Distr.: General
26 April 2024

English

Economic Commission for Europe

Conference of European Statisticians

Workshop on the International Recommendations on Statistics on Refugees, Internally Displaced Persons, and Statelessness

Geneva, Switzerland, 6 May 2024

Panel I: Using national population census and/or national household surveys to improve official statistics on refugees, IDPs and/or stateless persons through inclusion.

Estimating labour force participation of refugees by statistical matching of administrative and survey data

Note by German Federal Statistical Office*

Abstract

In the last decade, many European countries experienced a large influx of refugees particularly from Syria, Iraq and Afghanistan and more recently of people in refugee-like situations particularly from Ukraine. Accordingly, their integration on the labour market is a major focus in host countries.

In Germany, official statistics use the annual Microcensus survey to shed light on the integration of immigrants. While the Microcensus provides a detailed socio-economic picture of immigrants in general, refugees and people in refugee-like situations are neither reliably identified nor comprehensively covered. At the same time, refugees and people in refugee-like situations can be identified and are comprehensively registered in the administrative data of the Central Register of Foreigners. However, the administrative data does not provide information on socio-economic characteristics, housing and living conditions.

Therefore, merging administrative and survey data has the potential to close data gaps by combining reliable information on legal residence status with a wide range of socio-economic characteristics. As the survey data and the administrative data do not share a common personal identifier and information for probabilistic linkage (name, date of birth and address) is not available, this working paper presents a statistical matching approach to combine both data sources. The algorithm is based on a machine learning model that predicts probabilities of being a refugee or in a refugee-like situation for survey respondents and multiple imputation in order to assess the uncertainty in statistical inference.

*Prepared by Marieke Smilde-Becker (Marieke.Smilde-Becker@destatis.de), Jan Eberle (Jan.Eberle@destatis.de)

I. Introduction

1. In many European countries, the integration of people seeking humanitarian protection (refugees and people in refugee-like situations) on the housing market, the labour market and in civil society is a major focus of public debate in light of the high levels of immigration in 2015/2016 (particularly from Syria, Iraq and Afghanistan) and 2022 (particularly from Ukraine). Data on people seeking protection, which can be used to map integration progress and derive implications for integration policy, is therefore of central importance. In Germany, official statistics use the annual Microcensus (MC) survey to shed light on the integration of immigrants and people with a history of immigration. While the survey provides a detailed socio-economic picture of immigrants, those seeking protection are neither reliably identifiable nor comprehensively covered. [1] However, people seeking protection are comprehensively registered and can be identified clearly by means of their residence status in the administrative data of the Central Register of Foreigners (CRF). However, the CRF does not provide any reliable information on socio-economic characteristics, housing and living conditions.
2. Merging the MC survey data with administrative data from the CRF can therefore create synergies. Reliable information on residence status can be combined with a wide range of socio-economic characteristics. This in turn can make an important contribution to closing data gaps regarding the integration of people seeking protection without additional surveys.
3. The MC and CRF data available at the German Federal Statistical Office neither share a common personal identifier nor sufficient information for probabilistic linkage (e.g. via name, date of birth and address). Therefore, this paper presents a statistical matching algorithm in order to combine the two data sources. From a data protection perspective, this procedure offers the advantage that no individual persons need to be identified and linked in the different data sources.

II. Statistical matching

4. In contrast to linking information about identical entities, statistical matching can be formulated as an imputation problem in which a target variable for an individual in a recipient dataset is predicted based on information about similar individuals - statistical twins - in a donor dataset (Cielebak and Rässler, 2019).
5. In this use case, the problem is formally described using Figure 1. The information on protection seekers from the Central Register of Foreigners (donor) is used to estimate the probability of being a person seeking protection for respondents in the Microcensus (recipient). For the prediction, we use a selection of common variables (C). The aim is to analyse the unobserved bivariate distributions of refugee status (R) and employment status (E). Implicitly, this procedure assumes conditional independence of this unobserved bivariate distribution given the common variables C.

[1] The Microcensus records the reason for migration based on a self-assessment by the respondents. If several reasons apply, respondents are asked to state the main reason.

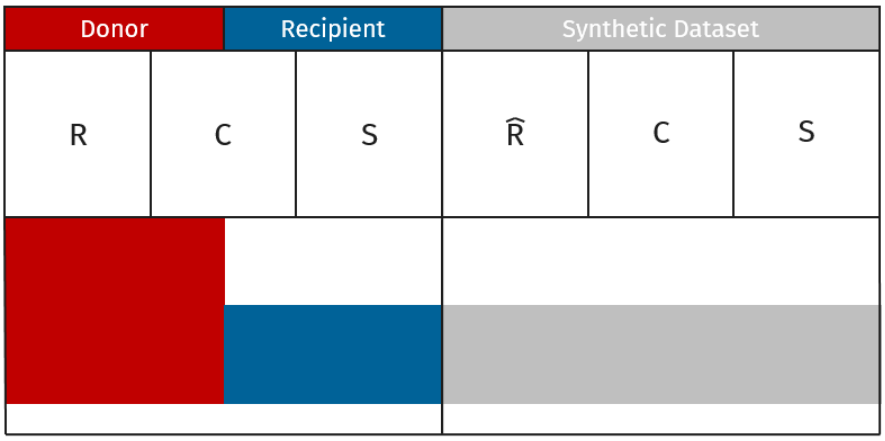
Figure 1
Statistical matching

(1) Donor (CRF) with common variables C and refugee status R

(2) Recipient (MC) with common variables C and employment status E

(3) Prediction $\widehat{R} = F(C)$

(4) Synthetic data set with C, E and $\widehat{R}$



III. Data sources

6. The Central Register of Foreigners (CRF) is one of the largest administrative registers in Germany. It contains information on all foreign nationals who are permanently staying in Germany, which usually means for more than three months. For statistical purposes, the German Federal Statistical Office receives an annual data extract for compiling statistics on foreigners. This extract contains residence status information required to identify persons seeking protection in Germany. Persons seeking protection are foreign nationals who are in Germany for humanitarian reasons and are recorded in the CRF with a corresponding legal residence status (Eberle, 2019). People seeking protection are therefore a subset of the foreign population registered in the CRF. In the following CRF data as of 31st of December 2021 is used. By then, 11.8 million foreign nationals were registered in the CRF, 1.9 million of whom were people seeking protection.

7. The Microcensus (MC) provides statistical information on the population structure, the economic and social situation of the population, families and households, employment, education and training, housing situation and health. The Microcensus is a representative random sample of all inhabitants. Each sampling unit has the same probability of being included in the sample. One-stage cluster sampling is used, in which areas (selection districts) are drawn in which all households are surveyed. Results are extrapolated on the basis of key population figures.

IV. Statistical matching algorithm

A. Harmonizing the data sources

8. First, the data sources must be harmonised for statistical matching as far as possible. Therefore, studying the metadata is crucial. Special attention should be given to the common variables used for matching. In this case, the common variables available are gender, age, age at immigration, year of immigration, citizenship, marital status and regional district.[2] After harmonising the common variables, both data sets are restricted to a more comparable target population. The MC main target population is people living in private households. Therefore, newly arrived refugees may be excluded from selection in the MC if they are temporarily living in temporary accommodations. Moreover, in the MC they may also not be fully included in the population in shared accommodation because the temporary accommodation provided to them by the authorities - vacant commercial or retail spaces, disused military barracks or repurposed school gymnasiums - are not always one of the known addresses of shared accommodation. In the CRF data, the population in private households is approximated by restricting the data to entries that are in the responsibility of local immigration offices, thereby excluding entries that are currently in the responsivity of initial reception centres. The resulting harmonised CRF data set still includes 11.6 million people, of which 1.9 million are asylum seekers.

B. Training the model

9. In order to optimally predict refugee status in the Microcensus, various machine learning algorithms are trained and evaluated with the harmonised CRF data. For this purpose, the CRF data is divided into training data (90 % or 10.5 million entries) and test data (10 % or 1.2 million people).

10. Records of persons with a citizenship, with a percentage of people seeking protection of less than one percent were removed from the training data, in order to increase the proportion of protection seekers in the training data. The algorithm uses this training data (5.4 million entries) to learn the model parameters. In addition, the optimal hyperparameters are determined from the training data using cross-validation.[3, 4] Once the best hyperparameters have been found and the corresponding model parameters have been learned, the resulting model is used to predict the protection seeker status in the test dataset. Actual and predicted results can then be compared within the test data.

11. For the specific task at hand, a C5.0 classification tree (Kuhn, 2013) showed the best performance. Classification trees divide the observations into groups (leaves) that are as homogeneous as possible with as few branches as possible. They can make two types of predictions. First, they can provide binary predictions, i.e. a negative or positive decision on

[2] Place of residence in the Microcensus. In the administrative CRF data the responsible immigration authority approximates the place of residence.

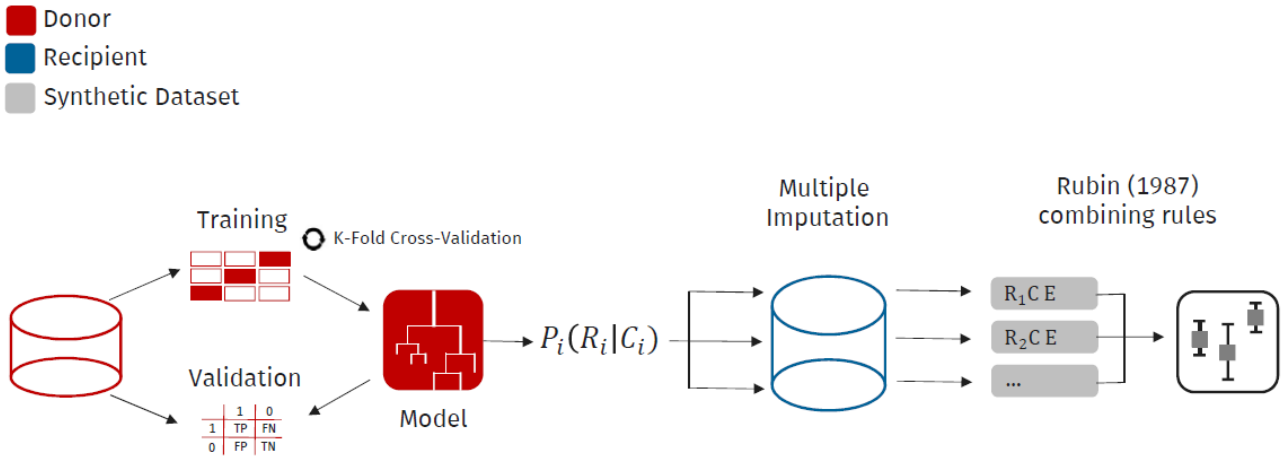
[3] Hyperparameters are parameters that are defined before training. For example, the C5.0 algorithm is given the number of individual classification trees to be learned in sequence in order to achieve the final model (boosting).

[4] The training data is further divided into K=10 subsets of equal size using K-Folds cross-validation. Performance metrics for different hyperparameter settings are determined by successively using each of these subsets for testing, while using the other subsets for training.

whether someone is classified as a person seeking protection or not. In this case, the majority of predicted outcomes in each leaf of the tree determine the same status for all observations in that leaf. Second, the model can provide the probability P(R|C) for each observation, i.e. the probability of being a person seeking protection. This probability is determined by the proportion of predicted positive and negative outcomes in each leaf of the tree. In the next step, the probability P(R|C) is used as the starting point for multiple imputation.

Figure 2

Statistical matching process



C. Predicting refugee status

12. In order to account for the uncertainty of the prediction from the estimation model, refugee status (R) is not only predicted once in the recipient data source, but 10 times. This means that 10 potentially different refugee statuses are assigned to each respondent in the Microcensus data. The specific binary status is drawn on the basis of the probability P(R|C) predicted from the model. This procedure is known as multiple imputation and it allows to assess how much uncertainty or variance we are adding through the imputation to the variance already present in our data.'

D. Inference

13. For statistical inference, the R package "mice" can be used to calculate the pooled expected value, the pooled variance and the associated degrees of freedom of the Student's t-distribution from the 10 estimates for the expected value and the variance. Following Rubin's (1987) combining rules in, a 95% confidence interval can be calculated for the expected value. Intuitively, smaller confidence intervals occur if the 10 predicted refugee statuses match, i.e. the model predicts a very high or very low probability that a person has a refugee status.

V. Evaluation

14. Careful evaluation is of central importance for a statistical matching project. The underlying assumption, the conditional independence of the unobserved bivariate distribution of the matched variables, cannot be verified. Nevertheless, there are properties of a statistical matching procedure that can be evaluated. First, this includes evaluating the quality of the estimation of the target variables. However, good performance in predicting the target variable is only a necessary and not yet a sufficient condition for reliable results. A decisive indicator of successful statistical matching is that central properties observed in the donor data set, such as univariate and bivariate distributions and correlation structures, can be reproduced in the recipient data set (Rässler, 2019).

A. Evaluating the estimation model

15. A confusion matrix is usually used to evaluate the quality of a classification algorithm. This matrix compares the predicted and actual values of the target variables in the test data. Metrics such as accuracy, sensitivity and specificity can be derived from this matrix to evaluate the model performance.

Table 1

Confusion matrix

Confusion matrix	Refugee status = 1 (true)	Refugee status = 0 (false)
Predicted refugee status = 1 (true)	True Positive	False Positive
Predicted refugee status = 0 (false)	False Negative	True Negative

(1) Accuracy = (TP+TN) / (TP+FP+FN+TN)

(2) Precision = TP / (TP+FP)

(3) Specificity = TP / (TP+FN)

16. In this case, the target variable is a binary indicator that shows whether a person recorded in the Central Register of Foreigners is seeking humanitarian protection or not. With around 16% positive and 84% negative observations, the target variable is clearly unevenly distributed in data. The prediction of exclusively negative results would therefore already provide a high accuracy of 84 %. When it comes to predicting rare events, the precision of the model is more informative and is therefore also used for parameter optimization within the classification algorithm in this case study.

Table 2

Model evaluation

Confusion matrix	Refugee status = 1 (true)	Refugee status = 0 (false)
Predicted refugee status = 1 (true)	141.767	42.606
Predicted refugee status = 0 (false)	43.879	934.879

(1) Accuracy = 0.93

(2) Precision = 0.77

(3) Specificity = 0.76

17. Further, the importance of the different predictors should also be considered as part of the evaluation of the estimation model. For classification trees, variable importance can for example be measured by the proportion of branches in which a predictor is used to split the data or the proportion of data in whose assignment to a leaf of the tree a predictor was involved. According to those measures, citizenship and year of entry are the most important predictions of the model. This result is intuitive: People with Syrian, Afghan and Iraqi citizenship who entered the country between 2014 and 2016 accounted for 38 % of those seeking protection at the end of 2021.

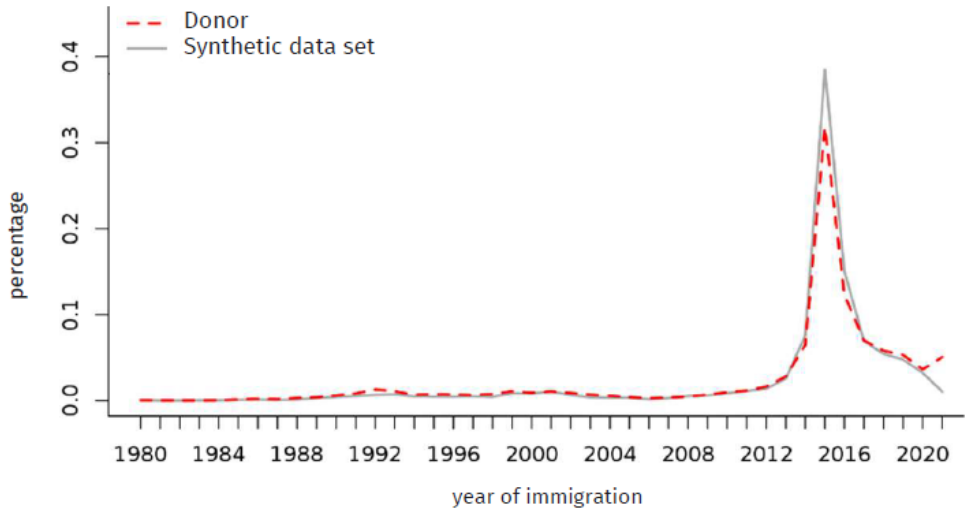
B. Evaluating the synthetic dataset

18. At the end of 2021, 1,862,000 protection seekers living in private households were registered in the Central Register of Foreigners. After applying statistical matching, 1,770,000 protection seekers are estimated in the Microcensus. It should be noted that the estimation procedure was optimized with regard to precision. False positive estimates are therefore evaluated more negatively than false negative predictions. In general, the aim of statistical matching is typically not to reproduce the total number of the target population, but to analyse unobserved distributions within the target population. Therefore, the evaluation of the reproduction of distributions from the donor dataset is particularly central (Rässler, 2019).

19. In this context, it is relatively easy to check whether the univariate marginal distributions from the donor data have been retained in the synthetic data set after statistical matching. Figure 3 depicts the distribution of people seeking protection by year of entry in the donor data and the synthetic Microcensus data.

Figure 3

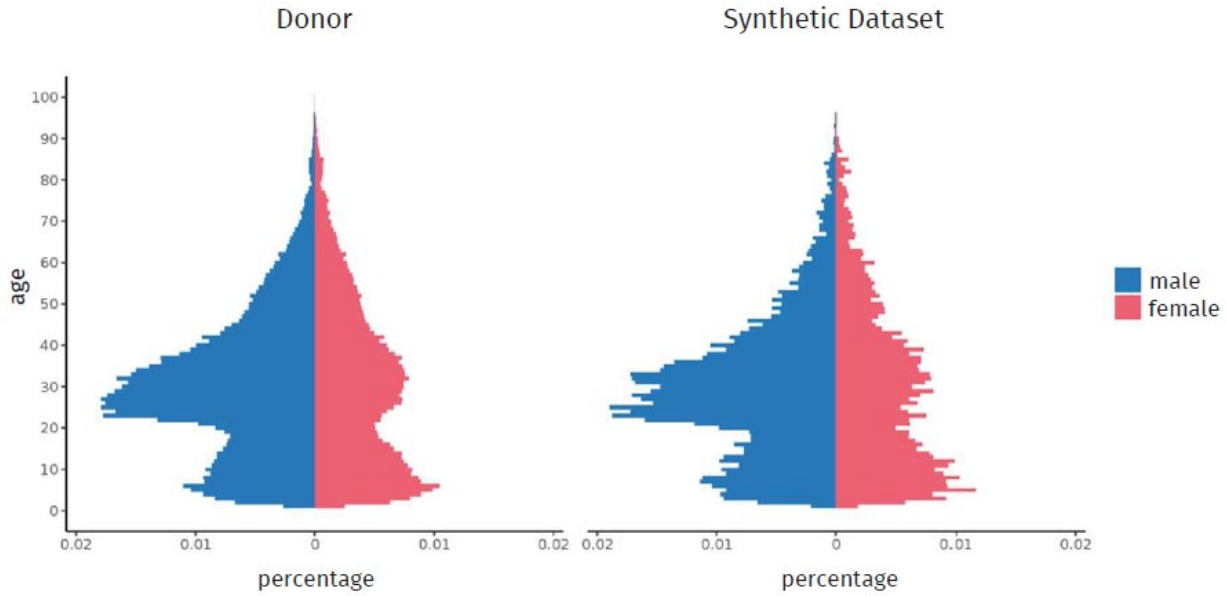
People seeking protection by year of immigration



Hellinger distance: 0,114
Kolmogorov-Smirnov [p-value]: 0.292

20. To assess whether a distribution has been successfully obtained after imputation, a Hellinger distance can also be calculated in addition to the visual impression. The Hellinger distance measures the similarity between two distributions on a scale between 0 and 1, where 0 indicates perfect similarity and 1 perfect dissimilarity. The aim is to achieve values of no greater than 0.05. However, values around 0.1, as in the case of the distribution of first entry years, are still considered acceptable in the literature (Eurostat, 2013).
21. Another way of comparing the distribution of donor and recipient data sets is by means of a Kolmogorov-Smirnov test. In this test, the null hypothesis states that the two distributions are the same. The alternative hypothesis, on the other hand, states that there is a significant difference between the two distributions. The p-value of 0.292 is above the previously defined significance level of 0.05, which is why the null hypothesis cannot be rejected.
22. In addition to univariate distributions, bivariate distributions can also be considered. The visual impression in Figure 4 confirms the preservation of the central characteristics in the age and gender structure of those seeking protection in the synthetic data set. The shape of the population pyramids is clearly similar. The Hellinger distance and the Kolmogorov-Smirnov test also indicate a preservation of the marginal distribution with regard to the age variable by gender at the univariate level.

Figure 4
People seeking protection by age and gender



Male:

Hellinger distance: 0.035

Kolmogorov-Smirnov [p-value]: 0.322

Female:

Hellinger distance: 0.037

Kolmogorov-Smirnov [p-value]: 0.124

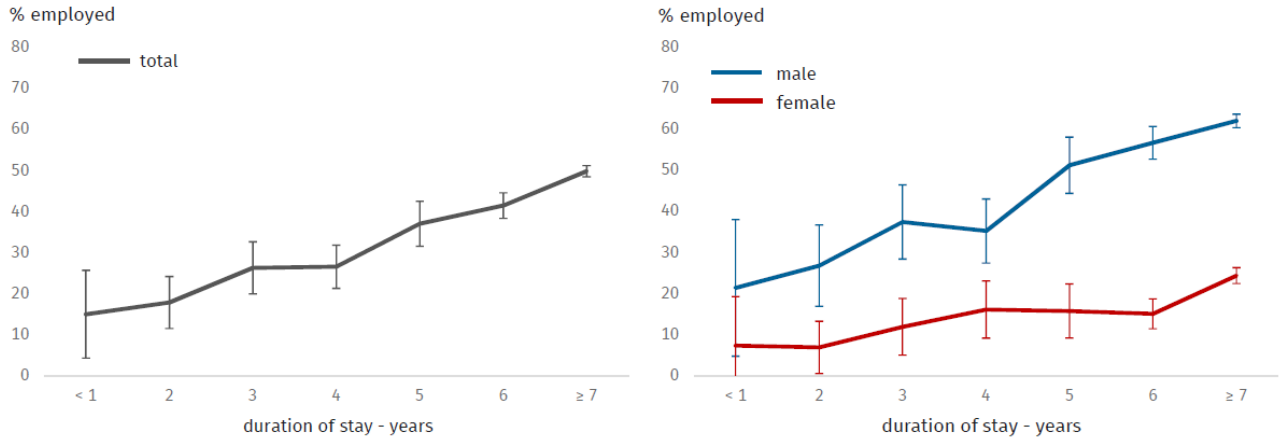
VI. Results

23. The synthetic Microcensus data enriched by the refugee status can be used to answer a variety of socio-economic questions about people seeking protection. For example, the Microcensus provides information on labour force participation, the economic sectors of those in employment, net income and school-leaving qualifications.

24. With regard to labour force participation, in the year of arrival the predicted refugees and people in refugee like situations show a low labour force participation and at the same time large confidence intervals. As the duration of stay increases, the employment rate increases, especially for men. Around a quarter of predicted people seeking protection living in Germany for between 2 to 3 years in 2021 were employed - men more frequently (37.5 %) than women (11.9 %). After a stay of between 5 to 6 years the proportion 56.8 % for men and 15.1 % for women. Large 95 % confidence intervals for refugees with a duration of stay of less than one year are related to a lower number of observations.

Figure 5

Employment rate of predicted people seeking protection in 2021 by year of arrival



VII. External validation

25. The results obtained from statistical matching can be validated with external results. In a joint project, several German research institutes conducted a targeted survey of newly arrived refugees.[5] The CRF was used as the selection frame for the survey. In the first wave of this targeted survey, a total of 4,500 refugees were interviewed.[6] The survey is designed as a longitudinal survey in which people are interviewed repeatedly since 2016. Results on labour market participation are published in Brücker et al. (2023).

26. Both the statistical matching estimate and the targeted survey find increasing employment rates as the duration of stay progresses and that this increase is much more pronounced for men than for women.

27. The employment rate of refugee women and men in the first year after moving to Germany is initially 7 % in the survey. Statistical matching results in a 95 % confidence interval between 4 % and 26 %. Although the result of the survey lies within the confidence interval of the statistical matching estimate, the confidence intervals for new arrivals are too large to be interpreted meaningfully.

28. Six years after immigration, the proportion of refugees in employment in the targeted survey is 54 %. Estimates based on statistical matching put the figure at between 38 % and 45 %. The survey finds that 67 % of men are employed at this time, and 23 % of women. The

[5] The survey is a joint project of the Institute for Employment Research (IAB), the Socio-Economic Panel (SOEP) of the German Institute for Economic Research (DIW) and the Federal Office for Migration and Refugees (BAMF).

[6] The definition of refugees in the IAB-BAMF-SOEP survey largely corresponds to the definition of people seeking protection in official statistics. It includes people who are still in the asylum process, people who have already been granted protection status and people whose asylum application has been rejected but who are still in Germany.

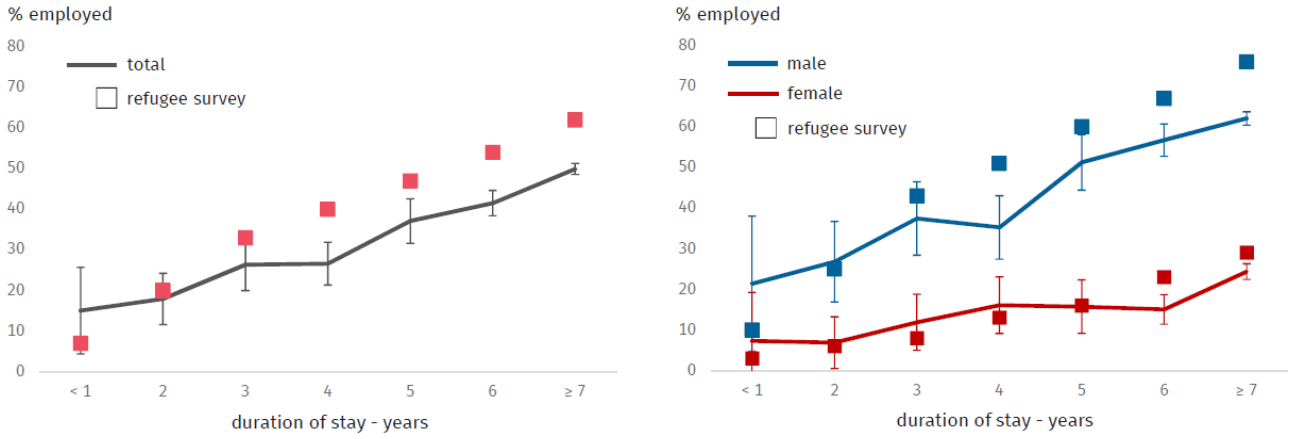
estimates from the synthetic Microcensus extended by refugee status are between 53 % and 61 % for men and between 11% and 19% for women.

29. After seven years or more in Germany the refugee survey estimates that overall 62 % of refugees living in Germany are employed, 76 % for men and 29 % for women. The statistical matching confidence interval ranges from 49 % to 51 % for both genders combined and from 60 % to 64 % for men and from 22 % to 26 % for women.

30. Figure 6 visually compares the estimates obtained from statistical matching with the results obtained from the refugee survey. Estimates for refugees with a duration of stay of less than one year cannot be interpreted meaningfully due to large confidence intervals. Apart from that, statistical matching estimates of employment rates are consistently lower for refugees after a duration of stay of three years. While estimates for females match quite well, statistical matching predicts lower rates of employment especially for males with longer durations of stay. The extent to which these differences can be attributed to a bias in statistical matching or a bias in the longitudinal refugee survey, for instance unemployed male refugees having a higher probability to drop out of the sample (attrition bias), cannot be assessed in the context of this paper.[7]

Figure 6

External validation



VIII. Conclusion

31. Statistical matching allows to analyse the relationship of variables that are not jointly observed and is therefore a potential alternative to increasing response burden in surveys. This working paper provides estimates for labour force participation of people seeking protection in Germany in 2021 obtained by statistically matching Microcensus survey data with administrative data from the Central Register of Foreigners.[8]

[7] Brücker et. al (2023) report to take countermeasures to mitigate sample attrition bias.
[8] Results only refer to people living in private households.

32. Further, analyses for further reporting years indicate that the results up to and including 2021 are relatively stable. However, initial analyses for 2022 show limits to statistical matching. In 2022 people who fled Ukraine after the Russian attack were not fully covered in the Microcensus. In the Central Register of Foreigners registration was more complete (Federal Statistical Office, 2023). This unequal coverage of Ukrainians in both data sources makes meaningful harmonization impossible, which means that a necessary condition for statistical matching cannot be met.
33. Another caveat is the uncertainty of estimates obtained by statistical matching. To a certain degree uncertainty can be considered and communicated by means of multiple imputation and by publishing confidence intervals rather than point estimates. However, unbiased results rely on the assumption of conditional independence, which cannot be verified. Therefore, careful evaluation is important: First, evaluation of the estimation model and second evaluation of the resulting synthetic data set. Compared to traditional surveys, in which all relevant variables are recorded together, there is additional risk of systematic bias in an estimate generated using statistical matching. Accordingly, statistical matching is an additional option in the toolbox of National Statistical Offices, which is particularly suitable for explorative analysis.

Literature

Brücker, H.; P. Jaschke, Y. Kosyakova and E. Vallizadeh (2023). Entwicklung der Arbeitsmarktintegration seit Ankunft in Deutschland: Erwerbstätigkeit und Löhne von Geflüchteten steigen deutlich (IAB-Kurzbericht 13/2023), Nürnberg.

Cielebak, J. and S. Rässler (2019). Data Fusion, Record Linkage und Data Mining, S. 423-439.

Eberle, J. (2019): Schutzsuchende. Ein Konzept zur Ermittlung der Zahl an Ausländerinnen und Ausländern, die sich aus humanitären Gründen in Deutschland aufhalten In Wirtschaft und Statistik, 1/2019

Eurostat, European Commission (2013). Statistical Matching: A Model-based Approach for Data Integration, Methodologies & Working papers, Luxembourg.

Kuhn, M., Johnson, K. (2013). Classification Trees and Rule-Based Models. In: Applied Predictive Modeling. Springer, New York.

Statistisches Bundesamt (2023). Bevölkerung und Erwerbstätigkeit - Statistik über Schutzsuchende, Qualitätsbericht 2022, Wiesbaden.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons, New York.