbook of Statistical Data Editing and Imputation. New Jersey, John Wiley & Sons Inc.

Zhang, L.-C. and C. Hendriks (2012). Micro integration of register-based census for dwelling and household. Work session on statistical data editing. Oslo, Norway.

List of figures

List of tables

Statistics Norway

Postal address:
PO Box 8131 Dept
NO-0033 Oslo

Office address:
Akersveien 26, Oslo
Oterveien 23, Kongsvinger

E-mail: ssb@ssb.no
Internet: www.ssb.no
Telephone: + 47 62 88 50 00

ISBN 978-82-537-8956-9 (electronic)

Statistisk sentralbyrå
Statistics Norway

Design: Siri Boquist