Creating a synthetic database for research in migration and subjective well-being

GERGELY MÁRK BAGÓ, ZOLTÁN CSÁNYI,
ANNA SÁRA LIGETI

2019 edition

STATISTICAL
WORKING PAPERS

eurostat

Creating a synthetic database for research in migration and subjective well-being

GERGELY MÁRK BAGÓ, ZOLTÁN CSÁNYI,
ANNA SÁRA LIGETI

2019 edition

Manuscript completed in June 2019

The European Commission is not liable for any consequence stemming from the reuse of this publication.

Luxembourg: Publications Office of the European Union, 2019

Copyright for cover photo: © Shutterstock/Alex Erofeenkov

Theme: Population and social conditions
Collection: Statistical working papers

The information and views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

PDF ISBN 978-92-76-08618-5 ISSN 2315-0807 doi: 10.2785/375543 KS-TC-19-003-EN-N

Abstract

While recent advances in migration studies call for targeted research on migration aspirations and the cognitive mechanisms of voluntariness when making choices among alternatives of behaviour, detailed information on (potential) migrants' subjective characteristics, including cognitive evaluations of life and life conditions is still scarce in official statistics. With the aim of establishing good practices for statistical matching estimation methods for official migration statistics in Hungary on the one hand, and encouraging migration researchers to create and use synthetic data to get insights into the subjective characteristics of potential migrants on the other, this paper reports on the methodological details of a statistical matching experiment that combines the International Migration and the Subjective Well-being complementary questionnaires of Hungarian Microcensus 2016. In the experiment, eight variants of non-parametric hot deck methods were performed to create synthetic data sets and evaluated both from a methodological and an analytical point of view.

Authors:
Gergely Márk Bagó (Hungarian Central Statistical Office, Methodology Department)
Zoltán Csányi (Hungarian Central Statistical Office, Population Census and Demographic Statistics Department)
Anna Sára Ligeti (Hungarian Central Statistical Office, Population Census and Demographic Statistics Department)

Keywords: Synthetic migration data, Statistical Matching, Migration behaviour, Subjective well-being, Hungarian Microcensus

Table of content

1 Introduction

Since a growing number of national migration policies aim at influencing behaviour related to different aspects of migration([1]), the growing need for producing high quality measurements of both migration behaviour and the socio-economic contexts in witch migration takes place is apparent. In this respect, beyond its role in describing and evaluating trends of migration and development, data production might also serve research purposes that support policy makers in identifying the factors of migration behaviour that "might be influenced by policy measures" (Carling and Collins, 2018). While most recent advances in migration studies call for targeted research on migrants' aspirations (Carling and Schewel, 2018) and the cognitive mechanisms of voluntariness when making choices among alternatives of behaviour (Erdal and Oeppen, 2018), detailed information on potential migrants' subjective characteristics and cognitive evaluations of their circumstances is still scarce in official statistics.([2])

Seeking to fill in this gap, this paper reports on the methodological details of our experiment of creating synthetic data: a statistical – or synthetic – matching experiment was carried out to combine data from two complementary questionnaires of Hungarian Microcensus 2016, thus allowing for the joint analysis of migration behaviour and subjective variables that are rarely collected together. Using information from the Microcensus 2016 basic questionnaires, we combined:

- the International Migration data set (with detailed questions on past experiences of migration and future migration aspirations);
- Subjective Well-being data set (that contains information on different aspects of satisfaction, trust in institutions, mental states, optimism about the future).

([1]) See United Nations Inquiry among Governments on Population and Development at https://esa.un.org/poppolicy/inquiry.aspx
([2]) See Stiglitz, Sen and Fitoussi (2009) on why subjective measurements in official statistics is crucial for policy makers.

In the exercise, eight variants of non-parametric hot deck methods were performed and evaluated both from a methodological and an analytical point of view. On the one hand, we expect that the results will contribute to establishing good practices in statistical matching methods for official migration statistics in Hungary,(3) and on the other, that the experiment would encourage migration researchers to create and use synthetic data to exploit the resulting analytical potential.

This paper – instead of entering in the details of analytical results – focuses on the methodological questions of creating and evaluating synthetic data. First, in Chapter 2, we offer a short overview of the implications of using Microcensus data for the matching exercise. Then in Chapter 3, we review the methodological details of the distance hot deck techniques used to perform statistical matching that will be followed in Chapter 4 by an evaluation of the results (both methodological and analytical perspectives). Finally, the concluding remarks close the paper.

[3] See Willekens (2017) on the role of estimation methods in official migration statistics.

2 The Microcensus data set

In 2016 October, with the aim of tracking social trends between full-scope censuses, the Hungarian Central Statistical Office (HCSO) carried out Microcensus, a population survey based on an unusually large sample covering 10 per cent of the Hungarian households (for more details, see HCSO, 2018). Apart from the basic questionnaires on dwellings and personal information, selected households were asked to fill in one of the following complementary surveys on a) international migration, b) subjective well-being, c) social stratification, d) occupational prestige, and e) health problems (see Table 1).

Table 1: The structure of basic and complementary questionnaires and sample sizes in Microcensus, 2016

Basic questionnaires	Complementary questionnaires
Personal Questionnaire (N= 815.521) Dwelling Questionnaire (N= 406.023)	International Migration (N=41.367)
	Subjective Well-being (N=51.281)
	Social Stratification (N=101.165)
	Occupational Prestige (N=43.480)
	Health Problems (N=68.196)

With its basic and complementary questionnaires, the Microcensus data set fulfils the conditions needed to combine the data sub sets and create synthetic data (see D'Orazio et al. 2006):

- detailed methodological information available on the sampling design and the variables of data collection;
- common methodology for data collection across the basic and complementary questionnaires;
- the sub-samples for complementary questionnaires belong to the same target population;
- data collection was carried out in the same period;
- similar distributions of common key variables;
- lack of data (missing data) negligible in both datasets.

As regards the data sub sets involved in the exercise, these were a) the International Migration data set with information on respondents' migration aspirations („seriousness” of migration plans, planned duration of stay, planned destinations, main activity expected, skills (mis)match expected) as well as on their past migration experiences (duration and number of stay(s), destinations, main activity abroad, skills (mis)match abroad, motivations for return, remittances sent home from abroad); and b) the Subjective Well-being data set on respondents' levels of satisfaction (overall life satisfaction, satisfaction with household's financial situation, personal incomes, work-life balance, social relations, living conditions, etc.) and on their mental/emotional states, trust in institutions or optimism about the future.

It should be noted that while in case of the Subjective Well-being survey, respondents were selected randomly from the resident population older than 15, in contrast, members of households aged 16-64, affected by migration (i.e. household members staying abroad having past migration experiences or future migration aspirations) were more likely to be selected for the International Migration survey (Dickmann and Ligeti, 2018). Further, similar to Census 2011, in determining the target population, Microcensus used the concept of usual residence. That is, those staying abroad for more than a year were not included in the sample of the Personal questionnaire. (On the distributions of respondents by main socio-demographic variables see Annex 1.)

3 Statistical matching and hot deck variants applied

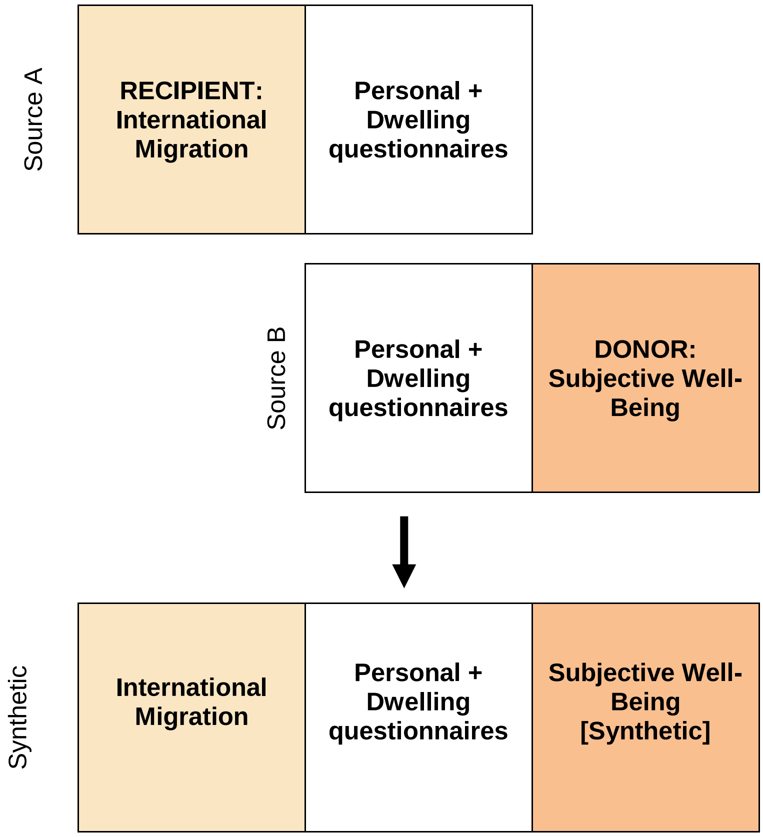
Statistical matching, in accordance with the UNECE Data Integration Guide (UNECE, 2018) “involves the integration of data sources with usually distinct samples from the same target population, in order to study and provide information on the relationship of variables not jointly observed in the data sets”. That is, the statistical matching exercise resembles an “imputation problem of the target variables from a donor to a recipient survey” (Leulescu and Agafitei, 2013)([4]). A scheme of our experiment can be seen in Table 2: synthetic data from the donor Subjective Well-being data set was created to complement information from the recipient, International Migration data set using variables of the basic Personal and Dwelling questionnaires. Donor variables include principal components of satisfaction, mental/emotional states, trust in institutions and a variable of ‘optimism about the future’ (see Annex 2). Given the uniform data collection of Microcensus and its structure of basic and complementary questionnaires, no reconciliation of the definitions, time references and sample frames was needed for the experiment.

In order to find the most suitable donors to be matched with recipients, we performed a total of eight variants of the Nearest neighbour, Random distance and Rank hot deck methods using R package([5]). The results of each approach were evaluated and compared with each other in terms of imputation quality (methodological perspective). Their coherence with relevant national empirical data was also examined (analytical perspective).

Table 3 gives an overview of the eight variants applied. While Nearest neighbour techniques aim at identifying the donors that are most similar to recipients in accordance with "a distance computed on the matching variables", in the case of Random hot deck methods, donors are selected randomly "from a suitable subset of all the available donors." Finally, when using Rank hot deck, donors are chosen based on a distance measured as the empirical cumulative distribution of one common variable (see D'Orazio, 2017).

([4]) See also De Waal (2015), D’Orazio et al. (2006), Serafino and Tonkin (2017)
([5]) R version 3.5.2 (2018-12-20) -- "Eggshell Igloo" Copyright (C) 2018 The R Foundation for Statistical Computing, Platform: x86_64-w64-mingw32/x64 (64-bit)

Table 2: Scheme of the experiment


Source A
RECIPIENT: International Migration
Personal + Dwelling questionnaires
Source B
Personal + Dwelling questionnaires
DONOR: Subjective Well-Being
Synthetic
International Migration
Personal + Dwelling questionnaires
Subjective Well-Being [Synthetic]

In using hot deck techniques, we considered some of the most common metrics – or distance functions – used in mathematical statistics. *Euclidean distances* appeared to be appropriate first due to the easy interpretation as the (dis)similarity of respondents along two or more variables. Second, *Mahalanobis distances,* which are useful in case of correlating matching variables with differing standard deviation, and third, *Manhattan distances,* based on the absolute differences of attribute values (coordinates), were also considered.

Table 3: Variants of hot deck methods, matching variables and donation classes used in the experiment

	Nearest neighbour distance hot deck			Random distance hot deck				Rank hot deck
	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8
Distance Method	Euclidean	Mahalanobis	Manhattan	Default	Euclidean *	Mahalanobis*	Manhattan *	**
Constraint	Unconstrained							
Matching variables	Age, partner, underage child, type of settlement							Age
Class variables (donation classes)	Sex, Education							
Donor variables	Satisfaction (PC), Mental / emotional state (PC), Trust in institutions (PC), Optimism about the future							

Notes: * The 10 closest donors (k=10) are retained. The donors are selected from the subset with equal probability (not weighted random hot deck).
** Empirical cumulative distribution of Age.

In choosing matching variables for the computation of distances, our starting point was that a relatively strong correlation of the selected variables with both subjective well-being and migration aspirations is needed. As a result, age, having a partner, having underage children and the type of settlement were established as matching variables (See Annex 3). Further, donation classes of sex and education were created, that is distance functions were run within groups defined by concordant categories of the sex and educational attainment of respondents. In the exercise no constraints were applied on how many times a donor can be used. While unconstrained matching generally results in minor distances, a possible disadvantage is that, potentially biasing the results, the same donors can be over selected. Table 4 shows that the maximum number of the same donors' recurrences was the highest – 28 – using the Rank hot deck variant, that we considered relatively low.

Table 4: Maximum number of recurrences of the same donors using hot deck variants

	Nearest neighbour distance hot deck			Random distance hot deck				Rank hot deck
Distance Method	Euclidean	Mahalanobis	Manhattan	Default	Euclidean *	Mahalanobis*	Manhattan *	**
Maximum number of recurrences	22	21	18	9	26	27	27	28

Notes: * The 10 closest donors (k=10) are retained. The donors are selected from the subset with equal probability (not weighted random hot deck).
** Empirical cumulative distribution of Age.

Finally, it should be noted that the independence of donor and recipient variables conditional on the common variables is required to assure the accuracy of results. However, usually lacking the information to confirm whether conditional independence holds or not, the statistical matching exercise is often carried out – including our case – under the Conditional Independence Assumption (CIA). Overcoming this assumption became one of the central questions of methodological research in statistical matching (see e.g. Leulescu and Agafitei, 2013, Donatiello et al. 2014, D'Orazio et al. 2017).

4 Evaluation of the results

One sample t-tests were used to examine differences between the mean distributions of donor and synthetic subjective well-being variables: no significant differences were found. (Please note that mean distribution of donor variables was used as a reference.) Then, frequencies were also compared. Figure 1 shows the means of the donor and the eight variants of synthetic variables (Satisfaction principal component, Mental/emotional states principal component, Trust in institutions principal component, and Optimism about the future) by age groups. While most variants fit well with the original, that of the rank hot deck variant apparently differs from the donor in all cases, while the random distance hot deck default variant differs significantly in the case of the mental/emotional states principal component. These variants were excluded from further analyses.

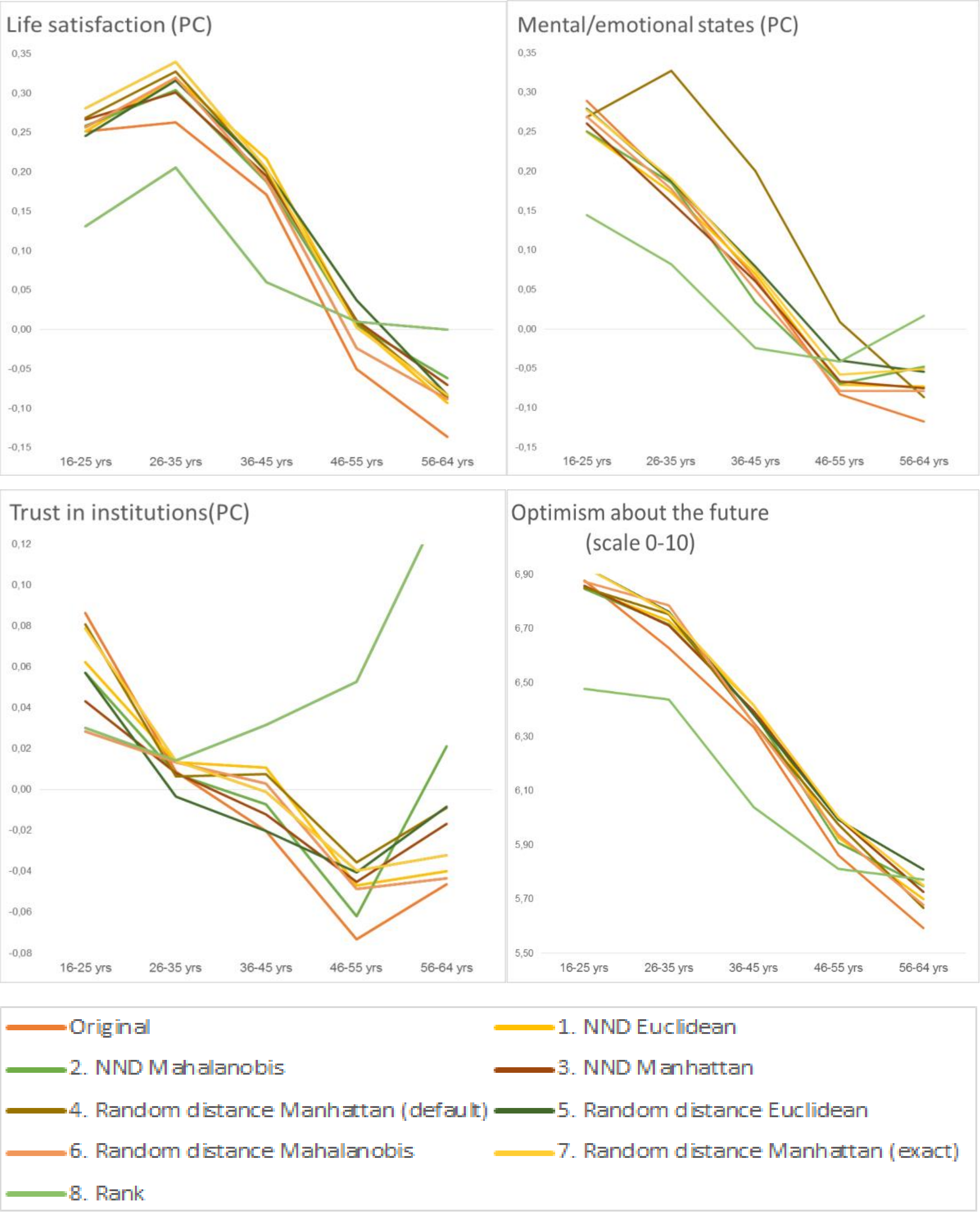
As a second step, we compared data with relevant previous research carried out in Hungary in recent years. While relatively abundant literature is available on who is aspiring to migrate, less is known on how different aspects of migration aspirations are related to subjective characteristics. In these regards, national inquiries([6]) suggest that the less satisfied are more prone to leave (see Gödri, 2016; Sik and Szeitl, 2016). The Microcensus Subjective well-being data set also supports this in the case of almost each satisfaction variable but satisfaction with one's health that – obviously related to age – affects negatively the development of migration aspirations (see Annex 2.).

Aspiring migrants are quite similar in terms of happiness and satisfaction with social relations, in accordance with Gödri (2016), Sik and Szeitl (2016) and the Subjective Well-being data set as well. Results on the variable 'optimism about the future' however diverge. While Gödri (2016) found no significant relations of this variable and migration aspirations, Sik and Szeitl (2016) found a positive relation. In contrast, in the Subjective well-being data set aspiring migrants are less optimistic. An explanation for these uncertainties is that some respondents might refer to the future in the sending countries and others to that in destinations. Each of the sources mentioned above suggest that aspiring migrants with more "serious" migration plans – who most probably would refer to destinations – are more optimistic. Interestingly, regarding the expected duration of stay the relationship is inverted U-

([6]) For international research in this field, see e.g. Ivlevs (2015), Migali and Scipioni (2018), Otrachshenko and Popova (2014), Skoglund and Csányi (2019).

shaped. Contrarily to the persons who would never return, those planning to leave for a longer period, but not forever, are more optimistic about the future.

Figure 1: Mean values of the original and the synthetic donor variables by age groups



Accordingly, in this phase of the experiment we expected that the synthetic variables resulting from the remaining six hot decks variants would have a significant effect on the seriousness of migration aspirations: Satisfaction, Mental/Emotional states and Trust in institutions would decrease the probability of developing more well-defined plans while trust in future would increase it. In the analysis we put special emphasis on the latter. Table 5 shows how synthetic data fulfil these expectations, on the basis of binary regression models of the seriousness of migration aspirations on a scale of "considering the possibilities of migration”; “seriously thinking about migrating” and “having already taken the migration decision” (the dependent variable was whether having already taken the migration decision or not). Results supporting previous national research are indicated in the table by light orange colour code.

As regards the Nearest neighbour methods, while in the case of Euclidean and Manhattan variants, the effects of all four variables were significant, optimism about the future did not fit our expectations. Contrarily, in the case of the Nearest neighbour Mahalanobis variant, the effects of mental/emotional states and trust in institutions were found to be not significant, however optimism about the future affected the seriousness of migration aspirations as expected. Further, R square was the highest in this case. For these reasons, the Mahalanobis variant was considered as the one among Nearest neighbour methods that gives the most accurate results.

Among Random distance methods, the Manhattan variant was first discarded since only Mental/emotional states affected migration aspirations as predicted. While optimism about the future fulfilled our expectations using both Euclidean and Mahalanobis variants, the effects of all four synthetic variables were significant only in the case of the former (however satisfaction had a contrary effects to that expected). Due to that – despite the Mahalanobis variant produced higher R square values – we deemed the Euclidean variant to be most accurate Random distance method.

Table 5: Effects of the synthetic variables on the seriousness of migration aspirations (population aged 16-64)

	NND Euclidean	**NND Mahalanobis**	**NND Manhattan**	**Random distance Euclidean**	**Random distance Mahalanobis**	**Random distance Manhattan**
Satisfaction	0.976**	0.936**	1.032**	1.037**	0.948**	1.022**
Mental / emotional states	0.977**	0.992	0.965**	0.922**	0.993	0.960**
Trust in institutions	0.979**	0.994	0.970**	0.972**	1.016**	1.007
Optimism about the future	0.993**	1.006*	0.986**	1.013**	1.029**	0.996
Nagelkerke R Square	**0.198**	**0.204**	**0.189**	**0.184**	**0.201**	**0.187**

Notes: ** p<0.01 *p<0.05

5 Concluding remarks

Results of the matching experiment are promising. Joint distributions of the variables involved suggest that most hot deck distance variants were effective in creating synthetic data. Further, through a comparison with the findings of previous national research, the Nearest neighbour Mahalanobis and the Random hot deck Euclidean variants appeared to be the most appropriate for creating synthetic data sets for in-depth analyses. In doing such comparison, advantages of the synthetic data also came to the fore: the diversity of variables whose analysis has become now available offers possibilities for researchers to reach new levels of analytical depth.

Not only relations of a wide range of subjective well-being and migration behaviour (past experiences/future aspirations) variables can be examined now, but it has also become clear that the inclusion of even more synthetic variables from other Microcensus complementary data sets (or others) is possible using statistical matching techniques. The potential of using synthetic data for research purposes and the importance of providing new evidences for migration theorization cannot be overemphasized. Beyond a better comprehension of how subjective characteristics and migration aspirations are interrelated, the possibility of examining aspirations of circular or multiple migrations is just one example of the most promising future research lines.

Though the Microcensus data set – with common data collection methods, sampling and variables across the basic and complementary subsets – was particularly well-suited for statistical matching, experimenting with other data sources in migration research should not be ignored. Despite its limitations, matching techniques have become widely accepted among social statisticians. As far as producers of official statistics are concerned, along with the growing popularity of disseminating experimental statistics on diverse economic and societal topics, data producers are further encouraged to use estimations in population and migration statistics.

References

Carling, J. and Collins, F. (2018), „Aspiration, desire and drivers of migration”. *Journal of Ethnic and Migration Studies*. 44(6): 909–926.

Carling, J. and Schewel, K. (2018), „Revisiting aspiration and ability in international migration”. *Journal of Ethnic and Migration Studies*. 44(6): 945–963

De Waal, T. (2015), „Statistical matching: Experimental results and future research questions.” *CBS Discussion Papers*, 2015/19

Dickmann, A. and Ligeti, A. (2018), *Mikrocenzus 2016, Nemzetközi vándorlás* (Microcensus 2016, International migration), Hungarian Central Statistical Office

Donatiello, G., D'Orazio, M.,Frattarola, D., Rizzi, A., Scanu, M., and Spaziani, M. (2014), „Statistical Matching of Income and Consumption Expenditures.*" International Journal of Economic Sciences,* Vol. III(3): 50 – 65.

D'Orazio, M. (2017), *Statistical Matching and Imputation of Survey Data with StatMatch*. Available at: https://cran.r-project.org/web/packages/StatMatch/vignettes/Statistical_Matching_with_StatMatch.pdf

D'Orazio, M., Di Zio, M. and Scanu, M. (2006), *Statistical Matching: Theory and Practice*. Wiley, Chichester

D'Orazio, M., Di Zio, M. and Scanu, M. (2017), "The use of uncertainty to choose matching variables in statistical matching." *International Journal of Approximate Reasoning*. 90: 433-440

Erdal, M.B. and Oeppen, C. (2018), „Forced to leave? The discursive and analytical significance of describing migration as forced or voluntary”. *Journal of Ethnic and Migration Studies*. 44(6): 981–998.

Gödri, I. (2016), "Elvándorlási szándékok – álmok és konkrét tervek között: A migrációs potenciál jellemzői és meghatározó tényezői a 18–40 évesek körében Magyarországon" (Migration intentions - dreams and concrete plans: Characteristics and determinants of Migration Potential among the 18-40 years old population in Hungary), *Research Reports* No.98, Hungarian Demographic Research Institute

HCSO (2018), *Microcensus 2016, Characteristics of the population and dwellings*. Hungarian Central Statistical Office. Available at: http://www.ksh.hu/mikrocenzus2016/book_2_characteristics_of_population_and_dwellings

Ivlevs, A. (2015), "Happy Moves? Assessing the Link between Life Satisfaction and Emigration Intentions." *IZA Discussion Paper Series* No. 9017.

Leulescu, A. and Agafitei, M. (2013), "Statistical matching: a model based approach for data integration", *Eurostat Methodologies and Working Papers*

Migali, S. and Scipioni, M. (2018), "A global analysis of intentions to migrate." European Comission, *JRC Technical Reports*

Otrachshenko, V. and Popova, O. (2014), "Life (Dis)Satisfaction and the Intention to Migrate: Evidence from Central and Eastern Europe". *Journal of Socio-Economics* 48: 40–49.

Serafino, P. and Tonkin, R. (2017), "Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey." *Eurostat Statistical Working Papers*

Sik, E. and Szeitl, B. (2016), „Migration intentions in contemporary Hungary." In: Blaskó, Zs. – Fazekas, K. (eds) *The Hungarian Labour Market 2016*. Institute of economics, Centre for Economic and Regional Studies, Hungarian Academy of Sciences Budapest, 2016, pp. 55-60. Available at: http://www.econ.core.hu/file/download/HLM2016/TheHungarianLabourMarket_2016_onefile.pdf

Skoglund, E. and Csányi, Z. (2019), „Quantitative analysis of the objective and subjective aspects of youth migration in the Danube region" *YOUMIG Working Papers* No.3. Available at: http://www.interreg-danube.eu/uploads/media/approved_project_output/0001/32/1f04dd4d6ee3459935876d76137f00984ee07c05.pdf

Stiglitz, J.M., Sen, A. and Fitoussi, J. (2009), *Report by the commission on the measurement of economic performance and social progress*. Paris: Commission on the measurement of economic performance and social progress.

UNECE (2018), *A Guide to Data Integration for Official Statistics*. Version 2.0. Available at: https://statswiki.unece.org/display/DI/Guide+to+Data+Integration+for+Official+Statistics

Willekens, F. (2017), *Evidence-based monitoring of international migration flows in Europe*. Keynote Speech at DGINS, Budapest, 2017. Available at: http://www.ksh.hu/dgins2017/papers/dgins2017_keynote_frans_willekens.pdf

Annex

Annex 1: Distributions of respondents by main socio-demographic variables, % (age 16-64)

	International Migration dataset	Subjective Well-being dataset
Sex		
Male	55	62
Female	45	38
Age group		
16-25 years old	20	16
26-35 years old	28	19
36-45 years old	26	25
46-55 years old	15	21
56-64 years old	10	20
Educational attainment		
Primary or lower	13	18
Vocational	20	27
Secondary	35	34
Tertiary	32	21
Underage children		
No	67	68
Yes	33	32
Type of settlement		
Small town	22	30
Town	31	36
County seat	18	17
Budapest	29	18
Partner		
No	44	38
Yes	56	62

Annex 2: Items of the Principal Components of life satisfaction, mental/emotional states, and trust in institutions by means of the aspiring and non-aspiring migrants (age 16-64, Subjective well-being dataset)

	Aspiring migrants	Non-aspiring migrants
Life satisfaction (Scale 0-10)		
How satisfied are you with your physical health?	7,2	6,9
How satisfied are you with your personal relationships?	7,2	7,4
How satisfied are you with your apartment/housing?	6,5	7,0
How satisfied are you with your quality of living conditions?	6,4	6,9
How satisfied are you with your life as a whole?	6,1	6,6
How satisfied are you with your workplace / school circumstances?	6,0	6,4
How satisfied are you with your free time?	5,4	5,7
How satisfied are you with your financial situation of your household?	5,3	5,9
How satisfied are you with your income?	4,4	5,2
Mental/emotional states (scale 1-5)		
How often do you feel lonely?	3,7	3,9
How often do you feel angry?	3,2	3,4
How often do you feel despondent?	3,2	3,3
How often do you feel stressed?	3,0	3,2
How often do you feel calm?	2,5	2,4
How often do you feel happy?	2,2	2,2
Trust in institutions (scale 0-10)		
Trust in the political system	2,6	3,8
Trust in the legal system	3,4	4,4
Trust in the police	4,3	5,3
Trust in the army	4,7	5,5

Annex 3: Means of the subjective well-being variables and distributions of potential migrants by the main socio-demographic variables (age 16-64, Subjective Well-being dataset)

	Life satisfaction PC (means)	**Mental/emotional states PC (means)**	**Trust in institutions PC (means)**	**Optimism about the future (means on scale 0-10)**	**Potential migrants (%)**
TOTAL	**0.04**	**0.03**	**-0.02**	**6.23**	**8.5%**
Sex					
Male	0.07	0.04	-0.03	6.24	8.7%
Female	0.00	0.00	0.00	6.21	8.2%
Age group					
16-25 years old	0.20	0.26	0.08	6.88	17.7%
26-35 years old	0.21	0.15	0.01	6.63	12.4%
36-45 years old	0.12	0.03	-0.02	6.33	8.0%
46-55 years old	-0.11	-0.12	-0.07	5.86	5.3%
56-64 years old	-0.20	-0.15	-0.05	5.59	2.5%
Educational attainment					
Primary or lower	-0.35	-0.11	-0.13	5.84	7.9%
Vocational	-0.12	-0.05	-0.10	5.87	6.8%
Secondary	0.12	0.08	0.01	6.35	9.8%
Tertiary	0.41	0.15	0.15	6.85	9.2%
Underage children					
No	0.01	0.00	-0.02	6.16	9.2%
Yes	0.10	0.08	-0.01	6.39	7.0%
Type of settlement					
Small town	-0.04	0.01	-0.01	6.09	6.9%
Town	0.05	0.02	0.00	6.23	7.4%
County seat	0.09	0.04	0.03	6.31	10.0%
Budapest	0.14	0.05	-0.09	6.41	12.0%
Partner					
No	-0.04	-0.07	-0.02	6.19	12.5%
Yes	0.09	0.08	-0.01	6.26	6.1%

p<0.001

Annex 4: Effects of the synthetic variables on the seriousness of migration aspirations (population aged 16-64)

	1 NND Euclidean	2 NND Mahalanobis	3 NND Manhattan	5 Random distance Euclidean	6 Random distance Mahalanobis	7 Random distance Manhattan (exact)
Male [a]	1.088**	1.069**	1.104**	1.001	1.077**	1.110**
Age [a]	0.998**	0.996**	0.998**	0.996**	0.999**	0.996**
Educational attainment (Ref: University, college, etc. with degree) [a]						
Primary education or lower	1.304**	1.016	1.299**	1.381**	1.200**	1.199**
Vocational education	1.431**	1.183**	1.360**	1.458**	1.341**	1.188**
Secondary education	1.246**	1.088**	1.273**	1.330**	1.301**	1.165**
Type of settlement (Ref: Budapest) [a]						
Small town	1.296**	1.243**	1.456**	1.493**	1.415**	1.321**
Town	1.026	1.018	1.065**	1.108**	1.069**	0.930**
County seat	0.967*	0.907**	1.073**	0.993	0.939**	0.949**
Single (with no partner)	1.053**	1.131**	1.101**	1.019	1.124**	1.072**
No underage children [a]	1.339**	1.299**	1.409**	1.418**	1.354**	1.356**
Occupation (Ref: Not employed) [a]						
Management, professionals, office	0.870**	0.777**	0.836**	0.890**	0.887**	0.969
Commercial and services occupations	1.183**	1.224**	1.127**	1.146**	1.130**	1.313**
Industry, construction industry, agriculture	1.195**	1.205**	1.122**	1.198**	1.167**	1.385**
Not home owner [a]	1.054**	1.106**	1.095**	1.122**	1.159**	1.082**
Knowledge of foreign languages [a]	1.452**	1.388**	1.443**	1.425**	1.344**	1.419**
Migration experience (own) [a]	6.245**	6.305**	5.761**	6.104**	6.251**	5.765**
Migration experience (household member) [a]	2.009**	1.818**	1.587**	1.667**	1.924**	1.912**
Migration aspiration: even forever [b]	1.989**	2.052**	2.054**	2.045**	2.146**	1.992**
Consumption (Ref: Low consumption) [c]						
High consumption	1.072**	0.924**	0.980	1.051**	1.105**	1.067**
Average consumption	1.031*	0.885**	0.928**	1.009	1.018	0.979
Satisfaction [c]	0.976**	0.936**	1.032**	1.037**	0.948**	1.022**
Mental / emotional states [c]	0.977**	0.992	0.965**	0.922**	0.993	0.960**
Trust in institutions [c]	0.979**	0.994	0.970**	0.972**	1.016**	1.007
Optimism about the future [c]	0.993**	1.006*	0.986**	1.013**	1.029**	0.996
Constant	*0.032***	*0.044***	*0.033***	*0.029***	*0.024***	*0.035***
Nagelkerke R Square	*0.198*	*0.204*	*0.189*	*0,194*	*0.201*	*0.187*

** p<0.01 *p<0.05

[a] common variables

[b] variables from the International migration dataset

[c] variables form the Subjective well-being dataset (donor variables)

KS-TC-19-003-EN-N

Creating a synthetic database for research in migration and subjective well-being

While recent advances in migration studies call for targeted research on migration aspirations and the cognitive mechanisms of voluntariness when making choices among alternatives of behaviour, detailed information on (potential) migrants' subjective characteristics, including cognitive evaluations of life and life conditions is still scarce in official statistics. With the aim of establishing good practices for statistical matching estimation methods for official migration statistics in Hungary on the one hand, and encouraging migration researchers to create and use synthetic data to get insights into the subjective characteristics of potential migrants on the other, this paper reports on the methodological details of a statistical matching experiment that combines the International Migration and the Subjective Well-being complementary questionnaires of Hungarian Microcensus 2016. In the experiment, eight variants of non-parametric hot deck methods were performed to create synthetic data sets and evaluated both from a methodological and an analytical point of view.

For more information
https://ec.europa.eu/eurostat/

Publications Office
of the European Union

ISBN 978-92-76-08618-5