SOSYOLOJİ ARAŞTIRMALARI DERGİSİ
JOURNAL OF SOCIOLOGICAL RESEARCH

Geliş Tarihi / Submited: 31.01.2022
Kabul Tarihi / Accepted: 19.04.2022

Cilt / Volume 25 Sayı / Issue 2 (Nisan/2022-April/2022) Araştırma Makalesi / Research Article

ISSN/e-ISSN: 2148-9947

EFFECT OF COMPLEX SAMPLE DESIGN ON DETERMINING COMMON VARIABLES IN STATISTICAL MATCHING METHOD FOR SOCIAL RESEARCH[1]

Cengiz Özkan[2] & Ahmet Sinan Türkyılmaz[3]

Abstract

It is of great importance for researchers to find out different ways of accessing microdata, due to the ever-increasing demand for data and the expectation of reducing the response burden and costs at the same time. In this sense, statistical matching methods have been used extensively to produce new data using existing microdata of surveys and registers recently. It has an increasing application area in social studies such as poverty, deprivation, the effects of newborn on the economic situation of the household, indebtedness and demography, due to the gradual improvement of the micro estimation levels. Selection of matching variables among common variables, at this point, is a critical step in terms of the quality of the microdata to be reached. In the study, while selecting the common variables in order to estimate consumption expenditures by using Statistics on Income and Living Conditions (2018) and Household Budget Survey (2018), weights were added to Hellinger Distance and Spearman2 applications as a new approach. In addition, the effects of design variables (stratum and cluster) were also included in the processes, taking into account the complex structure of both samples. Adding household level weights and design variables to the statistical processes changed the selected or unselected common variables dramatically.

Keywords: Statistical Matching, Data Fusion, Common Variables, Turkey.

[1] This article is based on a part of the PhD thesis entitled "Evaluation of Statistical Matching Methods" preparing by Cengiz Özkan, at Hacettepe University, Institute of Population Studies, Department of Social Research Methodology, Ankara, Turkey

[2] PhD Student, Hacettepe University, Institute of Population Studies, Department of Social Research Methodology, cengizozkan@tuik.gov.tr, ORCID: 0000-0001-7427-0431

[3] Professor, Hacettepe University, Institute of Population Studies, Department of Social Research Methodology, aturkyil@hacettepe.edu.tr ORCID: 0000-0002-2783-932X

SOSYAL ARAŞTIRMALAR İÇİN İSTATİSTİKSEL EŞLEŞTİRME YÖNTEMİNDE ORTAK DEĞİŞKENLERİN SEÇİMİNE KARMAŞIK ÖRNEKLEM TASARIMININ ETKİSİ

ÖZ

Sürekli artan veri talebi ile birlikte, cevaplayıcı yükünün ve maliyetlerin düşürülmesi gerekliliğinin aynı anda tezahür etmesi nedeniyle, araştırmacılar için mikro veriye ulaşmanın farklı yollarının bulunması giderek daha büyük önem arz etmektedir. Bu anlamda, istatistiksel eşleştirme yöntemi, mevcut çalışmaları kullanarak yeni verilerin üretilmesi için son dönemlerde yoğun olarak kullanılmaktadır. Mikro seviyede yapılan tahmin düzeylerinin giderek iyileştirilmesi nedeniyle, yoksulluk, yoksunluk, doğumun hanenin ekonomik durumuna etkileri, borçluluk ve demografi gibi sosyal araştırmalarda da artan bir uygulama alanına sahip olmaktadır. Bu noktada, ortak değişkenlerin arasından eşleşme değişkenlerinin seçimi süreci, ulaşılacak mikro verinin kalitesi açısından kritik bir aşamadır. Çalışmada Gelir ve Yaşam Koşulları Araştırması (2018) ile Hanehalkı Bütçe Anketi (2018) verileri kullanılarak tüketim harcaması tahmini yapılması amacıyla ortak değişkenlerin seçimi yapılırken, yeni bir yaklaşım olarak Hellinger Distance ve Spearman2 uygulamalarına ağırlıklar eklenmiştir. Ayrıca araştırmaların karmaşık yapıları dikkate alınarak tasarım değişkenleri olan tabaka ve küme bilgilerinin etkileri de süreçlere dâhil edilmiştir. Tasarım değişkenlerinin ve ağırlıkların istatistiksel süreçlere dâhil edilmesi seçilen ve seçilmeyen ortak değişkenler açısından önemli değişikliklere neden olmuştur.

Anahtar Kelimeler: İstatistiksel Eşleştirme, Veri Birleştirme, Ortak Değişkenler, Türkiye.

INTRODUCTION

Merging data sets coming from different surveys or administrative data in order to get a new variable which is not available at the same time in both data sets explains the general frame of data matching procedure. There are some procedures including harmonization of microdata, identifying variables and merging records corresponding to the same units (households, customers, patients, products, revenues etc.) from two or more databases. The method enables the researcher to exploit or to reach more variables from the available data sets. Designing a new survey, pre-test procedures of surveys, training of interviewers, data collection period and analysis of microdata take long time and cost high. Instead of these long and costly surveys, producing demanded variables from completed surveys or registers using data matching methods is more rational and time saving.

Because getting variables from available data sets has many advantages, new sub-methods and solutions have emerged with the increasing request especially in the past decades (De Waal, 2015). As a result of this rapid development, data matching procedures were diversified as data fusion, statistical matching, record linkage etc. Even so statistical matching could be categorized under the headings parametric approach, nonparametric approach and mixed method. Record linkage could also be categorized under the headings object identifier matching, unweighted matching of object characteristics, weighted[4] matching of object characteristics and probabilistic record linkage.

Micro matching method's essential issue is to fuse variables using matching variables which are almost same and available in existing data sets. They are generally called as X and selected from both data sets according to their similarity in terms of reference period, definition of units, classification etc. Besides, Y and Z variables are unique, and they are available only in one of the data sets respectively. The main purpose, most particularly in non-parametric micro matching method, is to procure a complete set of data including X, Y and Z at micro level. Contingency table or a regression coefficient may be the outputs of these processes especially at macro or mixed matching level.

Record linkage, in other words object matching is a new research field same as statistical matching and the aim is to identify the records in data sets representing the same entity.

Each method has various and complex implementing procedures, nevertheless micro matching or statistical matching and the elimination processes of common variables are our focus. To glance at surveys for this reason, it has seen that Household Budget Survey has two sub-modules. Individual module consists of 66 questions and the household module consists of 130 questions. Income and Living Conditions Questionnaire, in a like manner, has three sub-modules. Individual module consists of 66 questions, the individual record module consists of 10 and the household module consists of 65 questions. Eliminating and selecting variables from hundreds of data requires a series of statistical operations. The ultimate goal is to obtain a synthetic[5] micro file which consist of variables X, Y and Z jointly. This file is used for further social and economic researches such as poverty, deprivation etc. Quality of the micro file depends on elimination procedures. Final variables remained after this elimination period, are named as matching variables and used to get synthetic file. As too much matching variables cause many statistical problems such as misleading findings, matching noise[6] etc., number of common variables should be reduced to three or four variables. Synthetic data set of X, Y,

[4] Weighting is a method using assigned values for each unit in the datasets according to their significance or reliability.
[5] Synthetic refers to micro files gained with imputation methods.
[6] Matching noise refers to differences between the observed values and imputed values.

Z gained with three or four matching variables has more accurate values in the sense of convergence so as to use for further social or economic researches.

The core and objective of the study is based on three research questions:
- "Which factors are effective on the selection period of the matching variables among common variables in statistical matching?",
- "How complex sample design effect the selection of the matching variables among common variables?"
- "Do the variables chosen by traditional methods differ from those chosen with design variables?"

1. LITERATURE AND THEORETICAL FRAMEWORK

1.1. Literature

Data matching methods, both statistical matching and record linkage, do not back long. Initial academic struggles in data fusion area to use it for social researches dated back to 1972. Okner merged basically 1967 Survey of Economic Opportunity and 1966 Tax File in order to produce income distribution with regard to demographic characteristics. In spite of the ease with which one could get an estimation of total personal income of United States currently, there were not any register or official statistics on the size distribution of such income or any cross-classifications of personal income by typical demographic characteristics of the population. The new micro analytic implementation was performed so as to generate a set of comprehensive household income dataset to use for social research.

Kum and Masterson (2008) proved that statistical matching method could be used for medical researches. 2001 Survey of Consumer Finances containing many elements of wealth at the household level and Annual Demographic Survey of Current Population Survey data sets used to match. They aimed to get a measure of economic wellbeing with high representation.

D'Orazio et al. (2006) have summarized the classifications of these approaches as macro and micro; and parametric, nonparametric and mixed methods. D'Orazio carried out many statistical matching implementations in his publications (2001, 2011, 2013, 2015, 2017) mostly in the field of household surveys. Many packages including fusing and hot deck R codes especially in the statmatch[7] had written by D'Orazio and his publications contain comprehensive examples about the methods. Social surveys of European Union were matched to compare many social and economic indicators by country.

Zacharias (2014), in his study, named "Time Deficits and Poverty" used TURKSTAT microdata of Household Budget Survey (HBS) and Time Use Survey (TUS) for social research. Time spent on household production for each individual aged 15 years and older in TUS was transferred into HBS data. Poverty measures calculated by national offices generally do not contain time deficits. They assume that all households and individuals have time sufficiently to join to the needs of household members and underestimate both the scope and the depth of poverty. Their models consider intrahousehold disparities in time allocation unlike neoclassical model.

[7] Statmatch is an add-on package for R environment including functions to implement statistical methods.

Ahi (2015), in his master thesis, matched two surveys (SILC, HBS) to estimate variables on the basis of Classification of Individual Consumption According to Purpose's (COICOP) 12 main expenditure groups for households. The share of the main expenditure groups such as health, education, transportation and food has been estimated to analyze current social and economic situation of the households in Turkey.

Uçar (2017), analyzed the effect of a new-born on household poverty. Consumption expenditure transferred from Household Budget Survey to a longitudinal[8] survey (SILC) using non-parametric micro matching. Longitudinal statistical matching caused many complications with regard to reference period, weights, calibration, population in and out by years and deflation rates about revenues. In spite of everything, micro fusion method could be used for a demography thesis so as to find out relationship between poverty and fertility. While nonparametric micro matching method was used to generate synthetic data, Rensens' calibration method was used for complex sample design and Rassler method was used for validation. Economic indicators were used along with fuzzy measures of poverty and deprivation index in a comparative way.

Kim (2018) searched how in a best way to facilitate a small overlap of units in a data fusion situation if data consists of categorical variables. Combined estimator which is a combination of conditional independence assumption and direct estimators was developed in his paper as a new approach from small area estimation. Netherland Population Census data (2011) was divided into 3 parts randomly to get new data sets. Occupation and education level variables was used only in one sample. In other words, donor and recipient sample had only one variable respectively. 36 different experiments were carried out altering sample size of auxiliary data C, number of matching variables and total sample size of A and B. Expectation maximization algorithm estimator gave better results than combined estimator. The main aim was to get occupation and education information at micro level.

Öztürk (2019), aimed to evaluate non-parametric statistical matching methods (random, rank and nearest neighbor distance hot deck methods) in her master thesis. 2014-2015 Time Use Survey of Turkey and 2014 Life Satisfaction Survey of Turkey were used. Household level weights were used in logistic regressions. Constrained nearest neighbor distance approach and rank hot deck approach expected to provide more accurate result but implementations showed the opposite. Random hot deck especially 'min' option and nearest neighbor hot deck provided better results. In the dissertation, relationship between social indicators such as going to the cinema and theater, watching TV, using social media etc. and demographic indicators was investigated.

All researchers mentioned above, aimed mainly to generate a micro file in order to use it for following social and economic research effectively looking for the best data matching method. Since there are no study evaluating elimination procedures of statistical matching method in literature, studies closest to the subject are summarized instead of preliminary procedures of the approach.

[8] Longitudinal (or panel) survey is a research design involving repeated observations of the same variables of households over determined periods of time (annually, quarterly, monthly). The survey is repeated annually for four years to the selected households. Survey selected for any year in these four years is called cross-sectional.

1.2. Theoretical Framework

Both quantitative and qualitative social science research preferred to collect data needed from small sized surveys. They also favored large scaled field researches when there is a necessary situation. Field researches which have large scaled sample size include detailed information but could be performed only in long periods such as population census, household researches etc. Small sized surveys, on the other hand, can be more flexible in terms of timeliness but have not got comprehensive information. It is also possible to encounter representativeness issues. In addition to mentioned drawbacks, everlasting information demand which is more comprehensive and detailed, in very short periods and at high quality level compelled researchers and national statistical offices to find new and alternative methods.

Registers (administrative data) were the first source to produce data from available information even if they are not designed for statistical purposes initially. Surveys carried out for other purposes were also considered to be used for data matching methods. These existing data sources enabled to produce broader and new outputs by use of data fusion and record linkage methods. To summarize, better quality data, faster publication periods, lower costs for national statistical institutes and reduced response burden are fundamental contributions of data matching methods. These are also main objectives of national statistical offices.

Liking theory and notion of social distance is important in the sense of the source of data. The concept of liking theory is mainly about interaction between interviewer and respondent. According to liking theory, respondents would like to interact interviewers who have similar characteristics (Vercruyssen et al. 2017). Not only socio-demographic characteristics but also attitudes, religiousness and background could also improve liking among individuals (Byrne, 1971). Social distance, on the other hand, implies the differences between individuals in terms of social class, ethnicity, age and gender (Katz, 1942). When the social distance is considered within surveys, interviewers and respondents can differ in terms of age, gender, social class, and educational levels. Therefore, according to liking theory and social distance concept, similarity or dissimilarity between interviewers and respondents may have considerable effects on building rapport for interviews (Saraç, 2021). Obtaining the needed data through questionnaires instead of statistical matching or similar methods, causes various measurement errors and bias[9]. The relationship between the interviewer and the respondent can also create bias.

Theoretical frame of statistical matching is substantially based on combining experiments and combining sample studies. (Cochran 1937), aimed to combine separate sources so as to research in the field of crop yields using ANOVA[10] methods and much later than the experiments, methodological studies emerged for combining sample surveys. There were three main differences between combining experiments (CX) and combining sample (CS). CS procedures need too much attention during preparation and coordination phases. It is a great deal of starting with a good planning especially for multinational surveys contrary to national multidomain surveys which have a coordination naturally. The second difference of these applications that make up the theory of the SM method is that while CX concentrates on experiments, CS brings surveys into focus especially on the probability sampling and simple random selection of subjects. Final point of separation is about statistical analysis period. Contrary to CX, comprehensive analysis of survey method including joint

[9] Bias refers to inclination or prejudice for or against one person or group.
[10] ANOVA is an analysis tool used in statistics and means analysis of variance.

analysis, similarity and comparability is used intensely. Based on these studies, Leslie Kish, in 1999, described the notion as "*theory of combining populations*" including different types of cumulation of rolling samples' data "sample reported at regular intervals for time periods that overlap with preceding time periods" (Kish, 1990). Alexander (2001) also suggested that combining data from different countries or unions had its fundamental problems to experiment.

The process of gaining a definite ground for the theoretical infrastructure has reached a certain stage with the study of all the sub-headings of the subject in the course of time. Especially in the first studies, the idea of using the existing data more quickly and effectively came to the fore. In fact, it is based on the idea of saving time, reducing employee costs and survey expenses for the benefit of the public. However, with the methodological improvements made as a result of the statistical analyzes on the data quality, it has been fully established in a theoretical framework. Since misleading findings of exact statistical matching were evaluated by analogy, improvements in the validity procedures allowed it to sit on a more solid ground theoretically. Today, studies in this field are entirely aimed at improving the methodology of the statistical matching in general. In addition to the holistic perspective, the theoretical infrastructure of each sub-method is developed up to the distinction between social and economic studies.

2. METHODOLOGY

2.1. Data Sources

Finding convenient data source to get and use for the matching process is main difficulty of the research. Generally, several social survey results are accessible to use, and matching studies are largely done using two social surveys. Registers are both complicated to use and difficult to access. Therefore, two household surveys intended to use "Household Budget Survey (2018)" and "Statistics on Income and Living Conditions (2018)" and both micro data of surveys obtained from Turkish Statistical Office (TURKSTAT).

Regulations of TURKSTAT imposes strict rules about micro data demand, confidentiality of data, ethical issues and usage. Micro data is classified as A and B group. A group data can be utilized only in the institution with the assigned computer. Time deficit to analyze and match micro data sets is a problematic issue for this type of data such as population and housing studies. B group data is suitable for external use. Therefore, SILC and HBS data are preferred. Confidentiality of data is guaranteed by contract. Micro data sets cannot be shared by no means and statistical estimations cannot be done at regional basis. As stratum and cluster information enable us to produce regional based estimations, this information is provided after long negotiations with only alias codes instead of real variables.

2.1.1. Household Budget Survey (HBS)

Household Budget Survey is collected to produce information about consumption expenditure and income. Geographic coverage of the survey is all Turkey. Stratified two-staged cluster sampling method is used. Diaries are given to household members (14+ years old) so as to record individual consumption expenditures daily. Household Budget Survey consists of 8 tables and 3 separate sub-data sets as microdata level. These are individual data set, household data set and consumption expenditure data set. Individual data set consists of 66 questions and the household data set consists of 130 questions. Consumption expenditure data set consists of 4 subtitles. Classification of

consumption expenditure is based on COICOP (Classification of Individual Consumption by Purpose). Data set has an identifier named “unitno” enabling to link subsets of data (TURKSTAT, 2018a).

2.1.2. Statistics on Income and Living Conditions (SILC)

Income and Living Condition Survey is a longitudinal research but it is possible to use it as a cross-sectional survey collecting for many economic and social purposes. Determining distribution of income in the country, number of poor people and regional distribution of them, personal income transitions, material deprivation, general living conditions of people are the main goals of the survey to answer. Economic activities are recorded by 18 subtitles according to NACE Rev.2 economic activity classification. Geographic coverage of the survey is all Turkey. Stratified two-stage clustered sampling approach is used and final sampling unit is household. Face to face computer assisted personal interview and administrative registers for data editing and missing information were both used.

Micro data of Statistics on Income and Living Conditions questionnaire consists of 9 separate tables and 3 separate sub-data sets as micro data level. These sub-data sets are individual data set which has information about only 15+ years old of household members, individual register data set including information about all household members and household data set. The individual data set consists of 66 questions, the individual record data set consists of 10 questions and the household data set consists of 65 questions. Data sets are connected with the help of 2 identifier variables named as “fertid” and “bülten” (TURKSTAT, 2018b).

2.2. Dataset Preparation and Common Variables

The data preparation process is the first level involving intensive sequences of implementations. Definitions of variables, contents of them, reference periods of surveys are checked and response categories including different answers are synchronized (Uçar, 2016). Selection of common variables (X), and selection of unique variables Y and Z is carried out at this phase named as harmonization period.

Harmonization period consists of bringing into line the definition of statistical units, harmonization of reference period of surveys or registers, controlling of coverage of population for both surveys, controlling of classification of economic activity, adjusting for missing data and measurement errors and derived variables which have to be created (Laan, 2000). This period is performed to harmonize and compliance two data sets in order to use them for further processes.

Reference Person is defined differently in the two surveys. While household budget survey definition is referenced the member receiving the highest income in the household, Income and Living Conditions Survey definition is based on age and management and decision role in the household. Due to two different content of reference person, reassignment the reference person for the Income and Living Conditions Survey is done. In this sense, the reference person was reassigned with the SAS Enterprise program based on the column (FG140) containing the total income item in the data set. The data, which is 82.16% compatible before reassignment, has been made fully compatible after the process. As eight variables of common variables (X) are connected to the reference person, this reassignment process has enabled the matching quality to be increased.

Household Size is not available in the Income and Living Conditions Survey on the contrary to Household Budget Survey. Therefore, the household size variable is generated for Income and Living Conditions Survey making use of the individual register data set.

Harmonization of Classifications contains response categories of variables coded not in the same way. Response categories were created for reference person's age group and reference person's number of weekly working hours. The answers to the marital status question, which has different response categories, were harmonized. The answers to the education question were divided into subcategories. The answer of the reference person's economic activity of work and heating system of the dwelling question have been harmonized. Differences of response categories for ownership of mobile, computer, internet, washing machine, refrigerator, dishwasher, air conditioner and car were classified in a harmonized way.

Derivation of Variables is also very important issue in order to create and use for further processes. Demographic variables that are important and necessary to be used in data matching procedures were created in both data set. These are mainly about number of elderly, women, adult, children and employed persons in the household.

Harmonization of Household Income was a problematic issue. Although the sub-items of the income variable are the same in both surveys, income variable in SILC refers to the preceding year. Having tried many different ways to solve the problem, TURKSTAT CPI (Consumer Price Index) was used to bring into compliance income variables.

Choice of Donor and Recipient depends mostly on sample size of the surveys but it may alter according to target of study (D'Orazio, 2017). Surveys with smaller sample size is generally recipient and larger sample sized survey is donor. This approach prevents us from syntax errors occurring in hot deck procedures in R Studio. Nevertheless, Income and Living Conditions Survey is assigned as the recipient and Household Budget Survey is assigned as the donor data set. This phase is compulsive to reduce common variables in the Hellinger distance, spearman and regression applications.

Choice of Target Variables which means Y and Z variables, is also necessary for further stages. Y is income variable in the Income and Living Conditions Survey (recipient) and Z is household consumption expenditure in the Household Budget Survey (donor). These variables should be assigned elaborately to use in the statistical applications.

At the end of the harmonization period, common variables (X) and matching variables have to be determined according to multicollinearity. There are still 39 common variables and it is too much to match data sets effectively.

Table 1. List of the selected common variables and abbreviations

HSIZE	Household Size
NUM_CHI	Number of children (0-17) in the household
NUM_ADU	Number of adults (18-64) in the household
NUM_ELD	Number of elderly (65+) in the household
NUM_WOM	Number of women in the household
ALL_ADU	All household members are adults
ALL_ELD	All household members are elderly
ALL_WOM	All household members are women
NUM_EMP	Number of employed people
NUM_EMP_INC	Number of individuals with employee income

NUM_SELF_EMP_INC	Number of individuals with self-employed income
NUM_RET_INC	Number of individuals with retired income
REF_SEX	Reference person's sex
REF_AGE	Reference person's age group
REF_MAR	Reference person's marital status
REF_EDU	Reference person's education
REF_PRO	Reference person's professional status
REF_OCC	Reference person's occupation
REF_ECO	Reference person's economic activity of work
REF_WHRS	Reference person's number of weekly working hours
DWE	Dwelling type
TENURE	Tenure status
RENT_CAT	Current rent related to occupied dwelling
ROOM_NUM	Number of rooms
TOT_AR	Total space available to the household (m2)
HEAT_SYS	Heating system of the dwelling
BATH	Bath or shower in dwelling
TOILET	Indoor flushing toilet for sole use of household
PIPED_WAT	Piped water
HOT_WAT	Hot water
MOBILE	Mobile
COMP	Computer
INTERNET	Internet
WASH_M	Washing machine
REFRIG	Refrigerator
DISH_W	Dishwasher
AIR_CON	Air conditioner
CAR	Car
DIS_INC_CAT	Total disposable household income

2.3. Statistical Methods

Regression analysis is extensively used to reduce the number of selected common variables as sole method. As we have household weights for both surveys, “HB40 for Income and Living Conditions Survey” and “FACTOR for Household Budget Survey”, these variables are utilized in Hellinger Distance, spearman2 and regression analysis as a new technique to observe and evaluate the effect on the elimination period. Design variables are also benefitted as a new approach to investigate how complex sample designs effect the selection period.

2.3.1. Hellinger Distance

Hellinger Distance is a mathematical formulation developed by Ernst Hellinger in 1909 and takes final values between 0 and 1 representing similarity of variables. Probabilities of the response categories is fundamental for the formula. While zero indicates exact similarity, one indicates no similarity between the same variables of donor and recipient sample. Because Hellinger Distance

method is easy to calculate similarity and does not need information about sample design, it is very useful and common.

Formula 1. Hellinger Distance Formula

$$HD\,(\mathrm{D},\mathrm{R}) = \sqrt{\frac{1}{2}\sum_{i=1}^{K}\left(\sqrt{\frac{n_{Di}}{N_D}} - \sqrt{\frac{n_{Ri}}{N_R}}\right)^2}$$

D: Donor (Household Budget Survey)
R: Recipient (Income and Living Conditions Survey)
K: Total number of the cells
nDi: The frequency of response categories in Household Budget Survey
nRi: The frequency of response categories in Income and Living Conditions Survey
N: Total size of the contingency table.
In the academic literature for calculation results of Hellinger Distance method, variables having 5 percentages and above is not accepted for further analysis because there is no similarity between them. Therefore, variables excessing that cutoff value are considered incompatible for ongoing periods.

2.3.2. Spearman2[11]

Even if the Hellinger Distance method eliminates several common variables, there are generally still too many variables and having so much variables might lead to undesirable noise effecting synthetic data sets of statistical matching. Additional approach to select matching variables from remained common variables is spearman2 method which computes squares of Spearman's rho rank according to type of variables. Hmisc package in R studio was installed for further analysis processes.

Table 2. Types of variables in SILC and HBS

VARIABLES	TYPE OF DATA
X Common Var.	Categoric
Y Household Income Var. (SILC)	Continuous
Z Consumption Expenditure Var. (HBS)	Continuous

Spearman2 applied for both data sets separately so as to get two tables including adjusted rho2 values for each variable.

spearman2(Y~var1+var2+…, data= a)

spearman2(Z~ var1+var2+…, data=b)

Spearman2 procedure, used for second elimination method to reduce unnecessary variables and find out variables which have more explanatory power, is calculated using unweighted data invariably as Hellinger Distance is. This situation may cause that some variables left or out. So weighted calculation was used to avoid from that problem. Package wCorr and function weightedCorr were used in R

[11] Spearman method is a rank correlation and introduced by Charles Spearman in 1904.

program. The function could be used only with numeric categories so response categories of ref_whrs, tot_ar and dis_inc_cat were recategorized accordingly using numeric instead of ranks.

weightedCorr(x=data$var1,
y=data$var2, method = c("Spearman"), weights = data$weight)

2.3.3. Regression Analysis

Linear regression or logistic regression is performed prevalently according to type of dependent variables (categoric or continuous). As weights are generally ignored in regression analysis made for statistical matching, similar to the Hellinger Distance and spearman2 calculation period, in this study they were used as a new method and both weighted and unweighted regressions were run after dummy variables[12] created.

Finding matching variables processes from common variables normally ends up at this phase, selected variables approved after regression analysis can be easily used in non-parametric or parametric matching processes. However, effect of the design variables on the selection period will be investigated.

Effect of design variables, since it is thought to affect the common variable selection decision, is important for this study. Thus, in this stage, cluster and stratum information was included in the analysis process along with remained variables in order to observe the effect of design variables.

3. RESULTS

3.1. Results of Hellinger Distance Calculations

The first analysis results of Hellinger Distance pointed out that nine common variables have values out of range as seen in the figure below. If we do not insert household weights in the HD calculations, these nine variables will not use for following processes.

As mentioned in the methodology section, variables with a score under 5 percentages accepted as convenient for the following phases. First unweighted results in the figure 1 indicate that nine variables excessing that cutoff value are considered incompatible for ongoing periods. In another word, they do not have any similarity between them.

The same procedures are repeated with weights. When household weights named as "HB040" and "FAKTOR" included in calculation of response categories' percentages (nDi and nRi), mean value was decreased from 3.2 to 2.4 and four of the nine variables became reusable for following processes. Figure 2 exhibits the weighted results.

[12] Dummy variables refer to variables taking only the value 0 or 1 to indicate the absence or presence of some categorical effect.

Figure 1. Hellinger distance results of the common variables (unweighted)

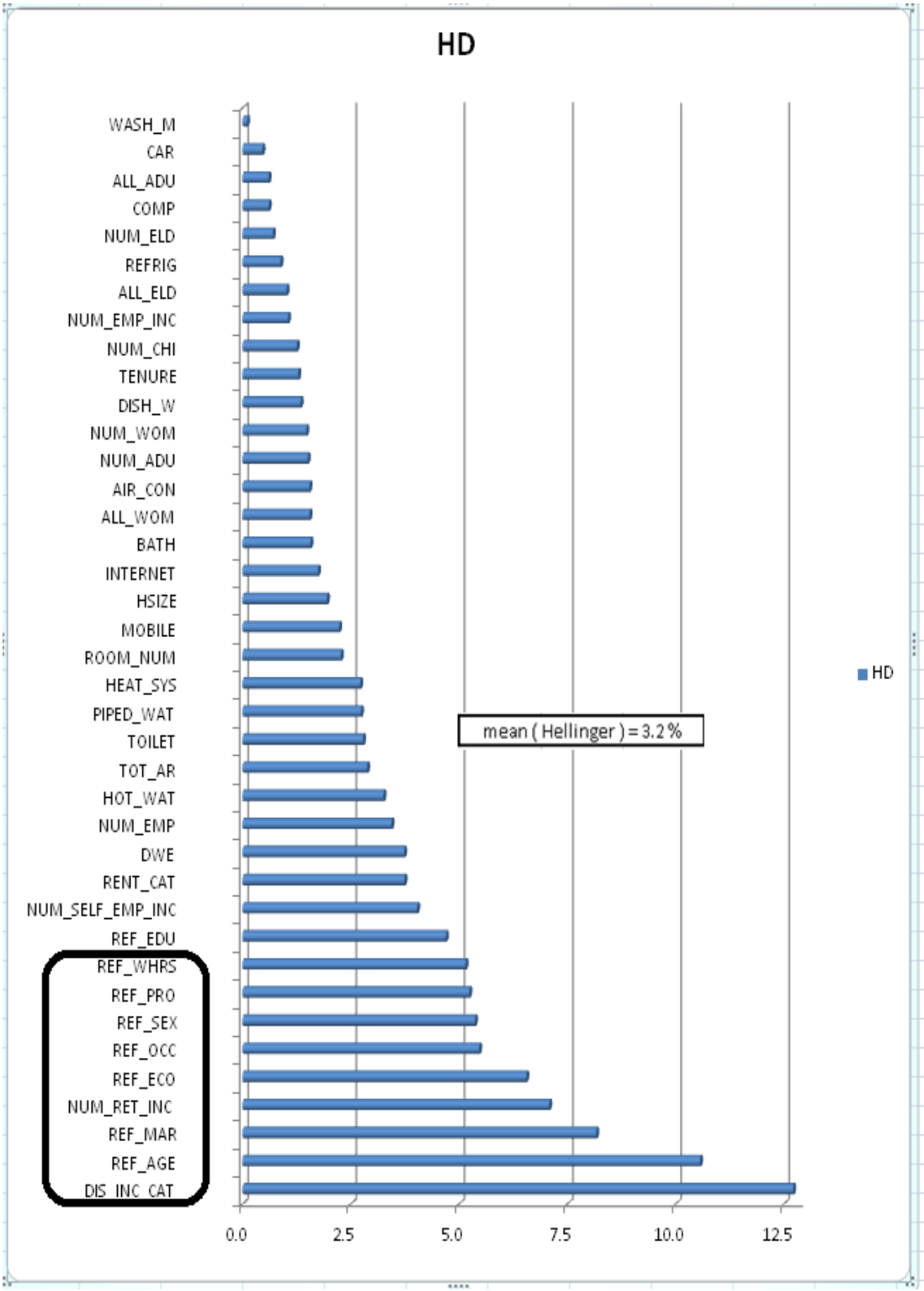
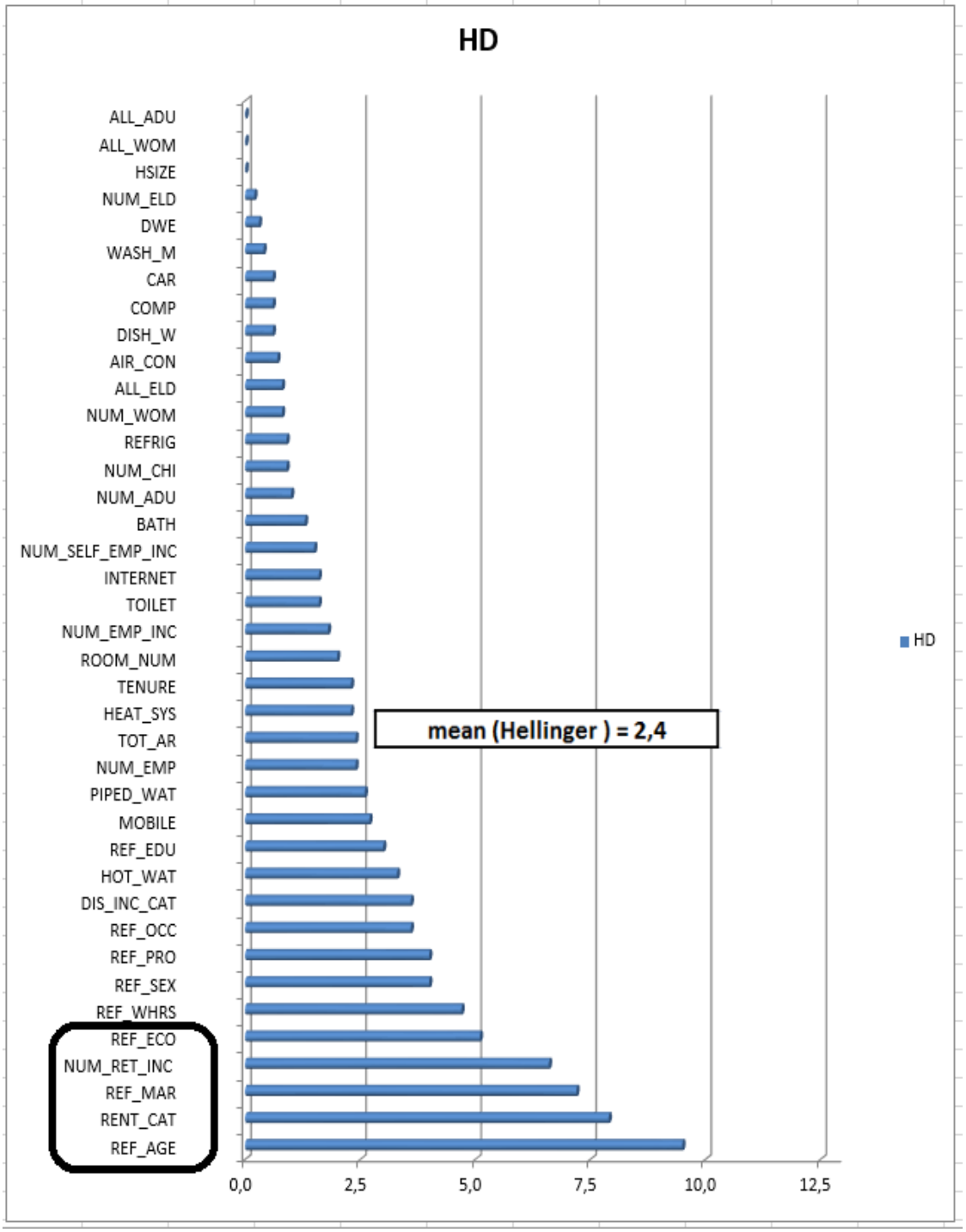


Figure 2. Hellinger distance results of the common variables (weighted)



Reference person's number of weekly working hours, reference person's sex, reference person's occupation and reference person's professional status are proper to use owing to the new approach in the statistical matching. Table 3 shows weighted and unweighted scores of four variables.

Table 3. Weighted and unweighted scores of the 4 variables

NAME OF VAR.	WEIGHTED HD SCORE	UNWEIGHTED HD SCORE
REF WHRS	4,7	5,2
REF SEX	4,0	5,4
REF PRO	4,0	5,2
REF OCC	3,6	5,5

3.2. Results of Spearman2 Calculations

Unweighted calculation of adjusted rho2 values are represented in the table below. As variables scored over ten percent indicate strong explanatory power, eleven variables scored over ten percent in both data set could be used for further stages. These are disposable income categories, reference person's education, number of employed people, heating system of the dwelling, number of individuals with employee income, internet, number of adults (18-64) in the household, dwelling type and ownership of computer, dishwasher and car. Excluding disposable income categories for both surveys which have highest scores, reference person's education status has highest value for Income and Living Conditions Survey. On the other side, ownership of car has highest value for Household Budget Survey.

Table 4. Adjusted rho2 values (unweighted)

Spearman rho^2 Response variable:YINCOME

	rho2	F	df1	df2	P	Adjusted rho2	n
REF_WHRS	0.127	876.51	4	24063	0	0.127	24068
REF_SEX	0.018	433.54	1	24066	0	0.018	24068
REF_PRO	0.057	1443.22	1	24066	0	0.057	24068
REF_OCC	0.010	239.42	1	24066	0	0.010	24068
DIS_INC_CAT	0.951	93031.86	5	24062	0	0.951	24068
HOT_WAT	0.083	2173.54	1	24066	0	0.083	24068
REF_EDU	0.260	8466.15	1	24066	0	0.260	24068
MOBILE	0.042	1063.32	1	24066	0	0.042	24068
PIPED_WAT	0.007	178.46	1	24066	0	0.007	24068
NUM_EMP	0.189	5595.07	1	24066	0	0.189	24068
TOT_AR	0.131	908.45	4	24063	0	0.131	24068
HEAT_SYS	0.166	4787.77	1	24066	0	0.166	24068
TENURE	0.013	326.58	1	24066	0	0.013	24068
ROOM_NUM	0.115	3142.46	1	24066	0	0.115	24068
NUM_EMP_INC	0.153	4347.93	1	24066	0	0.153	24068
TOILET	0.040	992.87	1	24066	0	0.040	24068
INTERNET	0.200	6023.60	1	24066	0	0.200	24068
NUM_SELF_EMP_INC	0.001	18.75	1	24066	0	0.001	24068
BATH	0.021	518.38	1	24066	0	0.021	24068
NUM_ADU	0.130	3582.57	1	24066	0	0.130	24068
NUM_CHI	0.003	65.69	1	24066	0	0.003	24068
REFRIG	0.015	377.61	1	24066	0	0.015	24068
NUM_WOM	0.018	432.46	1	24066	0	0.018	24068
ALL_ELD	0.075	1949.86	1	24066	0	0.075	24068
AIR_CON	0.030	744.43	1	24066	0	0.030	24068
DISH_W	0.193	5744.91	1	24066	0	0.193	24068
COMP	0.219	6756.05	1	24066	0	0.219	24068
CAR	0.176	5130.40	1	24066	0	0.176	24068
WASH_M	0.034	855.80	1	24066	0	0.034	24068
DWE	0.125	3453.41	1	24066	0	0.125	24068
NUM_ELD	0.026	639.91	1	24066	0	0.026	24068
HSIZE	0.052	1330.35	1	24066	0	0.052	24068
ALL_WOM	0.067	1740.08	1	24066	0	0.067	24068
ALL_ADU	0.004	87.41	1	24066	0	0.004	24068

Spearman rho^2 Response variable:ZCONSUMPTION

	rho2	F	df1	df2	P	Adjusted rho2	n
REF_WHRS	0.062	194.15	4	11823	0.0000	0.061	11828
REF_SEX	0.036	438.52	1	11826	0.0000	0.036	11828
REF_PRO	0.020	242.92	1	11826	0.0000	0.020	11828
REF_OCC	0.004	48.22	1	11826	0.0000	0.004	11828
DIS_INC_CAT	0.483	2212.81	5	11822	0.0000	0.483	11828
HOT_WAT	0.051	629.19	1	11826	0.0000	0.050	11828
REF_EDU	0.153	2134.97	1	11826	0.0000	0.153	11828
MOBILE	0.028	336.41	1	11826	0.0000	0.028	11828
PIPED_WAT	0.003	40.53	1	11826	0.0000	0.003	11828
NUM_EMP	0.116	1547.11	1	11826	0.0000	0.116	11828
TOT_AR	0.093	301.99	4	11823	0.0000	0.092	11828
HEAT_SYS	0.144	1987.70	1	11826	0.0000	0.144	11828
TENURE	0.001	11.59	1	11826	0.0007	0.001	11828
ROOM_NUM	0.076	968.72	1	11826	0.0000	0.076	11828
NUM_EMP_INC	0.105	1386.09	1	11826	0.0000	0.105	11828
TOILET	0.049	607.29	1	11826	0.0000	0.049	11828
INTERNET	0.195	2864.84	1	11826	0.0000	0.195	11828
NUM_SELF_EMP_INC	0.000	0.00	1	11826	0.9815	0.000	11828
BATH	0.015	178.42	1	11826	0.0000	0.015	11828
NUM_ADU	0.117	1562.42	1	11826	0.0000	0.117	11828
NUM_CHI	0.012	138.50	1	11826	0.0000	0.011	11828
REFRIG	0.010	118.41	1	11826	0.0000	0.010	11828
NUM_WOM	0.024	286.37	1	11826	0.0000	0.024	11828
ALL_ELD	0.075	956.62	1	11826	0.0000	0.075	11828
AIR_CON	0.039	485.29	1	11826	0.0000	0.039	11828
DISH_W	0.154	2148.06	1	11826	0.0000	0.154	11828
COMP	0.181	2612.60	1	11826	0.0000	0.181	11828
CAR	0.203	3005.93	1	11826	0.0000	0.203	11828
WASH_M	0.022	259.87	1	11826	0.0000	0.021	11828
DWE	0.117	1562.45	1	11826	0.0000	0.117	11828
NUM_ELD	0.039	483.93	1	11826	0.0000	0.039	11828
HSIZE	0.060	761.20	1	11826	0.0000	0.060	11828
ALL_WOM	0.049	610.44	1	11826	0.0000	0.049	11828
ALL_ADU	0.000	4.09	1	11826	0.0432	0.000	11828

Weighted calculation of adjusted rho2 values indicates that two variables having proper values for the unweighted spearman2 calculation received values outside of the specified ranges. Therefore, heating system and dwelling type variables did not use for the further analysis.

Table 5. Adjusted rho2 values (weighted)

VARIABLES	**SILC**	**HBS**	
REF_WHRS	0,07553	0,04815	NA
REF_SEX	0,01725	0,0455	NA
REF_PRO	0,0647	0,0284	NA
REF_OCC	0,00903	0,00349	NA
DIS_INC_CAT	**0,94095**	**0,48399**	**
HOT_WAT	0,07044	0,04297	NA
REF_EDU	0,23539	0,12932	**
MOBİLE	0,04649	0,02755	NA
PIPED_WAT	0,0059	0,00331	NA
NUM_EMP	**0,21147**	**0,13859**	**
TOT_AR	0,12319	0,08644	NA
HEAT_SYS	0,1372	0,09354	NA
TENURE	0,02024	0,00353	NA
ROOM_NUM	0,112	0,07147	NA
NUM_EMP_INC	**0,16431**	**0,10428**	**
TOILET	0,03523	0,03987	NA
INTERNET	**0,20128**	**0,17621**	**
NUM_SELF_EMP_INC	0,00151	0,00056	NA
BATH	0,01746	0,01302	NA
NUM_ADU	**0,15268**	**0,13425**	**
NUM_CHI	0,00452	0,01343	NA
REFRIG	0,01526	0,00689	NA
NUM_WOM	0,02848	0,03397	NA
ALL_ELD	0,08658	0,08186	NA
AIR_CON	0,03236	0,03289	NA
DISH_W	**0,17833**	**0,12872**	**
COMP	**0,21597**	**0,16444**	**
CAR	**0,16582**	**0,19915**	**
WASH_M	0,03467	0,01831	NA
DWE	0,10973	0,08951	NA
NUM_ELD	0,02686	0,03209	NA
HSIZE	0,07038	0,07547	NA
ALL_WOM	0,07655	0,06244	NA
ALL_ADU	0,00226	0,00008	NA

Only nine variables can be used for further analysis. Disposable income categories, reference person's education, number of employed people, number of individuals with employee income, internet, number of adults (18-64) in the household and ownership of computer, dishwasher and car. Adding weighting procedure to calculation of Spearman2 leads to change the matching variables to be used in the following periods.

3.3. Regression Results

Results of both the Hellinger Distance and spearman2 show that household level weights could significantly change the elimination period. In the additional third step to reduce matching variables to a reasonable number so as to avoid errors caused by introducing too much matching variables into the statistical matching processes, household level weights are used too. Table 6 indicates that which variables get appropriate values in which analysis. Ownership of computer, car, dish washer and disposable income categories are matching variables according to regression results.

Table 6. Regression results (weighted and unweighted)

REGRESSION	**LINEAR**				**LOG LINEAR**				**FREQ**
SAMPLE	**SILC**	**HBS**	**SILC**	**HBS**	**SILC**	**HBS**	**SILC**	**HBS**	
VARIABLES	**WEIGHTED**		**UNWEIGHTED**		**WEIGHTED**		**UNWEIGHTED**		
NUM_EMP	X	X			X	X	X		5 / 8
DWE	X		X			X		X	4 / 8
COMP	X	X	X	X	X	X	X	X	**8 / 8**
DISH_W	X	X		X	X	X	X	X	**7 / 8**
CAR	X	X	X	X	X	X	X	X	**8 / 8**
DIS_INC_CAT	X	X	X	X	X	X	X	X	**8 / 8**
INTERNET		X		X		X		X	4 / 8
NUM_ADU					X	X		X	3 / 8
REF_EDU						X		X	2 / 8

When design variables are attached the regression analysis, different results are observed. Unlike traditional analyses, 2 variables (number of adults and ownership of internet) that were not included in the previous regressions were found as significant for Household Budget Survey. Four final variables (ownership of computer, ownership of dish washer, ownership of car and disposable income categories) found proper to match similar to the previous results.

The same analysis was carried out for SILC. When complex sample design took into account, 6 variables obtained sufficient results for matching. Four of them are the same variables found by traditional methods but number of adults and number of employed people are the variables found as a result of consideration of complex sample design. Table 7 shows regression results with design variables.

Table 7. Regression results (with design variables)

VARIABLES	**SILC_DV**	**HBS_DV**
NUM_ADU	**x**	**x**
NUM_EMP	**x**	
NUM_EMP_INC		
REF_EDU		
COMP	**x**	**x**
INTERNET		**x**

DISH_W	x	x
CAR	x	x
DIS_INC_CAT	x	x

Studies in the literature are limited to certain patterns for Hellinger Distance, spearman2 and regressions. Here, an additional contribution has been made in terms of adding weights and design variables to each of the elimination methods. Adding weights in Hellinger Distance and spearman2 and regressions with design variables are innovations of the study. Although it is not the subject of the article, it has been observed that the validation of statistical matching results made with the matching variables obtained as a result of including the design variables, provide accurate information at micro level.

4. CONCLUSION AND DISCUSSIONS

Linear regression analysis results, calculated Hellinger Distance and spearman2 percentages indicate that weights and design variables have significant effects on the choosing phase of the statistical matching method separately. Variables excluding for next stage calculations due to their analysis scores (>%5 for HD and <%10 for spearman2) could be utilized after these factors included the calculations as a new method. This situation means that studies on this field may ignore some matching variables for not using design variables in procedures. Advantage of using design variables in the elimination processes is to generate more accurate estimations. On the other hand, shortcoming of this approach, design variables are very difficult to obtain.

Evaluating the processes in terms of weights, four fundamental variables about reference person related to demographic and labor force indicators could be added owing to weighted and recalculated percentages of response categories. While these four vital indicators included and reused, on the contrary, two variables excluded due to weighted recalculation of spearman2 method. Common variables having not representativeness to be matching variables deducted from the list and variables with high correlation used for regression analysis.

Regression analysis with traditional approaches firmly showed that four variables were final regressors to be used for matching phases. When complex sample design considered, “number of adults” variable found out as common variables for SILC and HBS. Besides, number of employed people and ownership of internet variables became useable variables for SILC and HBS respectively.

Although statistical matching method offers a very wide usage opportunity, it is still not used widely enough. It can be used in sociological researches such as immigration, economic and social studies on immigrants, where it is difficult to reach sufficient and comprehensive data. Different registers or surveys of Immigration Department, Ministry of Interior, Address Based Population Registration System etc. can be exploited to find out current sociological situation of immigrants in Turkey. Sociological and economic solution proposals can be implemented more accurately and quickly by considering the results of this research. It will be also beneficial for researchers who want to work in this field to consider design variables in terms of data quality.

ÖZET

Sosyal araştırma yöntemlerinde, özellikle hanehalkı çalışmalarında son dönemlerde yoğun bir kullanım alanına ulaşan veri eşleştirme çalışmaları zamanla yeni istatistiksel uygulamaları da bünyesine dâhil etmektedir. Bu çalışmada gelir ve yaşam koşulları araştırması 2018 yılı verileri ile hanehalkı bütçe anketi 2018 yılı verileri kullanılarak gelir ve yaşam koşulları mikro veri setinde mevcut olmayan hanehalkı tüketim harcaması değişkeninin bütçe anketinden istatistiksel eşleştirme yöntemiyle aktarılması sağlanmıştır. Eşleştirme kalitesini belirleyen en önemli etken olan ortak değişkenlerin seçimi süreci klasik yöntemlerle yapılmış olup bu yöntemlere ilaveten ağırlık ve tasarım değişkenleri de sürece ilk kez dâhil edilmiştir.

Ortak değişken seçiminden sonraki süreçlerde parametrik ya da parametrik olmayan yöntemlerin uygulanmasında genel bir uygulama silsilesi mevcut olduğundan süreçlere yeterli bir şekilde müdahale yapılması çok fazla mümkün olamamaktadır. Ancak hâlihazırdaki tüm değişkenlerin elenip ortak değişkenlerin tespit edilmesinden sonraki eşleştirme değişkenlerinin seçimi ise yeniliklere açık olan bir alandır. Tabakalı, iki aşamalı küme örneklemesi ile hanelerin seçiminin yapıldığı iki anket çalışmasında da tasarım değişkenleri ve hane ağırlık bilgileri ilk kez dikkate alınarak değişken seçim süreçlerine olan etkisi veri kalitesinin artırılması yönünde değerlendirilmiştir.

Veri setleri değerlendirilip aralarında korelasyon olan değişkenler belirlendikten sonra kalan 39 değişken için ilk olarak Hellinger Distance yöntemine göre hesaplama yapılmış olup 9 değişken temsiliyet yeteneği yeterli olmadığından kapsam dışına alınmıştır. Ancak her iki ankete ait ağırlıklar SPSS programı aracılığıyla kullanılarak, cevap kategorilerinin oranları yeniden hesaplanmıştır. Bu hesaplama sonucunda ilk değerlendirmelere göre kullanılmaması gereken 4 değişken yüzde 5 eşik değerinin altına inmesi nedeniyle sonraki süreçler için kullanılabilir hale gelmiştir. Yenilenmiş ve hane ağırlıkları dâhil edilmiş Hellinger Distance hesabı sonucu oluşan oranlar ile referans kişinin haftalık çalışma saati, referans kişinin cinsiyeti, referans kişinin çalışma durumu ve referans kişinin çalışma bilgisi gibi önemli demografik ve ekonomik faaliyet değişkenlerinin izleyen süreçlerde kullanıma uygun olduğu tespit edilmiştir.

Ortak değişkenlerin kategorik bir veri tipine sahip olduğu, hedef değişkenler olan Y ve Z değişkenlerinin ise sürekli (continuous) bir veri tipi yapısına sahip olduğu durumlarda eleme süreçlerinde kullanılabilecek bir yöntem olan spearman2 metodu da ilk olarak geleneksel bir şekilde yani hiçbir ağırlık bilgisi formüle eklenmeden hesaplanmıştır. Burada sadece 11 değişkenin istatistiksel eşleştirme süreçleri için uygun olduğu, kalan 23 değişkenin ise kullanılamayacağı sonucu ortaya çıkmıştır. R Studio programı ile wCorr paketi bünyesindeki weightedCorr fonksiyonu kullanılarak spearman2 hesabına hane ağırlıkları dâhil edilmiştir. İlk defa kullanılan bu yöntem ile yapılan yeni hesaplamalarda sadece 9 değişkenin referans değer olan yüzde 10 ve üzeri seviyelerde değer aldığı görülmüştür. İlk hesaplamaların aksine, hanede kullanılan ısıtma sitemi şekli ile oturulan evin hangi tip olduğu ile ilgili olan 2 temel değişkenin bu yeni yaklaşım sayesinde eşleştirme değişkeni olarak kullanılamayacağı tespit edilmiştir. Dolayısıyla ağırlık bilgisinin bu aşamada da önemli bir etkiye sahip olduğu görülmektedir.

Hedef değişkenlerinin tipine göre uygulanacak doğrusal ya da lojistik regresyon analizi, değişken seçiminde tek başına veya bu çalışmada uygulandığı üzere birkaç aşamadan sonra nihai seçim amacıyla kullanılabilen bir metot olarak karşımıza çıkmaktadır. Burada öncelikli olarak her iki anket verisi için doğrusal regresyon analizi ağırlıklı ve ağırlıksız olarak uygulanmıştır. Daha sonra hedef değişkenler için log alınarak ağırlıklı ve ağırlıksız olmak üzere SPSS ve SAS Enterprise üzerinden regresyonlar gerçekleştirilmiştir. Sonuçlar incelendiğinde 8 farklı uygulamada 4 değişkenin nihai değişken olarak kullanılabileceği anlaşılmıştır. Bunlar hanede bilgisayara sahip olma durumu, hanede

bulaşık makinesine sahip olma durumu, hanede araç sahibi olma durumu ile harcanabilir gelir kategorileri değişkenleri olarak ön plana çıkmaktadır.

Temsil yeteneği ve korelasyon katsayısı yüksek olan bu 4 değişken ile istatistiksel eşleştirme süreçlerine devam edilebilmesi mümkün olmakla birlikte, bu çalışmada tasarım değişkenleri olan tabaka ve küme (blok) bilgilerinin regresyon analizi sürecine dahil edilerek olası etkileri gözlemlenmek istenmiştir. Bölgesel tahmin yapmaya imkân verebileceği için gelir ve yaşam koşulları için küme bilgileri; hanehalkı bütçe anketi için ise tabaka ve küme bilgileri sanal kodlar ile Türkiye İstatistik Kurumu' ndan temin edilmiştir.

Hanehalkı Bütçe Anketi verilerine tasarım değişkenleri eklenerek yapılan analizler sonucunda ağırlıklı ve ağırlıksız olarak yapılan regresyon analizlerinden farklı sonuçlara ulaşılmıştır. Hanedeki 15-64 yaş arası birey sayısı ile internet sahipliği değişkenlerinin, bu hesaplamalarda yüksek temsiliyete sahip olduğu için, eşleştirme değişkenleri olarak kullanılabilme imkânı doğmuştur. Daha önceki analizlerde nihai eşleşme değişkeni olarak seçilen dört değişkenin bu hesaplamada da uygun oldukları tekrar test edilmiştir.

Gelir ve Yaşam Koşulları Araştırması verilerine tasarım değişkenleri eklenerek yapılan analizler sonucunda da ağırlıklı ve ağırlıksız olarak yapılan regresyon analizlerinden farklı sonuçlara ulaşılmıştır. Daha önceki hesaplamalarda farklı sonuçlar veren hanedeki yetişkin sayısı değişkeni, tasarım değişkenleri dâhil edildiğinde ortak değişken olma kriterlerini karşılamıştır. Ayrıca GYK için çalışan sayısı ve HBA için de internet sahipliği değişkenleri olumlu sonuçlar vermiştir. Geleneksel yöntemlerle ulaşılan dört değişkene ise bu yöntemlerle de ulaşılmıştır.

Dünyada ve Türkiye' de son yıllarda devam eden göçmen ve sığınmacılarla ilgili sosyolojik ve sosyo-ekonomik durumun tespitine yönelik araştırmalar için istatistiksel eşleştirme yöntemi yeni bir yaklaşım sunabilme kapasitesine sahiptir. İdari kayıtlar ve çeşitli amaçlarla derlenen anket verileri birleştirilerek alt kırılımlarda veri üretimi mümkündür. Proje destekli ve uzun süreli araştırmalarla elde edilen verilere, bu yöntemle daha az maliyetle ve daha hızlı bir şekilde ulaşılabilir. Göçmen ve sığınmacılarla ilgili araştırma yapanların, istatistiksel eşleştirme yöntemini kullanırken makalede bahsedilen şekilde anket verilerine ait tasarım değişkenlerini de dikkate alması, ulaşacakları bulguların kalitesi açısından da son derece faydalı olacaktır.

REFERENCES

Ahi, L. (2015). Veri Madenciliği Yöntemleri İle Ana Harcama Gruplarının Paylarının Tahmini., Hacettepe Üniversitesi, Yüksek Lisans Tezi.

Alexander, C. H. (2001), Stıll Rollıng: Leslie Kish's "Rolling Samples" and The American Community Survey.

Balin, M., D'ORAZIO, M., Di Zio, M., Scanu, M., & Torelli, N. (2009). Statistical Matching of Two Surveys with a Common Subset (No. 124). Working Paper.

Byrne, D. (1971). *The Attraction Paradigm.* New York: Academic Press.

Cochran W.G. (1937). Problems Arising in the Analysis of a Series of Similar Experiments. Supplement to the Journal of the Royal Statistical Society, 4, 102-118.

De Waal, T. (2015). Statistical matching: experimental results and future research questions. Statistics Netherlands.

D'orazio, M., Di Zio, M., & Scanu, M. (2001, June). Statistical Matching: a tool for integrating data in National Statistical Institutes. In Proc. of the Joint ETK and NTTS Conference for Official Statistics.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), Statistical Matching: Theory and Practice. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

D'Orazio, M. (2017). Statistical Matching and Imputation of Survey Data with StatMatch.

Katz, D. (1942). Do Interviewers Bias Poll Results? *Public Opinion Quaretrly*, 6(2), 248-268.

Kim, D. (2018) "Development of a statistical matching method with categorical data"

Kish, L. (1990), "Rolling Samples and Censuses", Survey Methodology, 16, 63-79.

Kum and Masterson (2008), Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Well-Being, The Levy Economics Institute of Bard College, Working Paper No:535.

Laan, P. van der. 2000. 'Integrating Administrative Registers and Household Surveys'. Netherlands Official Statistics, Vol. 15 (Summer 2000): Special Issue, Integrating Administrative Registers and Household Surveys, ed. P.G. Al and B.F.M. Bakker, pp. 7-15.

Okner, B. (1972), "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File", Annals of Economic and Social Measurement 1, pp. 325-342.

Öztürk, C. (2019), Nonparametric Statistical Matching Methods: An Application On Household Surveys in Turkey, master thesis, University of Hacettepe, Turkey.

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. New York: Springer. Rasner, A., J. R. Frick, and M. M. Grabka. 2011. Extending the Empirical Basis for Wealth Inequality Research Using Statistical Matching of Administrative and Survey Data. SOEP papers 359. Berlin: DIW.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation", Survey Methodology, 24, 171-183.

Saraç M. (2021). The Contribution of Rapport Between Interviewer and Respondent On Interview Quality from Non-Sampling Error Perspective: *Evidence from 2014 Research On Domestic Violence Against Women in Turkey.*

Turkstat, (2018a), Handbook for Household Budget Survey, Ankara.

Turkstat, (2018b), Handbook for Statistics on Income and Living Conditions Survey, Ankara.

Uçar, Baris, and Gianni Betti. (2016). "Longitudinal Statistical Matching: Transferring Consumption Expenditure from HBS to SILC Panel Survey." Papers of the Department, No. 739. Siena: Department of Economics, University of Siena. Available at: http://econpapers.repec.org/paper/usiwpaper/739.htm

Uçar, B. (2017), The Effect of a New Born on Household Poverty in Turkey: The Current Situation and Future Prospects by Simulations, PHD thesis, University of Hacettepe, Turkey.

Vercruyssen, A. Wuyts, C. & Loosveldt, G. (2017). The Effect of Sociodemographic (Mis)match between Interviewer and Respondents on Unit and Item Nonresponse in Belgium. *Social Science Research*, 67, 229-238.

Zacharias, A., Masterson, T., Kim, K. (2014), "The Measurement of Time and Income Poverty in Korea". Economics Working Paper Archive, Levy Economics Institute, http://www.levyinstitute.org/pubs/rpr_8_14.pd