Distr.: General
29 April 2024

English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Migration Statistics
Geneva, Switzerland, 7−8 May 2024
Item 6 of the provisional agenda
Improvements in use of administrative data for migration statistics

Preliminary results of a method for finding reference persons to family immigrants to Norway

Note by Statistics Norway*

Abstract

In 2009, a registration scheme for citizens of countries in the EEA replaced the residence permit requirement. As a result, a significant amount of information regarding immigrants from countries in this area has been lost, including concerning the identity of the reference persons to family immigrants. Through a project financed by the Ministry of Labour and Social Inclusion (AID), Statistics Norway is currently developing a method to impute reference persons for this group. The paper will present an outline of the methodology and data sources of the method in its present form (April 2024), along with some preliminary results from testing. Based on the latter results, the paper will identify some areas in which further development of the method can be directed. As of now, the method utilizes a limited selection of administrative data related to kinship and household members of family immigrants. Questions concerning the application of these data in the method, as well as considerations regarding precision (individual versus family/household-level), will be among those discussed.

*Prepared by Christian Sørlien Molstad

I. Introduction: Background and aim of the project

1. Family immigration is one of five main categories[1] in Statistics Norway's statistics for reasons for immigration among non-Nordic citizens. It constitutes the single most common type of immigration in the period since 1990 (Molstad et al., 2022). Information regarding the so-called reference persons, the person to whom the family immigrant is immigrating, have proven useful to understand this type of immigration by providing individual-level context.
2. The implementation of Directive 2004/38/EC into the EEA-agreement and introduction of the registration scheme in 2009 meant that immigrants from EEA countries are no longer required to obtain residence permits. One consequence has been loss of information regarding the reference persons of family immigrants. The data loss has for the most part affected so-called family reunifications, i.e. family immigration where the family relation existed before the reference person was resident in Norway. For family establishment, where the family immigration entails establishment of a new relationship (usually marriage), imputation using administrative data on spouses has mitigated the data loss.
3. The loss of data on reference persons has led to discontinuation of the delivery of several cross tables supplied through the annual monitor for family immigration and the introduction of severe limitations to analyses utilising this information. In order to rectify these gaps, a project was initiated in 2022 with funding from the Ministry of Labour and Social Inclusion (AID). In the project we aim to utilise administrative data to impute reference person for family immigrants for whom such data are missing. The goal is to successfully impute data for family immigrants that arrived from the year the registration scheme was introduced, 2009, until 2020.
4. A more detailed outline of the background of the project, as well of a brief discussion of initial considerations, have been presented in Molstad (2022). This paper will present an outline of the methodology in its present form (April 2024), as well as some preliminary results from tests using data for family immigrants for whom we have reference persons. In the closing sections of the paper, the results will be briefly discussed and some points for further development and use of the methodology presented.

II. Data sources and methodology

A. Core principles

5. In the method presented in this paper we have as far as possible opted for parsimony, balancing the attempt to measure the missing relations as precisely as possible while simultaneously applying a) a minimum of variables using b) as simple principles as possible. This "economical" justification is based on several considerations.
6. One concerns the practical aspect of data management. Register data is by default reliant on administrative routines and the collection and storing of data in the relevant agencies (see section B in this chapter). Changes in these practices can therefore affect the quality and/or characteristics of the available data, as illustrated by the loss of information regarding reference persons to family immigrants due to changes in the requirements for residence for EEA-citizens. For an imputation method using register data, the exposure to such changes

[1] The others being labour, refuge, education and other.

will to a significant degree be determined by the number and characteristics of the chosen data sources. Applying data from fewer sources entails in principle less exposure to potential changes, and correspondingly less vulnerability.

7. Another consideration concerns the potential use of the imputed values in future statistical analyses. Data on reason for immigration are utilised extensively in analyses on topics related to immigration and integration. In some of these analyses, information on the reference persons of family immigrants have been used as a mean for subdividing these immigrants into groups. Traits of reference persons, like immigration category, sex and (for immigrants) reason for immigration, have proven useful in explaining patterns among family immigrants with regards to such phenomena as naturalisation and structural integration (Arnesen & Molstad, 2024; Molstad et al., 2022). A method for imputation of reference persons will have to take into account that the application of the imputed data for analytical purposes can be severely curtailed by the choice of variables. A method drawing on a number of variables known to correlate with characteristics of reference persons could potentially produce more precise imputations. However, such imputed data would naturally be unusable for studies aimed at investigating these same relations.
8. In the imputation method we have therefore adopted a conservative approach to including variables, starting with a limited set of variables reflecting demographic characteristics and immigration. This does not preclude the possibility of introduction of biases in the imputed data that could affect results of analyses. The choice of reference person in the imputation method is, as we will see below, partly reliant on assumptions "favouring" one gender over another (mothers over fathers, time of arrival being equal). Instead of eliminating the potential of such biases, a parsimonious method for imputation, using a minimum of variables in a simple manner, makes for easier understanding of the sources of potential biases and limits for analyses. As we will argue in the discussion (chapter IV), one can furthermore argue that some of these limits may be bypassed using variables on household and/or family level.
9. A third consideration relates to probability. It has been argued that simpler hypotheses are more likely to be true (Swinburne, 2019), even if just because they introduce fewer assumptions and thereby fewer possibilities for error (Simms, 2024). This parallels the justification in the first consideration, concerning limiting the number of variables to limit exposure to potential changes in data sources (see above).
10. In the methodology presented in this paper we have hence aimed at keeping the number of assumptions to a minimum, adhering to William Ockham's maxim that "plurality should not be posited without necessity" (*Pluralitas non est ponenda sine necessitate*) (Ockham, 1990, p. xxi). As far as the aim of successfully predict reference persons permits us, we have attempted to make very broad and rudimentary categorisations and deduce reference persons through simple logical inferences, such as through transitive relations[2] (e.g. family members' relative arrival times to the country).

B. Data sources

11. The data on reasons for immigration is generated using information from two main data sources: the Aliens Register (UDB), administered by The Norwegian Directorate of

[2] A transitive (binary) relation R is defined as such if (and only if) "for all elements d, e, f of [the set] S: if ⟨d,e⟩ ∈ R and ⟨e,f ⟩ ∈ R then also ⟨d,f⟩ ∈ R" (Halbach, 2015, p. 9).

Immigration (UDI), and The Central Population Register (DSF/FREG),[3] administered by The Directorate of Taxes (Skatteetaten). Apart from this data, the method in its current form applies primarily two kinds of data, concerning *kinship* and *households*.

12. The data source of *kinship data* is primarily The Central Population Register (DSF/FREG). One challenge associated with the available kinship data concerns missing information regarding parents. This problem is most prevalent among immigrants. Table 1 shows percentage of missing data on fathers and mothers for persons resident in Norway as per 1.1.2021. As we can see, social security numbers are missing for a majority of immigrants. This is not necessarily a sign of low quality: for many immigrants, especially those arriving as adults, parents will neither have been or become resident in Norway and therefore not be registered in The Central Population Register. For immigrants arriving at a young age, the coverage tends to be better.[4] In general the gaps in the data are more substantial for fathers compared to for mothers, irrespective of immigration category.

Table 1
Percentage missing social security numbers for parents, resident population in Norway per 1.1.2021, by immigration category

	Mothers	Fathers
Immigrants	77,8	82,6
Norwegian-born to immigrant parents	0,2	4,8
Without immigrant background	7,0	8,6
Total	17,3	19,4

Source: Statistics Norway

13. *Household data* is based on information from three main data sources, namely The Central Population Register (DSF/FREG), The Cadastre (Matrikkelen) and The Central Coordinating Register for Legal Entities (Enhetsregisteret i Brønnøysund).
14. In its current form, the imputation method applies household data for the earliest data point after time of immigration, i.e. 1st of January of the subsequent year. For immigrants 18 years or older at the time of arrival, information on spouses/partners is collected from this data point.
15. Household data are also used to impute potentially missing parents for family immigrants that were under 18 years of age at arrival. In the case of missing data for either mother or father, household members of family immigrants who are a) of opposite age and b) within a 15-year

[3] A modernised version of the Central Population Register was introduced in 2020. In an intermediate period from 2020 to September 2022 substitute data similar to the data from DSF, the previous system, were delivered by The Directorate of Taxes. The files used in the methodology presented in this paper are based on data older than 2022.
[4] An example can be found in Molstad & Barstad (2023). Analysing intergenerational social mobility among persons born between 1979 and 1989, the authors used kinship data to connect such information as parents' income and education to the persons of interest. While coverage was generally limited among the immigrants, it was high among so-called "early arrivers" (immigrants arriving before 7 years of age), see table 3.1 on p. 21.

age range (older or younger) of the registered parent and c) who are not one of the other registered family members are imputed as possible parents. The imputed parents and data regarding their first date in Norway are included in the imputation method described below.

C. Outline of approach

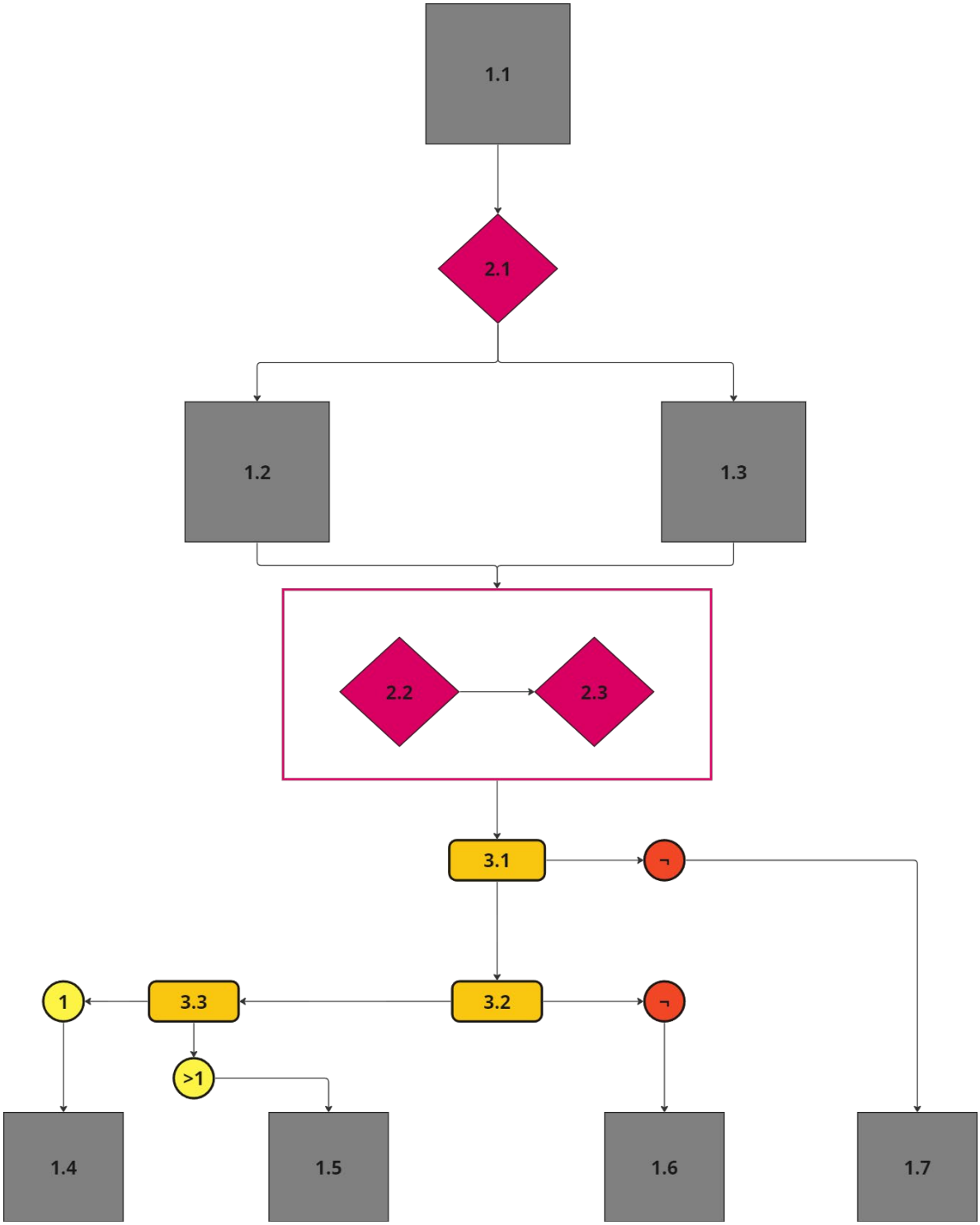
16. The method applies the presented data in two stages. Based on the age of the family immigrant at the time of arrival, kinship data and the date of arrival (first date)[5] of him/her and his/her family members (kin), 1) a typology based on different family constellations and sequences of arrival is constructed. Through logical deduction from this typology, 2) most likely reference persons are inferred (when possible).
17. These stages will be presented in the following subchapter separately.

1. Construction of typology

18. Figure 1 displays a graphical presentation of the approach for subdividing the family immigrants into a four-category typology. There are four main types of elements in the illustration in figure 1 and 2 (colour and numbering in brackets), denoting subcategories of the population (grey, 1), data sources (red, 2), choices (yellow, 3) and imputation categories (green, 4).
19. Using the primary person's age at the time of immigration, a fundamental distinction is made between persons below 18 years of age at the time of immigration (i.e. children) (1.2) and persons that were 18 years or older (adults) (1.3). This age distinction is introduced due to assumed different characteristics of family immigration involving children and adults. Family immigrants having arrived as children can be presumed to often come through reunification with parents. We may on the other hand assume adult family immigrants to often reunite with spouses or children.
20. These presumed differences form the basis for prioritisation of kinship relations in the subsequent categorisation of family immigrants into four main categories using kinship data (2.2) and the date of arrival (first date) of the child and family members (kin) (2.3).
21. Based on kinship data, the family immigrants are first split according whether there are kinship relations (i.e. family members) registered (3.1). If not (¬), the family immigrants are placed in a separate category (1.7).
22. Those with registered kinship relations are further split according to whether the family immigrants have family members arriving at an earlier date (i.e. earlier date of arrival) than themselves (3.2). Those that do not (¬) are separated into an own category (1.6).
23. The remaining family immigrants, that do have registered kinship relations and of who at least one of these have arrived earlier than the family immigrant, are divided according to the number of kin having arrived earlier (3.3). Family immigrants with more than one (>1) such kin are split from those with only one (1).

[5] For family members born in Norway this date of "arrival" will be their birth date.

Figure 1
Approach for constructing of typology



24. The distinction between family immigrants arriving as children and adults is applied in this part of the process. For children, priority is given to parents, meaning that if at least one parent is found in the available data, only this relation (and that of the second parent, if available) will be counted in choice 3.3. Only if no parents are registered, other kinship relations are used to distinguish between those with one or more family members with earlier arrival date. Corresponding priority is given to spouses and children in the case of adult family immigrants.
25. Through the procedure described above, the following four categories are constructed:
26. **A)** *Family immigrants with registered kinship relations, and with only one family member with an earlier or equal first date (1.4)*. In principle, these family immigrants only have one potential reference person, namely the family member arriving to Norway before them. We may therefore assume this group to constitute the "easiest" in term of imputation.

    B) *Family immigrants with registered kinship relations, and with more than one family member with an earlier or equal first date (1.5)*. Due to there being more than one family member arriving earlier than the family immigrant, there are several potential reference persons for the family immigrants in this category. The imputation will therefore necessitate a choice between "candidates" based on specified criteria.

    C) *Family immigrants with registered kinship relations, but where (the) family member(s) have later first date(s) (1.6)*. With no family members arriving earlier or at the same time as the family immigrant, there is in principle no potential reference person. However, assuming a degree of error in the first date of the family immigrant and/or his/her kin, there is still some room for imputation using available kinship relations. In cases in which there are more than one registered kinship relation, a choice between "candidates" based on criteria will be necessary.

    D) *Family immigrants with no registered kinship relations (1.7)*. No registered kinship relations means that no potential reference person can be inferred with the method stipulated in this paper.

2. Imputation

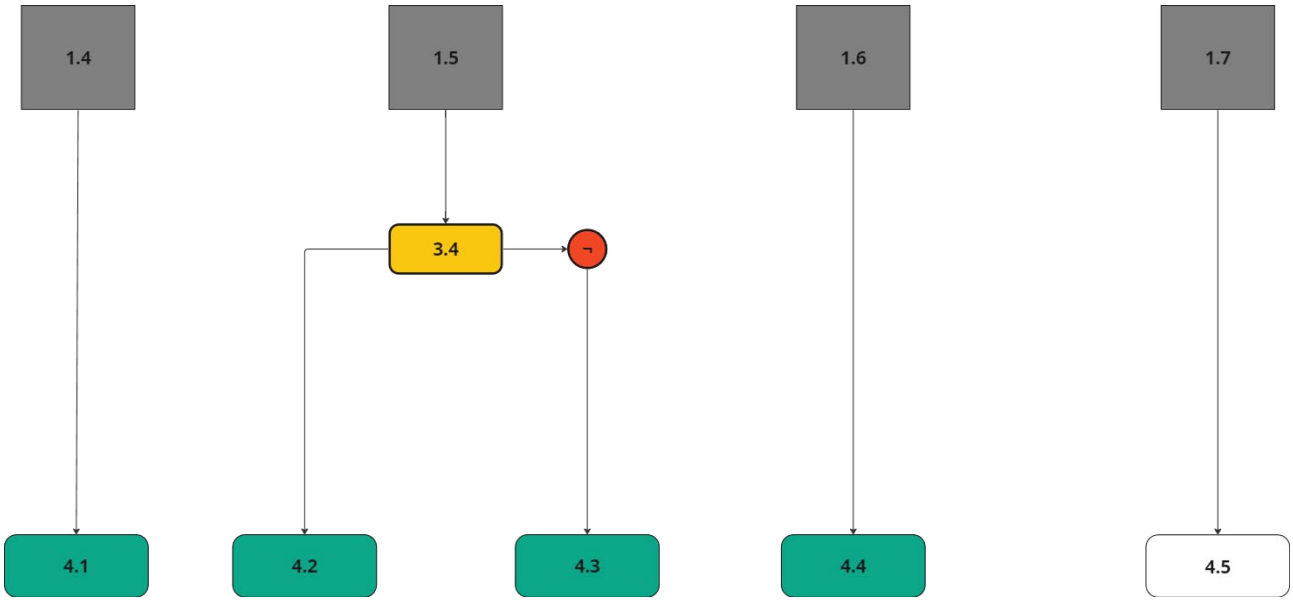
27. From these subcategories imputation may be conducted based on logical inferences (figure 2). Among the simplest are those drawn for the members of subcategory 1.4, for whom there is only one family member for whom there is an earlier or equal date of arrival compared with the family immigrant. In these cases, imputation means picking this family member as the most likely reference person (4.1).
28. Equally straightforward is the procedure for subcategory 1.7. With no registered kinship relations available, the reference persons for these family immigrants are bound to remain missing (4.5).
29. For the family immigrants in the subcategories 1.5, for whom more than one eligible family member is available, introduction of further criteria is required (choice 3.4). One such criteria is relative arrival date among the family members. There are several options for how to implement such a criterion, of which two main ones stand out: picking the family member arriving at the earliest date vs the family member with the arrival date closest in time compared to the family immigrant.
30. Which principle should be chosen as criterion? Choosing the earliest arrival could be justified with the assumption that this family member is the "pioneer" to which the rest of the family is

immigrating. However, this assumption does not take into consideration the possibility that the distance in time between the arrival of the earliest family member and the arrival of the family immigrant (and other family members) is vast. The immigration of one or more of the kin of the family immigrant could in principle be completely unrelated. It can be argued that the possibility of the immigrations being unrelated increases with greater distance and that immigrations more proximate in time therefore are more likely to be related. In the present version of the imputation model, we have however chosen to favour earlier arrivals, selecting the family member arriving first relative to the family immigrant and other family members (4.2).

31. In those cases where there is no difference in arrival date between eligible family members (¬), an alternative criterion will have to be introduced for choosing. Here we have chosen to introduce different criteria based on the age of arrival of the family immigrant. For children the mother is given priority (4.3). However, this logic only applies if parents are available, the family with the earliest first date is selected in the priority of first siblings, then grandparents (4.2). For adults spouses are prioritised over children (4.3). If spouses and children are not available, the first other family member is chosen in a similar priority as for children (4.2).

32. The principle is followed in the case of family immigrants in subcategory 1.6, who has arrived before other family members. For these we choose the family member with the earliest first date (4.4).

Figure 2
Principles for imputation of reference persons



III. Preliminary results

A. The test population

33. To test the imputation method, we use data for 18 895 family immigrants who a) arrived in the period 2004-2020 and b) for whom information on reference person is not missing on the statistical file for reasons for immigration. The population is one of ten randomly drawn[6] groups from the total population of all family immigrants for whom we have data on reference persons in the period on the statistical file for reason for immigration.
34. In the test our primary interest is the results for the 4 142 persons with first citizenship from a country in the EEA (table 2). Results for non-EEA-citizens will however also be referenced, as it provides an indication for the broader predictive power of the method.
35. Table 2 compares the characteristics of the test population and of the family immigrants for whom reference persons are missing in the period 1990 to 2020 (henceforth referred to as main population). As we can see the test population consists of a majority of citizens from outside the EEA, 78 percent. In the main population, the relation is opposite. This is not surprising, given that the loss of data on reference persons is most extensive among citizens from EEA countries.
36. Among EEA-citizens, we furthermore see different distributions regarding gender and age at arrival. More than two thirds of the family immigrants in the test population were adults at the time of immigration. Less than one third of the family immigrants for whom we lack data on reference persons were 18 years or older. This may partly be due to the low degree of missing reference persons in the case of family establishments, due to the imputation already done for this type of family immigration (see Molstad (2022), p. 7). Most, if not practically all, of the family immigrants coming to the country to marry a person already resident in the country will have been at least 18 years of age.
37. The test and main population are in other words not comparable on an aggregate level. In order to evaluate the "success" of the imputation method, we must therefore break down the results according to group level characteristics.

[6] The groups were generated using the RAND function in SAS, see SAS Help Center (2024).

Table 2
Descriptive statistics, test and main population

	Test population, number	Test population, percentage	Main population, number	Main population, percentage
Non-EEA citizens	14 753	78,1	9 844	17,1
EEA citizens	4 142	21,9	47 685	82,9
Whereof (EEA citizens):				
Children (<18 years old)	*1 303*	*31,5*	*34 919*	*73,2*
Adults (18 years or older)	*2 839*	*68,5*	*12 766*	*26,8*
Male	*998*	*24,1*	*20 067*	*42,1*
Female	*3 144*	*75,9*	*27 618*	*57,9*
Family reunion	*3258*	*78,7*	*47 646*	*99,9*
Family establishment	*884*	*21,3*	*39*	*0,1*
Total	18 895	100	57 529	100

Source: Statistics Norway

B. Results of imputation

39. Table 3 shows the results of the imputation method when applied on the test population. Absolute numbers and percentages indicate number of family immigrants selected into the relevant imputation categories and the percentage correct imputations.

40. As we can see, the great majority (83 percent) of the imputations are correct, in the sense that the selected family member matches the reference person registered on the statistical file. The degree of correct imputations varies greatly according to the imputation category. As expected, we see the highest percentage among family immigrants with only one family member with an earlier or equal first date (4.1). For these family immigrants, 90 percent of the imputations are correct. In cases where there are several family members with earlier or equal first date and there were difference(s) in the first dates among the family members (4.2), the percentage correct imputations is almost the same, 86. For this category we chose the family member with the earliest first date, a prediction that seems to generate overall fairly accurate results.

41. In cases we did not have a difference in first date(s) to go by (4.3), and we prioritized mothers (for children) and spouses (for adults), the percentage correct imputations was significantly lower. 73 percent of the family immigrants in this category were assigned the correct reference person.

42. Relatively few family immigrants did not have family members registered in the administrative records (4.4) or only family members that had later first dates than the family immigrant (4.5). These family immigrants constituted only 2 percent respectively of the total number of the test population. In the former case, where the family member with first date was chosen, only very few (6 percent) of the imputations were correct.

43. The overall pattern in other words confirms what we expect to see with regards to the relative precision of the imputation categories. As we move from left to right, the level of predictive power drops, most markedly in those categories where we are forced to deviate from the principle of picking a family member that has a first date earlier than a) the family immigrant and b) other eligible family members.

Table 3
Number of persons and percentage correct imputations, by imputation category and characteristics

	4.1	4.2	4.3	4.4	4.5	Total
Non-EEA citizens	3 796 (89,7)	9 525 (88,8)	724 (54,8)	330 (3,9)	378 (0,0)	14 753 (83,2
EEA citizens	927 (91,3)	2 331 (76,7)	742 (89,8)	90 (12,2)	52 (0,0)	4 142 (80,0)
Whereof (EEA citizens):						
Children (<18 years old)	*184* *(92,4)*	*747* *(31,6)*	*372* *(79,6)*	*0* *(0,0)*	*0* *(0,0)*	*1 303* *(54,0)*
Adults (18 years or older)	*743* *(91,0)*	*1 584* *(98,0)*	*370* *(100,0)*	*90* *(12,2)*	*52* *(0,0)*	*2 839* *(91,9)*
Male	*199* *(91,5)*	*534* *(51,3)*	*238* *(81,5)*	*11* *(9,1)*	*16* *(0,0)*	*998* *(65,2)*
Female	*728* *(91,2)*	*1 797* *(84,3)*	*504* *(93,7)*	*79* *(12,7)*	*36* *(0,0)*	*3 144* *(84,6)*
Family reunion	*683* *(89,8)*	*1 726* *(68,5)*	*742* *(89,8)*	*66* *(15,2)*	*41* *(0,0)*	*3 258* *(75,9)*
Family establishment	*244* *(95,5)*	*605* *(100,0)*	*0* *(0,0)*	*24* *(4,2)*	*11* *(0,0)*	*884* *(94,9)*
Total	4 723 (90,0)	11 856 (86,4)	1 466 (72,5)	420 (5,7)	430 (0,0)	18 895 (82,6)

Source: Statistics Norway

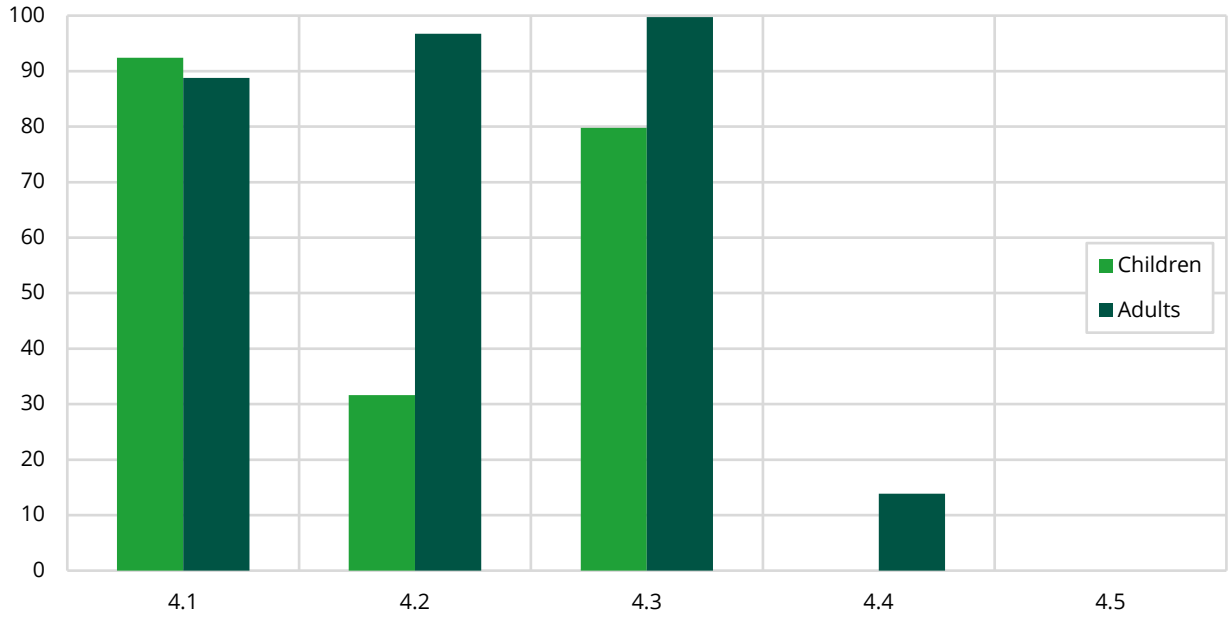
44. However, this pattern is less clear when the results are broken down on group level. While the left to right drop in precision is still evident among non-EEA citizens, it is less so among citizens from inside the EEA. The percentage of correct imputations is consistently high (90 percent or above) among the family immigrants from the EEA-area in category 4.1, irrespective of background characteristic. The differences are more significant in category 4.2 and 4.3. Picking the family member with the earliest first date leads to very different outcomes: for adult family immigrants from the EEA, it leads to 98 percent correct imputations, while it for children results in a meagre 32 percent of "hits".

45. A possible reason behind these differences is suggested by the results for family reunions and family establishments. In situations where EEA citizens have arrived Norway to marry a person already resident in the country and there are more than one eligible family member to choose from, the imputation method selects the right reference person in 100 percent of the cases. The corresponding percentage for EEA citizens arriving through family reunification is 69 percent. The better results for family establishments can be explained by the large number of already imputed values in this group on the data file. A high degree of similar matches is not surprising given that we prioritise the same type of persons, i.e. spouses (see chapter I), in parts of our imputation method for adults (4.3).

46. As family immigrants coming to Norway through marriage are overwhelmingly 18 years or older, it is therefore tempting to attribute the more favourable results for adults to the inclusion of family establishments in the test population. However, as figure 3 illustrates, this seems not to be the full reason behind the differences between children and adults. Considering only family reunifications, we see the high percentage of correct imputations among adult family immigrant persisting. Prioritizing the spouse in the case of similar first dates among family members (4.3) results in a near perfect prediction. We see a similar tendency among children with regards to mothers. Prioritizing mothers when first dates are equal leads to 80 percent correct imputations.

47. Giving primacy to an earliest first date, when this is possible (4.2), has very different consequences for children and adults. While for children it is associated with low precision (32 percent, as mentioned above), it means almost perfect prediction for adults (98 percent). The differing percentages of "hits" may be due to a common cause: mothers are selected in only 20 percent of the cases in this category for children, while spouses are picked for nearly all, 98 percent, of the adults. The fewer correct imputations among children is in other words potentially due to mothers being systematically registered as reference persons even when there are one or several more other family members present in the country.

Figure 3
Percentage correct imputations, EEA citizens coming through family reunification, by imputation category and age at time of immigration



IV. Discussion

48. There are two main options for how to approach improvement of the described method. One possibility is to further simplify the selection criteria in the cases where there are several eligible candidates for reference person. As we have seen in the previous chapter, there seems to be a tendency toward picking mothers for children and spouses for adults being associated with high precision. Selecting mothers and spouses for children and adults where eligible mothers and spouses are available would reduce the number of steps and imputation categories in the method. However, doing this would possibly mean accepting a significant number of wrong imputations. Assuming the family reunifications involving EEA citizens in the test population being fairly representative of the family immigrants on the statistical file with missing reference persons, the method would select the wrong family member 20 percent of the time for children similar to those categorised in 4.3. The inaccuracies could furthermore be considerably higher on (sub)group level, severely limiting the use of the imputed data. Making the method more sophisticated in order to make it more precisely select the correct family member among several eligible candidates could therefore be fruitful.

49. This would not necessarily require adding new data to the imputation process. In further development of the method, household data could for example be used more extensively to either discriminate among potential reference persons or to suggest (impute) missing family members. Imputation of potentially missing fathers and mothers is already done for children with only one registered parent (see subchapter II.C.). We see that a majority (69 percent overall, 79 percent in the case of EEA citizens) of the family immigrants for whom the imputed value is wrong live in the same household as the reference person registered on the statistical file.

50. However, even an eventual imputation method achieving near perfect prediction across all (sub)groups will always be associated with uncertainty. While developing an imputation method with the help of test populations, we always risk moulding it (the method) too closely to their (the test populations') characteristics. Furthermore, given that we may not fully know how representative the test population is compared to the population we aim to impute data for, we may never be completely sure of the validity of the predictions of our method. That is of course unless new data become available that be used to directly verify the predictions.

51. Both demonstrably imprecise predictions in testing and the perennial question of representativeness can however arguably be bypassed. Precision has in this paper been measured by individual level correspondence between the information about the reference person registered on the statistical file and the information imputed through the outlined method. If the identity of the persons referred to does not correspond, the prediction is deemed incorrect. However, this does not mean that the imputed data is of equal value to missing data. We may know with reasonable certainty that the person imputed as the reference person is a family member. Analyses of immigration and integration often take a family and/or household perspective as this often is the arena of many of the processes and the decision making.[7] Identifying which family and/or household the family immigrant arrived/"belong" to may therefore be more significant than identifying the exact family member was registered in the administrative records as the reference person. For this reason, adding variables indicating household and/or family affiliation to the finished file could mitigate the shortcomings of the individual level output of the imputation method.

V. Conclusion

52. The method presented in this paper has tried to balance methodological parsimony in terms of the number and use of variables with the need to achieve high precision in predicting reference persons for family immigrants. The preliminary results from the current version of the method are mixed. Among family reunifications involving EEA citizens, who most probably resemble the family immigrants with missing data on reference persons the most, there is substantial variation in the precision of the method, especially on group level.

53. Further work on the method could mean simplifying the method further. This would quite possibly lead to improved precision for important groups of family immigrants, such as for those arriving as children. This is one of the biggest groups among the family immigrants we lack data on reference persons for. Another possibility is development of a more fine-grained method for discriminating between potential reference persons, hopefully resulting in selection which is (even) more precise. Whichever approach is chosen, it would most probably be beneficial to add variables indicating family and/or household affiliation. These could provide a basis for making statistics and analyses for which the imputed data otherwise would be unsuited for.

[7] This is especially the case in research within the New Economics Migration (NELM)-approach (De Haas, 2010).

VI. References

Arnesen, R., & Molstad, C. S. (2024). *Overgang til norsk statsborgerskap 1977-2021* (2024/3). Statistisk Sentralbyrå. https://www.ssb.no/befolkning/innvandrere/artikler/overgang-til-norsk-statsborgerskap-1977-2021

De Haas, H. (2010). Migration and development: A theoretical perspective. *International migration review*, *44*(1), 227-264.

Halbach, V. (2015). *The Logic Manual*. Oxford University Press.

Molstad, C. S. (2022). *Finding family relations: About quality issues regarding family immigration statistics and their potential solution using administrative data* United Nations Economic Commission for Europe (UNECE): Conference of European Statisticians, Group of Experts on Migration Statistics, Geneva. https://unece.org/sites/default/files/2022-10/WP.8_Molstad_Norway_SA_ENG_1.pdf

Molstad, C. S., & Barstad, A. (2023). *Sosial mobilitet blant innvandrere og norskfødte med innvandrerforeldre* (2023/28). Statistisk Sentralbyrå. https://www.ssb.no/befolkning/innvandrere/artikler/sosial-mobilitet-blant-innvandrere-og-norskfodte-med-innvandrerforeldre

Molstad, C. S., Gulbrandsen, F. B., & Steinkellner, A. (2022). *Familieinnvandring og ekteskapsmønster 1990-2020* (2022/03). Statistisk Sentralbyrå. https://www.ssb.no/befolkning/innvandrere/artikler/familieinnvandring-og-ekteskapsmonster-1990-2020

Ockham, W. (1990). *Philosophical Writings*. Hackett Publishing Company.

SAS Help Center. (2024). *RAND Function*. Retrieved 12.04 from https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/lefunctionsref/p0fpeei0opypg8n1b06qe4r040lv.htm

Simms, C. (2024). *Occam's razor*. New Scientist. Retrieved 14.04.2024 from https://www.newscientist.com/definition/occams-razor/#:~:text=14th%E2%80%93century%20friar%20William%20of,should%20prefer%20the%20simpler%20one.

Swinburne, R. (2019). *Simplicity as Evidence of Truth*. Marquette University Press.