HACETTEPE UNIVERSITY

INSTITUTE OF POPULATION STUDIES

AN EVALUATION OF STATISTICAL MATCHING METHODS: AN APPLICATION ON TURKEY INCOME AND LIVING CONDITIONS SURVEY AND HOUSEHOLD BUDGET SURVEY

Cengiz ÖZKAN

Department of Social Research Methodology

PhD Thesis

Ankara

July 2022

HACETTEPE UNIVERSITY

INSTITUTE OF POPULATION STUDIES

AN EVALUATION OF STATISTICAL MATCHING METHODS: AN APPLICATION ON TURKEY INCOME AND LIVING CONDITIONS SURVEY AND HOUSEHOLD BUDGET SURVEY

Cengiz ÖZKAN

Supervisor: Prof. Dr. Ahmet Sinan TÜRKYILMAZ

Department of Social Research Methodology

PhD Thesis

Ankara

July 2022

APPROVAL PAGE

An Evaluation of Statistical Matching Methods: An Application on Turkey Income and Living Conditions Survey and Household Budget Survey

Cengiz Özkan

This is to certify that we have read and examined this thesis and in our opinion it fulfils the requirements in scope and quality of a thesis for the degree of Doctor of Philosophy in Social Research Methodology.

Jury Members:

Member (Chair): …………………………………………..
Prof. Dr. Yaprak Arzu ÖZDEMİR
Gazi University, Faculty of Sciences, Department of Statistics

Member (Supervisor): …………………………………….....
Prof. Dr. A. Sinan TÜRKYILMAZ
Hacettepe University, Institute of Population Studies, Department of Social Research Methodology

Member: …………………………………………………....
Assoc. Prof. Berna Burçak Başbuğ ERKAN
Middle East Technical University, Faculty of Arts and Sciences, Department of Statistics

Member: …………………………………………………....
Assoc. Prof. Dr. Alanur Çavlin BİRCAN
Hacettepe University, Institute of Population Studies, Department of Demography

Member: …………………………………………………....
Assoc. Prof. Dr. İlknur YÜKSEL-KAPTANOĞLU
Hacettepe University, Institute of Population Studies, Department of Social Research Methodology

Member: …………………………………………………....
Asst. Prof. Dr. Tuğba ADALI
Hacettepe University, Institute of Population Studies, Department of Social Research Methodology

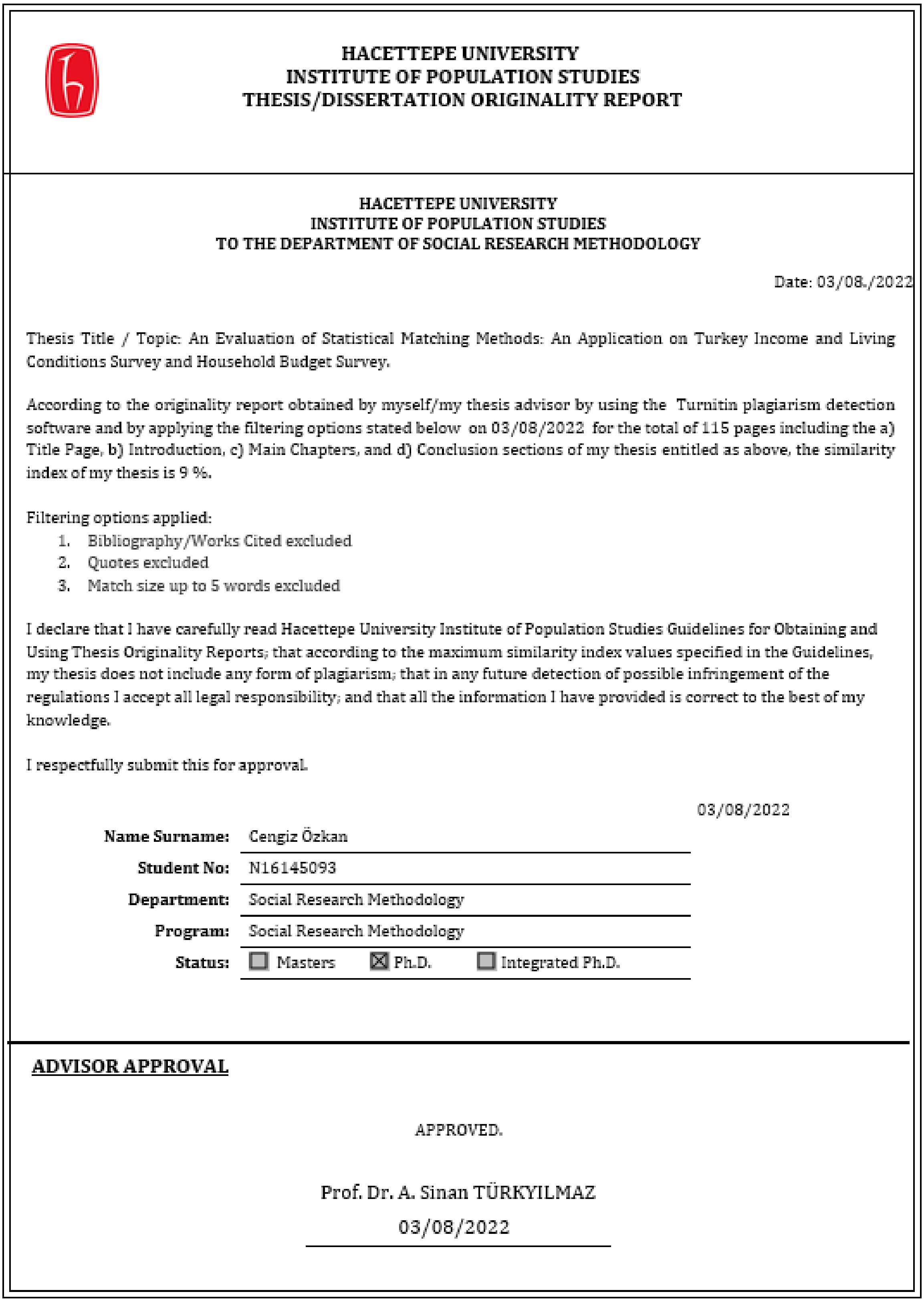
This thesis has been accepted by the above-signed members of the Jury and has been confirmed by the Administrative Board of the Institute of Population Studies, Hacettepe University.

………………......……………………………

…/…/ 2022

Prof. Dr. İsmet KOÇ
Director

HACETTEPE UNIVERSITY INSTITUTE OF POPULATION STUDIES THESIS ORIGINALITY REPORT



HACETTEPE UNIVERSITY
INSTITUTE OF POPULATION STUDIES
THESIS/DISSERTATION ORIGINALITY REPORT

HACETTEPE UNIVERSITY
INSTITUTE OF POPULATION STUDIES
TO THE DEPARTMENT OF SOCIAL RESEARCH METHODOLOGY

Date: 03/08./2022

Thesis Title / Topic: An Evaluation of Statistical Matching Methods: An Application on Turkey Income and Living Conditions Survey and Household Budget Survey.

According to the originality report obtained by myself/my thesis advisor by using the Turnitin plagiarism detection software and by applying the filtering options stated below on 03/08/2022 for the total of 115 pages including the a) Title Page, b) Introduction, c) Main Chapters, and d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is 9 %.

Filtering options applied:

1. Bibliography/Works Cited excluded
2. Quotes excluded
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Institute of Population Studies Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

03/08/2022

Name Surname: Cengiz Özkan

Student No: N16145093

Department: Social Research Methodology

Program: Social Research Methodology

Status: □ Masters ☒ Ph.D. □ Integrated Ph.D.

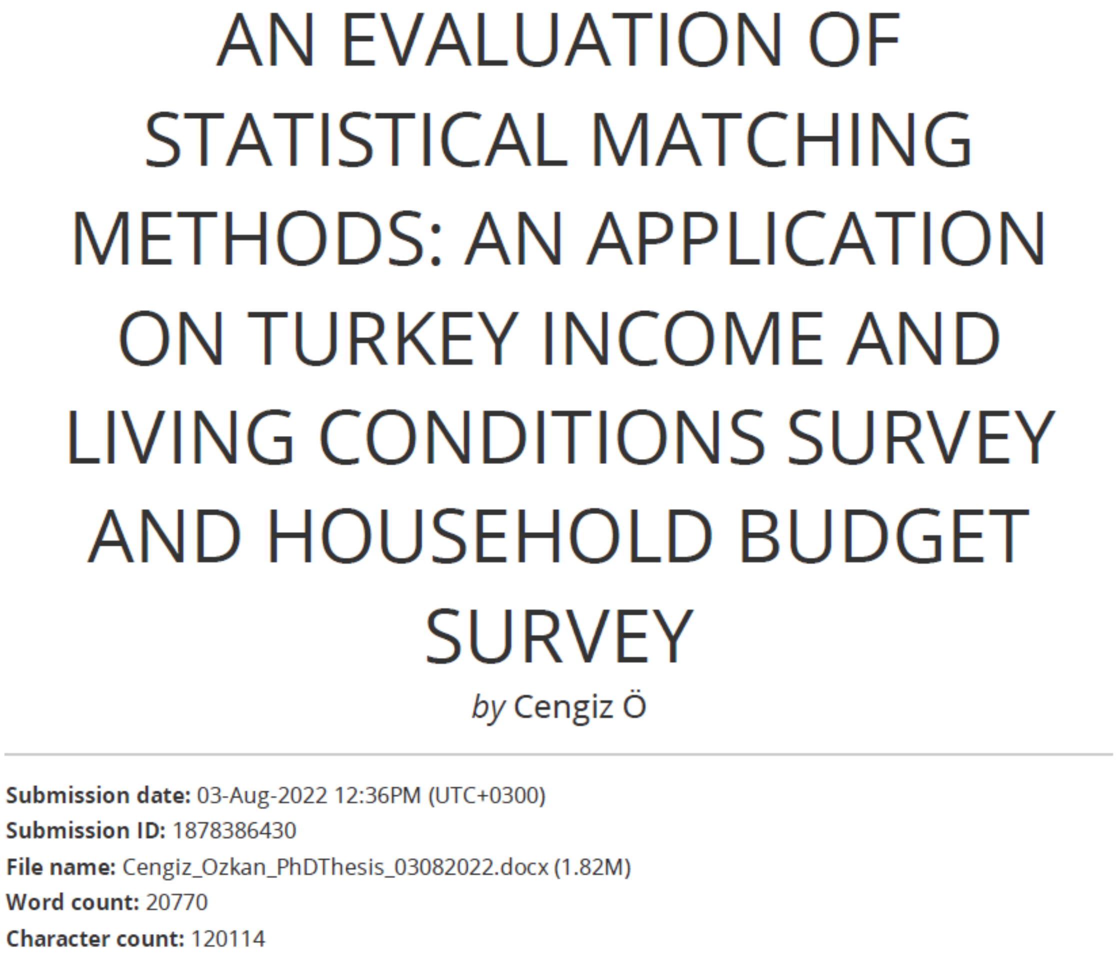
ADVISOR APPROVAL

APPROVED.

Prof. Dr. A. Sinan TÜRKYILMAZ

03/08/2022

SIMILARITY INDEX PAGE FROM TURNITIN PROGRAM



AN EVALUATION OF STATISTICAL MATCHING METHODS: AN APPLICATION ON TURKEY INCOME AND LIVING CONDITIONS SURVEY AND HOUSEHOLD BUDGET SURVEY

by Cengiz Ö

Submission date: 03-Aug-2022 12:36PM (UTC+0300)
Submission ID: 1878386430
File name: Cengiz_Ozkan_PhDThesis_03082022.docx (1.82M)
Word count: 20770
Character count: 120114

AN EVALUATION OF STATISTICAL MATCHING METHODS: AN APPLICATION ON TURKEY INCOME AND LIVING CONDITIONS SURVEY AND HOUSEHOLD BUDGET SURVEY

ORIGINALITY REPORT

9%	8%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.openaccess.hacettepe.edu.tr:8080 Internet Source	3%
2	acikbilim.yok.gov.tr Internet Source	1%
3	Submitted to University of Surrey Student Paper	1%
4	Eva Endres, Paul Fink, Thomas Augustin. "Imprecise Imputation: A Nonparametric Micro Approach Reflecting the Natural Uncertainty of Statistical Matching with Categorical Data", Journal of Official Statistics, 2019 Publication	<1%
5	Francesco D. d’Ovidio, Paola Perchinunno, Laura Antonucci. "Data Integration Techniques for the Identification of Poverty Profiles", Social Indicators Research, 2020 Publication	<1%

ETHICAL DECLARATION

In this thesis study, I declare that all the information and documents have been obtained on the basis of the academic rules and all audio-visual and written information and results have been presented according to the rules of scientific ethics. I did not do any distortion in data set. In case of using other works, related studies have been fully cited in accordance with scientific standards. I also declare that my thesis study is original except the cited references. It was produced by myself in consultation with my supervisor (Prof. Dr. A. Sinan TÜRKYILMAZ) and written according to the rules of thesis writing of Hacettepe University Institute of Population Studies.

…………………….
Cengiz ÖZKAN

DECLARATION OF PUBLISHING AND INTELLECTUAL PROPERTY RIGHTS

I declare that I give permission to Hacettepe University to archive all or some part of my master/PhD thesis, which is approved by the Institute, in printed (paper) or electronic format and to open to access with the following rules. With this permission, I hold all intellectual property rights, except using rights given to the University, and the rights of use of all or some parts of my thesis in the future studies (article, book, license, and patent).

I declare that the thesis is my original work, I did not violate rights of others and I own all rights of my thesis. I declare that I used texts with the written permit which is taken by owners and I will give copies of these to the University, if needed.

As per the "Regulation on the Online Availability, Arrangement and Open Access of Graduate Theses" of Council of Higher Education, my thesis shall be deposited to National Theses Center of the Council of Higher Education/Open Access System of H.U. libraries, except for the conditions indicated below;

o The access to my thesis has been postponed for 2 years after my graduation as per the decision of the Institute/University board.(1)
o The access to my thesis has been postponed for …. month(s) after my graduation as per the decision of the Institute/University board.(2)
o There is a confidentiality order for my thesis.(3)

03/08/2022
……………………
Cengiz ÖZKAN

----

i Regulation on the Online Availability, Arrangement and Open Access of Graduate Theses

(1) Article 6.1. In the event of patent application or ongoing patent application, the Institute or the University Board may decide to postpone the open access of the thesis for two years, upon the proposal of the advisor and the assent of the Institute Department.

(2) Article 6.2. For theses that include new techniques, material and methods, that are not yet published articles and are not protected by patent and that can lead to unfair profit of the third parties in the event of being disseminated online, the open access of the theses may be postponed for a period not longer than 6 months, as per the decision of the Institute or the University Board upon the proposal of the advisor and the assent of the Institute Department.

(3) Article 7.1. The confidentiality order regarding the theses that concern national interest or security, the police, intelligence, defense and security, health and similar shall be issued by the institution certified the thesis*. The confidentiality order for theses prepared pursuant to the cooperation protocol with institutions and organizations shall be issued by the University Board, upon the proposal of the related institutions and organizations and the assent of the Institute or the Faculty. The theses with confidentiality order shall be notified to the Council of Higher Education.

Article 7.2. During the confidentiality period, the theses with confidentiality order shall be kept by the Institute or the Faculty in accordance with the confidentiality order requirements, in the event of termination of the confidentiality order the thesis shall be uploaded to Thesis Automation System.

□ Shall be issued by the Institute or Faculty Board upon the proposal of the advisor and the assent of the Institute Department.

ACKNOWLEDGEMENTS

It is a great honor to be a part of Department of Social Research Methodology at Hacettepe University Institute of Population Studies (HUIPS) as a doctoral graduate student. I have learned a lot in this field during the thesis process that will help me in my following studies and business area.

I would like to express my sincere gratitude to my supervisor Prof. Dr. Ahmet Sinan Türkyılmaz who made this thesis possible with his tremendous contributions. During the last period, especially in the corona days and in the busy pace of business life, he motivated and encouraged me every time I despaired about finishing my thesis. I also thank him for guiding me to choose this crucial and popular topic contributing to my future plans.

I am also grateful to thesis monitoring committee member Assoc. Prof. Dr. Alanur Çavlin Bircan for her positive contributions accelerating the processes about the article and making the thesis more comprehensive and understandable. Besides, I would like to thank the other thesis monitoring committee member Asst. Prof. Dr. Tuğba Adalı, for her insightful comments and suggestions to improve the study and helped to fulfillment of the thesis. I would also like to express my gratitude for their carefully studied suggestions for the article. I would like to thank other jury members Prof. Dr. Yaprak Arzu ÖZDEMİR, Assoc. Prof. Dr. Berna Burçak Başbuğ ERKAN and Assoc. Prof. Dr. İlknur Yüksel Kaptanoğlu for their precious feedback and comments during completion process of dissertation and my colleague Cansu Öztürk for her support related to non-parametric codes and survey design variables.

Finally, I would like to thank my valuable teachers Assoc. Prof. Dr. Burkay Genç and Dr. Pelin Çağatay, whose lectures I had the chance to take during my doctoral education.

ABSTRACT

The dissertation aims to evaluate the effectiveness of statistical matching methods from a comparative perspective. Since the studies in the literature mostly focus on non-parametric micro methods, it is aimed to conduct a study that deals with macro, micro, mixed, parametric and non-parametric methods in a holistic and comparative way as well as to observe the effects of different donor classes, and interventions in the sample size. In addition, it is aimed to expand the procedures regarding the selection processes of matching variables by including survey design variables and weights for the first time and to re-evaluate their effectiveness. With the inclusion of options in the processes, it is also aimed to observe the efficiency of matching between methods, to determine the practical limitations and to test the issues that are open to intervention.

Applications were made on the selection of matching variables and statistical matching methods using the 2018 datasets of the Turkey Statistics on Income and Living Conditions and Household Budget Survey, which have complex sample design features. After the survey data were harmonized, parametric, non-parametric and mixed methods were applied at the macro and micro levels, considering the mentioned breakdowns to produce the outputs. Imputation procedure, random hot deck, rank hot deck and nearest neighbor distance hot deck were used in non-parametric micro methods.

Statistical matching methods, which allow the production of high quality, faster, lower cost and timeliness data by using existing data sources such as administrative records and survey data, also have the potential to provide positive contributions in terms of theoretical statistical approaches such as reducing the response burden and interviewer bias. The method is also used for demography studies that aim to find the correlation between poverty and fertility. The results show that weighted and unweighted micro matching applications provide us with highly accurate and reliable estimations. Although the limitations of the mixed methods regarding the size of the observations have been determined, it has been observed that they are effective in producing quality synthetic data. Parametric methods, on the other hand, did not give the expected quality results on data integration.

Key words: Data matching, statistical matching, SILC, HBS

ÖZET

Bu tez, istatistiksel eşleştirme yöntemlerinin etkinliğinin karşılaştırmalı bir bakış açısıyla değerlendirilmesini amaçlamaktadır. Literatürdeki çalışmalar daha çok non-parametrik mikro yöntemler üzerine odaklandığından, makro, mikro, mixed, parametric ve non-parametrik yöntemleri bütüncül ve karşılaştırmalı olarak ele alan bir araştırmanın yapılmasının yanı sıra farklı donor sınıflarının ve örneklem büyüklüğünde yapılacak müdahalelerin etkilerinin gözlemlenmesi amaçlanmıştır. Ayrıca eşleşme değişkenlerinin seçim süreçlerine dair prosedürlerin, tasarım değişkenleri ve ağırlıkların ilk kez dâhil edilerek genişletilmesi ve etkinliklerinin yeniden değerlendirilmesi amaçlanmıştır. Opsiyonların da süreçlere dâhil edilmesi ile yöntemler arası eşleştirmenin etkinliğinin gözlemlenmesi, uygulamaya dönük sınırlılıkların belirlenmesi ve müdahaleye açık konuların test edilmesi hedeflenmiştir.

Karmaşık örneklem tasarımı yapılarına sahip Türkiye Gelir ve Yaşam Koşulları Araştırması ile Hanehalkı Bütçe Araştırması 2018 yılı veri setleri kullanılarak eşleştirme değişkenlerinin seçimi ve istatistiksel eşleştirme yöntemleri üzerine uygulamalar yapılmıştır. Anket verileri uyumlu hale getirildikten sonra, çıktıların üretilmesi için bahsedilen kırılımlar dikkate alınarak makro ve mikro düzeyde parametrik, parametrik olmayan ve karma yöntemler uygulanmıştır. İmputasyon prosedürü, random hot deck, rank hot deck ve nearest neighbor distance hot deck, parametrik olmayan mikro yöntemlerde kullanılmıştır.

İdari kayıtlar ve anket verileri gibi mevcut veri kaynakları kullanılarak yüksek kalitede, hızlı, daha düşük maliyetli ve zamanlılık ilkesine uygun veri üretimine imkân veren istatistiksel eşleştirme yöntemleri aynı zamanda cevaplayıcı yükünün ve anketör yanlılığının azaltılması gibi teorik istatistiki yaklaşımlar açısından da olumlu katkılar sağlayacak potansiyele sahiptir. Yöntem, yoksulluk ve doğurganlık arasındaki ilişkiyi bulmayı amaçlayan demografi çalışmalarında da kullanılmaktadır. Sonuçlar, ağırlıklı ve ağırlıksız mikro eşleştirme uygulamalarının bize son derece doğru ve güvenilir tahminler sağladığını göstermektedir. Karma yöntemlerin gözlem büyüklüğü ile ilgili sınırlılıkları tespit edilmiş olsa da kaliteli sentetik veri üretimi açısından etkin oldukları gözlemlenmiştir. Parametrik yöntemler ise veri entegrasyonu açısından beklenen kalitede sonuçlar vermemiştir.

Anahtar kelimeler: Veri eşleştirme, istatistiksel eşleştirme, GYK, HBA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ABPRS	Address Based Population Registration System
BHATT	The Bhattacharyya Coefficient
CAPI	Computer Assisted Personal Interview
CATI	Computer Assisted Telephone Interview
CE	Combined Estimator
CIA	Conditional Independence Assumption
CMM	Conditional Mean Matching
CPI	Consumer Price Index
DBPD	Draws Based on a Predictive Distribution
DRL	Deterministic Record Linkage
EM	Expectation Maximization
HBS	Household Budget Survey
HD	Hellinger Distance
IRS	Internal Revenue Service
MAR	Missing at random
MCAR	Missing Completely at Random
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MN	Markow Networks
MS	Moriarty and Scheuren
MSE	Mean Square Error
NND	Nearest Neighbor Hot Deck
NUTS	Nomenclature of Territorial Units for Statistics
PHC	Population and Housing Census
PPS	Probability Proportional to Size

PRL	Probabilistic Record Linkage
RL	Record Linkage
RND	Random Hot Deck
SAS	Statistical Analysis Software
SEO	Survey of Economic Opportunity
SILC	Statistics on Income and Living Conditions
SM	Statistical Matching
SPSS	Statistical Package for the Social Sciences
TURKSTAT	Turkish Statistical Office
TVD	Total Variation Distance

CHAPTER 1. INTRODUCTION

Everlasting and growing demand for timely, less costly and high-quality statistics in the field of economic and social life has forced researchers, national officers and decision-makers to apply experimental methods. Use of administrative data, integration of data sets and new statistical methods have been tried result of this demand. The main issue is to create new variables by using existing statistics such as surveys and administrative data. As studies progressed, data matching began to take shape on two main methods: statistical matching (SM) and record linkage (RL). Both SM and RL methodologies are basically built on merging two data sets and exploiting from the available sources to gain variables but they are completely different approaches in terms of application. Completely different from the purpose of statistical matching method to match *similar unit*[1], record linkage, named in literature also as object or record matching aims to match the *same units* represented in different files. The similar unit notion, in practice, is not identical and mostly refers to different units.

The hypostasis in the statistical matching (micro or data fusion[2]) approach is to retile necessary or required information from 2 different micro data sets by means of common variables denoted as X (Van Der Putten et al., 2002). The system is basically constructed on three variables named X, Y, and Z. Common variables, as the name suggests, are available in both data sets at the same time even if they might have different response categories. Donor and recipient data sets can contain too many X variables which are non-representative for the model. Procedure steps to reduce the number of Xs and to dig out proper matching variables will be explained in chapter 4. Y and Z are unique variables and available in only one data set A or B. The situation is not an obligation and observed values are useful for validity checks having completed the matching procedures. Although the accepted view in the literature in the selection of donor and recipient data is that the data with a larger sample size should be donor, small-sized data sets might choose as donor. The situation may cause

[1] Unit refers to an observation or measurement for which data are collected such as people, consumer products, travelers and taxpayers, etc.

[2] *Data fusion* is a set of methodologies for transferring information from independent sources.

complications in hot deck and parametric matching procedures, especially in some phases of correlation matrix.

Briefly, target variables (Y, Z) aren't jointly observed in any sample survey. While Y variable is possible to observe only in sample A, Z variable, on the other hand, is observed only in sample B, Xs are observed in both data sets and A and B samples do not overlap (Waal, 2015). SM estimates the joint distribution function of the target variables and involves synthetic or micro data of three variables in non-parametric and mixed matching, on the other hand, provides correlation matrix, contingency tables and regression coefficient in parametric macro and micro matching.

RL can be classified substantially as deterministic record linkage (DRL) and probabilistic record linkage (PRL) approaches. Object identifier matching, unweighted and weighted matching of object characteristics, Fellegi and Sunter (1969) and Jaro (1972) approach could be viewed in sub-details. Object identifier matching is applied when we have sufficient or good quality same unique key variable in both datasets. Unweighted matching of object characteristics method consists of preparation and application steps. That takes many iterations to obtain a cut-off[3] value for metrics and it is an important step to reach the goal of getting enough candidate matches as we have no identity number. In the weighted matching of object characteristics method, possible candidate matches are accessed using weights. Probabilistic record linkage may also be considered as weighted matching if an explicit use of probabilities is available. PRL has complex procedure steps and makes implementations in different phases aiming for pairs of records in order to classify them as links, possible links, or non-links. To summarize, RL method has various and complex implementation procedures, but as our focus is on statistical matching, the subject of RL is covered theoretically.

SM is a way of faster publishing of new outputs at micro level through reducing survey expenditures and timeliness pressures. Besides the reduction of survey preparation,

[3] Cut-off value is a threshold value associated with a quantity which is a choice of the researcher.

field organization and personnel costs, reducing the response burden of respondents are the other benefits of SM increasing the reputation of the institutions. The flexible structure of the method has even been used for a demography study that aims to find out the correlation between poverty and fertility. Parallel to this increasing interest, the number of programs and packages that carry out SM applications with large-scale data has increased rapidly. By using these programs (SPSS, R Studio and SAS Enterprise), very comprehensive and detailed analyses of household and individual data containing approximately 36,000 records were made. The data come from Income and Living Conditions Survey (2018) and Household Budget Survey (2018) of Turkey. Since survey design variables (cluster, stratum and household level weights) can be provided for both data, this information is also intensely included in the processes of the analyses. In this study, it was aimed to observe the effects of different variations on the results with changes in the number of matching variables, experiments with different variable groups by creating donor classes, and interventions in the sample size. In the fourth chapter, more detailed information about these different applications and their results is given.

In parallel with the fundamental motivation mentioned up to now, the dissertation has three main objectives:

(1) Examining the statistical matching methods both methodologically and practically,

(2) Observing the effect of survey design variables while selecting matching variables among common variables,

(3) To investigate and compare the accuracy of the estimations obtained by parametric, non-parametric and mixed methods.

Since the subject and objectives of the dissertation are completely related to the quantitative approach, the methodology and literature chapter cover studies made with a similar approach. The quantitative approach is also adopted when findings of the statistical matching are interpreted in the result chapter of the thesis.

This thesis consists of mainly four chapters. The initial chapter is designated as the introduction providing essential information on key issues of the dissertation: statistical matching and record linkage. The general concept and workflow of the SM and data used for analysis are explained briefly. In addition to the essential concepts of the data matching methodology and its evolution extending to years, data sources and their general structure are mentioned in short.

The second chapter reviews the literature taking into account certain integrity from the earliest examples to the present to understand the concept of the subject. The data matching literature, to put it more clearly and understandably, has been reviewed in detail, from its early rudimentary pattern to the more organized and methodologically established standards it is today. Nevertheless, it is aimed to convey the studies of the recent decades as much as possible observing the development over time too. Although limited in number, national studies have also been tried to be reflected in the literature section. Each study that has been written on this subject before has been researched, considering conveying information that was not mentioned in previous studies to the literature section.

In addition to the literature part of the second chapter, theories relevant to the subject and theoretical framework regarding thesis interests such as Kish's "*theory of combining populations"* are mentioned in this chapter. The emergence process of the idea of data matching spanned nearly a hundred years, the formation of the needs that put this idea into action, sociological and statistical infrastructure stages helped shape the theoretical framework in this process and the oldest studies on this subject have been researched and included in the relevant section to conceptualize.

The third chapter of the thesis is divided into three subsections. The first section is about data sources and harmonization and provides detailed information about SILC and HBS data sets from the acquiring of them to the type of variables and sampling design. Diagnosis of both data sets spanned nearly three weeks, the structure of individual and household level records and data set preparation procedures are explained in the chapter. Individual and personal registers in SILC and HBS micro data sets, household records and structure and size of them were explained. Besides, an abstract of the essential harmonization procedures and processes specifically applied to the data sets we have such as reference person and household size, changes for alignment in response categories and derivation of common variables is explained in the chapter three. Different methods including the deflation process related to the harmonization of income variables for different years and its results are also mentioned in this section.

The second section of the chapter three of the thesis is dedicated to the preliminary processes of the SM method. There are three main aims in the chapter. The initial purpose is the determination of donor and recipient data sets accordingly since it is important for hot deck procedures and data quality. Statistical matching from larger sample-sized data to smaller one, from smaller sized data set to larger one and matching between equal sized records were investigated in the sense of availability, complications and quality of synthetic data set. The other objective is the identification of target and common variables. The identification of unique variables was also described at this stage. The final phase is to select the matching variables from common variables with maximum representation. As the selection of the matching variables is an elimination process, it has three main steps including statistical calculations. In the first step, having determined all common variables, Hellinger distance and spearman2 formulas both weighted and unweighted applied to reduce the number of them. In the last phase, weighted and unweighted regression analysis was run to minimize the number of variables to avoid matching noise and bias.

The last part of the chapter three of the thesis includes the core of statistical matching methodologies and practical applications of macro and micro matching methods. Parametric micro, parametric macro, non-parametric micro, non-parametric macro and mixed matching methods are explained in terms of methodologically and program codes with validation procedures. Statistical notations were expressed in detail to demonstrate how the mechanism behind the code system works. Since the detailed R codes are added to the appendix section, program codes related to options such as “rot”, “min”, “exact”, “constrained”, “unconstrained” etc. were represented briefly in this chapter in order to avoid repetition.

Chapter four of the thesis is devoted to findings of the statistical matching applications (for all sub-methods and options of parametric and non-parametric methods) in the context of the existing literature and related theories. Finally, recommendations expected to be implemented in subsequent studies about the statistical matching procedures are listed.

CHAPTER 2. LITERATURE AND THEORETICAL FRAMEWORK

2.1. Literature

Statistical matching method can be considered relatively new and initial academic works in data fusion field dating back to 1972. Okner (1972), merged two datasets “1967 Survey of Economic Opportunity” and “1966 Tax File” with a view to produce income distribution related 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 at this date. The new micro analytic implementation was indispensable to generate a set of comprehensive household income data which was lacking. The information available in both data sets was selected from Survey of Economic Opportunity (SEO) and Tax File to merge the files and combined. The family registers in SEO were the base for the study and information selected from the records was systematically imputed to each family row from the 1966 Tax File. The new merged file containing both demographic and income data was created in this manner after over a year of study. Criticism was mainly about bias in SEO income items and comparing distributions of them with the corresponding distributions from the Internal Revenue Service (IRS) sample to correct where disparities occurred advised academically (Sims, 1972).

Rodgers (1984), in his paper “An Evaluation of Statistical Matching”, stated that “Validity of the outputs of synthetic micro file dramatically depends on the accuracy of underlying assumptions about relationships between variables that are unique to each input file. Simulations of statistical matching procedures on samples from populations with known characteristics provide the basis for an evaluation of the usefulness of statistical matching, and for choosing among various matching techniques. Simulations frequent and substantial errors in estimates of bivariate and multivariate relationships between variables taken from two matched files.” Rodgers assumes that procedures should have developed with theoretical structures and as a result of this situation, solid methodologic justifications and empirically supportive techniques may be needed to overcome quality-based issues and misleading findings in the lack of them.

Laan (2000) in “Integrating Administrative Registers and Household Surveys” explains how data matching methods can change the organizational structure of the national offices as a result of the growing demand for coherent statistical information with the pressure of lower staff costs and response burden. Approaches for economic and social statistics were summarized as for data extracted from different isolated sources, comprehensive sample surveys, linked or isolated accounting systems and integrated micro databases with strong and weak aspects. The necessary harmonization procedures containing a set of rules before the micro integration process were classified as nine steps for the first time.

D’Orazio et al. (2006) may be considered as the first and most comprehensive study summarizing the theory and classification of all sub-methods. Besides macro and micro approaches, parametric, nonparametric and mixed methods have been explained. Having published the book, several publications and presentations were carried out by D’Orazio until today. Household surveys and randomly selected sample surveys from R database were used in these publications. The researcher was focused on the nonparametric matching area and R Studio implementations of the hot deck methods in time and wrote some of the packages in the field of micro fusing and hot deck especially in the statmatch[4] function. D’Orazio has also published articles containing auxiliary sample using, uncertainty and conditional independence assumption /CIA) with comprehensive examples in R environment. Statistical notifications in the methodology chapter of the dissertation were fundamentally based on the approach of D’Orazio.

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

Kum and Masterson (2008) demonstrated that the statistical matching method could be used for medical research. 2001 Survey of Consumer Finances (SCF) and Annual Demographic Survey of Current Population Survey (ADS) data sets were used to match. In addition to challenges and difficulties encountered during statistical matching procedures, especially the distribution of weights, was examined. Propensity score statistical matching procedure which is a constrained SM method used for data sets. SCF containing many elements of wealth at the household level was used as donor data set and ADS which is an annual survey containing information about income and demographics was used as recipient data set. Since the aim is to maximize the explanatory power and validity of the method depends significantly on the explanatory power of the Xs, the elimination period of common variables (X) in the logit model to estimate propensity scores made conscientiously. A representative measure of economic wellbeing in response to their requirements, and as a result of that necessity, statistical matching procedures preserving marginal distribution of the data sets were needed. The researchers have observed that in case of violation of the method under the conditional independence assumption, outputs may be compromised due to divergent series of joint distributions. They also assumed that outputs of the merging data sets which are representative at the national level and state level might be representative to some extent. Comparing conditional distributions of the imputed values in synthetic micro files and donor data sets is an insufficient but necessary way of checking the quality status of the output of statistical matching in the lack of auxiliary information or sample.

Zacharias et al. (2014), presented statistical matching methodology using a two-dimensional poverty measure for Turkey. TURKSTAT microdata of Household Budget Survey (HBS) and Time Use Survey (TUS) were exploited to fuse time spending on own production for each member aged 15 years and older in TUS. Poverty measures calculated by national offices generally do not contain time deficits. As their assumption is household members have time sufficiently to contribute the requirements of them, they underestimate both the scope and the depth of poverty. Their models consider intrahousehold disparities in time allocation unlike the neoclassical model. Estimation of time deficits for household members aged from 18 to 70 years old was the focus of the study because they make up 95 percent of the employed population in 2006 data. SM procedure is practiced with reference to estimated propensity scores which derive from stratum and common variables. An observed value in the donor data was matched with the same or nearest neighbor based on the rank of their propensity scores for each recipient. In the matching procedure, a penalty weight was determined to the propensity score according to the size and ranking of the coefficients of strata variables not used in a particular matching round. The quality of statistical matching is evaluated traditionally by comparing the marginal and joint distributions of the variables of interested in the donor data and the synthetic file.

Kim (2018), focused on how an effective way to generate a small overlap of units can be constructed in a statistical matching procedure if available data sets have only categorical variables. The innovation in the dissertation was fundamentally about three estimators. In addition to the conditional independence assumption estimate and direct estimators, a new estimator was developed by merging these two estimators named combined estimator. Netherland Population Census data (2011) was selected as the only data source for the whole research. Population data is divided into three parts using simple random sampling method to generate new data sets. Occupation and education level variables which are target variables were removed from the data sets separately. To put it clearly, donor and recipient data had not these two variables at the same time. Several experiments were carried out such as altering the size of auxiliary sample data C, changing the number of X variables and total sample size of A and B. Unlike expectations, EM algorithm estimator gave better results than CE. Mean square error (MSE) was used as a quality and validation object. Variance of A∪B, size of overlap C, and different selection methods for Xs are evaluated as a kind of quality scale. The essential assumption in the approach is that conditional independence is an adopted and beneficial method, which is probably when the information in common variables is quite enough and it is predictive of the behavior of the target variables. The estimation of the joint probability can be evaluated as estimating a `deviance' compared to the model with conditional independence assumption. Deviances mentioned could be estimated using the small overlap, by the observation of the disparity between the DE and the CIA estimator. On condition that the size of deviances is considerably enough, they will be represented in the joint distribution estimation.

Newger (2018), aimed to research how Markow networks (MN) could be used in SM procedures specific to the categorical target and common variables. Joint probability distribution in (MN) is reconstituted to compliance SM methodology. A and B datasets, which are disjoint, simulated for applying MN approach using IsingFit package in R software. Validation was based on four levels of Rassler and aimed to preserve; individually existing information in data sets, joint distributions, correlation structure and marginal distribution of variables. As a result of the work, it's seen that the conditional independence assumption is connected with MNs very strongly because of the structure of joint probability distribution.

Statistical matching methods, when national researches are examined, are a very new area in Turkey and the limited number of papers and dissertations belong to the last decade. There are only a few studies performed explicitly using the statistical matching method. Studies were conducted on surveys, which generally focus on social research, and mostly non-parametric methods were used.

Ahi (2015), matched cross-sectional SILC and HBS data of Turkey belonging 2012 and 2011 years respectively in order to gain variables on the basis of Classification of Individual Consumption According to Purpose's (COICOP) twelve main expenditure groups for households with data mining methods. Complementary usage type of the statistical matching method was utilized and three approaches, parametric, non-parametric and mixed method, were compared in the sense of twelve main expenditure groups. When the results were examined, it was observed that the non-parametric method gave more consistent results.

Albayrak and Masterson (2017), matched different years' data (2005, 2008, 2009, 2012 HBS and SILC of Turkstat) to research consumption behavior and indebtedness of households and inequality for Turkey using estimated propensity scores developed by Kum and Masterson in 2010. Ratios of mean and median household expenditure, comparing the mean values of the transferred variables by income deciles and Gini coefficients for per capita evaluated as an element of validation. Because of the inconsistency between income variables in the sense of reference periods of them in the different microdata sets, they suggested matching HBS with the reference of (t-1) as a solution method. Small differences in the Lorenz curves, comparison of density functions, conditional distributions, mean values, Gini coefficient and indicators for population subgroups demonstrate that the overall quality of the statistical matching is sufficient.

Uçar (2017), compared and matched two longitudinal surveys (four years) unlike the ordinary statistical matching applications done so time for the purpose of analyzing the effect of a newborn on household poverty. Consumption expenditure information for each year was transferred from HBS to SILC using non-parametric micro matching but the structure of the surveys caused many complications especially with regard to the reference period of surveys, household level weights, harmonization and calibration of data, the population in and out by years and deflation rates about household's revenues. Nevertheless, micro-level data fusion method could be used for a demography study to find out the correlation between poverty and fertility. Because two surveys have stratified and clustered sample design, Renssens' calibration method to deal with complex sample design and Rassler's validation method was used. Survey design variables were attached to current data sets. Calibration of the data sets was done by linear, raking and poststratification methods. Micro fusion by Renssen's approach was not resulted as expected. Thus, non-parametric micro matching method and nearest neighbor hot deck were applied to produce synthetic data covering target and common variables together.

Öztürk (2019), focused on categorical variables and evaluated non-parametric statistical matching approach using 2014-2015 Time Use Survey of Turkey and 2014 Life Satisfaction Survey of Turkey. Even if the number of common variables is less compared to other social surveys, Hellinger Distance calculation method was used to reduce the number of common variables. Then, logistic regression was run as target variables have binary response categories. Hot deck methods were applied in R environments both weighted and unweighted. In the first analysis, 8 common variables were reduced to five and then finally four (age, sex, marital status and number of rooms) according to the logistic regression results. Even though expectations were constrained nearest neighbor distance hot deck and rank hot deck approaches would provide more accurate estimation levels, implementations revealed that random hot deck method's 'min' option and nearest neighbor distance hot deck provided better results. "Rot" option did not give accurate results along with these.

2.2. Theoretical Framework

Multi-population and periodic surveys have grown more widespread and significant in terms of design and operation. Only in recent decades have the necessary vast resources, both financial and technical, been assembled, and the great worth of both has been acknowledged. The focus of development for both sorts of designs has been on survey comparisons. Combining survey statistics is still possible, desired and practiced because of the coordination and harmonization required for comparisons. But until recently, the combinations of surveys have been achieved and presented largely without a theoretical/methodological framework, and often briefly and initially by (Kish, 1999) as one of the first literature. The theoretical frame of SM is substantially based on combining experiments and combining sample studies. Cochran (1937) aimed to combine separate sources to research in the field of crop yields using ANOVA 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 samples (CS). CS procedures need too much attention during preparation and coordination phases. It is a great deal of starting with good planning especially for multinational surveys contrary to national multidomain surveys which have coordination naturally. The second difference between 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 SRS of subjects. The final point of difference is about the statistical analysis period. Contrary to CX, comprehensive analysis of survey method including joint analysis, similarity and comparability are used intensely in CS.

National offices conduct large-scale field researches over long periods. They also conduct small-size surveys more frequently such as monthly, quarterly or annually. Registers in other words administrative data may be used by them. Despite all these large data sources, the everlasting information demand in short periods for microdata, which includes separate variables in different studies, has accelerated CS and finally SM studies. After Leslie Kish defined the concept as "*theory of combining populations*" including different types of *accumulation of rolling samples*'[5] data "sample reported at regular intervals for time periods that overlap with preceding time periods", CX and CS methodology evolved over time in the statistical matching direction.

Although there are more recent theories and concepts covering bias and cost aspects, its sociological and theoretical framework is not very clear. The process of gaining a definite ground for the theoretical infrastructure has not reached a certain stage. Theoretical framework will be more understandable and explicit with the study of all the sub-headings of the subject in the course of time. Today, a few studies in this field are superficially aimed at improving the methodological sides of statistical matching in general. In addition to the holistic perspective, the theoretical infrastructure of each sub-method has to be developed up to the distinction between social and economic studies.

In the statistical matching applications, especially in the earlier 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 and reducing employee costs and survey expenses for the benefit of the public. Contributions on methodological aspects could not be developed at the same pace as practices. However, as misleading findings of exact statistical matching were evaluated by analogy, improvements in the validity procedures will be allowed it to sit on a more solid ground theoretically.

[5] Rolling sample is a panel sample design concept planned for several purposes (Alexander, 2001).

CHAPTER 3. METHODOLOGY

The methodology chapter consists of three sub-headings. The structure of SILC and HBS data sets and their harmonization procedures are mentioned briefly in the Chapter 3.1. Since the harmonization process methodologically includes a series of certain operations, it has been classified in various stages by adhering to the theoretical operation rules. Preliminary processes of operational procedures are explained in the second section of the methodology chapter. In addition to the traditional methods and statistical calculations used in the selection stages of the matching variables, the effect of the complex sample design of the samples is included in the processes. Statistical matching methods were performed for parametric, non-parametric and mixed approaches considering micro and macro matching methods.

3.1. Data Sources and Harmonization

During the planning phase, detailed research was conducted on various data sets, including the 2011 Population and Housing Census of Turkey (PHC), economic surveys about enterprises and mortality statistics (2009-2018). Data dissemination procedure of Turkish Statistical Institute (TURKSTAT) classifies microdata sets in two types: A and B group microdata. A group microdata sets can only be used at a data research center and on designated computers in TURKSTAT. B group microdata can be used outside of TURKSTAT. The data research center is designed for researchers to use restricted data. Since Population and Housing Census data are subject to the A group microdata procedure, work permits are granted only in data research centers of TURKSTAT. On the other hand, only a 5% subsample of the data can be used by researchers. However, the idea of working with 2011 PHC data could not be put into practice because it coincided with a time span when the pandemic period was and the data research center was closed for use due to bans. As a result, HBS and SILC questionnaires were preferred because they are subject to the B group microdata procedure and can be used outside of data research centers without limitations. Although it was not possible to provide survey design variables during the first application process, as a result of the applications made during the analysis studies, stratum and cluster information could be obtained provided that alias codes were assigned. Therefore, the analyzes were repeated to include the provided survey design variables.

Ethical concerns, as data sets have economic, social and demographic information of respondents, are very important for researchers and TURKSTAT. Data sets contain information on both household and individual basis. SILC has very detailed income data on individuals. HBS contains comprehensive data on household consumption patterns. Therefore, confidentiality is a vital and mandatory issue for all sides of the research. A contract was signed between TURKSTAT and the researcher, and data confidentiality is guaranteed officially in this way. The contract obliges two things:

1- The researcher cites the institutional microdata he used while publishing the results obtained from the study,

2- The researcher may not reproduce, give to third parties, sell or transfer the micro data set he has received.

Another ethical concern is about block and cluster variables. Estimation level of SILC is NUTS II and whole Turkey for HBS. Microdata of SILC and HBS can enable the researcher to produce estimations in unpublished levels in case of the researcher has survey design variables. Thus, it is an institutional principle or policy not to share survey design variables with third parties. This matter has been a long-lasting problem despite the commitment. Finally, the issue has been overcome by providing the data with alias codes. Statistical estimations, ethically, will not be made on a regional basis under no circumstances.

3.1.1. Statistics on Income and Living Conditions (SILC)

Type of survey design: SILC is a four-year panel survey (or longitudinal[6]) conducted annually for determining the living standards of Turkey since 2006 regularly. In the process until 2006, HBS questionnaires were used to procure estimations on income

[6] Longitudinal or panel survey refers to a research design involving repeated observations of the same variables (people, household, etc.) over short or long periods of time such as annually, quarterly or monthly. The same households are interviewed annually in SILC for four years.

distribution. In SILC, the rotational design was used. In other words, %75 of households intended to be remained in the sample from the year (t) to the next year (t+1). %25 of the households in the related year were replaced by new households which are selected statistically.

Objectives: Owing to the panel survey structure of SILC, it is possible to monitor changes on individual basis over time, while analysis of income, poverty, social exclusion and other living conditions can be made on an annual basis from its cross-sectional structure. Information is collected on housing, economic situation, social exclusion, real estate ownership, education, demography, health status, working status and income status.

Geographical coverage: Coverage of the SILC is all settlements within the territory of Turkey.

Sampling method: Stratified, two-staged and clustered sampling method is used. The first selection is from Address Based Population Registration System (ABPRS). Blocks consist of approximately 100 dwellings are selected by PPS (probability proportional to size) method. These are named as the primary sampling units. Then, households (twelve for urban and eight for rural) are selected from these blocks.

Sampling unit: Household is the sampling unit and all household members residing in Turkey are covered apart from immigrants and people living in prisons, military facilities, nursing homes, childcare centers, hotels and private hospitals.

Estimation level: Statistical estimations about poverty, education, etc. are produced for the whole of Turkey. Since the design of SILC provides cross-sectional and panel estimates, cross-section estimations are obtained from the sample applied in the relevant year, panel estimations are generated from the sample continued in consecutive years. Turkey, NUTS-I and NUTS-II level estimations are produced according to the results of cross-sectional research. Within the scope of the longitudinal research, it is aimed to produce estimates for the country in general with panel data of 2, 3 and 4 years.

Questionnaires: SILC questionnaire consists of nine different modules. These are forms for registering and monitoring to households and individuals, questionnaires of household and individuals, and questionnaires of agriculture and modules (see appendix A for cover page of SILC).

Weights: The SILC longitudinal weights are generated by taking into consideration the non-responses and base weights of the individuals who participate in the panel over the corresponding year. They are created by assigning 2, 3 and 4-year multiplier factors to the base weights of the individuals.

Classifications: It is based on NACE Rev.2 for economic activities and ISCO 08 for occupational status.

Mode of survey: Computer assisted personal interviewing (CAPI) method is used as data collection method.

Data collection: The time period of data compilation starts in March and is scheduled to finish in July. Interviewers aim to complete the whole interviews with respondents between (t+3) and (t+7) months. Standardized practice is to interview all selected households for 4 times in a year.

Size of microdata: Micro data set of SILC 2018 includes information from 27,068 households and 81,178 individuals.

3.1.2. Household Budget Survey (HBS)

Type of survey design: Household Budget Survey is a cross-sectional survey that is compiled on an annual basis. Although the history of Household Budget Survey dates back to 1987 with different names, regularly repeated surveys start in 2002 as a consequence of the continuous development and socio-economic transformation of the country.

Objectives: The survey basically aims to measure socio-economic indicators, consumption structures and income levels of the households and individuals to observe whether or not socio-economic policies are realized. National income calculations and base year weights of CPI are counted as some of the usage areas but HBS collected for monitoring consumption patterns and their change in time, determination of the minimum wages, poverty threshold and living standards of households, etc.

Geographical coverage: Coverage of the Household Budget Survey is all settlements within the territory of Turkey as SILC is.

Sampling method: Sampling method is stratified, two-staged and clustered sampling method. The first selection is from Address Based Population Registration System (ABPRS) and blocks consisting of approximately 100 dwellings are selected by PPS method. Then, households (twelve for urban and eight for rural) are selected from these blocks.

Sampling unit: Household is the sampling unit and all household members residing in Turkey are covered apart from elderly houses, rest homes, military facilities and hospitals with specific features, nursery and nomadic population.

Estimation level: The estimation level of the Household Budget Survey is whole of Turkey.

Questionnaires: Household Budget Survey consists of eleven sub-modules containing information on the basis of individual and household (see appendix A for cover page of HBS).

Weights: The results of the survey are weighted using the most recent projection of population. After ABPRS is established, population projections have been renewed according to the most recent registers. Weights based on this national and regional projections have been used.

Classifications: Classification is fundamentally based on COICOP for good and service expenditure and it contains 12 main expenditure groups such as health, education, transportation, etc., NACE Rev.2 for economic activities and ISCO 08 for occupational status are also used.

Mode of survey: Household Budget Survey has a mixed survey mode. In addition to diaries, computer assisted personal interviewing (CAPI) method is used as data collection method.

Data collection: Household Budget Survey, contrary to many other surveys, is compiled throughout the whole year. Diaries have been given to households in order to record the whole consumption expenditures of them daily. Compilation of data starts with consumption expenditures and household's own production recorded in diaries during the last 12 months.

Size of microdata: Micro data set of HBS 2018 includes information from 11,828 households and 40,688 individuals.

3.1.3. Data Set Preparation and Harmonization

Okner (1974) and Kish (1965) proposed pioneer but partial suggestions about the data preparation processes which actually is an intensive and long sequence of operations. Rasner et.al. (2007) have also explained data set preparation phases for registers. The process includes the examination of definitions and contents of all variables and response categories. Determination of common and target variables should be done at this stage. Response categories should be harmonized, especially as a result of determining common variables. Therefore, the need to define the principles and details of the harmonization process more comprehensively has arisen. Laan (2000), summarized micro-integration processes in 9 initial steps:

> *“a. harmonization of units: are the statistical units defined uniformly in all sources? (special reference to comparability in space and time);*
> *b. harmonization of reference periods: do all data refer to the same period or the same point in time?*
> *c. completion of populations (coverage): do all sources cover the same target population?*
> *d. harmonization of variables: are corresponding variables defined in the same way? (special reference to comparability in space and time);*
> *e. harmonization of classifications: are corresponding variables classified in the same way? (special reference to comparability in space and time);*
> *f. adjusting for measurement errors (accuracy): after harmonizing definitions, do the corresponding variables have the same value?*
> *g. adjusting for missing data (item non-response): do all the variables possess a value?*
> *h. derivation of variables: are all variables derived using the combined information from different sources*
> *i. checking overall consistency: do the data meet the requirements imposed by identity relations?”*

Common variables used for matching procedures do not have missing items. Data sets have the same definition of household. Microdata sets of SILC and HBS do not have known measurement errors (Turkstat, 2018a, Turkstat 2018b). Complications were mainly about variables and reference periods. In this sense, the transactions needed for consistency are made within the scope of these problematic harmonization steps after diagnosing the micro data sets.

Table 3.1. Basic Survey Information on HBS and SILC

INFORMATION	SILC	HBS
Sample Size (Individual)	81,178	40,688
Sample Size (Household)	24,068	11,828
Individual Data Set (Number of Questions)	66	66
Individual Register Data Set (Number of Questions)	10	
Household Data Set (Number of Questions)	65	130
Consumption Expenditure Data set (Number of Subsets)		4

3.1.3.1. Diagnosis of Data

In addition to the various guides, microdata of Statistics on Income and Living Conditions questionnaire was provided as 9 separate tables and 3 separate sub-datasets. These are:

1- Individual data set (information about only 15+)
2- Individual register data set (information about all household)
3- Household data set

The individual data set consists of 66 questions, the individual record data set consists of 10 and the household data set consists of 65 questions. Data sets are merged with the help of 2 variables named fertid and bulten.

HBS data is also provided as 8 tables and 3 separate sub-datasets with various dictionaries providing information about variables. These are:

1- Individual data set
2- Household data set
3- Consumption expenditure data set

In terms of variables, it is seen that the individual data set consists of 66 questions and the household data set consists of 130 questions. Consumption expenditure data set consists of 4 subsections. Classification of consumption expenditure on COICOP (Classification of Individual Consumption by Purpose) basis has been made. It is possible to link data with a variable named unitno. Sample sizes and number of questions of subsets can be seen in the Table 3.1.

3.1.3.2. Harmonization of Reference Person

Reference persons have some specific characteristics and these could be used as matching variable. When both questionnaires were examined, it was observed that there was a difference in the definitions of the reference person as a concept. For HBS, the definition is "A member of the household who receives the highest income", while the definition for SILC is "The household member over a certain age who has a say in the management of the household and plays the most active role in the legal, social and economic planning and decision process of the household." in the form. The definition in SILC may be considered as more traditional description contrary to HBS. Since the definition difference will have an effect on the selection procedures to be made, it has become necessary to harmonize the definitions of these two data by examining them. The reference person in SILC was revised in accordance with the definition of HBS. In this sense, the head of household was reassigned with the SAS Enterprise program based on the variable (FG140) containing the total income item in the data set. The new variable, which is 82.16% compatible when compared to the first one, has been made fully compatible in this way. Since 8 variables (approximately 20 percent of all common variables) among the common variables (X) are related to the reference person, this assignment has enabled the matching quality to be increased.

3.1.3.3. Harmonization of Household Size

Even though Household Budget Survey includes household size data, microdata of SILC, on the other hand, does not have the same variable in this distinction. Since SILC data has identification numbers for individuals and households, this variable is obtained for SILC through the individual register data set and these ID numbers.

3.1.3.4. Harmonization of Classifications and Response Categories

Common variables and their response categories may be classified in a different way. Therefore, in some variables which ask the same questions but are coded in a different way, response categories were reorganized accordingly. The reference person's age group and reference person's number of weekly working hours were categorized in this sense. The answers to the marital status question, which has different response categories, were also harmonized. The answers to the education question were divided into subcategories. The answer to the reference person's economic activity of work and heating system of the dwelling question has been harmonized. Differences in response categories for ownership of mobile, computer, internet, washing machine, refrigerator, dishwasher, air conditioner and car were classified in a harmonized way (see appendix C for distributions and harmonized response categories of the variables).

3.1.3.5. Derivation of Variables

Existing data sets can allow the derivation of new and needed variables. As it is aimed to observe how the derived variables give results in the elimination processes, eleven variables are created in both data sets using available micro data sets. These are the "number of children, number of adults, number of elderly, number of women, households all members are adults, elderly and women, number of employed people, number of individuals with employee income, number of individuals with self-employed income and number of individuals with retired income". These variables are included in HD, spearman2 and regression processes along with others.

3.1.3.6. Harmonization of the Reference Periods

SILC and HBS surveys refer to the same year, however, the collecting period of surveys and reference time of some variables are different for both surveys. SILC has different reference periods regarding income, unemployment, demographic indicators and dwelling. Income variables planning to use as Y (one of the target variables) in SILC survey refers to the preceding year (2017) contrary to HBS. Since both surveys have cross-sectional survey types, the selected solution for reconciliation was to inflate income variables including "HG110" column using Consumer Price Index value of the related year.

3.2. Preliminary Stages

Statistical matching method consists of four preparatory stages to do before starting data matching applications. The first stage is to choose target variables in SILC and HBS data. These variables are unique for each data set and used for further procedures. The second stage of the preliminary phase is the determination of the common variables that are present in both sets of data. The next step is to determine which data set is donor or recipient data, according to the requirements of the researcher and limitations of the data matching procedures. The final step focuses on the elimination of common variables to clarify which of them could be used as matching variables in the following stages.

3.2.1. Choice of Variables (Y, Z)

The preparatory stages of SM applications begin with the determination of the target variables Y and Z. These should be uncommon and unique variables in both datasets. Therefore, income variable from SILC data is Y, and consumption expenditure from HBS is Z in the analyses. Fused variable (Z) depends heavily on the goal of the data matching. As consumption expenditure is wanted to be added to SILC microdata sets, Z variable was selected accordingly. Choice of Y variable depends on Z variable. Y variable should be related to the concept of Z variable. Because consumption and income are relevant variables, income variable was selected as Y variable.

The choice of Y and Z variables is entirely related to the needs of the research. There are no specific or certain rules especially for the selection of Y variable. In this study, it is aimed to assign the Z variable to the SILC data in the synthetic file on a micro basis and the selections were made in this way. However, inverse matches have also been applied experimentally and the results have been observed.

In case of a lack of suitable data sources or the aim of the researcher is to investigate the statistical matching method experimentally, only one data set can be divided into two parts randomly. Variables are removed from the data set and Y and Z target variables could be determined artificially. Since the researcher has observed values, validation procedure is performed by using them. In addition to the mentioned advantages, auxiliary information or auxiliary sample can be easily created in this type of data source. A third data set is created from the same sources and this sample is used to test the quality of the synthetic data set.

3.2.2. Determination of Common Variables

Common variables (Xs) are the list of variables compiled in the same or similar way in both data sets, which can be aligned with the operations to be made in the response categories or classifications, and as a result, form a pool where the selection of matching variable could be made in the SM applications. At this point, variables with these characteristics are determined in the previously diagnosed data sets. 15 variables were aligned in terms of response categories to use as common variables. Apart from the derived variables, as a result of the examinations, it was determined that 13 variables could be used without the need for any harmonization process. Finally, 39 variables present in both surveys have been determined as common variables.

Determination of common variables is a very important step but it is not so critical for making survey errors. The pool of variables (Xs) can be kept as wide as possible. The most relevant and predictive variables can be determined from this wide pool. Moreover, since common variables will be subjected to elimination processes in the next stages, final matching variables will be selected accordingly. They will be selected according to their explanatory power.

In the Table 3.2., it can be seen a list of selected and derived common variables. All subsequent procedures of selection and data matching will be based on these variables in the list.

Table 3.2. Selected and Derived Common Variables

ABBREVIATIONS	VARIABLES
“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”[7]

[7] Turkstat (2018a), Turkstat (2018b) and Uçar and Gianni (2016) are used for variable names and abbreviations.

3.2.3. Assigning Donor and Recipient Data Sets

The third step of the data set preparation process is to choose which data set is donor or recipient. Donor data set is used for providing the assignment of Z values to the recipient data set during the matching processes and imputation phase. Assignment of Z values in the recipient data set is performed by donor data.

Larger sample sized data sets are generally preferred as donor data set since the value of variables would be imputed repeatedly. (D'Orazio, 2006). Besides, selecting smaller sized data set as donor data set may cause syntax errors preventing to proceed in non-parametric matching codes. However, assignment procedure can be determined in both directions for the purpose of the research.

Looking at this dissertation specifically, since consumption expenditure of household is aimed to fuse (or transfer) to synthetic data, smaller sized data set (HBS) is assigned as donor data set and larger sample sized data (SILC) is assigned as recipient data. Nevertheless, the effectiveness of the method was tested by matching data sets in both directions.

3.2.4. Elimination of Common Variables

Common variables should be thought of as a repository or framework containing variables that are considered representative for further processes. It is not possible to use all variables in data sets, and the use of more than necessary variables causes unnecessary noises in the models. In this respect, statistical methods including a series of elimination processes should be applied to determine as fewer final variables as possible.

Possible calculation methods to eliminate common variables are Hellinger Distance, spearman and regression analysis. These methods are used for comparing similarity of distribution of variables. Depending on the number of the common variables, one or more of them are applied (D'Orazio, 2013). However, as a new approach in this thesis, all three methods are applied both weighted and unweighted. In addition to these innovations, survey design variables of SILC and HBS samples are used to observe the effect in the elimination period of the common variables.

3.2.4.1. Hellinger Distance

Selection of matching variables can be done by many prominent and effective methods such as *χ^2, K-S, runs test,* etc. which are more complicated and need sample design of donor and recipient surveys. HD, on the other hand, is easy to use and the method calculates a value between 0 and 1 representing similarity by using response categories of the variables.

The HD value calculation is given based on the Formula 3.1.

$$HD\ (D, R) = \sqrt{\frac{1}{2}\Sigma_{\mathrm{i}=1}^{\mathrm{k}}\left(\sqrt{\frac{\mathrm{n_{Di}}}{\mathrm{N_D}}} - \sqrt{\frac{\mathrm{n_{Ri}}}{\mathrm{N_R}}}\right)^2} \qquad (3.1)$$

D: Donor data set

R: Recipient data set

K: Total number of cells

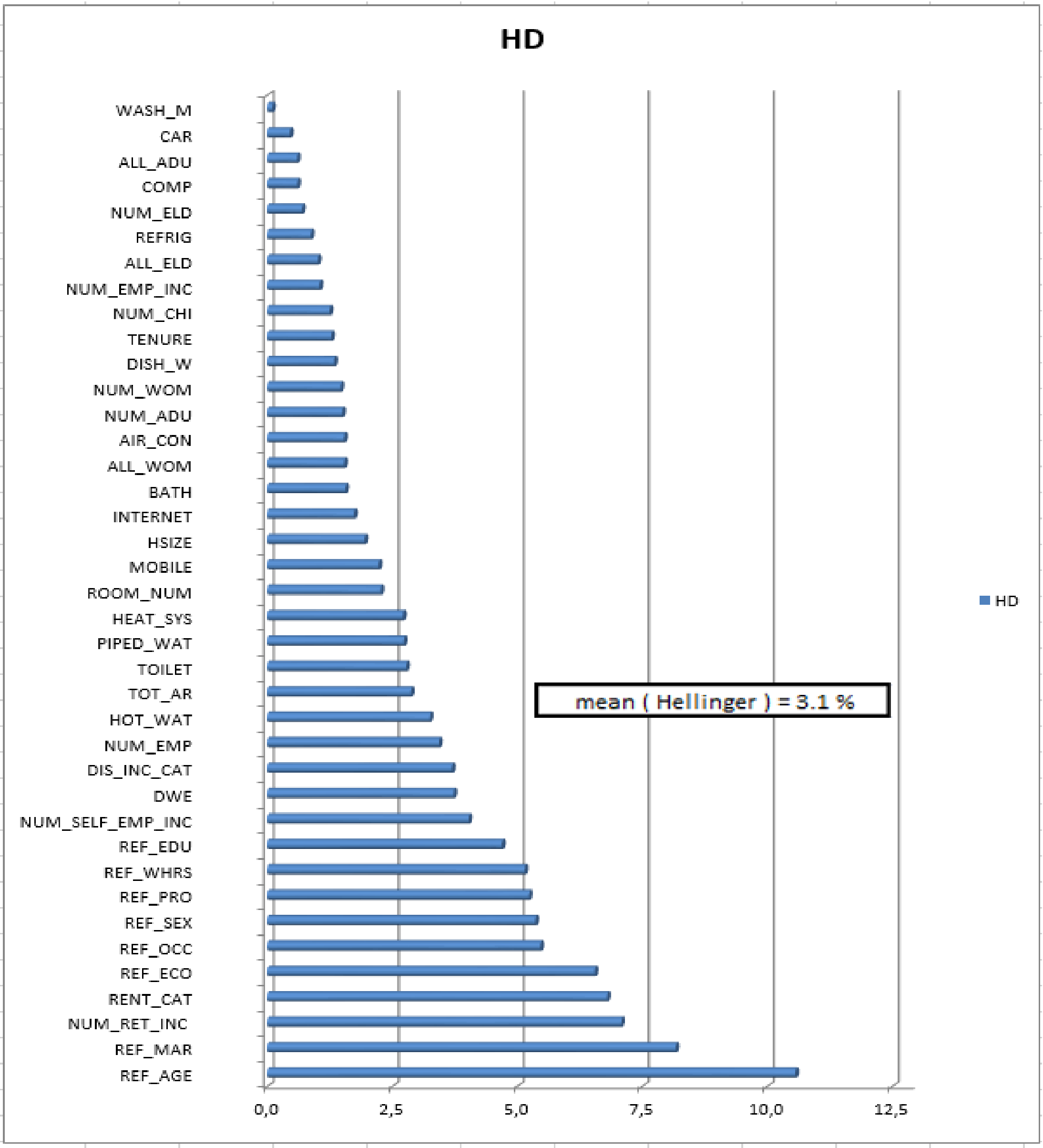
$\mathrm{n_{Di}}$: The frequency of response categories in donor

$\mathrm{n_{Ri}}$: The frequency of response categories in recipient

N: Total size of the contingency table.

In literature, >= 5 % is accepted as a cutoff value to exclude unfit variables. Variables with < 5 % scores are selected as possible matching variables (Webber and Tonkin, 2013). So nine variables of Xs have HD values higher than 5% as a result of the unweighted procedure. In the Figure 3.1., outputs of unweighted Hellinger Distance calculations can be seen.

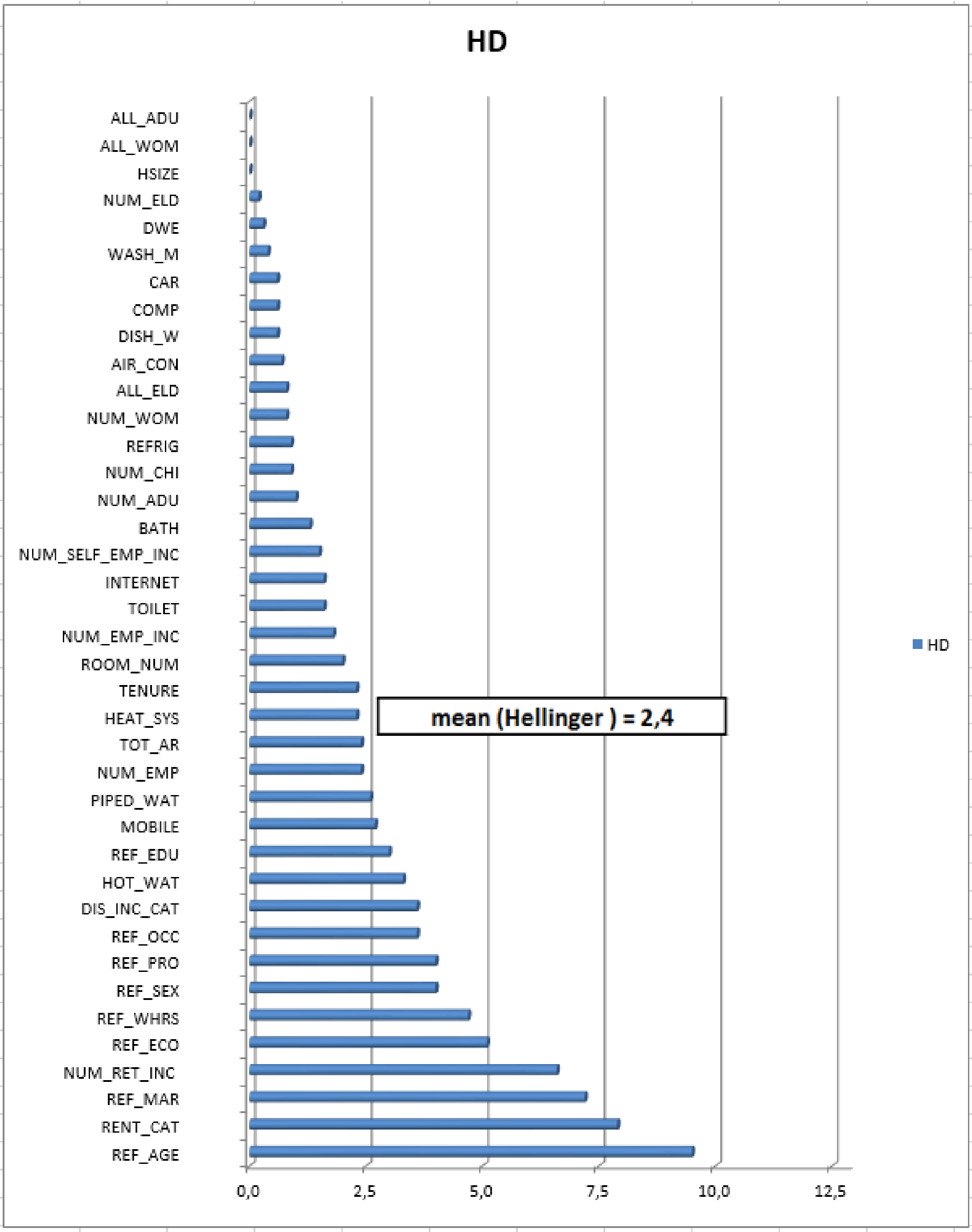
Figure 3.1. Unweighted Hellinger Distance Scores of Common Variables



Researches related to SM make use of HD calculation to obtain final matching variables or to narrow common variables without using sample weights invariably. In this study, household level sample weights were used in HD calculations and for further analysis to reach more accurate results.

Percentages of the response categories of common variables are recalculated in SPSS by considering household level weights. New weighted outputs are seen in the Figure 3.2.

Figure 3.2. Weighted Hellinger Distance Scores of Common Variables



Results of the weighted Hellinger Distance calculation indicate that only five variables remain out of the process contrary to unweighted scores. Besides, the mean value of all scores decreased from 3.1 to 2.4. Decrease in the mean value of total HD scores means that there is an overall improvement in the new findings.

In addition to the general evaluation, when viewed on a variable basis, four variables that do not have HD scores under cut-off value in the unweighted HD calculation, could also be used for further analysis. These are reference person's number of weekly working hours, reference person's sex, reference person's occupation and reference person's professional status. Weighted and unweighted scores of these four variables can be seen in the Table 3.3. The largest proportional and numerical change was observed especially in the “occupational status of the reference person” variable among these four variables.

Table 3.3. Weighted and Unweighted Scores of 4 Variables

VARIABLES	WEIGHTED HD	UNWEIGHTED HD
REF_WHRS	4.7	5.2
REF_SEX	4.0	5.4
REF_PRO	4.0	5.2
REF_OCC	3.6	5.5

3.2.4.2. Spearman

Common variables were reduced from 39 to 35 to obtain final matching variables but this level needs more elimination processes because too many variables may cause undesired noise affecting SM results.

Another method to eliminate Xs is a calculation method named spearman2[8] in R studio “Hmisc” package. There are not certain rules or hierarchy to start or go on with any of three methods. Since the type of Xs are categorical and the type of target variables (Y and Z) are continuous, they all are admissible for the requirements of the function. According to the generally accepted rates in the literature, the evaluation of the results is that the variables with a value of 10 percent or more are appropriate (Harrell, 2016).

[8] “Spearman2 computes the square of Spearman’s rho rank correlation and a generalization of it in which x can relate non-monotonically to y. This is done by computing the Spearman multiple rhosquared between (rank(x), rank(x)2) and y. When x is categorical, a different kind of Spearman correlation used in the Kruskal-Wallis test is computed (and spearman2 can do the Kruskal Wallis test). This is done by computing the ordinary multiple R2 between k-1 dummy variables and rank(y), where x has k categories.” (Harrell, 2016)”

Adjusted rho2 values represented in the Table 3.4. show that eleven variables scored over ten percent indicate strong explanatory power in both data sets. These variables could be used in subsequent processes. 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 in the household, dwelling type and ownership of computer, dishwasher and car. Excluding disposable income categories for both surveys which have the highest scores, reference person's education status has the highest value for SILC. On the other side, ownership of car has the highest value for HBS.

Table 3.4. 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 spearman: Spearman2 calculation was also implemented using household level weights in order to minimize unnecessary variables and discover variables having strong explanatory power contrary to traditional approaches not using weights. R program's "wCorr" and "weightedCorr" functions, which enable only with numeric categories, were utilized. Therefore, variables that are not numerical types such as "ref_whrs", "tot_ar" and "dis_inc_cat" were recategorized accordingly. Weighted and unweighted functions are based on the Formula 3.2 and Formula 3.3.

```
spearman2(Y~var1+var2+…, data= a)
spearman2(Z~ var1+var2+…, data=b)                                    (3.2)

weightedCorr (x=data$var1, y=data$var2)
method = c("Spearman"), weights = data$weight)                       (3.3)
```

Results represented in the Table 3.5. indicate that unexpected conditions would occur compared to the unweighted method of spearman2. In the HD calculation, variables suitable for the unweighted method had values out of the specific ranges. However, as a result of weighted HD calculations, it was observed that four of them could be reused for the next stages. But in this process, 2 common variables scored over 10%, received values outside of the specified ranges for weighted spearman2 calculation. Thus, variables about heating system and dwelling type could not be used for the further analysis periods.

After evaluating both weighted and unweighted spearman2 scores, 9 of 34 variables could be used in subsequent stages: disposable income categories, reference person's education, number of employed people, number of individuals with employee income, internet, number of adults in the household and ownership of computer, dishwasher and car. Adding household level weights into the HD and spearman2 procedures as a new approach changes the results dramatically. Dwelling type variable eliminated after weighted spearman2 contrary to unweighted calculations.

Table 3.5. Adjusted Rho2 Values (Weighted)

NO	VARIABLES	SILC	HBS
1	REF_WHRS	0.07553	0.04815
2	REF_SEX	0.01725	0.04550
3	REF_PRO	0.06470	0.02840
4	REF_OCC	0.00903	0.00349
5	**DIS_INC_CAT**	**0.94095**	**0.48399**
6	HOT_WAT	0.07044	0.04297
7	**REF_EDU**	**0.23539**	**0.12932**
8	MOBİLE	0.04649	0.02755
9	PIPED_WAT	0.00590	0.00331
10	**NUM_EMP**	**0.21147**	**0.13859**
11	TOT_AR	0.12319	0.08644
12	HEAT_SYS	0.13720	0.09354
13	TENURE	0.02024	0.00353
14	ROOM_NUM	0.11200	0.07147
15	**NUM_EMP_INC**	**0.16431**	**0.10428**
16	TOILET	0.03523	0.03987
17	**INTERNET**	**0.20128**	**0.17621**
18	NUM_SELF_EMP_INC	0.00151	0.00056
19	BATH	0.01746	0.01302
20	**NUM_ADU**	**0.15268**	**0.13425**
21	NUM_CHI	0.00452	0.01343
22	REFRIG	0.01526	0.00689
23	NUM_WOM	0.02848	0.03397
24	ALL_ELD	0.08658	0.08186
25	AIR_CON	0.03236	0.03289
26	**DISH_W**	**0.17833**	**0.12872**
27	**COMP**	**0.21597**	**0.16444**
28	**CAR**	**0.16582**	**0.19915**
29	WASH_M	0.03467	0.01831
30	DWE	0.10973	0.08951
31	NUM_ELD	0.02686	0.03209
32	HSIZE	0.07038	0.07547
33	ALL_WOM	0.07655	0.06244
34	ALL_ADU	0.00226	0.00008

3.2.4.3. Regression Analysis

Regression analysis could be used as an elimination method in the selection procedure of the common variables. Kleinbaum et.al. (1997) has explained the concept:

> *"Regression analysis is a statistical tool for evaluating the relationship of one or more independent variables X_1, X_2, ..., X_i to a single, continuous variable Y. It is most often used when the independent variables cannot be controlled, as when they are collected in a sample survey or other observational study. Nevertheless, it is equally applicable to more controlled experimental situations. In practice, regression analysis is appropriate for several possibly overlapping situations."*

Although the number of common variables has been significantly limited in the processes up to now (from 39 to 9) the number of variables at this level is still too high for the statistical matching. Even if it is not possible to give an exact number of matching variables, Augurzky and Schmidt (2001) emphasize that parameters in the models should not be much. Bryson et al. (2002) also stress that an over-parameterized model may cause an increase in the variance of estimations. Therefore, an additional third step to reduce them to an acceptable level should be applied. Linear regression can be applied when the type of the dependent variables Y or Z are continuous. Logistic regression, on the other hand, could be performed when the type of the dependent variable Y or Z is binary. Therefore, it is a method changing according to the type of dependent variables. Ignoring sample weights are common in this step similar to Hellinger Distance and spearman2 calculation period. But as in all other selection processes, having created dummy variables for nine common variables, regression analyses were carried out both weighted and unweighted. Dependent variables are income and consumption expenditure variables for SILC and HBS respectively.

Table 3.6. Weighted and Unweighted Regression Results of Selected Variables

VARIABLES	LINEAR REGRESSION				LOG_LINEAR REGRESSION				FREQ.
	WEIGHTED		UNWEIGHTED		WEIGHTED		UNWEIGHTED		
	SILC	HBS	SILC	HBS	SILC	HBS	SILC	HBS	
NUM_EMP	*	*	*		*		*		5/8
COMP	*	*	*	*	*	*	*	*	8/8
DISH_W	*	*	*	*	*	*	*	*	8/8
CAR	*	*	*	*	*	*	*	*	8/8
DIS_INC_CAT	*	*	*	*	*	*	*	*	8/8
INTERNET		*		*		*		*	4/8
NUM_ADU					*	*	*	*	4/8
REF_EDU						*		*	2/8

Eight different regression models for selected nine variables are run in order to find out variables that have more explanatory power compared to other common variables (see appendix B for t values and R square results). These are:

- Ownership of computer
- Ownership of dishwasher
- Ownership of car
- Disposable income categories

These four common variables are evaluated as matching variables. According to regression results, number of individuals with employee income variable is not suitable for any weighted and unweighted model (Table 3.6.).

Effect of survey design variables will be examined in the next step in the sense of selected and unselected common variables. Besides, variables selected as a result of this process but not selected in the analyzes made with traditional methods will be used in the matching procedures. With this comparison, the effectiveness of the results of the statistical matching with the completely neglected common variables will also be investigated. It is aimed to observe the efficiency of selection procedures by using these variables in the statistical matching applications.

3.2.4.4. Effect of the Survey Design Variables

“Survey design variables are entities that can change the shape or properties of the model within a specified range during a sensitivity or optimization design study.” These are, for sample survey, information about stratum and cluster and are generally difficult to obtain from TURKSTAT. As these variables are statistically effective in the decision period of matching variables, it is aimed to include survey design variables in the elimination processes as a new approach for this study. Adding them into the regression analyses processes may result in some variables being used or not used for further processes. Outputs of new regressions, including cluster and stratum information in the models, indicate significant changes in the results as expected.

The results in the Table 3.7. should be evaluated from three different perspectives. The first perspective is to reach the same matching variables as the variables obtained by traditional methods. This situation will be demonstrated that these methods are effective. The second side is to reach different variables from the variables obtained by traditional methods. Both results will be tested in this case. The third perspective is not to reach variables obtained by traditional methods. Selected and unselected variables will be tested and each method will be evaluated according to SM results.

Two variables “number of adults and ownership of internet” previously considered as non-representative found as important variables for HBS contrary to the analysis done by without considering survey design variables. Selected four matching variables “ownership of computer, ownership of dish washer, ownership of car and disposable income categories” showed good results, similar to the previous results. When the survey design variables included the regression analysis for the SILC survey, six common variables gave better results to be matching variables. 4 of 6 variables are the same variables discovered by traditional methods that ignore survey design variables, yet, “number of adults and number of employed people” are found as a result of taking into account the complex sample design (see appendix B for regression results).

Table 3.7. Significant Variables According to the Regression Results

VARIABLES	SILC_DV	HBS_DV
NUM_ADU	*	*
NUM_EMP	*	
NUM_EMP_INC		
REF_EDU		
COMPUTER	*	*
INTERNET		*
DISH_W	*	*
CAR	*	*
DIS_INC_CAT	*	*[9]

3.3. Statistical Matching

The concept of statistical matching is based on two data sets (A, B) including three variables (X, Y, Z). Donor and recipient data sets have both candidate common variables X= (X_1, $X_{2,}$ …. Xp), besides each data has only unique variables Y and Z respectively (Figure 3.3.). Y variable is missing for data B and Z is for A. The target variables (Y, Z) and common variables Xs are aimed to be obtained in a synthetic file altogether especially in micro-level and non-parametric matching. From a general perspective, the object of the statistical matching can vary over a wide range from to probe inferences about the relationship between Y variables in A and B datasets, joint distribution of them to correlation matrix or contingency tables of variables particularly in the parametric matching method according to final goal of the researchers of national offices. The assumption is that A and B are independent samples and the distribution of them is identical. Estimation of $f(X;Y1;Y2)$ joint distribution is established on the assumption that observations are obtained from the same distribution. The mechanism is characterized and expressed elaborately as a presence of missing data and an absence of (X, Y, Z) synthetic data jointly (Alpman et al., 2001).

[9] Cells marked with * represent variables that are statistically representative according to the regression results.

Figure 3.3. Work Flow of the Statistical Matching

Y_1,Y_2,...Yq	X_1, X_2, ...Xq	NOT AVAILABLE	HBS DATA
NOT AVAILABLE	X_1, X_2, ...Xq	Z_1, Z_2,...,Zq	SILC DATA
SM APPLICATIONS			
⇩	⇩	⇩	
(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), ..., (Xq, Yq, Zq)			SYNTHETIC DATA

To express more clearly, the statistical matching algorithm has some assumptions regarding the structure of data sets and models. As the target and common variables are not observed jointly, it means that there may be different possible joint distributions that are not identifiable by marginal distributions of the variables meaning that the relationship is not proper for estimation. In addition to the inestimable type of the structure, variation of $f(X;Y1;Y2)$ models causes a testing issue on which is suitable or not. Despite the complications mentioned above, various statistical ways have been tried to overcome these problems.

Conditional independence assumption, is a common and heavily used assumption to cope with the unidentifiable model issue of A∪B suggesting that Y and Z are independent variables and these target variables are conditional on the common variables in the absence of a third auxiliary data (Dawid, 1979). The structure of the density function under CIA which assumes Xs have sufficient information to explain any possible relationship between Y and Z and Xs are closely related to Y and Z is based on the Formula 3.4.

$$f(X,Y,Z) = fY|\mathrm{X}\ (Y|\mathrm{X})\ f\ Z|\mathrm{X}(\mathrm{Z}|\mathrm{X})\ f(X)X_{,}\ \forall\ \mathrm{X}\ \in\ X,\mathrm{Y}\ \in\ Y,\mathrm{Z}\ \in\ Z \qquad (3.4)$$

CIA approach based on an assumption considered mostly a wrong assumption causing bias because of misspecification of the model and regression coefficients of target variables are null with zero mean and variance.

Auxiliary data or information, refers to a third source providing information about $f(X;Y;Z)$ to use in the model in case of donor and recipient data do not have sufficient information of them. It is an effective way in order not to depend on CIA or other assumptions regarding models. A third source which has (X, Y, Z) or only (Y, Z) jointly, to explain in more detail, common and target variables or only target variables could be observed at the same time, is a way of solution to implement auxiliary data procedures in SM. Since the additional data is coming from external sources and the logic of SM is fundamentally constituted on the lack of enough data, the approach is not applicable for all conditions. The link between proxy variables, i.e. jointly observed variables Y and Z that are predicted to be distributed similarly to Y and Z, may provide plausible values for the inestimable parameters (D'Orazio et al., 2001). The mechanism of the auxiliary information "as a small overlap of the units" can be seen in the Table 3.8.

Since it is not possible to get an external source to use as a small overlap in all cases, researchers working generally for the methodologic purposes, use available data sources divided into three parts. Randomly selected data sets (A, B, C) are used for data matching procedures. Selected variables were removed from A and B data sets. C data, as it has inestimable parameters and joint observations of common and target variables, is used as auxiliary data.

Table 3.8. Usage of Auxiliary Data

Y1a,Y2a,Y3a,...	X1a, X2a, X3a,...	NOT AVAILABLE	SILC DATA
Y1c,Y2c,Y3c,...	X1c, X2c, X2c,...	Z1c,Z2c,Z3c,...	OVERLAP DATA
NOT AVAILABLE	X1b, X2b, X3b,...	Z1b,Z2b,Z3b,...	HBS DATA

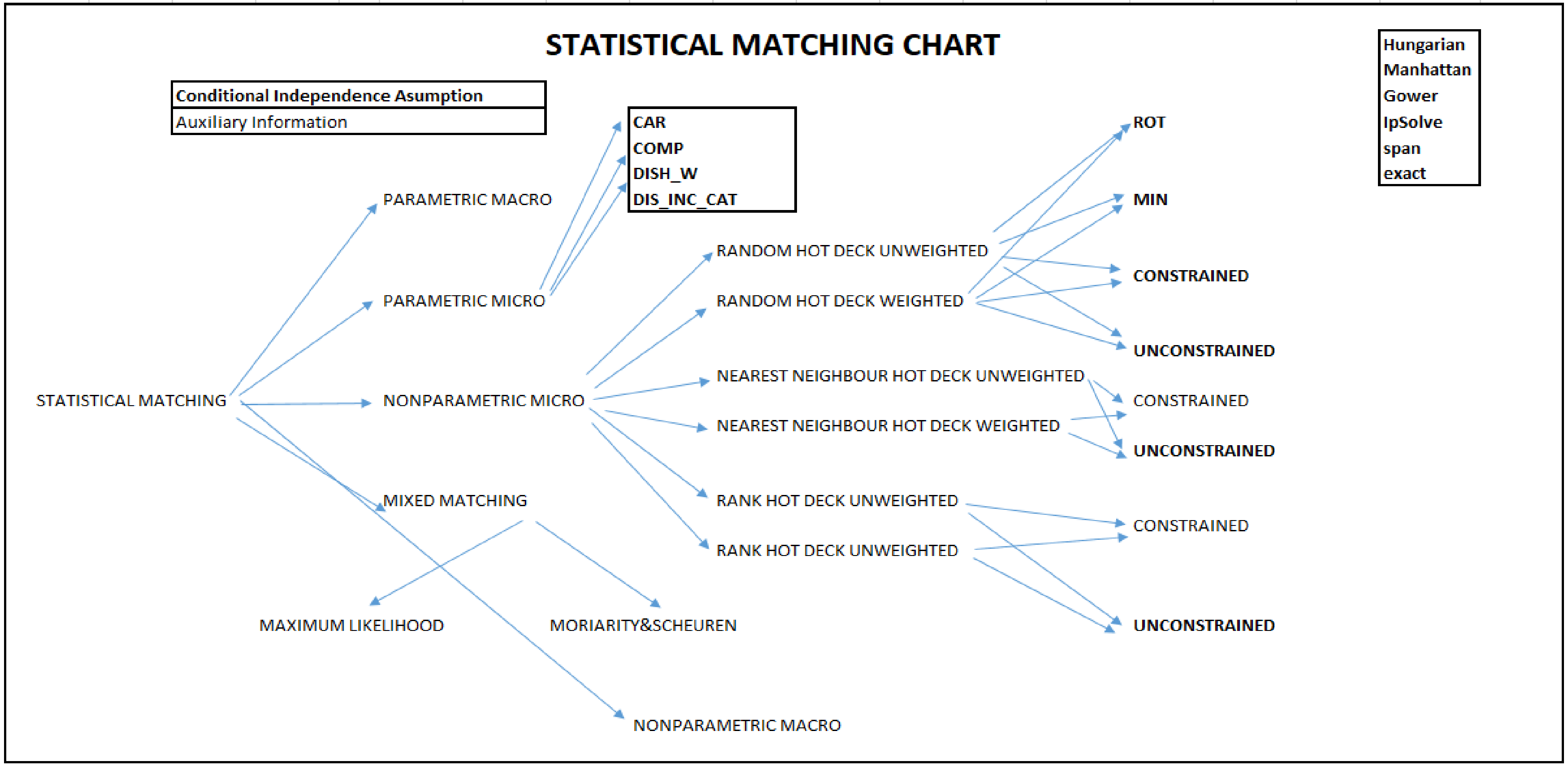
Uncertainty interval, is a notion denoting multiplicity of logically proper models which gives the existing information that could be presented by an uncertainty interval. When CIA assumption is identifiable and or not suitable and no auxiliary data exist currently, the conditions may cause uncertainty for the model. Estimations of marginal distributions for partially observed variables could be obtained from donor and recipient data, here HBS and SILC respectively, even though unknown true values were imputed. This situation limited the possible estimations causes that real parameters lie in an interval called “uncertainty interval”.

Having explained the general frame of statistical matching and basic concepts and notions[10] of the approach, each indispensable part of the method probed in detail, is figured out in the Figure 3.4. Basically they are classified as macro and micro methods, also titled as parametric, not parametric and mixed approaches. In the following subsections, the ordered methods will be explained both methodologically and the

[10] Statistical notions are based on the approach of D’Orazio (2006).

program codes of the applications and output of related approaches will be shown as summary results.

Figure 3.4. Statistical Matching Chart[11]



[11] All options given in bold have been applied.

3.3.1. Parametric Micro Approach

Obtaining a completed data set of matching and target variables together (X, Y, Z) is the fundamental goal of the predictive approach by imputing missing values of target variables. Having constructed a parametric model to estimate, missing consumption expenditure items of A and missing income items of B is predicted for A∪B by exploiting estimated joint distributions of currently observed variables. When examined in outline, two subcategories come into prominence in the parametric predictive family: conditional mean matching (CMM) and draws based on a predictive distribution (DBPD).

In CMM procedures, the determination of substitution mechanism depends basically on the expectation of missing items which have observations and information per unit. The necessary parameters related to target variables are estimated by MLE. Although it is very beneficial, the disability of the predictive approach of CMM is that distribution of predictive values of target variables has too much dependency and concentration on the expected values of them. On the condition that Y and Z variables are continuous, the substitution of missing items is done by expectation of variable which is not available given observed variables. Imputation procedure is based on the Formula 3.5.

$$\bar{z}_a = E \langle Z|X = x_a \rangle = \int_z z\, f z|x(z \mid x_a;\ \theta z \mid x\,)dz, \qquad a = 1, \dots, n_A,$$

$$\bar{\mathrm{y}}_{\mathrm{b}} = E \langle Y|X = x_b \rangle = \int_y y\, f y|x(y \mid x_b;\ \theta y \mid x\,)dy, \qquad b = 1, \dots, n_B, \qquad (3.5)$$

Draws based on a predictive distribution approach were developed by taking into account the drawbacks of the CMM method. Under the assumption that the missing mechanism of a partially observed sample is MCAR or MAR, the data generating distribution f (X; Y1; Y2) is preserved in a better way by imputing missing values with a random draw from a predictive distribution. (Kim, 2018). The method, referring to a stochastic regression imputation, is convenient when X, Y and Z are multinormal.

Having expressed the statistical notions of the parametric micro approach, R codes of the method in Statmatch package implemented on SILC and HBS data are demonstrated below to practice stochastic regression imputation in R.

Type of data and variables in R environment is significant and the first procedure is to regulate related data sets as data frame. Besides, matching variables are regulated as numeric according to the needs of the regression function in order not to take syntax errors in correlation matrix of A and B. Structures of data sets are represented in the Table 3.9. and Table 3.10.

Table 3.9. Structure of SILC Data

data.frame':	24068 obs. of 9 variables:
$ COMP	“: Factor w/ 2 levels "1","2": 2 2 1 2 1 1 1 2 1 2 ...
$ DISH_W	: Factor w/ 2 levels "1","2": 1 1 1 2 1 1 2 2 1 1 ...
$ CAR	: Factor w/ 2 levels "1","2": 2 1 2 2 1 1 2 2 1 2 ...
$ DIS_INC_CAT	: Factor w/ 6 levels "1","2","3","4",..: 2 3 4 2 6 4 4...
$ YINCOME	: num 1643 2088 3487 1898 6994 ...
$ BIRIMNO/BULTEN NO	: num 2 3 4 5 6 7 8 9 10 11 ...
$ HANE_AGIRLIK	: num 1693 1017 1307 1099 1483 ...
$ HSIZE	: num 1 2 3 2 5 4 5 1 2 2 ...
$ REF_EDU	: num 1 2 2 1 4 3 2 1 2 2 ...”

Table 3.10. Structure of HBS Data

data.frame'	11828 obs. of 9 variables:
$ COMP	“: Factor w/ 2 levels "1","2": 2 1 2 1 2 1 2 2 2 1 ...
$ DISH_W	: Factor w/ 2 levels "1","2": 2 2 2 1 2 1 1 2 1 1 ...
$ CAR	: Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 2 2 2 ...
$ DIS_INC_CAT	: Factor w/ 6 levels "1","2","3","4",..: 2 4 5 1 6 6 3 3 3 4 ...
$ ZCONSUMPTION	: num 1022 1938 3753 777 2354 ...
$ BIRIMNO/BULTEN NO	: num 1800001 1800002 1800003 1800004 1800005 ...
$ HANE_AGIRLIK	: num 3225 2591 1660 1850 2509 ...
$ HSIZE	: num 1 4 2 2 3 4 1 2 3 3 …
$ REF_EDU	: num 2 2 2 2 2 2 2 3 2 2 ...”

As “cor” function, computing correlation between the variables, needs Xs as numeric, four variables are adapted to the situation for parametric micro approach. Regression is run in A data for Y vs X.

```
> reg.yx <- lm(YINCOME~CAR,data=AA)

> coefficients(reg.yx)
      (Intercept)     CAR
        9428.579   -2984.882
> outp <- summary(reg.yx)
> outp$sigma ‶residual sd s_Y|X″  4917.23
> predyB <- predict(reg.yx, newdata=BB)  ‶predicted values″
> impyB <- predyB + rnorm(nrow(BB), mean=0, sd=outp$sigma)
> BB$YINCOME <- impyB ‶filling Y in BB″
```

The same procedure is run in data B for YZ vs. X.

```
> reg.zx <- lm(ZCONSUMPTION~CAR, data=BB)
> coefficients(reg.zx)
  (Intercept)   CAR
  5697.353   -2625.395
> outp <- summary(reg.zx)
> outp$sigma ‶residual sd s_Z|X″ 3429.878
> predzA <- predict(reg.zx, newdata=AA) ‶predicted values″
> impzA <- predzA + rnorm(nrow(AA), mean=0, sd=outp$sigma)
> AA$ZCONSUMPTION <- impzA ‶fill in Z in AA″
```

Results could be seen using following codes. Concatenation of datasets and estimated var-cov of datasets can be seen in the Table 3.11. and Table 3.12. respectively.

```
> A∪B <- rbind(AA,BB) ‶concatenation AA ∪ BB″
> head(A∪B)
```

Table 3.11. Concatenation of Datasets

COMP	DISH_W	CAR	DIS_IC	Y_INC	WEIGHT	Z_CONS
2	2	2	1	897,17860	777,30255	258,13085
2	1	2	5	4859,00370	2061,54484	486,56654
2	1	1	3	2996,15614	2188,83267	1850,58927
2	2	2	3	2033,18156	381,55525	2463,58244
2	1	2	2	1603,53746	1021,23675	3407,05024
1	1	1	4	3190,17363	796,26690	3407,13116

```
> cor(A∪B) “estimated var-cov”
```

Table 3.12. Estimated Variance-Covariance of Datasets

	COMP	DISH_W	CAR	DIS_IC	Y_INC	Z_CONS
COMP	1.00000	0.33680	0.27536	-0.44399	-0.31355	-0.32664
DISH_W	0.33680	1.00000	0.26377	-0.41628	-0.20187	-0.16455
CAR	0.27536	0.26377	1.00000	-0.39667	-0.21963	-0.18031
DIS_IC	-0.44440	-0.41628	-0.39667	1.00000	0.41778	0.25517
Y_INC	-0.31355	-0.20187	-0.21963	0.41778	1.00000	0.09909
BIRIMNO	0.00838	0.01824	-0.00592	0.03139	-0.00004	-0.00232
WEIGHT	-0.11129	-0.07504	-0.00569	0.13904	0.07153	0.06022
HHSIZE	-0.13364	-0.07038	-0.14506	0.23538	0.08675	0.06926
REF_EDU	-0.11102	-0.33599	-0.29599	0.46551	0.27702	0.20201
Z_CONS	-0.32664	-0.16455	-0.18031	0.25517	0.09909	1.00000

Procedures can be repeated for other matching variables. Rassler (2002) offers to check marginal and joint distribution of variables for validation. Marginal distribution of the imputed variables indicates that results are coherent with the original distribution of the data as it is seen in the Figure 3.5. and Figure 3.6. Since household size and education information are important demographic indicators, these two variables are selected for comparison and validation.

Figure 3.5. Marginal Distribution of Parametric Micro Results of Household Size Variable

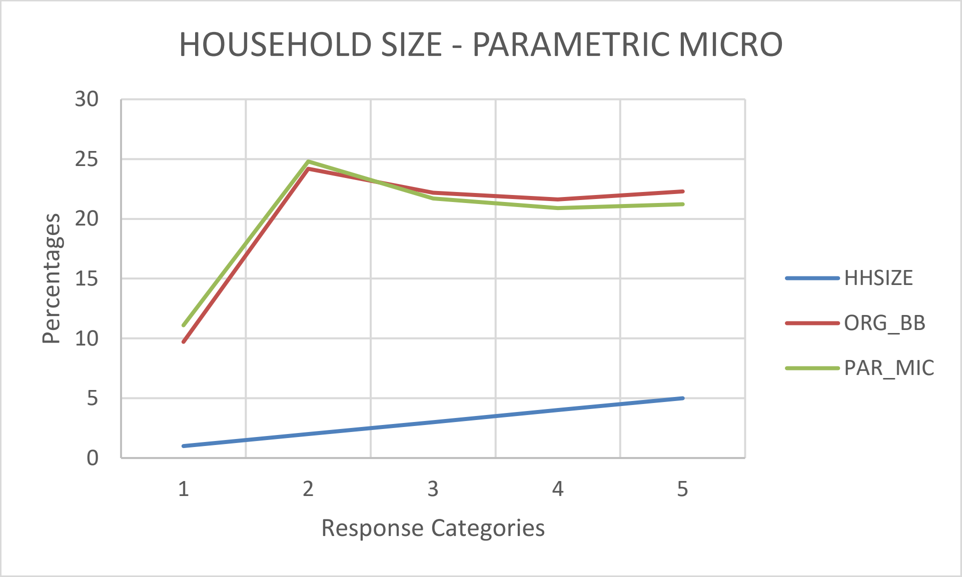
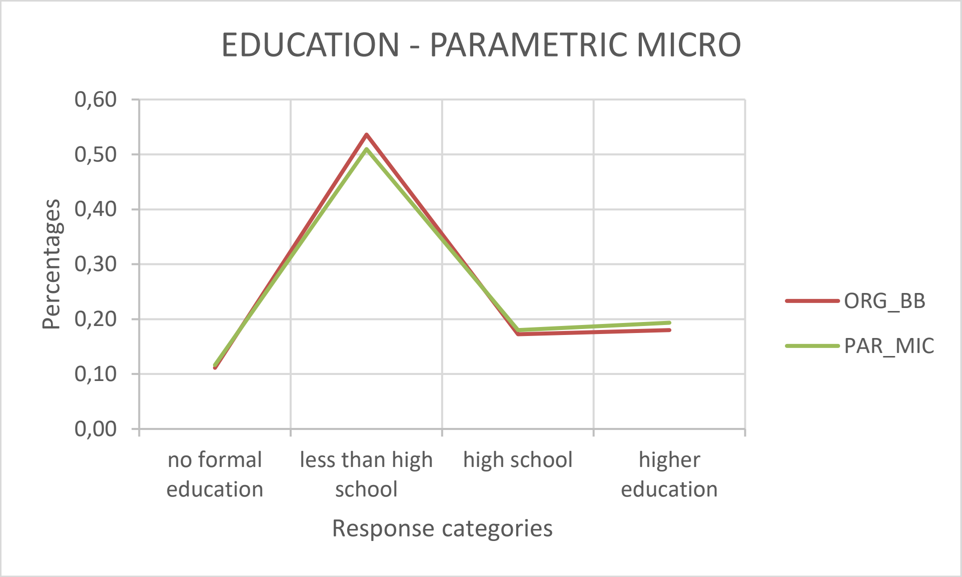


Figure 3.6. Marginal Distribution of Parametric Micro Results of Education Variable



3.3.2. Parametric Macro Approach

Models in the parametric approach of statistical matching are required to adoption of them for (X; Y; Z) explicitly, otherwise misspecification occurs and unreliable outputs came out. Maximum likelihood method is beneficial to obtain estimations in the parametric model.

$$f(x, y, z;\ \theta) = fx\ (x;\ \theta x)\ fyz|x\ (y, z|\text{x};\ \theta yz|x)_{,} \qquad (3.6)$$

Equation of density function depending on auxiliary information or CIA assumption is based on the Formula 3.6. Under CIA, observed likelihood function of A∪B is based on the Formula 3.7.

$$L\,(\theta \mid A \cup B\,) = \prod_{a=1}^{n_A} fxy(\,x_a, y_a; \theta) \prod_{b=1}^{n_B} fxz(\,x_b, z_b; \theta)$$

$$= \prod_{a=1}^{n_A} fy|x(\,y_a|x_a; \theta_{Y|X}) \prod_{b=1}^{n_B} fz|x(\,z_b|x_b; \theta_{Z|X})$$

$$x \prod_{a=1}^{n_A} fx(\,x_a; \theta_X) \prod_{b=1}^{n_B} fx(\,x_b; \theta_X) \qquad (3.7)$$

Despite the fact that missing items affect the data set A∪B, maximum likelihood estimates of the parameters θX, θY|X and θZ|X can be calculated directly from appropriate selections of complete data without using iterative procedures (D'Orazio, 2017).

In case of existence of a third file (C), computation of the function uses (n_A + n_B + $n_{C)}$ and (n_A + $n_{B)}$ on the conditions of auxiliary information. Type of the auxiliary source (sample or only parametric information[12]) designs the model and approaches to be implemented. When (X, Y, Z) observed together in C file, likelihood function is based on the Formula 3.8.

$$L\,(\theta \mid A \cup B \cup C\,) = \prod_{a=1}^{n_A} fx(\,x_a, \theta_X) \prod_{b=1}^{n_B} fx(\,x_b, \theta_X) \prod_{c=1}^{n_C} fx(\,x_c, \theta_X)$$

$$x \int_z \prod_{a=1}^{n_A} fyz|x(\,y_a, t|x_a; \theta_{YZ|X})dt$$

$$x \int_y \prod_{a=1}^{n_B} fyz|x(\,t, z_b, |x_b; \theta_{YZ|X})dt$$

$$x \prod_{a=1}^{n_C} fyz|x(\,y_c; z_c|x_c; \theta_{YZ|X}) \qquad (3.8)$$

[12] "*Parametric information* is a kind of knowledge about the structure of the current survey or data gained from previous samples or proxy variables".

R codes for parametric macro approach can be written both for ML and MS[13] approaches.

```
> MS Method
> x.mtc.MS<- c("CAR","COMP","DISH_W","DIS_INC_CAT")
> mix.MS <- mixed.mtc(data.rec=AA, data.don=BB,
              match.vars=x.mtc.MS,y.rec="YINCOME",
              z.don="ZCONSUMPTION", method="MS",rho.yz=0, micro=FALSE,
              constr.alg="lpSolve")
input value for rho.yz is 0
low(rho.yz)= -0.4172
up(rho.yz)= 0.9894
The input value for rho.yz is admissible

> names(mix.MS)
"rho.yz"  "mu"      "vc"      "cor"     "phi"     "res.var" "call"

> mix.MS$rho.yz
  start low.lim  up.lim    used
 0.0000 -0.4172  0.9894  0.0000

> mix.MS$mu “estimated means”
CAR       COMP     DISH_W     DIS_INC_CAT  YINCOME     ZCONSUMPTION
1.572292  1.593437 1.307472  4.171607    4729.138906  4206.196097
>mix.MS$vc “estimated var-cov matrix”
```

Estimated var-cov matrix of the mixed MS method is represented in the Table 3.13. (see Table 3.2. for the abbreviations of the variables).

Table 3.13. Estimated Variance-Covariance Matrix with Moriarty and Scheuren

VAR.	CAR	COMP	DISH_W	DIS_IC	Y_INC	Z_CONS
CAR	1000000	0.2753588	0.2637714	-0.3966697	-0.2872941	-0.3550150
COMP	0.2753580	1000000	0.3367984	-0.4443987	-0.3124438	-0.3236434
DISH_W	0.2637714	0.3367984	1000000	-0.4162832	-0.2500505	-0.2701376
DIS_IC	-0.3966697	-0.4443987	-0.4162832	1000000	0.5561880	0.4789605
Y_INC	-0.2872942	-0.3124438	-0.2500505	0.5561880	1000000	0.0000000
Z_CONS	-0.3550150	-0.3236434	-0.2701376	0.4789605	0.0000000	1000000

[13] MS is the abbreviation for Moriarty and Scheuren method used in multiple imputation procedures to determine upper and lower bounds (Moriarty and Scheuren, 2001).

```
> ML Method
> x.mtc.ML<- c("CAR","COMP","DISH_W","DIS_INC_CAT")


> mix.ML <- mixed.mtc(data.rec=AA, data.don=BB,
            match.vars=x.mtc.MS,y.rec="YINCOME",
            z.don="ZCONSUMPTION", method="ML",rho.yz=0, micro=FALSE,
            constr.alg="lpSolve")
input value for rho.yz is 0
low(rho.yz)= -0.4172
up(rho.yz)= 0.9894
The input value for rho.yz is admissible

> names(mix.ML)
"rho.yz"  "mu"      "vc"      "cor"     "phi"     "res.var" "call"
> mix.ML$rho.yz
  start low.lim  up.lim    used
 0.0000 -0.4172  0.9894  0.0000

> mix.ML$mu “estimated means”
CAR       COMP      DISH_W     DIS_INC_CAT  YINCOME     ZCONSUMPTION
1.572292     1.593437     1.307472     4.171607  4784.531225  4149.378999
>mix.MS$vc “estimated var-cov matrix”
```

Estimated var-cov matrix of the mixed ML method is represented in the Table 3.14. (see Table 3.2. for the abbreviations of the variables).

Table 3.14. Estimated Variance-Covariance Matrix with Maximum Likelihood

VAR.	**CAR**	**COMP**	**DISH_W**	**DIS_IC**	**Y_INC**	**Z_CONS**
CAR	10000000	0.2753588	0.2637714	- 0.3966697	- 0.2835569	- 0.3622077
COMP	0.2753588	10000000	0.3367983	- 0.4443986	- 0.3064922	- 0.3351010
DISH_W	0.2637714	0.3367984	10000000	- 0.4162831	- 0.2390964	- 0.2883291
DIS_IC	- 0.3966697	- 0.4443987	- 0.4162831	10000000	0.5447100	0.5032551
Y_INC	- 0.2835569	- 0.3064922	- 0.2390964	0.5447100	10000000	0.2934385
Z_CONS	- 0.3620774	- 0.3351010	- 0.2883291	0.5032555	0.2934385	10000000

Parametric macro approach provides us with only model parameters contrary to micro approaches enable to gain micro file including X, Y and Z together in a complete set of data.

3.3.3. Nonparametric Micro Approach

Among macro, micro and mixed matching method, non-parametric approach is a very common practice especially when the purpose is to create a synthetic data set of (X, Y, Z) instead of correlation matrix, model parameters, etc. This popularity has been increasing in recent decades due to the fact that it enables result-oriented micro level file production called “synthetic data set”. Since the objective of researchers and national officers focuses on the new data sets without field organization of survey, programs and functions including nonparametric micro applications have developed rapidly. The method, in essence, is a subject related to imputation procedures from donor to recipient data. The general concept is that recipient data has missing items of Z and donor data has these items; transferring these Z values according to SM procedures. As hot deck imputation methods do not need any parametric distribution or density function, non-parametric micro approach is a very applicable and effective way of filling the missing values in the recipient data. After diagnosing the A and B data sets, donor and recipient data are selected, harmonization and elimination procedures are applied, then random, rank or distance hot deck methods are selected to acquire synthetic data set of (X, Y, Z).

3.3.3.1. Random Hot Deck

Random hot deck is a very effective and frequently used method among imputation methods. Imputation processes involve the construction of donor and recipient data sets for each record as imputation classes using simple random selection from donor data. The group of prospective donors comprises of the units within the same class with y observed. Auxiliary variables are exploited for each recipient unit I in a certain imputation class. A random donor is chosen from among these prospective donors (by equal-probability sampling) and is then used to impute the recipient. The approach assumes that all auxiliary variables used to determine the imputation classes have exactly the same values for both the donor and the recipient. The donor is chosen entirely at random, subject to certain auxiliary variables (Memobust, 2014).

Random hot deck imputation procedures are used in non-parametric statistical matching applications extensively. There are a lot of experiments using the method to create a synthetic data set at micro level. D'Orazio (2017) has stated that:

> *"Random hot deck consists in randomly choosing a donor record (in the donor file) for each record in the recipient file. Sometimes the random choice is made within a suitable subset of units in the donor files. In particular, units of both the files are usually grouped into homogeneous subsets according to given common characteristics (units in the same geographical area, individuals with the same demographic characteristics, etc.); defined as donation classes. Thus, for an individual in a given geographical area, only records in the same area will be considered as possible donors. In general, the donation classes are defined using one or few categorical variables* ***X*** *chosen within the set of common variables in A and B."*

Random hot deck procedures of non-parametric micro approach are implemented at maximum variety. It is possible to use donor classes by grouped matching variables. Even if these techniques depend on a distance measure calculated by use of the matching variables, it is possible to make different interventions to the process with the help of changes to be made in the codes. There are several options in the hot deck procedure of Statmatch package. "Rot" option takes into account a subset of closest donors, "span" considers a proportion k of the closest possible donors, "exact" uses k closest donors and finally "min" works on the basis of minimum distance. Addition to these options, "k.dist" uses k which has equal or less distance from the recipient.

R codes of non-parametric micro approach can be expressed both for weighted and unweighted (see appendix D for detailed R codes).

```
> Unweighted random hot deck with donor classes

> group_1 <- c("CAR")
> group_2 <- c("CAR", "COMP")
> group_3 <- c("CAR", "COMP", "DISH_W")
> group_4 <- c("CAR", "COMP", "DISH_W", "DIS_INC_CAT") # donation
            classes #
> out.rnd_1<- RANDwNND.hotdeck(data.rec = AA, data.don = BB,
            don.class= group_1)

> fA.rnd_1 <- create.fused(data.rec=AA,
            data.don=BB,mtc.ids=out.rnd_1$mtc.ids,z.vars="ZCONSUMPTI
            ON")

> Weighted random hot deck with donor classes

> out.rnd_1w <- RANDwNND.hotdeck(data.rec=AA, data.don=BB,
            match.vars=NULL,don.class=group_1,
            weight.don="HANE_AGIRLIK")

> fA.rnd_1w <- create.fused(data.rec=AA, data.don=BB,
            mtc.ids= out.rnd_1w$mtc.ids, z.vars="ZCONSUMPTION")


> “rot” option

> rnd_opt_2 <- RANDwNND.hotdeck(data.rec=AA, data.don=BB,
           match.vars=X.mtc,don.class=group_5,dist.fun="gower"[14],
           cut.don="rot")
```

When joint distribution of Xs is examined, it indicates that the results of random hot deck method is significant and percentages of categories appear to be similar or very close (Table 3.15.).

[14] Gower distance is a number between 0 and 1 used to measure differences of 2 variables or records (Gower, 1971).

Table 3.15. Joint Distribution of Matching Variables

CAR-COMP-DISH_W / DIS_INC_CAT (ORIGINAL BB)						
CAR	**<=1000**	**>1000 and <=2000**	**>2000 and <=3000**	**>3000 and <=4000**	**>4000 and <=5000**	**>5000**
1	0.07	0.13	0.28	0.40	0.51	0.66
2	0.93	0.87	0.72	0.60	0.49	0.34
COMP						
1	0.06	0.09	0.20	0.37	0.48	0.66
2	0.94	0.91	0.80	0.63	0.52	0.34
DISH_W						
1	0.18	0.35	0.57	0.70	0.76	0.87
2	0.82	0.65	0.43	0.30	0.24	0.13
CAR-COMP-DISH_W / DIS_INC_CAT (SYNTHETIC RANDOM HD)						
CAR	**<=1000**	**>1000 and <=2000**	**>2000 and <=3000**	**>3000 and <=4000**	**>4000 and <=5000**	**>5000**
1	0.07	0.14	0.28	0.42	0.50	0.67
2	0.93	0.86	0.72	0.58	0.50	0.33
COMP						
1	0.07	0.12	0.23	0.37	0.50	0.70
2	0.93	0.88	0.77	0.63	0.50	0.30
DISH_W						
1	0.17	0.36	0.61	0.74	0.82	0.90
2	0.83	0.64	0.39	0.26	0.18	0.10

3.3.3.2. Rank Hot Deck

The rank hot deck imputation technique is used to fuse variables to the synthetic micro data set looking for the donor at a minimum distance from the provided recipient record. Distances are calculated using percentage points from the empirical cumulative distribution function of the unique common variable. To be more specific, the donor is picked so that the space between percentage points in the empirical distribution is as small as possible (Singh et al., 1993).

The empirical cumulative distribution function of the distribution of *X* in the recipient and donor file is formed respectively in the Formula 3.9.

$$\hat{F}_X^A(x) = \frac{1}{n_A} \sum_{a=1}^{n_A} I(x_a \; \leq x \,), \;\; x \; \in X)$$

$$\hat{F}_X^B(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I(x_b \; \leq x \,), \;\; x \; \in X) \tag{3.9}$$

R codes of the rank hot deck imputation procedure are generally expressed below (see appendix D for detailed preliminary codes).

```
> rank hot deck unweighted
> group.rnk <- c("CAR","COMP","DISH_W", "DIS_INC_CAT")
> rnk.a <- rankNND.hotdeck(data.rec=AA,
            data.don=BB,var.rec="X.mtc.a_n",
            var.don="X.mtc.a_n",don.class = group.rnk)

> fA.rnk.a <- create.fused(data.rec=AA,
            data.don=BB,mtc.ids=rnk.a$mtc.ids,z.vars="ZCONSUMPTION",
            dup.x=TRUE, match.vars=X.mtc_a)

> rank hot deck weighted

> rnk.wa <- rankNND.hotdeck(data.rec=AA,
      data.don=BB,var.rec="X.mtc.a_n", var.don="X.mtc.a_n",
      don.class = group.rnk,
      weight.rec="HANE_AGIRLIK",weight.don="HANE_AGIRLIK")

> fA.rnk.wa <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.wa$mtc.ids,z.vars="ZCONSUMPTION",
      dup.x=TRUE, match.vars=X.mtc_a)
```

Marginal distribution of imputed values for household size and education variables could be seen in the Table 3.16. and Table 3.17.

Table 3.16. Marginal Distribution of Imputed Values of Household Size Variable

HOUSEHOLD SIZE	ORIGINAL DATA	SYNTHETIC DATA (RANK)
1	9.7	11.1
2	24.2	24.8
3	22.2	21.7
4	21.6	20.9
5	22.3	21.2

Table 3.17. Marginal Distribution of Imputed Values of Education Variable

EDUCATION	ORIGINAL DATA	SYNTHETIC DATA (RANK)
1	11.1	11.6
2	53.6	50.9
3	17.3	18.0
4	18.0	19.3

3.3.3.3. Nearest Neighbor Distance Hot Deck

Nearest neighbor distance has been a common practice field in the very early experiments of SM applications. Distance functions are computed by use of variables, instead of the limitation that both data have identical scores on each auxiliary variable. The closest records in A data are used to impute B data. The mechanism of the imputation is based on the Formula 3.10.

$$d_{ab^*} = \left|\mathrm{x}_{\mathrm{a}}^{\mathrm{A}} - \mathrm{x}_{\mathrm{b}^*}^{\mathrm{B}}\right| = \min_{1\leq \mathrm{b} \leq \mathrm{nb}} \left|\mathrm{x}_{\mathrm{a}}^{\mathrm{A}} - \mathrm{x}_{\mathrm{b}}^{\mathrm{B}}\right|$$

(3.10)

While in the unconstrained distance hot deck approach, each record in the B file could be used more than one time, constrained distance hot deck approach gives only one access to each record in B as donor.

```
X.mtc_a <- c("CAR")
X.mtc_b <- c("CAR", "COMP")
X.mtc_c <- c("CAR", "COMP", "DISH_W")
X.mtc_d <- c("CAR", "COMP", "DISH_W", "DIS_INC_CAT")
```

```
group.nnd <- c("CAR","COMP","DISH_W", "DIS_INC_CAT")

> unweighted distance hot deck
> out.nnd_a <- NND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_a, don.class=group.nnd,
      dist.fun="Gower")

> fA.nnd_a <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=out.nnd_a$mtc.ids,match.vars = group.nnd,
      z.vars="ZCONSUMPTION")> weighted distance hot deck



> weighted distance hot deck
> out.nnd.wa <- NND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_a, don.class=group.nnd,
      dist.fun="Gower",weight.rec="HANE_AGIRLIK",weight.don="HANE_AG
      IRLIK")

> fA.nnd.wa <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=out.nnd.wa$mtc.ids,match.vars = group.nnd,
      z.vars="ZCONSUMPTION")
```

3.3.4. Nonparametric Macro Approach

It is possible to get output if the family F of distribution of interest is parametric. In fact, the multinomial distribution is quite versatile when the variables are categorical or discrete. Nevertheless, when the variables dealt with are continuous, there might not be enough information to limit F to a parametric family of distributions (e.g. the multinormal). Nonparametric approaches are preferred in such instances because they are unaffected by erroneous assumptions about F's parametric structure. There are two techniques to think about named as macro and micro. Micro technique to solve the statistical matching issue has been widely employed. Macro approach, on the other hand, has gotten little attention compared to micro matching (Donatiello et al., 2014).

Under CIA after factorizing, factors could be estimated by the Formula 3.11.

$$F_{YZ|X}(y,z|x = \ F_{Y|X}\ (y|x)F_{Z|X}\ (z|x)$$

$$\hat{F}_{Y|X}\ (y|x) = \frac{\sum_{a=1}^{nA} I\ (y_a \leq y) I\ (x_a = x)}{\sum_{a=1}^{nA} I\ (x_a = x)}$$

$$\hat{F}_{Z|X}\,(z|x) = \frac{\sum_{a=1}^{nB} I\,(z_b \leq z) I\,(x_b = x)}{\sum_{b=1}^{nB} I\,(x_b = x)} \qquad (3.11)$$

When X is categorical, as some categories of X are unable to observe, the empirical cumulative distribution function for the category of X is not possible to estimate. On the condition that availability of completed third data set C, marginal X distribution could be estimated by use of Kernels “$k$nn” method.

3.3.5. Mixed Matching Approach

Mixed matching method is a combination of parametric and nonparametric approaches to reach not only parameters but also micro-level synthetic file. The first step is the adaptation of a parametric model and the second phase is to apply a nonparametric hot deck option to obtain micro data set. Papers in literature cover the fictive and small sized data sets mostly as a part of reel data. During applications with large sized survey data, it has been seen that R Studio aborted the sessions due to the dimension of the matrix.

The idea behind the mixed approach which combines both parametric and non-parametric specifications is based on two pillars: parsimonious structure of parametric approach and non-parametric approach’s robust to model with less misspecification error. After fitting the model and estimating its parameters (θijk), nonparametric method evaluated as protective against misspecifications caused by models, applied. The mechanism starts with computing primary intermediate values in A and B, then final intermediate values computed and finally matching step made by ML or MS.

R codes for mixed matching approach could also be written both for ML and MS approaches.

```
> ML Method
> X.mtc <- c("CAR","COMP","DISH_W",“DIS_INC_CAT”)
> mix.ML <- mixed.mtc(data.rec=AA_MIX, data.don=BB_MIX,
            match.vars=X.mtc,y.rec="YINCOME", z.don="ZCONSUMPTION",
            method="ML",rho.yz=0, micro=TRUE, constr.alg="lpSolve")
> fill.ML <- create.fused(data.rec=AA_MIX,
          data.don=BB_MIX,mtc.ids=mix.ML$mtc.ids,
          z.vars="ZCONSUMPTION")
> cor(mix.ML$filled.rec)
> MS Method
> X.mtc <- c("CAR","COMP","DISH_W",“DIS_INC_CAT”)
> mix.MS <- mixed.mtc(data.rec=AA_MIX, data.don=BB_MIX,
          match.vars=X.mtc,y.rec="YINCOME", z.don="ZCONSUMPTION",
          method="MS",rho.yz=0, micro=TRUE, constr.alg="lpSolve")
> fill.MS <- create.fused(data.rec=AA_MIX,
          data.don=BB_MIX,mtc.ids=mix.MS$mtc.ids,
          z.vars="ZCONSUMPTION")
> cor(mix.MS$filled.rec)
```

Each method enables us with synthetic completed data set of X, Y and Z. When the distributions of synthetic and original data are examined, results of mixed matching are successful (Table 3.18. and Table 3.19.).

Table 3.18. Marginal Distribution of Mixed Matching of Household Size Variable

HOUSEHOLD SIZE	ORIGINAL DATA	SYNTHETIC DATA (MIXED)
1	0.10	0.12
2	0.24	0.25
3	0.22	0.22
4	0.22	0.20
5	0.22	0.21

Table 3.19. Marginal Distribution of Mixed Matching of Education Variable

EDUCATION	ORIGINAL DATA	SYNTHETIC DATA (MIXED)
1	0.11	0.12
2	0.54	0.53
3	0.17	0.17
4	0.18	0.19

3.3.6. Renssen Approach

Complex sample design issue could be solved by a few methods. In the study of (Rubin, 1986), empirical likelihood method was adopted for file concatenation as (Wu,

2004) did. Renssen method, however, provides us with two opportunities: taking into account complex structure of samples as much as possible and exploiting of auxiliary sample or information (Renssen, 1998). Before starting the procedures, the current situation of data sets is controlled to find out whether the two data sets are in good harmony. Statistical indices in the Table 3.18. are used to demonstrate similarity and dissimilarity (harmonization of data sets). Total variation distance (tvd) is a dissimilarity index that ranges from 0 to 1. Zero means completely similar and 1 means completely dissimilar. Overlap is the opposite of tvd. Bhattacharyya coefficient (Bhatt) measures the similarity too. It ranges from 0 (less similar)to 1 (more similar) (Xhava, 2015). Hell means Hellinger Distance (see 3.2.4.1.)

```
<-tt.AA <- xtabs(HANE_AGIRLIK~CAR+COMP+DISH_W+DIS_INC_CAT, data=AA)
<-tt.BB <- xtabs(HANE_AGIRLIK~CAR+COMP+DISH_W+DIS_INC_CAT, data=BB)
<-(prop.table(tt.AA)-prop.table(tt.BB))*100
<-comp.prop(p1=tt.AA, p2=tt.BB, n1=nrow(AA),n2=nrow(BB), ref=FALSE)
```

Table 3.20. Overlap of SILC and HBS

	tvd	overlap	Bhatt	Hell
SRS_2000	0.089	0.910	0.990	0.096
ORIGINAL	0.059	0.941	0.997	0.057

As seen in the Table 3.20., data sets, especially original ones, are in good harmony. R codes creating and attaching survey design variables are denoted below:

```
<-svyA <- svydesign(~1, weights=~HANE_AGIRLIK, data=AA)
<-svyB <- svydesign(~1, weights=~HANE_AGIRLIK, data=BB)
```

Uçar (2017) has stressed that:

> *"The calibration in the "harmonize" operation could be carried out with three different methods, namely, "linear", "raking" and "poststratify". "Linear" option could lead to negative weights. There is a risk of convergence and the calibration may not result in "linear" and "raking" methods. "Poststratification" on the other hand avoids the problem of convergence. The downside of "poststratification" is*

> *that it may produce final weights with a higher variation. With regard to x.tot, which refers to the total population, since the exact population totals of the variables are not known, x.tot is taken as "NULL". Otherwise, population totals for each variable were to be reached."*

R codes for harmonization operation are represented below.

```
<-outhzR <- harmonize.x (svy.A=svyA, svy.B=svyB,
form.x=~CAR:COMP:DISH_W:DIS_INC_CAT-1 ,cal.method="linear")
```

Outputs of calibrations in the Table 3.21. indicate that new calibrated values are not concluded as expected before even if there is a small improvement in the original data sets.

Table 3.21. Overlap of Calibrated SILC and HBS

	tvd	overlap	Bhatt	Hell
SRS_2000	0.148	0.854	0.980	0.140
ORIGINAL	0.053	0.946	0.997	0.053

In the next step, matching procedures of Renssen are implemented. The method enables us to put a third auxiliary sample into prosesses but it was not implemented as an auxiliary sample is not available. When an auxiliary sample, containing information of X,Y and Z jointly or observed values of Y and Z, is available, it is added to comb.samp function as survey.c after attaching survey design variables into the processes in order to exploit this information.

```
<- comb.samples(svy.A=outhzR$cal.A,svy.B=outhzR$cal.B,svy.C=NULL,
       y.lab="YINCOME",z.lab="ZCONSUMPTION",form.x=~CAR:COMP:DISH_W:
       DIS_INC_CAT-1,estimation="STWS",micro="TRUE")
```

Since the outputs are not as high quality as expected, Renssen method of data matching has only been applied experimentally.

CHAPTER 4. RESULTS

On the SILC and HBS data, all SM macro and micro approaches were applied. As expressed in the introduction of the thesis, parametric and macro methods provide us with model parameters. Thus, results of micro and mixed matching methods are the focus of the thesis. Random hot deck, nearest neighbor distance hot deck and rank hot deck methods were performed both weighted and unweighted. Additionally, rot, min, exact, constrained and unconstrained variations (options) were examined. Mixed approaches were also examined in both MS and ML. Their results were compared by HD calculation scores.

Parametric micro matching approach was implemented for “car” and “comp” matching variables separately. Results indicate that it is not an effective way to match compared to nonparametric micro and mixed approaches. Therefore, nonparametric micro and mixed methods are compared in donor class distinction. Donor classes created artificially are very beneficial when there are a lot of matching variables and computations are time-consuming (D’Orazio et al., 2006). Matching variables are divided into four donor classes. The first donor class (X.mtc_a) consists of only car variable. X.mtc_d consists of four matching variables (car, comp, dish_w and dis_inc_cat). Each donor class uses only the variables specified in its content during the statistical matching implementations.

a. X.mtc_a: ("CAR")
b. X.mtc_b: ("CAR", "COMP")
c. X.mtc_c: ("CAR", "COMP", "DISH_W")
d. X.mtc_d: ("CAR", "COMP", "DISH_W", "DIS_INC_CAT")

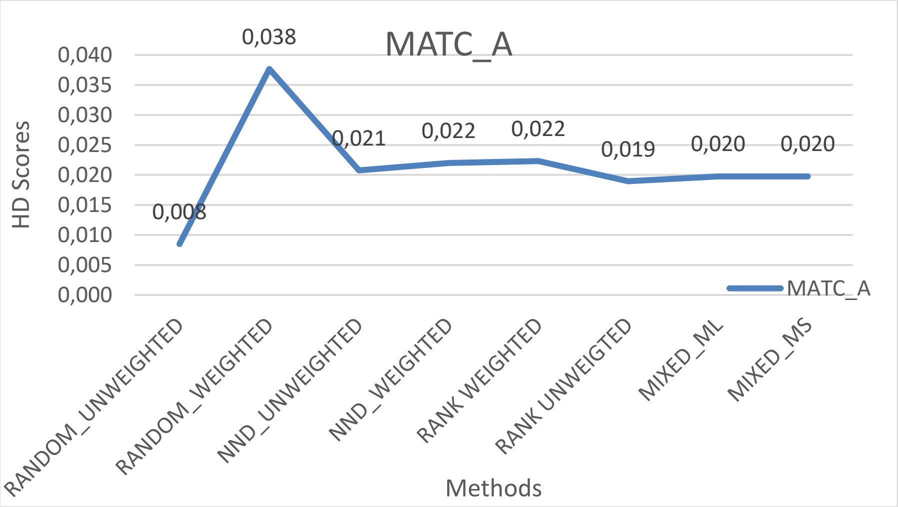
In addition to these donor classes, grouped variables used in the data matching procedure, are also significant in the results. Traditional statistical matching methods and their performances indicate that nonparametric micro and mixed procedures give better HD scores including random, rank and distance imputation methods.

Table 4.1. Weighted and Unweighted Hellinger Distance Scores of Selected Methods

METHODS	MATC_A	MATC_B	MATC_C	MATC_D
RANDOM_UNWEIGHTED	0.008	0.006	0.009	0.019
RANDOM_WEIGHTED	0.038	0.023	0.025	0.019
NND_UNWEIGHTED	0.021	0.020	0.020	0.019
NND_WEIGHTED	0.022	0.021	0.019	0.019
RANK WEIGHTED	0.022	0.022	0.022	0.022
RANK UNWEIGTED	0.019	0.018	0.020	0.020
MIXED_ML	0.020	0.020	0.020	0.020
MIXED_MS	0.020	0.020	0.020	0.020

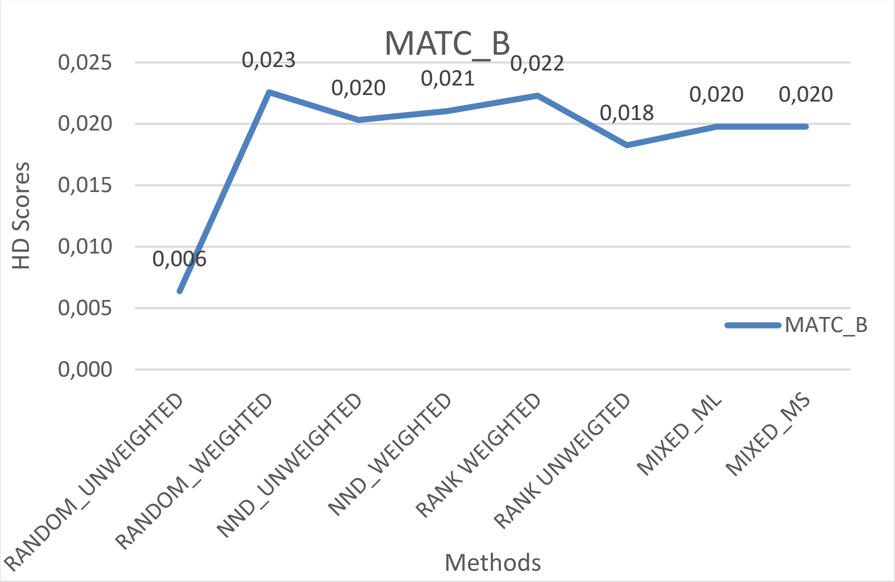
When random hot deck imputation method is implemented as unweighted, matching a, b and c combinations which consist of “car”, “comp” and “dish_w” variables in a way of represented in the Table 4.1. perform better than the other seven SM applications.

Figure 4.1. Matching A Results



While Figure 4.1., Figure 4.2. and Figure 4.3. demonstrate that unweighted random hot deck imputation performs better, Figure 4.4. represents that the method is less effective for matching only in “d” combination. Weighted random hot deck, on the other hand, performs the worst in “a” combination. The other six methods’ results are very close and all methods seem effective to gain micro file.

Figure 4.2. Matching B Results



“Car”, “comp” and “dish_w” combination of donor classes named x.matc.c has also better results. Although it was seen that random unweighted method was positively differentiated, the other seven methods gave similar results.

Figure 4.3. Matching C Results

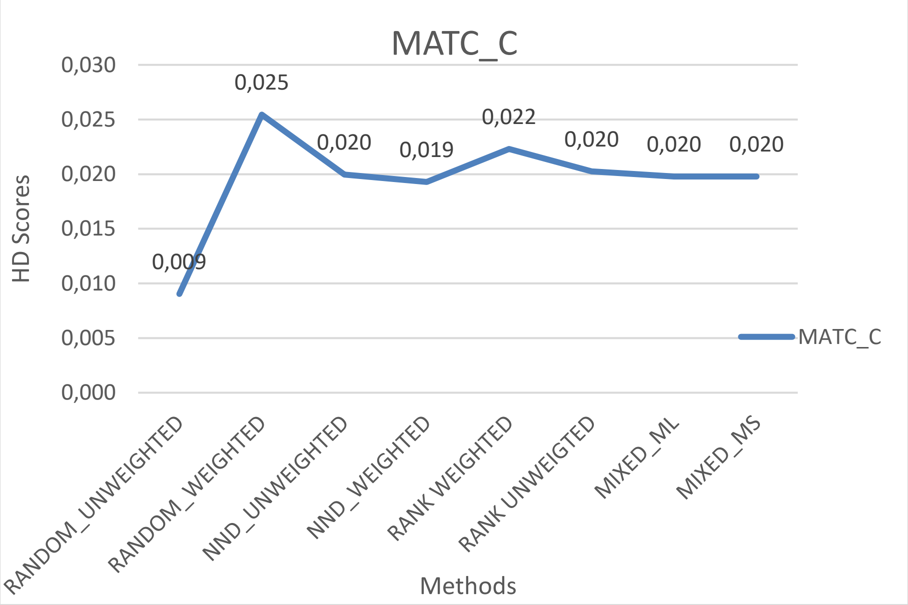
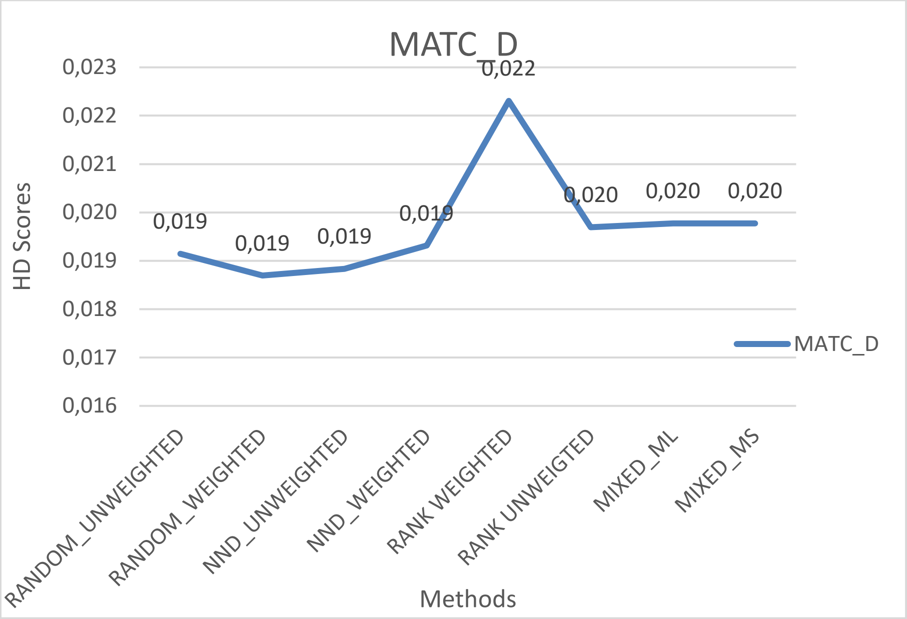


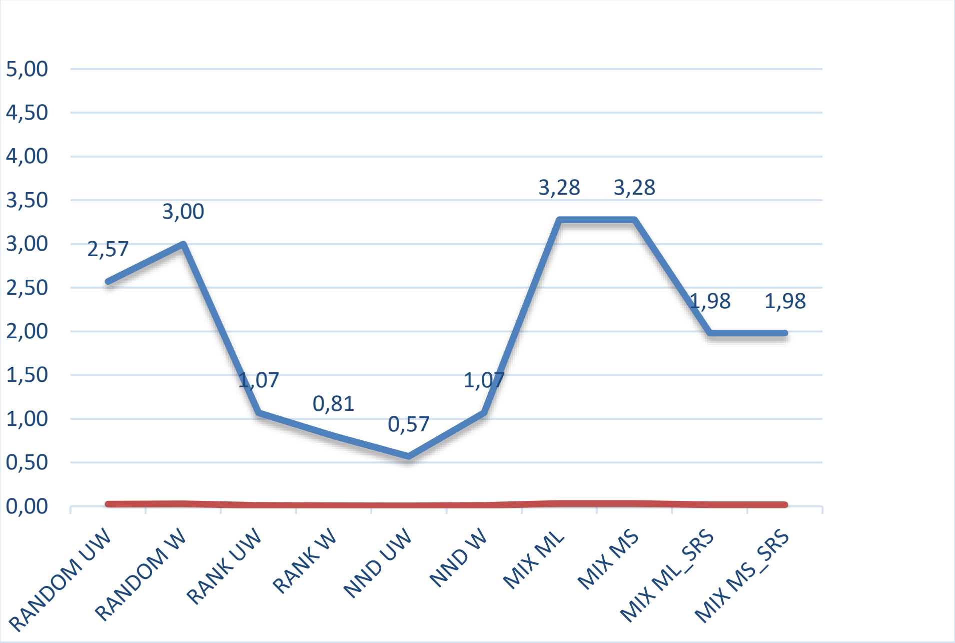
Figure 4.4. Matching D Results



When all four matching variables are utilized together, the outcomes are different from the other three classes. The worst result is obtained using the weighted rank hot deck imputation approach, while the best result is obtained using the random weighted hot deck method. When considered as a whole, each of the eight methods produces correct findings.

The same procedures were implemented without using income categories. Results show that nearest neighbor distance hot deck method performed the best with the HD score of %0,57. In the analysis period, mixed matching method (as MS and ML) is run for two separate data sets; randomly selected records and a subset selected by simple random selection method in SPSS with new house level weights. Analysis with records selected by SRS performed better than randomly selected ones.

Figure 4.5. Statistical Matching Results without Income Categories (%)



After traditional approaches, options “rot”, “min”, “exact”, “constrained” and “unconstrained” were applied for the related SM methods. The first analysis indicates that some combinations of them get over the cut-off value of %5 between 5.1 and 6.5, especially in random approach. Random unweighted exact matching got the highest value with %9.2. Weighted random hot deck methods’ “min” option took the best value for match B. Besides “exact” and “unconstrained” options of rank hot deck approach give convenient HD scores both weighted and unweighted as it’s seen in the Table 4.2. Thus, considering the results in the Figure 4.5., grouped variables were re-examined in order to find out better SM outputs for random hot deck method.

Table 4.2. Weighted and Unweighted Hellinger Distance Scores for Options

METHODS	MATC_ A	MATC_ B	MATC_ C	MATC_ D
RANDOM_WEIGHTED_MIN	0.0180	0.0155	0.0198	0.0183
RANK_UNWEIGHTED_EXACT	0.0156	0.0183	0.0209	0.0214
RANK_UNWEIGHTED_UNCONSTRAINED	0.0217	0.0218	0.0192	0.0227
NND_UNWEIGHTED_MIN	0.0213	0.0201	0.0187	0.0181
NND_UNWEIGHTED_ROT	0.0199	0.0183	0.0192	0.0190
NND_UNWEIGHTED_EXACT	0.0180	0.0183	0.0192	0.0190
NND_WEIGHTED_MIN	0.0194	0.0191	0.0200	0.0210
NND_WEIGHTED_ROT	0.0200	0.0206	0.0192	0.0216
NND_WEIGHTED_EXACT	0.0186	0.0203	0.0191	0.0214
NND_WEIGHTED_UNCONSTRAINED	0.0168	0.0187	0.0213	0.0200

Reconstruction of donor classes made a noticeable improvement in results as seen in the Table 4.2. Outputs were obtained below the threshold value in all categories. Rank unweighted hot deck “exact” option performed the best compared to other selected methods. Even if rank unweighted hot deck “unconstrained” option gave the worst score for “a” combination using “car” variable, the whole method performed in a reasonable range (Figure 4.6.)

Figure 4.6. Matching A Results II (%)

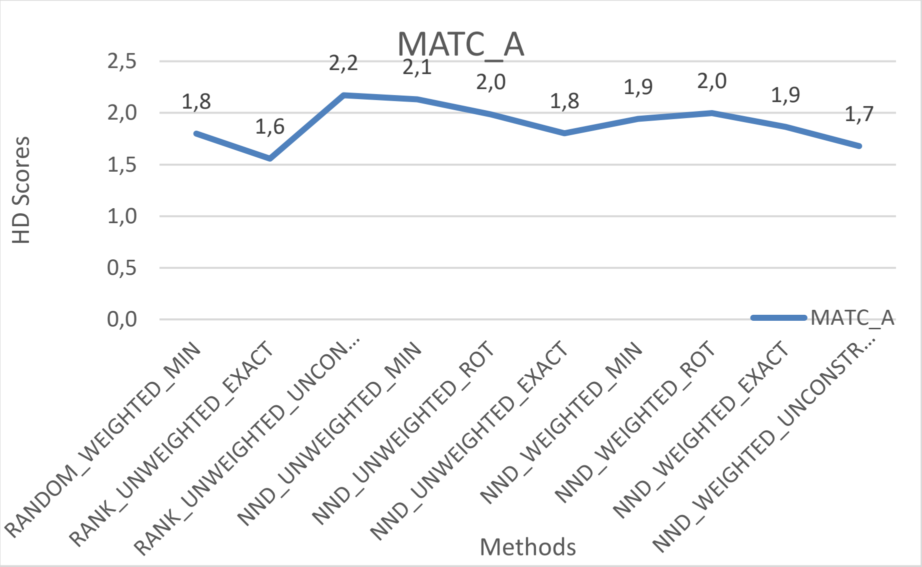
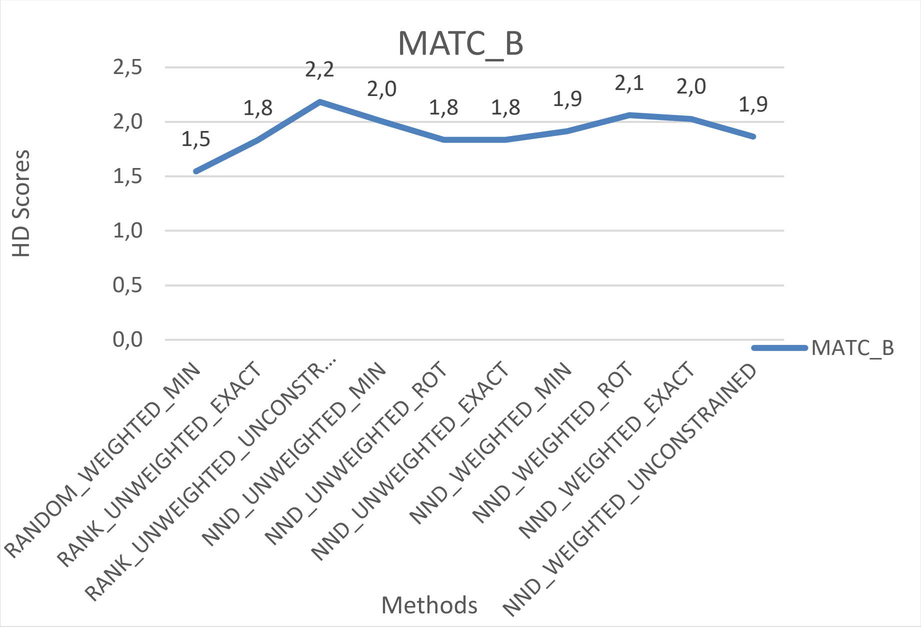


Figure 4.7. Matching B Results II (%)



“B” donor class performed also in a close range. As in “a” combination, rank unweighted hot deck “unconstrained” option gave the worst score for “b” combination too (Figure 4.7.). The third combination of matching took different values than before. Nearest neighbor distance hot deck imputation method gave better results both for “min” and “exact” options (Figure 4.8.).

Figure 4.8. Matching C Results II (%)

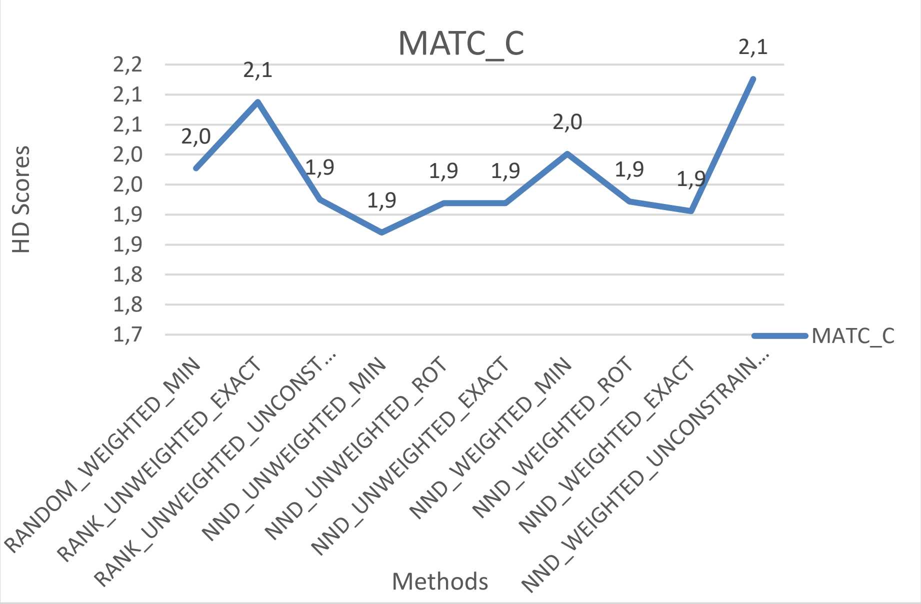


Figure 4.9. Matching D Results II (%)

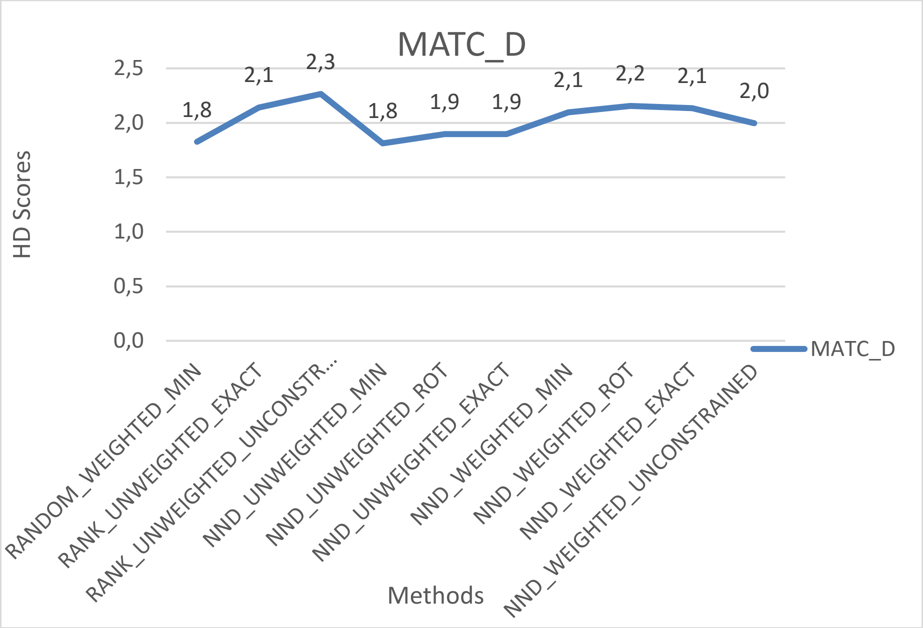
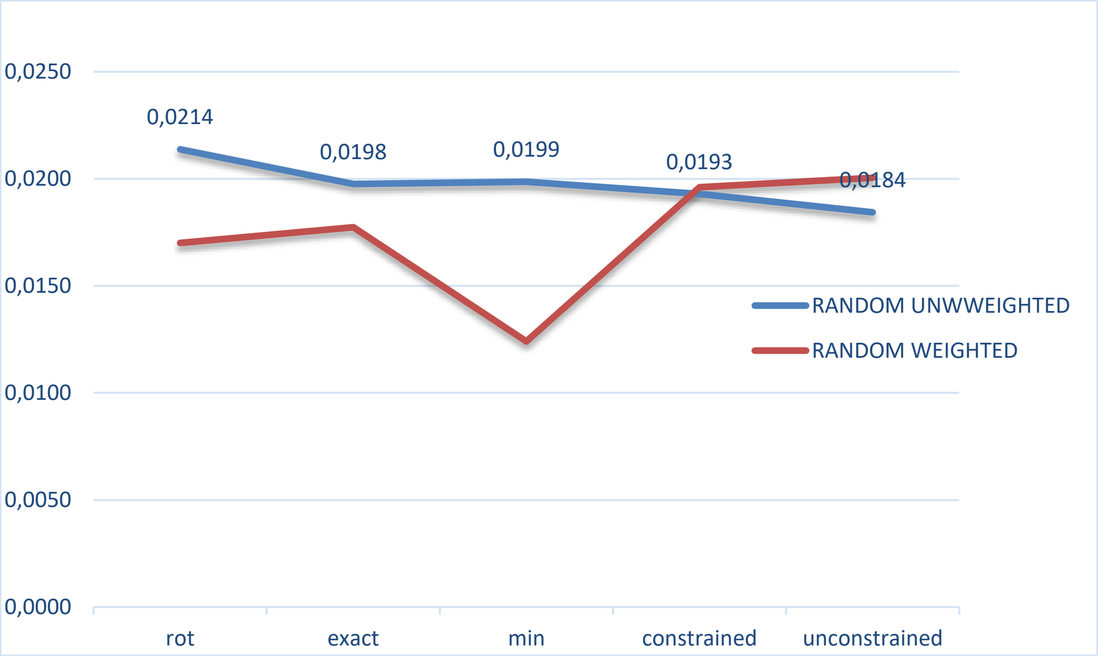


Figure 4.9. shows that donor classes of “d” gave very homogeneous results altering from 1.8 to 2.0. Nearest neighbor distance unweighted hot deck imputation method “min” option gave the best HD score compared to the other methods.

Figure 4.10. Match.vars=Null Results (%)

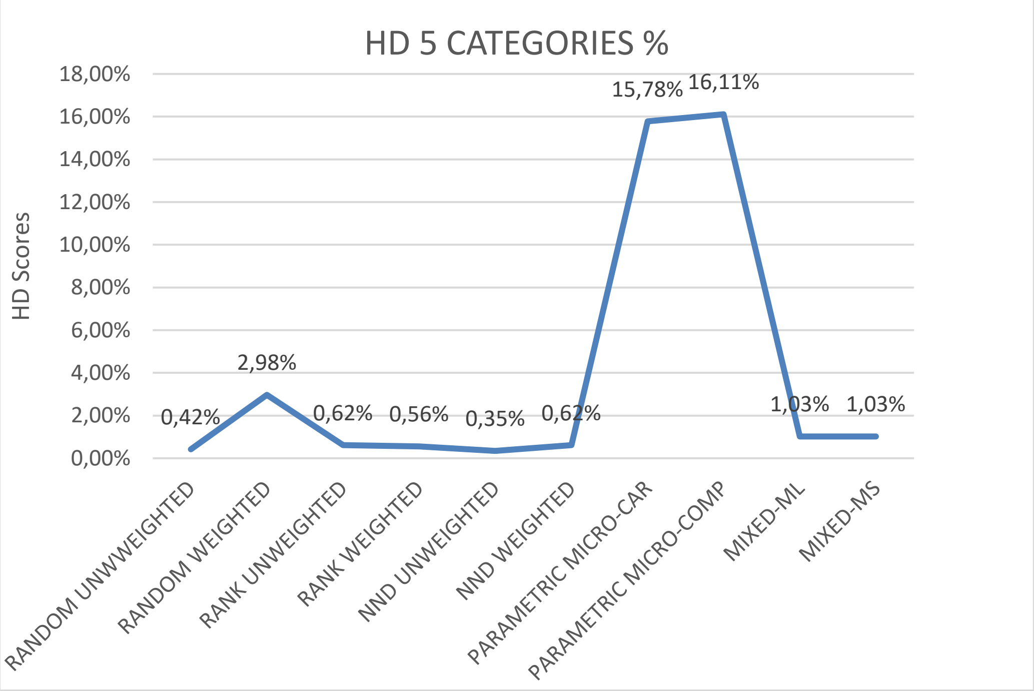


Random hot deck imputation procedure allows making matching applications in different ways. Statmatch package in R program and random hot deck techniques can be used to pick a donor wholly at random or with a probability proportional to a weight in case of the usage of match.vars=null. After R codes were revised according to the situation, the outputs seen in the Figure 4.10. have been highly promising.

Weighted “min” option gave the best result compared to the other nine experiments. Unweighted “rot” option, on the other hand” gave the worst result even if it is still under the cut-off value.

In cases of unique variables have a categorical structure, it is possible to influence the validation process because the response categories are open to research intervention. In order to observe the difference that this situation will create, the effects of the change in the response categories of the consumption variable in the original data set and the imputed data set were examined in the sense of the quality at the aggregated level. As seen in the Figure 4.11., each method gives better results than the first one except weighted random hot deck.

Figure 4.11. Hellinger Distance Results with Different Categories (%)



Common variables are eliminated and matching variables are selected as a result of a series of processes. Two experiments related to these variables were done in the final period. First of all, randomly selected variables are put into processes: household size, hot water availability and heating system of dwelling. Variables selected as proper for only one dataset as a result of survey design variables' usage in regressions, put into processes secondly. These are ownership of internet, number of adult and number of employed people in household. Here, it is aimed to measure both the effectiveness of the 3-step variable selection method and the effect of survey design variables on the microdata in the synthetic file.

When donor classes and grouped variables are used, none of the methods takes values under the cut-off value. In the light of these indicators, it proves that the harmonization and preparatory processes and the three-staged elimination mechanisms work reliably. Hellinger distance calculation, spearman2 procedures and regression analyses are capable of choosing common variables according to randomly selected variables' scores. Nevertheless, when variables are put into processes uniquely, matching variables selected by regression analyses considering the complex structure of samples, provide us with admissible percentages. In other words, outputs of regressions taking into account stratum and cluster information may use as matching variables.

CHAPTER 5. CONCLUSION AND DISCUSSION

The main focus of the dissertation is to evaluate the effectiveness of the statistical matching method at different levels by using the data of two social surveys that are conducted annually in Turkey. As a result of trying the all available macro and micro statistical matching methods by creating different donor classes, solutions were found for common problematic areas experienced in the application of each method and at the same time, the most effective sub-method on the basis of matching variable was tried to be discovered. Since it is a relatively new field and has not been studied sufficiently by national researchers, many problems have been encountered inevitably during the data transfer phase and solutions have been produced to contribute to the literature. The analyzes of all statistical matching categories were concluded and presented comparatively.

Integration of two large-sized data sets for the purpose of fusing consumption expenditure from HBS to SILC requires the use of advanced programs indispensably. R, SPSS and SAS Enterprise, thanks to their fast and effective results, enable us the opportunity to match data sets in many different sizes and numbers. In this respect, it has been possible to evaluate a large number of data matching methods in the content of the thesis. Estimations for the unavailable variables, provide us to observe them in a synthetic micro file altogether for each sub-method. Aggregated level results of these sub-methods indicate that micro and mixed approaches have good matching values and there are no substantial divergent at breakdowns probed. Despite promising results, additional procedures were put in place to monitor the validity of the method such as altering donor and recipient regardless of sample size of them, putting in unselected variables in the sense of HD, spearman and regression analysis, etc.

Complex sample design structure of HBS and SILC was also intended to incorporate into the preliminary procedures of the SM. Survey design variables were considered during the elimination operations of common variables contrary to the ordinary approach in literature. Household level weights, stratum and cluster information of samples were added to the Hellinger Distance calculation, spearman and regression analyses. It was observed that complex structure of the samples was very effective in terms of selection and elimination processes of the common variables. Since the quality of quantitative outputs could change according to selected matching variables, survey design variables should be added to the procedures by all means. Estimated values may have good matching results at aggregated level but including mentioned structure in the process provide us with more accurate and reliable values at disaggregated level. This perspective would assure the researcher in cases where the micro file is needed for subsequent studies. The situation also shows that the flexible structure of the SM method allows for different interventions to the processes. A negative aspect or shortcoming of the method is that these interventions such as adjustments in the response categories of common variables during the harmonization period, can create bias especially in applications where the categorical variables are used.

Alternative classification of the response categories was tried and compared with the first results in order to demonstrate the intervention-friendly nature of the data matching. Especially when evaluated on the basis of aggregated data, halving the number of categories leads to relatively better results in terms of validation. Even in parametric micro results, improvements up to 68.18% (according to HD scores for validation) are provided in terms of evaluating data quality, although it is above the desired threshold value. While better results were obtained in almost all sub-methods, outputs of random unweighted hot deck method were worse than the initial results unexpectedly. Evaluation of the effects when response categories of matching variables are reorganized is recommended for future studies.

The problem encountered in mixed matching was that the R functions were insufficient due to the increase in data size and/or process capacity of the computers. To solve this problem, the current limit was increased with the memory size and memory limit functions, but abort mission warning was still received. Experiments were made with sub-samples of various sizes roughly taken from the data and the limit consisting of the maximum number of data that the program could yield results was determined.

Then, a subsample was created using SPSS and SRS method and the weights were reassigned. This sub-sample was used in the analysis made with the mixed method.

Another subject that needed to be examined was whether the methods used in the selection of common variables were effective. In order to investigate this situation and to see the effects of matching variables that are not among the selected Xs, the randomly selected variables were matched and the results were compared. Especially in the elimination process, variables that did not give appropriate results were tried. In addition, the results of the regression analyses made as a result of taking into account the complex sample structure were also taken into account. The aim here is to gain new approaches in addition to measuring the effectiveness of traditional methods in the literature. Results indicate that traditional approaches have enough capacity to minimize the common variables. On the other hand, complex sample design and its components should be utilized during the elimination period. Household level weights, stratum and cluster information may be beneficial in regression analysis. As a result of the regressions made considering the complex sample structure, proper variables could be reached. Reducing common variables to matching variables, can be considered in this respect and should be examined in future studies detailed.

There are certain conventional approaches to the determination of donor and recipient data. It is recommended to use the data set with a larger sample size as a donor. This is not always possible in terms of the needs of the researcher. For this reason, two-way statistical matches were made regardless of sample sizes. As a result, it has been seen that effective and reliable outputs can be made regardless of the direction of matching by sample size.

Renssen method and its two sub-methods were investigated due to its capacity to allow micro-level data matching and exploiting of a third data. Since auxiliary sample or information usage of jointly observed X, Y, and Z or only Y and Z is not possible with available data sets, implementation procedures were explained in the related chapter. In addition to the auxiliary sample usage rules, micro-level matching procedures have been done. Contrary to expectations, overlap of data sets before and after Renssen calibration approach, indicate that Renssen method provides very small improvements in datasets in the sense of calibration. Therefore, the results related to this approach were not ranked among the methods evaluated.

This thesis aims not only fills the gap between comparative and applied studies between methods in the literature but also calls for further studies to investigate fields not fully clarified such as parametric approach. Parametric approaches have produced less stable outputs. Hot deck procedures, on the other hand, gave better results. Controls in synthetic files and other validation operations demonstrate that there are big differences in the sense of quality. Positive correlation between estimations and observed values in completed micro file indicates that parametric approach has to be developed (Linskens, 2015). An explanation for this situation is that regression model doesn’t capture the distribution of microdata as much as expected (Waal, 2015). However, new studies should be practiced in parametric matching field to find out problematic aspects to get more accurate and statistically usable outputs.

There are still research areas to be developed for future works. Validation or quality assessment techniques are relatively old methods. New and more appropriate methods should be developed to measure the quality of estimations. Besides, survey data or registers non-overlapping or not referring to the same population should be fused to investigate data quality. The applicability of big data using SM methods should be investigated.

The application of SM methods in areas that need rapid and timely data such as international migration movements will provide decision-makers with more effective intervention in problematic areas. Micro methods have the potential to provide needed data. These datasets can be created by fusing administrative records and survey data gathered for various purposes. With this approach, the data obtained from project-supported and lengthy researches can be accessed more rapidly and inexpensively. (Özkan and Türkyılmaz, 2022).

Statistical matching is a very effective way of merging data sets to obtain a synthetic subset of unavailable variables at present. It creates a very wide working field with all the different sub-methods mentioned above. The reason of non-parametric approach is getting widespread amongst all these is the capability to respond to ever-increasing microdata demand effectively. The flexible structure of SM, which can be shaped according to the needs of the researcher, is the most prominent reason for its widespread use.

CHAPTER 6. REFERENCES

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

Albayrak, Ö., & Masterson T. (2017). Quality of Statistical Match of Household Budget Survey and SILC for Turkey, Working Paper No. 885.

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

Alpman A., Gardes F., & Thiombiano N. (2017). Statistical Matching for Combining Time-Use Surveys with Consumer Expenditure Surveys: An Evaluation on Real Documents de Travail du Centre d'Economie de la Sorbonne.

Augurzky, B., & C. Schmidt (2001) "The Propensity Score: A Means to an End"", Discussion Paper No. 271, IZA.

Bryson, A., Dorsett, R., & Purdon S. (2002) "The Use of Propensity Score Matching in the Evaluation of Labour Market Policies", Working Paper No. 4, Department for Work and Pensions.

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.

Dawid, A.P. (1979) Conditional independence in statistical theory. Journal of the Royal Statistical Society, B **41**, 1–31.

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

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.

D'Orazio, M. (2013). Statistical Matching: Methodological Issues and Practice with R-statmatch.

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

Fellegi, I.P. & Sunter, A.B., (1969). A theory for record linkage, *Journal of the American Statistical Association,* 64, 1183-1210.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Harrell, F.E. (2016) Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. 2nd Edition. New York, Springer.

Jaro, M.A., (1972). A computer system for generalized record linkage under conditions of uncertainty, *Spring Joint Computer Conference*, 40, 523-530.

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

Kish, L. (1965). Survey Sampling, New York: John Wiley & Sons

Kish, L. (1999). Cumulating/Combining population surveys. Survey Methodology, 129-138.

Kleinbaum, C.D., Kupper, L.L., Muller, K.E., & Nizam, A., (1997). Applied Regression Analysis and Other Multivariable Methods, Third Edition.

Kum. H, & Masterson, T. (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.V.D. (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,7-15.

Linskens, S.J. (2015), Statistical Matching: A Comparison of Random and Distance Hot Deck. Report, Tilburg University, The Netherlands.

Memobust, (2014). Methodology of Modern Business Statistics, Memobust Handbook.

Moriarty C., Scheuren F. (2001). Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure, Journal of Official Statistics, Vol, 17, No:3, 2001, 407-422.

Newger, K. (2018). Statistical Matching of Categorical Data with Markov Networks, Ludwig-Maximilians-Universitat München, Department of Statistics.

Okner, B.A. (1972) Constructing a new data base from existing microdata sets: the 1966 merge file, Annals of Economics and Social Measurement 1 (3) 325–342.

Okner, B.A. (1974) Data matching and merging: an overview, Annals of Economic and Social Measurement 3 (2) 347–352.

Öztürk, C. (2019). Nonparametric Statistical Matching Methods: An Application On Household Surveys in Turkey.

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. New York: Springer.

Rasner, A., Himmelreicher, R. K., Grabka. M. M., & J. R. Frick (2007). Best of Both Worlds – Preparatory Steps in Matching Survey Data with Administrative Pension Records. The Case of the German Socio Economic Panel and the Scientific Use File Completed Insurance Biographies 2004. SOEP papers 70. Berlin, Deutsches Institut für Wirtschaftsforschung: 210.

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

Rodgers, W.L. (1984). An Evaluation of Statistical Matching, Journal of Business & Economic Statistics Vol. 2, No. 1 (Jan., 1984), pp. 91-102.

Rubin, D. B. (1986), "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations", Journal of Business and Economic Statistics, 4, 87- 94.

Sims, C.A. (1972). Comments (on Okner), Annals of Economic and Social Measurement 1 (3) 343–345.

Singh, A.C., Mantel, H., Kinack, M. & Rowe, G. (1993) Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology* 19, 59–79.

Turkstat (2018a). Handbook of Household Budget Survey,

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

Uçar, B., & Gianni B. (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.

Waal, T. (2015). Statistical matching: Experimental results and future research questions.

Webber, D. & Tonkin, R.P. (2013), Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditure and material deprivation, Eurostat Methodologies and working papers, Publications office of the European Union, Luxembourg.

Van der Putten, P., Kok, J. N., & Gupta, A. (2002). Data fusion through statistical matching.

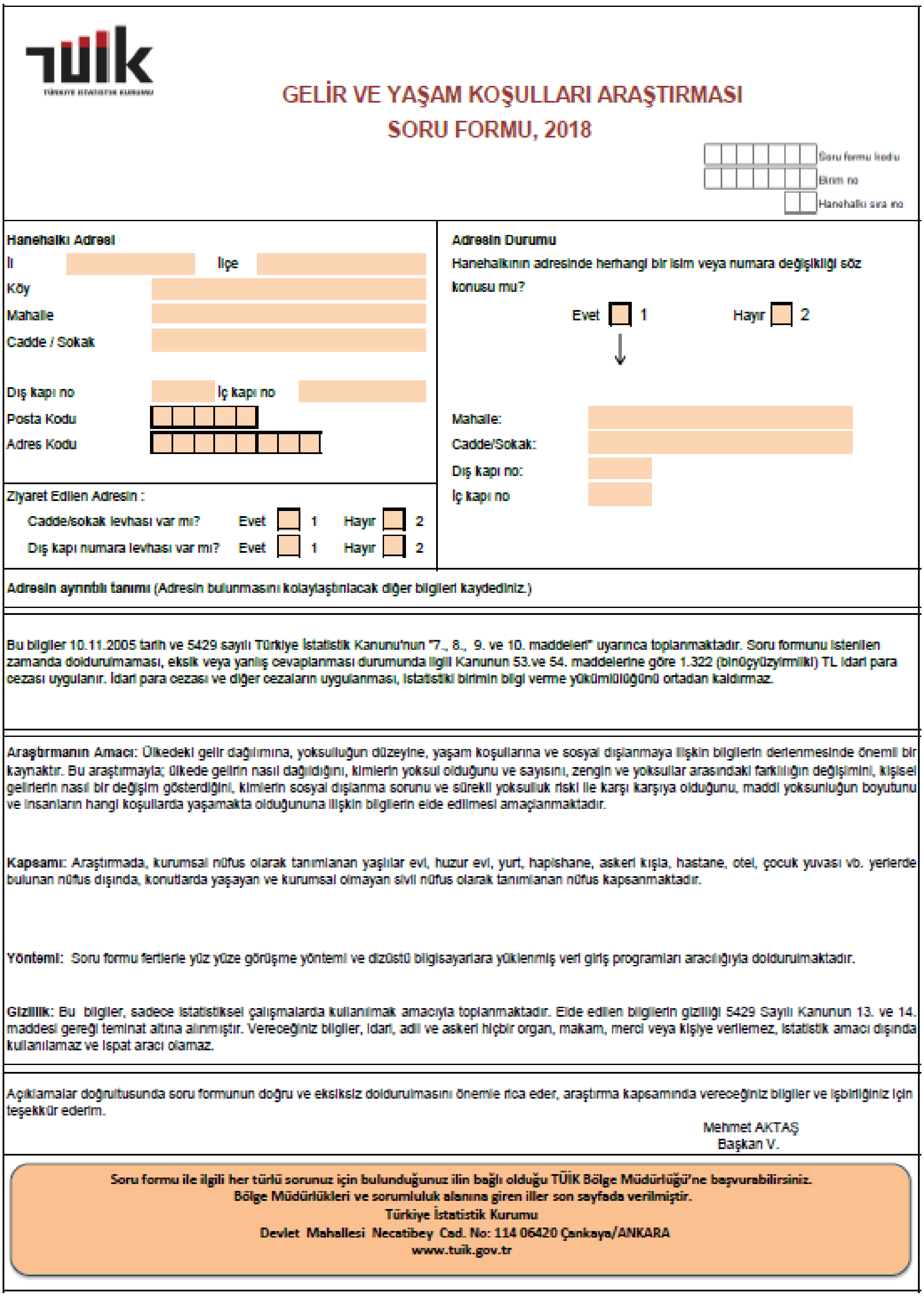
Wu, C. (2004). "Combining information from multiple surveys through the empirical likelihood method", The Canadian Journal of Statistics, 32, 112.

Xhava, E. (2015). Statistical Matching for Data Integration with Different Data Sources.

Zacarias A., Masterson T., Memiş E. (2014). Time Deficits and Poverty: The Levy Institute Measure of Time and Consumption Poverty for Turkey.

Appendix A. SILC and HBS Questionnaires (Cover Pages)

A.1. SILC Questionnaire



GELİR VE YAŞAM KOŞULLARI ARAŞTIRMASI
SORU FORMU, 2018

□□□□□□□ Soru formu kodu
□□□□□□□ Birim no
□□ Hanehalkı sıra no

Hanehalkı Adresi

İl ____ İlçe ____
Köy ____
Mahalle ____
Cadde / Sokak ____

Dış kapı no ____ İç kapı no ____
Posta Kodu □□□□□
Adres Kodu □□□□□□□□

Adresin Durumu

Hanehalkının adresinde herhangi bir isim veya numara değişikliği söz konusu mu?

Evet □ 1 Hayır □ 2

↓

Mahalle: ____
Cadde/Sokak: ____
Dış kapı no: ____
İç kapı no ____

Ziyaret Edilen Adresin :

Cadde/sokak levhası var mı? Evet □ 1 Hayır □ 2
Dış kapı numara levhası var mı? Evet □ 1 Hayır □ 2

Adresin ayrıntılı tanımı (Adresin bulunmasını kolaylaştırılacak diğer bilgileri kaydediniz.)

Bu bilgiler 10.11.2005 tarih ve 5429 sayılı Türkiye İstatistik Kanunu'nun "7., 8., 9. ve 10. maddeleri" uyarınca toplanmaktadır. Soru formunu istenilen zamanda doldurulmaması, eksik veya yanlış cevaplanması durumunda ilgili Kanunun 53.ve 54. maddelerine göre 1.322 (binüçyüzyirmiiki) TL idari para cezası uygulanır. İdari para cezası ve diğer cezaların uygulanması, istatistiki birimin bilgi verme yükümlülüğünü ortadan kaldırmaz.

Araştırmanın Amacı: Ülkedeki gelir dağılımına, yoksulluğun düzeyine, yaşam koşullarına ve sosyal dışlanmaya ilişkin bilgilerin derlenmesinde önemli bir kaynaktır. Bu araştırmayla; ülkede gelirin nasıl dağıldığını, kimlerin yoksul olduğunu ve sayısını, zengin ve yoksullar arasındaki farklılığın değişimini, kişisel gelirlerin nasıl bir değişim gösterdiğini, kimlerin sosyal dışlanma sorunu ve sürekli yoksulluk riski ile karşı karşıya olduğunu, maddi yoksunluğun boyutunu ve insanların hangi koşullarda yaşamakta olduğununa ilişkin bilgilerin elde edilmesi amaçlanmaktadır.

Kapsamı: Araştırmada, kurumsal nüfus olarak tanımlanan yaşlılar evi, huzur evi, yurt, hapishane, askeri kışla, hastane, otel, çocuk yuvası vb. yerlerde bulunan nüfus dışında, konutlarda yaşayan ve kurumsal olmayan sivil nüfus olarak tanımlanan nüfus kapsanmaktadır.

Yöntemi: Soru formu fertlerle yüz yüze görüşme yöntemi ve dizüstü bilgisayarlara yüklenmiş veri giriş programları aracılığıyla doldurulmaktadır.

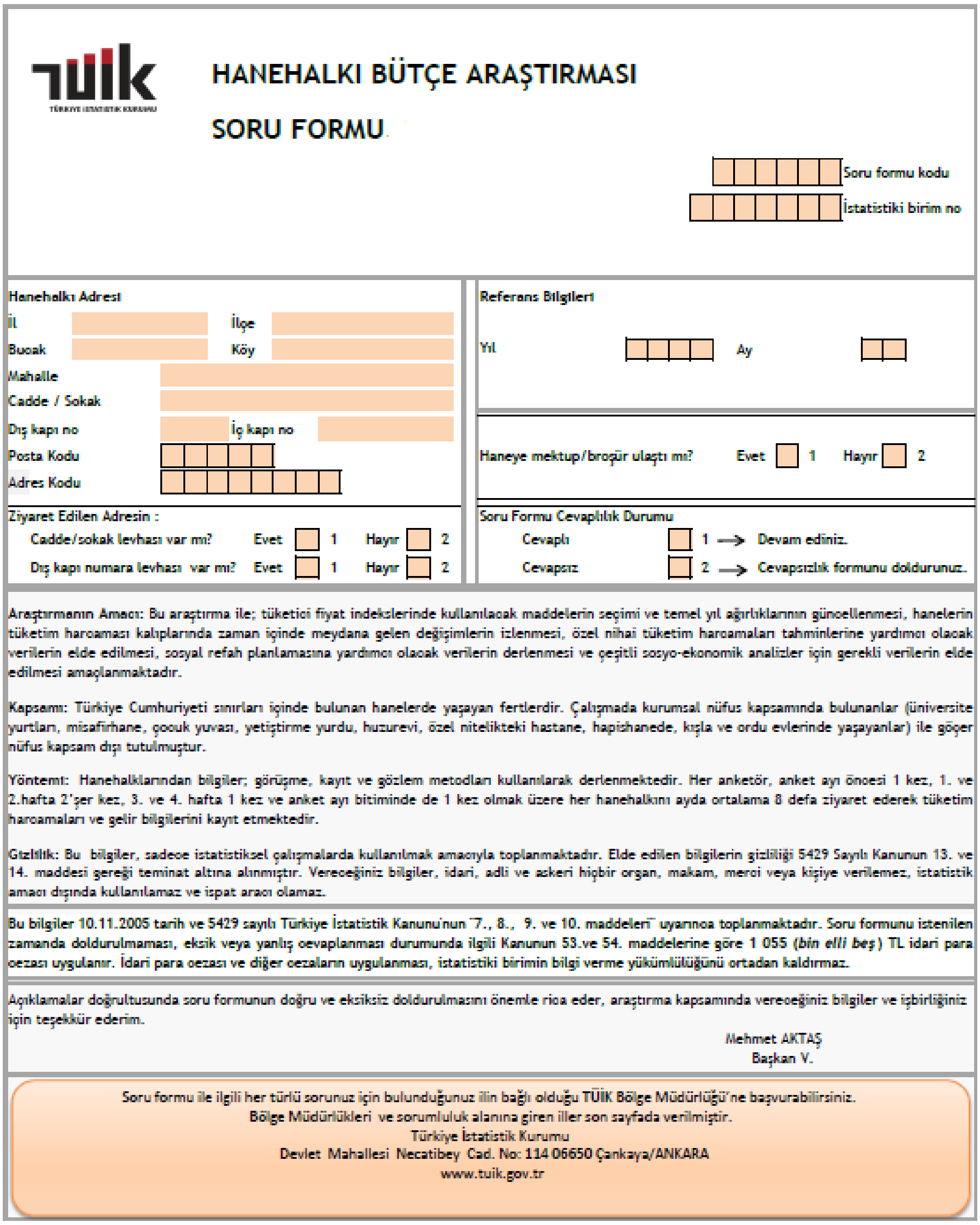
Gizlilik: Bu bilgiler, sadece istatistiksel çalışmalarda kullanılmak amacıyla toplanmaktadır. Elde edilen bilgilerin gizliliği 5429 Sayılı Kanunun 13. ve 14. maddesi gereği teminat altına alınmıştır. Vereceğiniz bilgiler, idari, adli ve askeri hiçbir organ, makam, merci veya kişiye verilemez, istatistik amacı dışında kullanılamaz ve ispat aracı olamaz.

Açıklamalar doğrultusunda soru formunun doğru ve eksiksiz doldurulmasını önemle rica eder, araştırma kapsamında vereceğiniz bilgiler ve işbirliğiniz için teşekkür ederim.

Mehmet AKTAŞ
Başkan V.

Soru formu ile ilgili her türlü sorunuz için bulunduğunuz ilin bağlı olduğu TÜİK Bölge Müdürlüğü'ne başvurabilirsiniz.
Bölge Müdürlükleri ve sorumluluk alanına giren iller son sayfada verilmiştir.
Türkiye İstatistik Kurumu
Devlet Mahallesi Necatibey Cad. No: 114 06420 Çankaya/ANKARA
www.tuik.gov.tr

B.1. HBS Questionnaire



HANEHALKI BÜTÇE ARAŞTIRMASI

SORU FORMU

□□□□□□ Soru formu kodu

□□□□□□□ İstatistiki birim no

Hanehalkı Adresi

İl ________ İlçe ________

Buoak ________ Köy ________

Mahalle ________

Cadde / Sokak ________

Dış kapı no ________ İç kapı no ________

Posta Kodu □□□□□

Adres Kodu □□□□□□□□

Referans Bilgileri

Yıl □□□□ Ay □□

Haneye mektup/broşür ulaştı mı? Evet □ 1 Hayır □ 2

Ziyaret Edilen Adresin :

Cadde/sokak levhası var mı? Evet □ 1 Hayır □ 2

Dış kapı numara levhası var mı? Evet □ 1 Hayır □ 2

Soru Formu Cevaplılık Durumu

Cevaplı □ 1 → Devam ediniz.

Cevapsız □ 2 → Cevapsızlık formunu doldurunuz.

Araştırmanın Amaoı: Bu araştırma ile; tüketici fiyat indekslerinde kullanılaoak maddelerin seçimi ve temel yıl ağırlıklarının günoellenmesi, hanelerin tüketim haroaması kalıplarında zaman içinde meydana gelen değişimlerin izlenmesi, özel nihai tüketim haroamaları tahminlerine yardımoı olaoak verilerin elde edilmesi, sosyal refah planlamasına yardımoı olaoak verilerin derlenmesi ve çeşitli sosyo-ekonomik analizler için gerekli verilerin elde edilmesi amaçlanmaktadır.

Kapsamı: Türkiye Cumhuriyeti sınırları içinde bulunan hanelerde yaşayan fertlerdir. Çalışmada kurumsal nüfus kapsamında bulunanlar (üniversite yurtları, misafirhane, çoouk yuvası, yetiştirme yurdu, huzurevi, özel nitelikteki hastane, hapishanede, kışla ve ordu evlerinde yaşayanlar) ile göçer nüfus kapsam dışı tutulmuştur.

Yöntemi: Hanehalklarından bilgiler; görüşme, kayıt ve gözlem metodları kullanılarak derlenmektedir. Her anketör, anket ayı önoesi 1 kez, 1. ve 2.hafta 2'şer kez, 3. ve 4. hafta 1 kez ve anket ayı bitiminde de 1 kez olmak üzere her hanehalkını ayda ortalama 8 defa ziyaret ederek tüketim haroamaları ve gelir bilgilerini kayıt etmektedir.

Gizlilik: Bu bilgiler, sadeoe istatistiksel çalışmalarda kullanılmak amaoıyla toplanmaktadır. Elde edilen bilgilerin gizliliği 5429 Sayılı Kanunun 13. ve 14. maddesi gereği teminat altına alınmıştır. Vereoeğiniz bilgiler, idari, adli ve askeri hiçbir organ, makam, meroi veya kişiye verilemez, istatistik amaoı dışında kullanılamaz ve ispat araoı olamaz.

Bu bilgiler 10.11.2005 tarih ve 5429 sayılı Türkiye İstatistik Kanunu'nun "7., 8., 9. ve 10. maddeleri" uyarınoa toplanmaktadır. Soru formunu istenilen zamanda doldurulmaması, eksik veya yanlış oevaplanması durumunda ilgili Kanunun 53.ve 54. maddelerine göre 1 055 (*bin elli beş*) TL idari para oezası uygulanır. İdari para oezası ve diğer oezaların uygulanması, istatistiki birimin bilgi verme yükümlülüğünü ortadan kaldırmaz.

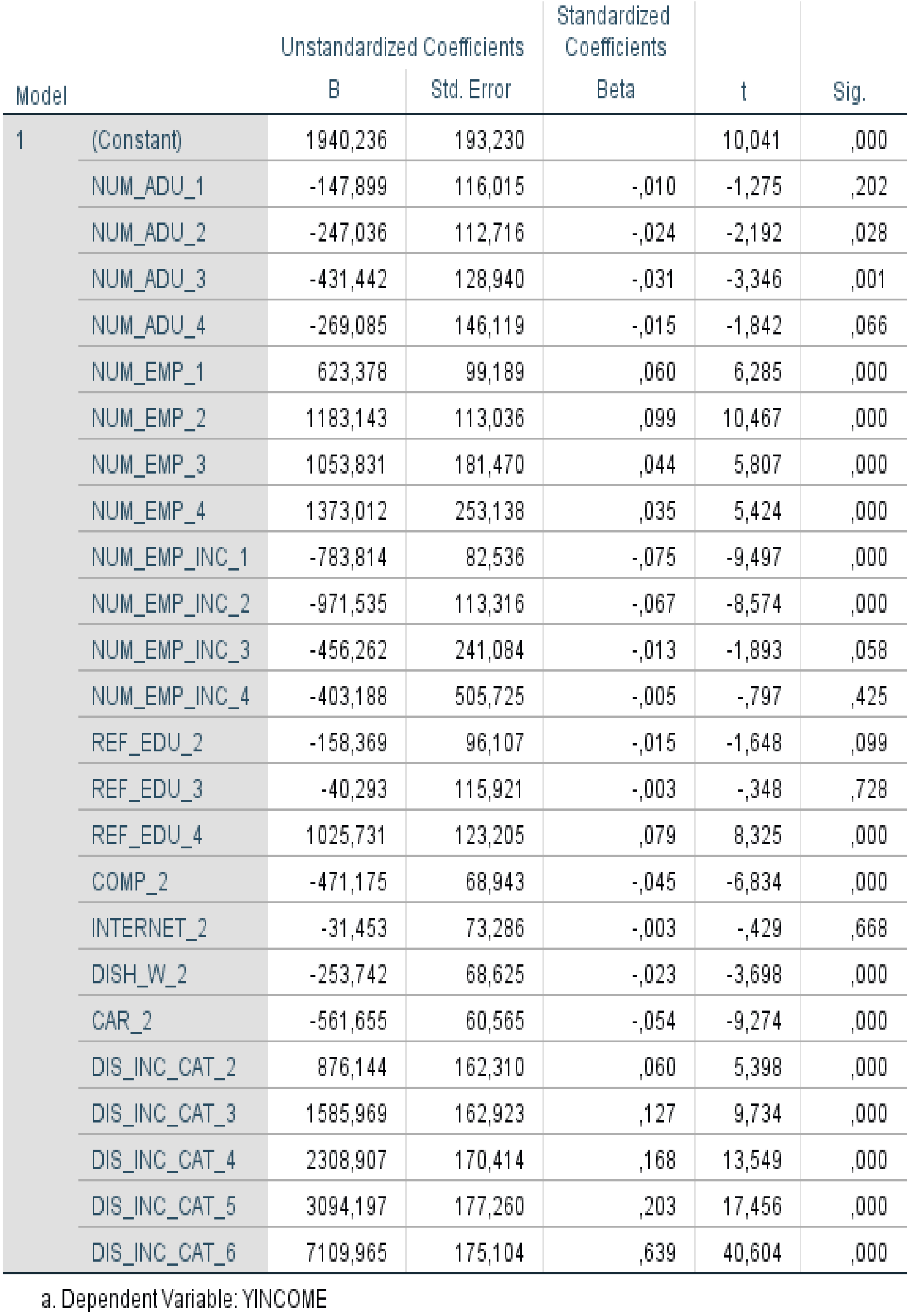
Açıklamalar doğrultusunda soru formunun doğru ve eksiksiz doldurulmasını önemle rioa eder, araştırma kapsamında vereoeğiniz bilgiler ve işbirliğiniz için teşekkür ederim.

Mehmet AKTAŞ
Başkan V.

Soru formu ile ilgili her türlü sorunuz için bulunduğunuz ilin bağlı olduğu TÜİK Bölge Müdürlüğü'ne başvurabilirsiniz.
Bölge Müdürlükleri ve sorumluluk alanına giren iller son sayfada verilmiştir.
Türkiye İstatistik Kurumu
Devlet Mahallesi Necatibey Cad. No: 114 06650 Çankaya/ANKARA
www.tuik.gov.tr

Appendix B. Regression Results

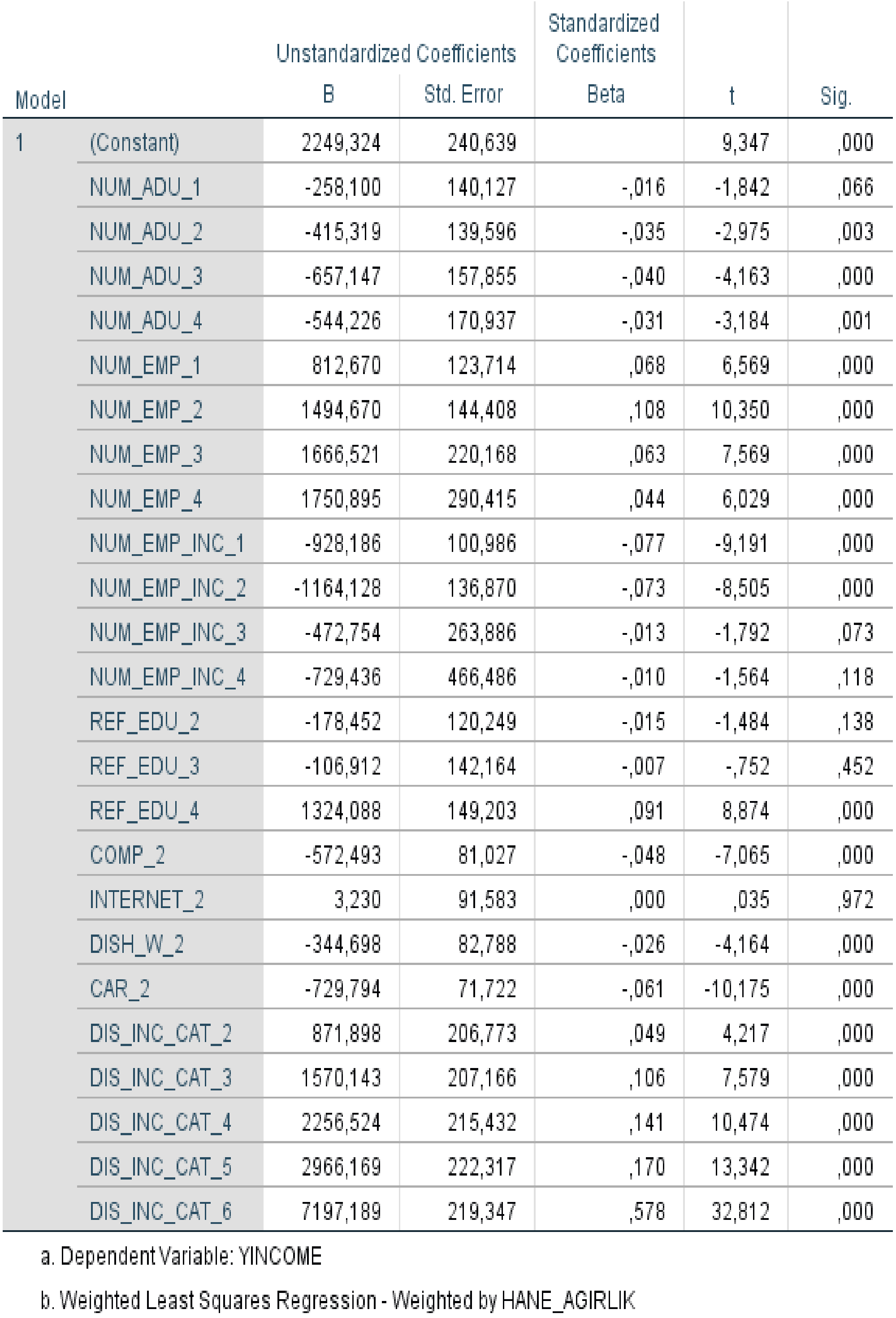
B.1. SILC Unweighted Linear Regression



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1940,236	193,230		10,041	,000
	NUM_ADU_1	-147,899	116,015	-,010	-1,275	,202
	NUM_ADU_2	-247,036	112,716	-,024	-2,192	,028
	NUM_ADU_3	-431,442	128,940	-,031	-3,346	,001
	NUM_ADU_4	-269,085	146,119	-,015	-1,842	,066
	NUM_EMP_1	623,378	99,189	,060	6,285	,000
	NUM_EMP_2	1183,143	113,036	,099	10,467	,000
	NUM_EMP_3	1053,831	181,470	,044	5,807	,000
	NUM_EMP_4	1373,012	253,138	,035	5,424	,000
	NUM_EMP_INC_1	-783,814	82,536	-,075	-9,497	,000
	NUM_EMP_INC_2	-971,535	113,316	-,067	-8,574	,000
	NUM_EMP_INC_3	-456,262	241,084	-,013	-1,893	,058
	NUM_EMP_INC_4	-403,188	505,725	-,005	-,797	,425
	REF_EDU_2	-158,369	96,107	-,015	-1,648	,099
	REF_EDU_3	-40,293	115,921	-,003	-,348	,728
	REF_EDU_4	1025,731	123,205	,079	8,325	,000
	COMP_2	-471,175	68,943	-,045	-6,834	,000
	INTERNET_2	-31,453	73,286	-,003	-,429	,668
	DISH_W_2	-253,742	68,625	-,023	-3,698	,000
	CAR_2	-561,655	60,565	-,054	-9,274	,000
	DIS_INC_CAT_2	876,144	162,310	,060	5,398	,000
	DIS_INC_CAT_3	1585,969	162,923	,127	9,734	,000
	DIS_INC_CAT_4	2308,907	170,414	,168	13,549	,000
	DIS_INC_CAT_5	3094,197	177,260	,203	17,456	,000
	DIS_INC_CAT_6	7109,965	175,104	,639	40,604	,000

a. Dependent Variable: YINCOME

B.2. SILC Weighted Linear Regression

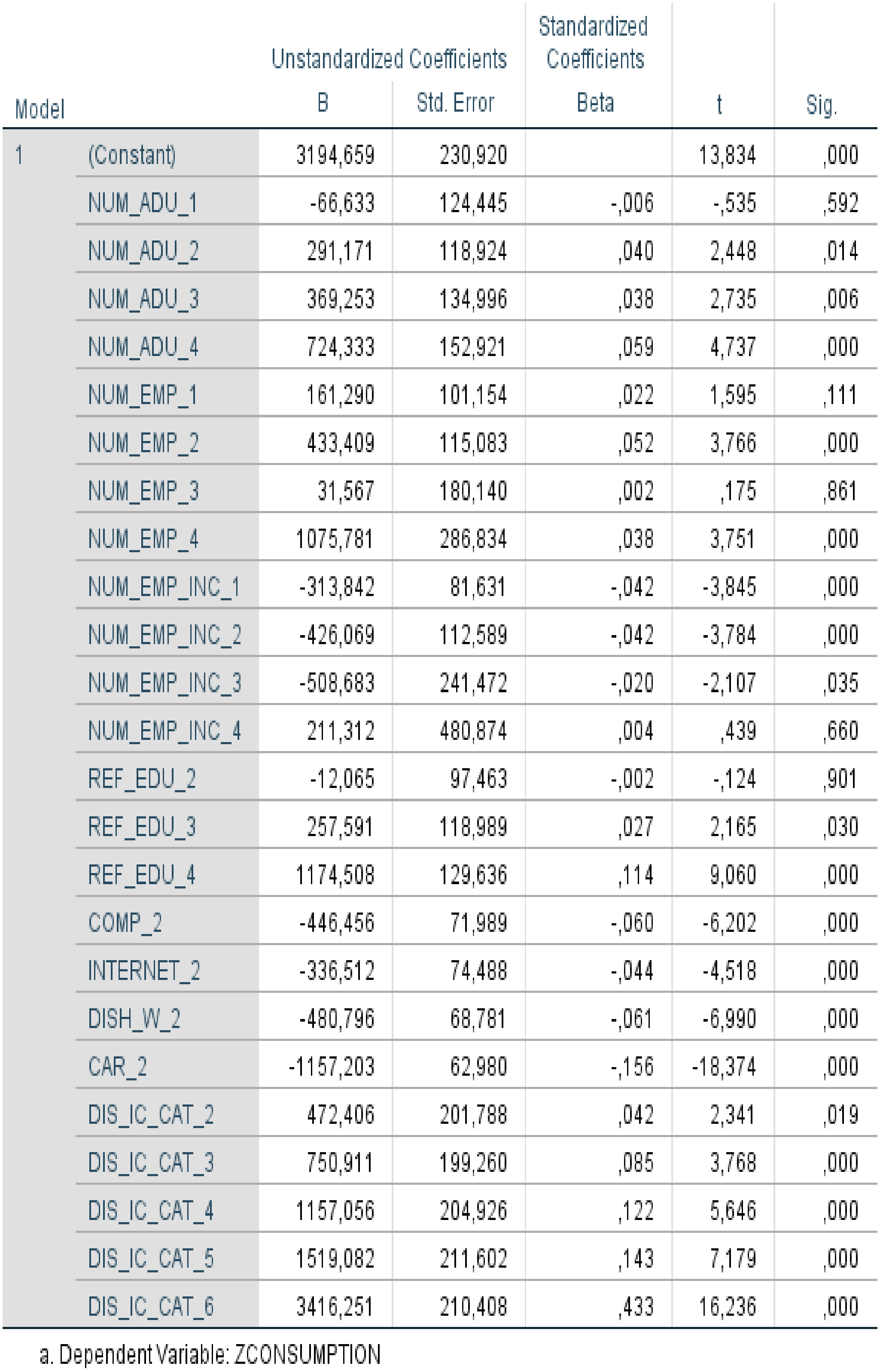


Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2249,324	240,639		9,347	,000
	NUM_ADU_1	-258,100	140,127	-,016	-1,842	,066
	NUM_ADU_2	-415,319	139,596	-,035	-2,975	,003
	NUM_ADU_3	-657,147	157,855	-,040	-4,163	,000
	NUM_ADU_4	-544,226	170,937	-,031	-3,184	,001
	NUM_EMP_1	812,670	123,714	,068	6,569	,000
	NUM_EMP_2	1494,670	144,408	,108	10,350	,000
	NUM_EMP_3	1666,521	220,168	,063	7,569	,000
	NUM_EMP_4	1750,895	290,415	,044	6,029	,000
	NUM_EMP_INC_1	-928,186	100,986	-,077	-9,191	,000
	NUM_EMP_INC_2	-1164,128	136,870	-,073	-8,505	,000
	NUM_EMP_INC_3	-472,754	263,886	-,013	-1,792	,073
	NUM_EMP_INC_4	-729,436	466,486	-,010	-1,564	,118
	REF_EDU_2	-178,452	120,249	-,015	-1,484	,138
	REF_EDU_3	-106,912	142,164	-,007	-,752	,452
	REF_EDU_4	1324,088	149,203	,091	8,874	,000
	COMP_2	-572,493	81,027	-,048	-7,065	,000
	INTERNET_2	3,230	91,583	,000	,035	,972
	DISH_W_2	-344,698	82,788	-,026	-4,164	,000
	CAR_2	-729,794	71,722	-,061	-10,175	,000
	DIS_INC_CAT_2	871,898	206,773	,049	4,217	,000
	DIS_INC_CAT_3	1570,143	207,166	,106	7,579	,000
	DIS_INC_CAT_4	2256,524	215,432	,141	10,474	,000
	DIS_INC_CAT_5	2966,169	222,317	,170	13,342	,000
	DIS_INC_CAT_6	7197,189	219,347	,578	32,812	,000

a. Dependent Variable: YINCOME

b. Weighted Least Squares Regression - Weighted by HANE_AGIRLIK

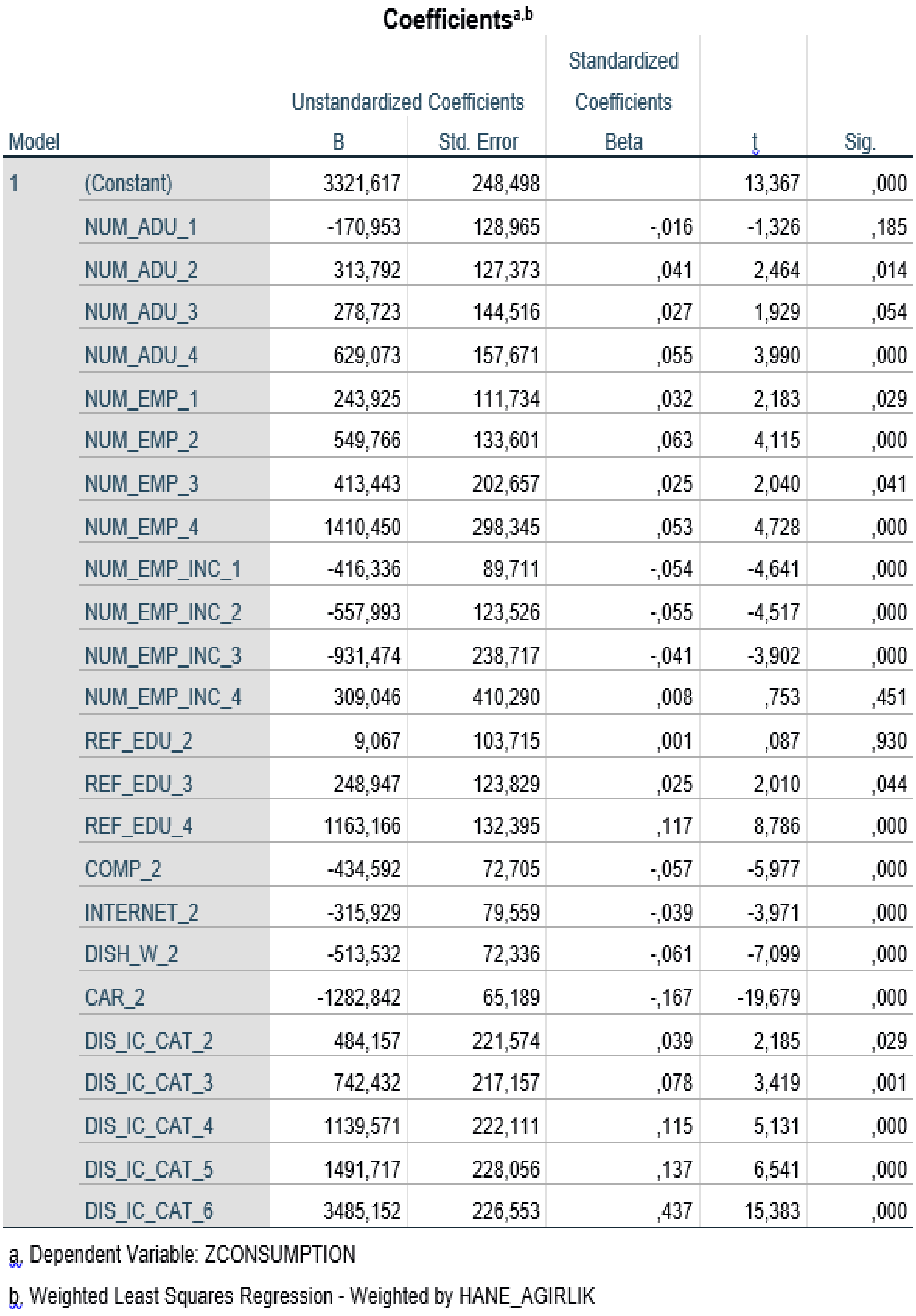
B.3. HBS Unweighted Linear Regression



Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3194,659	230,920		13,834	,000
	NUM_ADU_1	-66,633	124,445	-,006	-,535	,592
	NUM_ADU_2	291,171	118,924	,040	2,448	,014
	NUM_ADU_3	369,253	134,996	,038	2,735	,006
	NUM_ADU_4	724,333	152,921	,059	4,737	,000
	NUM_EMP_1	161,290	101,154	,022	1,595	,111
	NUM_EMP_2	433,409	115,083	,052	3,766	,000
	NUM_EMP_3	31,567	180,140	,002	,175	,861
	NUM_EMP_4	1075,781	286,834	,038	3,751	,000
	NUM_EMP_INC_1	-313,842	81,631	-,042	-3,845	,000
	NUM_EMP_INC_2	-426,069	112,589	-,042	-3,784	,000
	NUM_EMP_INC_3	-508,683	241,472	-,020	-2,107	,035
	NUM_EMP_INC_4	211,312	480,874	,004	,439	,660
	REF_EDU_2	-12,065	97,463	-,002	-,124	,901
	REF_EDU_3	257,591	118,989	,027	2,165	,030
	REF_EDU_4	1174,508	129,636	,114	9,060	,000
	COMP_2	-446,456	71,989	-,060	-6,202	,000
	INTERNET_2	-336,512	74,488	-,044	-4,518	,000
	DISH_W_2	-480,796	68,781	-,061	-6,990	,000
	CAR_2	-1157,203	62,980	-,156	-18,374	,000
	DIS_IC_CAT_2	472,406	201,788	,042	2,341	,019
	DIS_IC_CAT_3	750,911	199,260	,085	3,768	,000
	DIS_IC_CAT_4	1157,056	204,926	,122	5,646	,000
	DIS_IC_CAT_5	1519,082	211,602	,143	7,179	,000
	DIS_IC_CAT_6	3416,251	210,408	,433	16,236	,000

a. Dependent Variable: ZCONSUMPTION

B.4. HBS Weighted Linear Regression



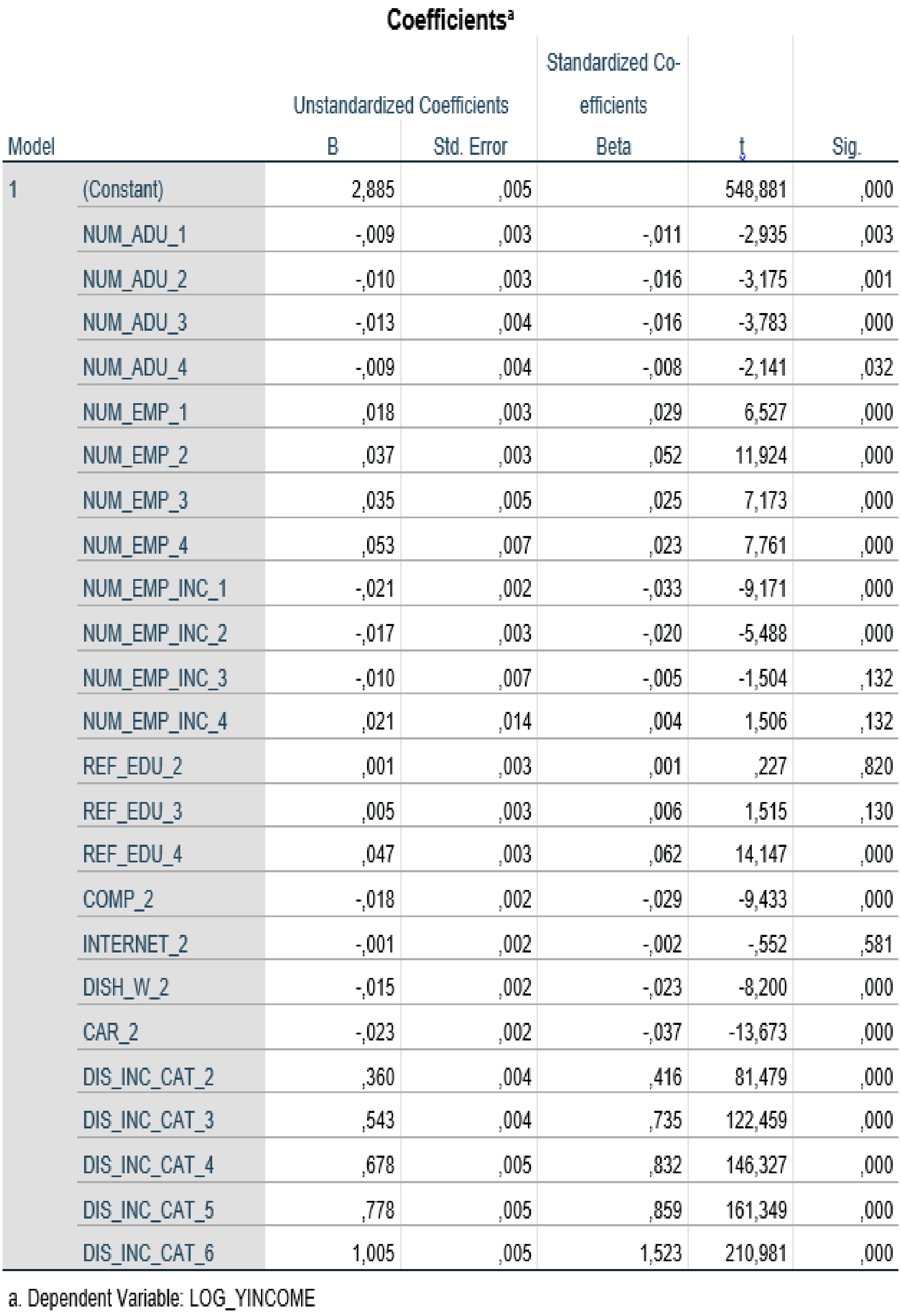
Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3321,617	248,498		13,367	,000
	NUM_ADU_1	-170,953	128,965	-,016	-1,326	,185
	NUM_ADU_2	313,792	127,373	,041	2,464	,014
	NUM_ADU_3	278,723	144,516	,027	1,929	,054
	NUM_ADU_4	629,073	157,671	,055	3,990	,000
	NUM_EMP_1	243,925	111,734	,032	2,183	,029
	NUM_EMP_2	549,766	133,601	,063	4,115	,000
	NUM_EMP_3	413,443	202,657	,025	2,040	,041
	NUM_EMP_4	1410,450	298,345	,053	4,728	,000
	NUM_EMP_INC_1	-416,336	89,711	-,054	-4,641	,000
	NUM_EMP_INC_2	-557,993	123,526	-,055	-4,517	,000
	NUM_EMP_INC_3	-931,474	238,717	-,041	-3,902	,000
	NUM_EMP_INC_4	309,046	410,290	,008	,753	,451
	REF_EDU_2	9,067	103,715	,001	,087	,930
	REF_EDU_3	248,947	123,829	,025	2,010	,044
	REF_EDU_4	1163,166	132,395	,117	8,786	,000
	COMP_2	-434,592	72,705	-,057	-5,977	,000
	INTERNET_2	-315,929	79,559	-,039	-3,971	,000
	DISH_W_2	-513,532	72,336	-,061	-7,099	,000
	CAR_2	-1282,842	65,189	-,167	-19,679	,000
	DIS_IC_CAT_2	484,157	221,574	,039	2,185	,029
	DIS_IC_CAT_3	742,432	217,157	,078	3,419	,001
	DIS_IC_CAT_4	1139,571	222,111	,115	5,131	,000
	DIS_IC_CAT_5	1491,717	228,056	,137	6,541	,000
	DIS_IC_CAT_6	3485,152	226,553	,437	15,383	,000

a. Dependent Variable: ZCONSUMPTION

b. Weighted Least Squares Regression - Weighted by HANE_AGIRLIK

B.5. SILC Unweighted Linear Regression (on the log of target var.)

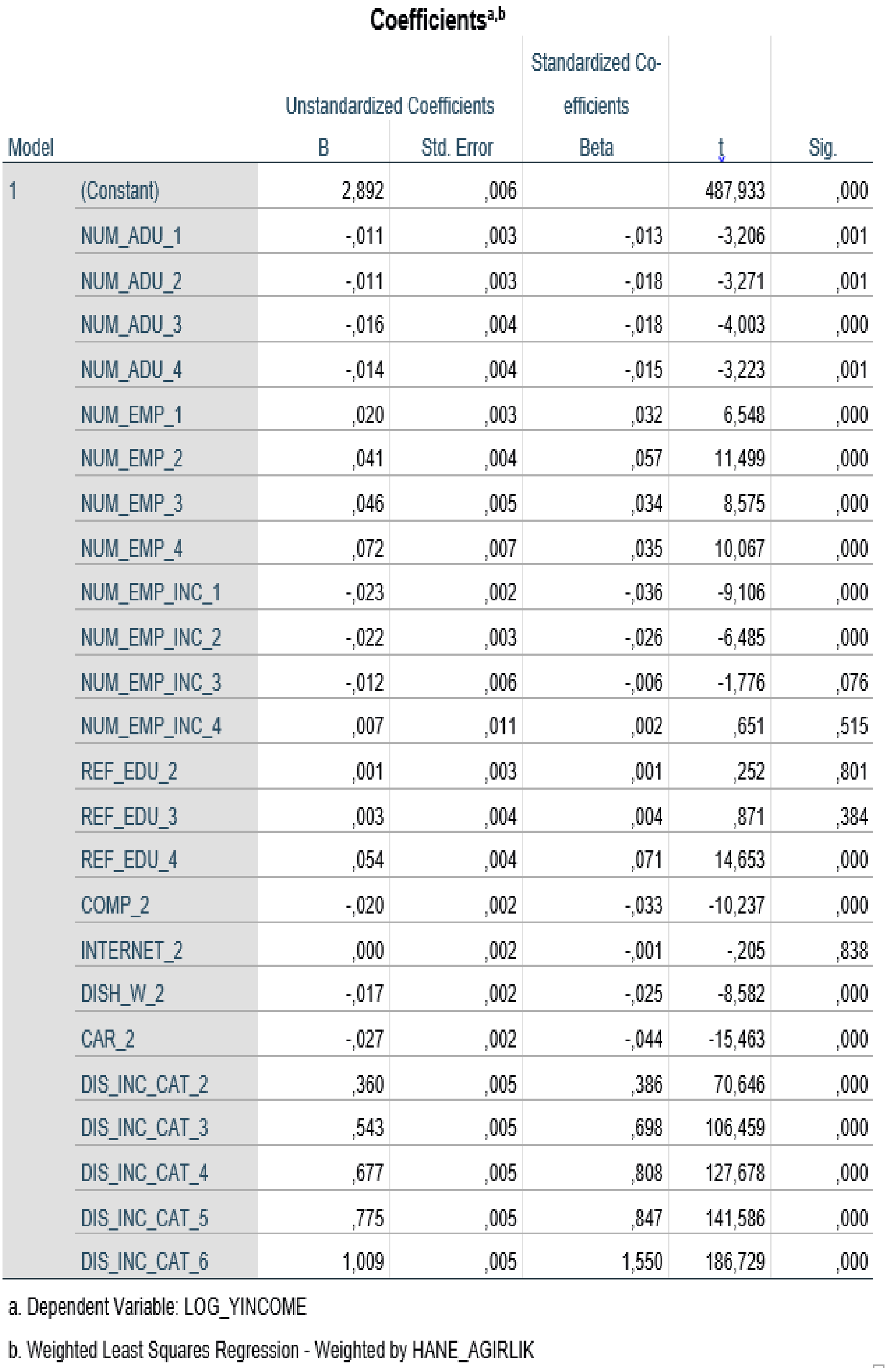


Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2,885	,005		548,881	,000
	NUM_ADU_1	-,009	,003	-,011	-2,935	,003
	NUM_ADU_2	-,010	,003	-,016	-3,175	,001
	NUM_ADU_3	-,013	,004	-,016	-3,783	,000
	NUM_ADU_4	-,009	,004	-,008	-2,141	,032
	NUM_EMP_1	,018	,003	,029	6,527	,000
	NUM_EMP_2	,037	,003	,052	11,924	,000
	NUM_EMP_3	,035	,005	,025	7,173	,000
	NUM_EMP_4	,053	,007	,023	7,761	,000
	NUM_EMP_INC_1	-,021	,002	-,033	-9,171	,000
	NUM_EMP_INC_2	-,017	,003	-,020	-5,488	,000
	NUM_EMP_INC_3	-,010	,007	-,005	-1,504	,132
	NUM_EMP_INC_4	,021	,014	,004	1,506	,132
	REF_EDU_2	,001	,003	,001	,227	,820
	REF_EDU_3	,005	,003	,006	1,515	,130
	REF_EDU_4	,047	,003	,062	14,147	,000
	COMP_2	-,018	,002	-,029	-9,433	,000
	INTERNET_2	-,001	,002	-,002	-,552	,581
	DISH_W_2	-,015	,002	-,023	-8,200	,000
	CAR_2	-,023	,002	-,037	-13,673	,000
	DIS_INC_CAT_2	,360	,004	,416	81,479	,000
	DIS_INC_CAT_3	,543	,004	,735	122,459	,000
	DIS_INC_CAT_4	,678	,005	,832	146,327	,000
	DIS_INC_CAT_5	,778	,005	,859	161,349	,000
	DIS_INC_CAT_6	1,005	,005	1,523	210,981	,000

a. Dependent Variable: LOG_YINCOME

B.6. SILC Weighted Linear Regression (on the log of target var.)



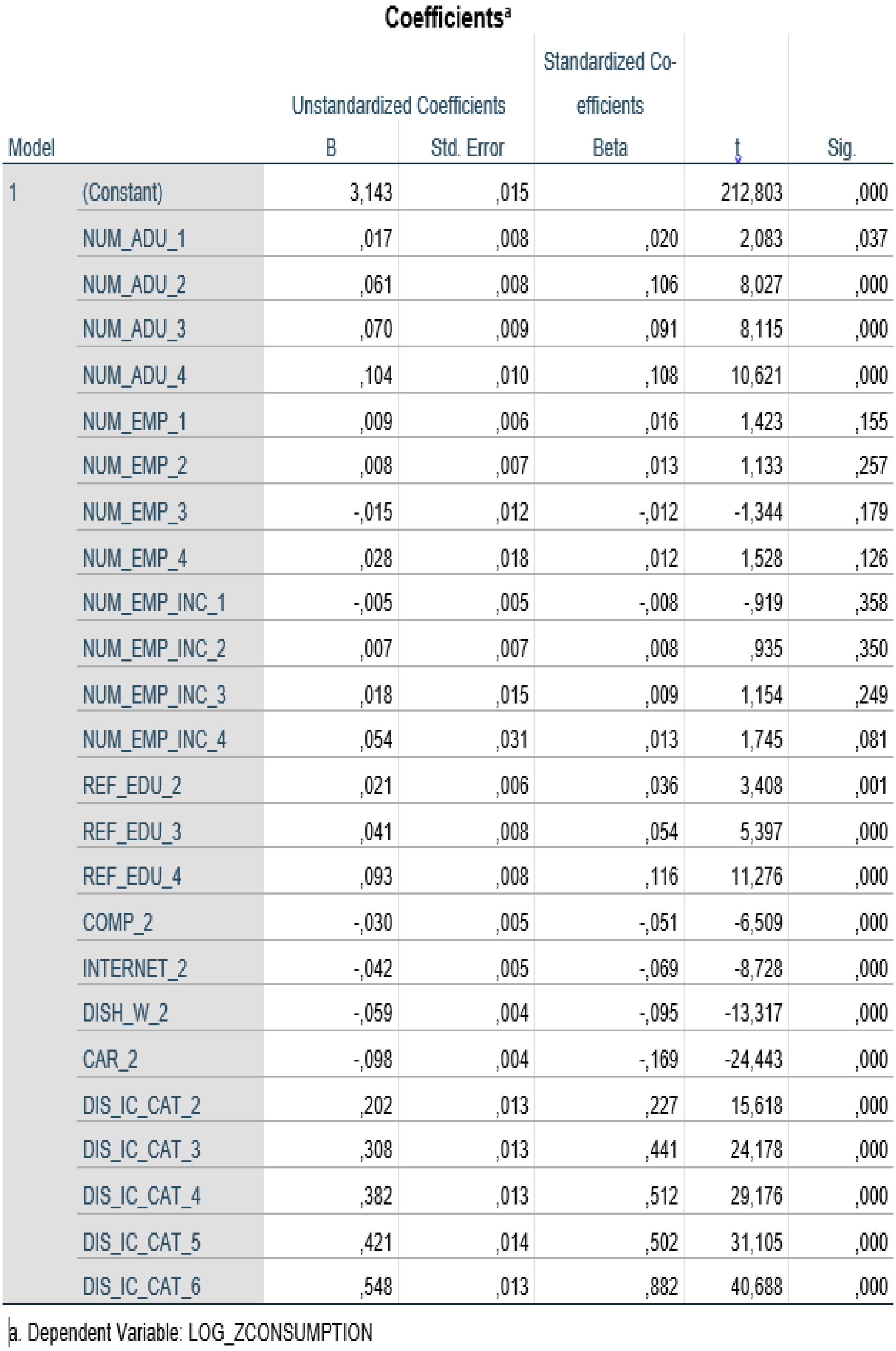
Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2,892	,006		487,933	,000
	NUM_ADU_1	-,011	,003	-,013	-3,206	,001
	NUM_ADU_2	-,011	,003	-,018	-3,271	,001
	NUM_ADU_3	-,016	,004	-,018	-4,003	,000
	NUM_ADU_4	-,014	,004	-,015	-3,223	,001
	NUM_EMP_1	,020	,003	,032	6,548	,000
	NUM_EMP_2	,041	,004	,057	11,499	,000
	NUM_EMP_3	,046	,005	,034	8,575	,000
	NUM_EMP_4	,072	,007	,035	10,067	,000
	NUM_EMP_INC_1	-,023	,002	-,036	-9,106	,000
	NUM_EMP_INC_2	-,022	,003	-,026	-6,485	,000
	NUM_EMP_INC_3	-,012	,006	-,006	-1,776	,076
	NUM_EMP_INC_4	,007	,011	,002	,651	,515
	REF_EDU_2	,001	,003	,001	,252	,801
	REF_EDU_3	,003	,004	,004	,871	,384
	REF_EDU_4	,054	,004	,071	14,653	,000
	COMP_2	-,020	,002	-,033	-10,237	,000
	INTERNET_2	,000	,002	-,001	-,205	,838
	DISH_W_2	-,017	,002	-,025	-8,582	,000
	CAR_2	-,027	,002	-,044	-15,463	,000
	DIS_INC_CAT_2	,360	,005	,386	70,646	,000
	DIS_INC_CAT_3	,543	,005	,698	106,459	,000
	DIS_INC_CAT_4	,677	,005	,808	127,678	,000
	DIS_INC_CAT_5	,775	,005	,847	141,586	,000
	DIS_INC_CAT_6	1,009	,005	1,550	186,729	,000

a. Dependent Variable: LOG_YINCOME

b. Weighted Least Squares Regression - Weighted by HANE_AGIRLIK

B.7. HBS Unweighted Linear Regression (on the log of target var.)

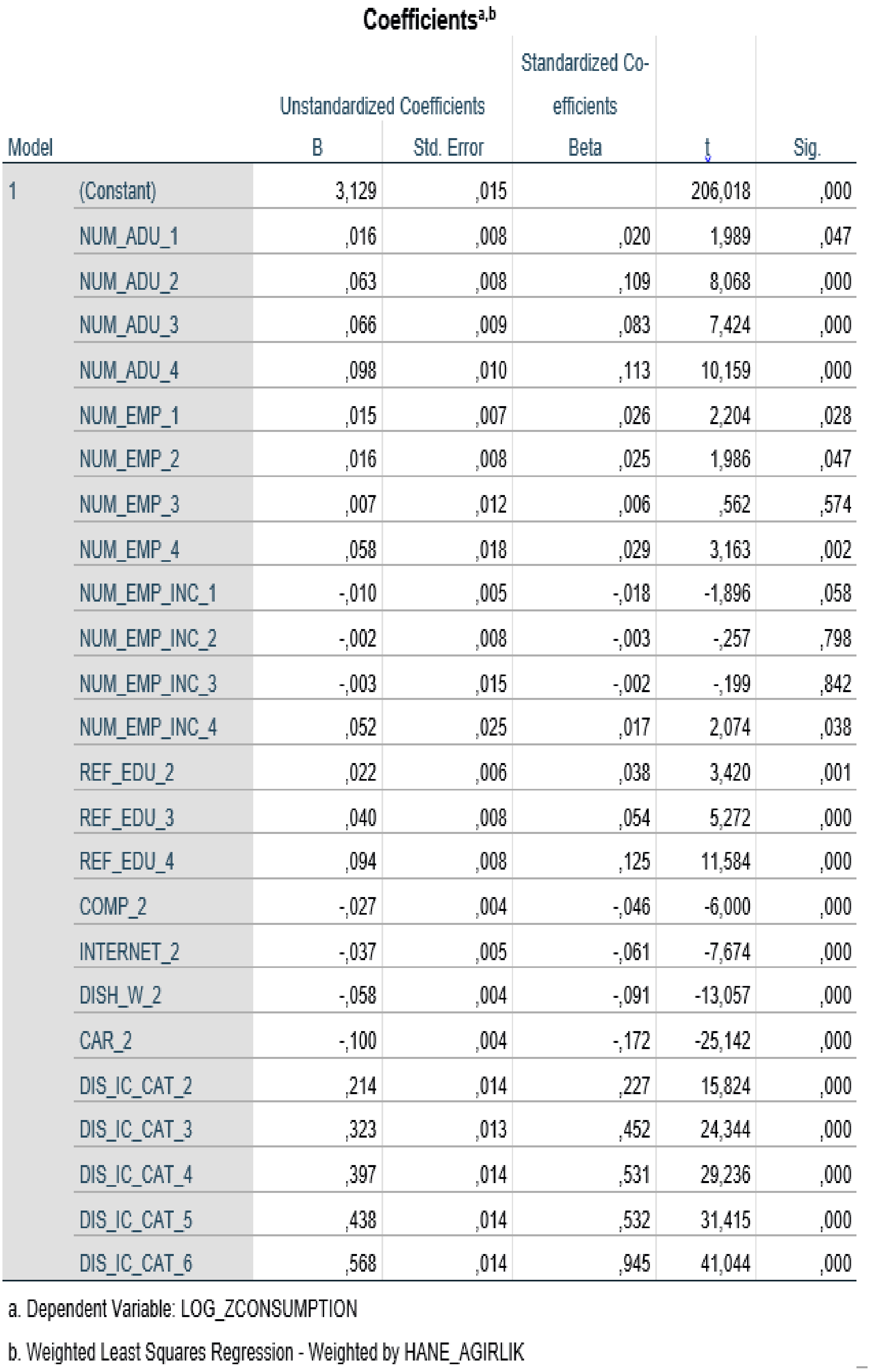


Coefficients[a]

		Unstandardized Coefficients		Standardized Co-efficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	3,143	,015		212,803	,000
	NUM_ADU_1	,017	,008	,020	2,083	,037
	NUM_ADU_2	,061	,008	,106	8,027	,000
	NUM_ADU_3	,070	,009	,091	8,115	,000
	NUM_ADU_4	,104	,010	,108	10,621	,000
	NUM_EMP_1	,009	,006	,016	1,423	,155
	NUM_EMP_2	,008	,007	,013	1,133	,257
	NUM_EMP_3	-,015	,012	-,012	-1,344	,179
	NUM_EMP_4	,028	,018	,012	1,528	,126
	NUM_EMP_INC_1	-,005	,005	-,008	-,919	,358
	NUM_EMP_INC_2	,007	,007	,008	,935	,350
	NUM_EMP_INC_3	,018	,015	,009	1,154	,249
	NUM_EMP_INC_4	,054	,031	,013	1,745	,081
	REF_EDU_2	,021	,006	,036	3,408	,001
	REF_EDU_3	,041	,008	,054	5,397	,000
	REF_EDU_4	,093	,008	,116	11,276	,000
	COMP_2	-,030	,005	-,051	-6,509	,000
	INTERNET_2	-,042	,005	-,069	-8,728	,000
	DISH_W_2	-,059	,004	-,095	-13,317	,000
	CAR_2	-,098	,004	-,169	-24,443	,000
	DIS_IC_CAT_2	,202	,013	,227	15,618	,000
	DIS_IC_CAT_3	,308	,013	,441	24,178	,000
	DIS_IC_CAT_4	,382	,013	,512	29,176	,000
	DIS_IC_CAT_5	,421	,014	,502	31,105	,000
	DIS_IC_CAT_6	,548	,013	,882	40,688	,000

a. Dependent Variable: LOG_ZCONSUMPTION

B.8. HBS Weighted Linear Regression (on the log of target var.)



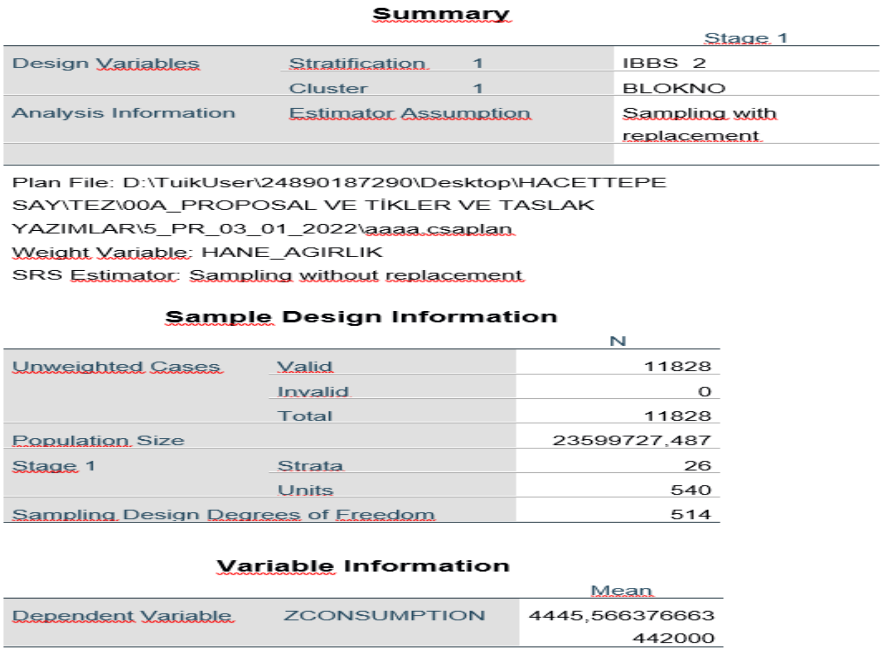
Coefficients[a,b]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	3,129	,015		206,018	,000
	NUM_ADU_1	,016	,008	,020	1,989	,047
	NUM_ADU_2	,063	,008	,109	8,068	,000
	NUM_ADU_3	,066	,009	,083	7,424	,000
	NUM_ADU_4	,098	,010	,113	10,159	,000
	NUM_EMP_1	,015	,007	,026	2,204	,028
	NUM_EMP_2	,016	,008	,025	1,986	,047
	NUM_EMP_3	,007	,012	,006	,562	,574
	NUM_EMP_4	,058	,018	,029	3,163	,002
	NUM_EMP_INC_1	-,010	,005	-,018	-1,896	,058
	NUM_EMP_INC_2	-,002	,008	-,003	-,257	,798
	NUM_EMP_INC_3	-,003	,015	-,002	-,199	,842
	NUM_EMP_INC_4	,052	,025	,017	2,074	,038
	REF_EDU_2	,022	,006	,038	3,420	,001
	REF_EDU_3	,040	,008	,054	5,272	,000
	REF_EDU_4	,094	,008	,125	11,584	,000
	COMP_2	-,027	,004	-,046	-6,000	,000
	INTERNET_2	-,037	,005	-,061	-7,674	,000
	DISH_W_2	-,058	,004	-,091	-13,057	,000
	CAR_2	-,100	,004	-,172	-25,142	,000
	DIS_IC_CAT_2	,214	,014	,227	15,824	,000
	DIS_IC_CAT_3	,323	,013	,452	24,344	,000
	DIS_IC_CAT_4	,397	,014	,531	29,236	,000
	DIS_IC_CAT_5	,438	,014	,532	31,415	,000
	DIS_IC_CAT_6	,568	,014	,945	41,044	,000

a. Dependent Variable: LOG_ZCONSUMPTION

b. Weighted Least Squares Regression - Weighted by HANE_AGIRLIK

B.9. HBS Regression Results with survey design variables



Summary

			Stage 1
Design Variables	Stratification	1	IBBS 2
	Cluster	1	BLOKNO
Analysis Information	Estimator Assumption		Sampling with replacement

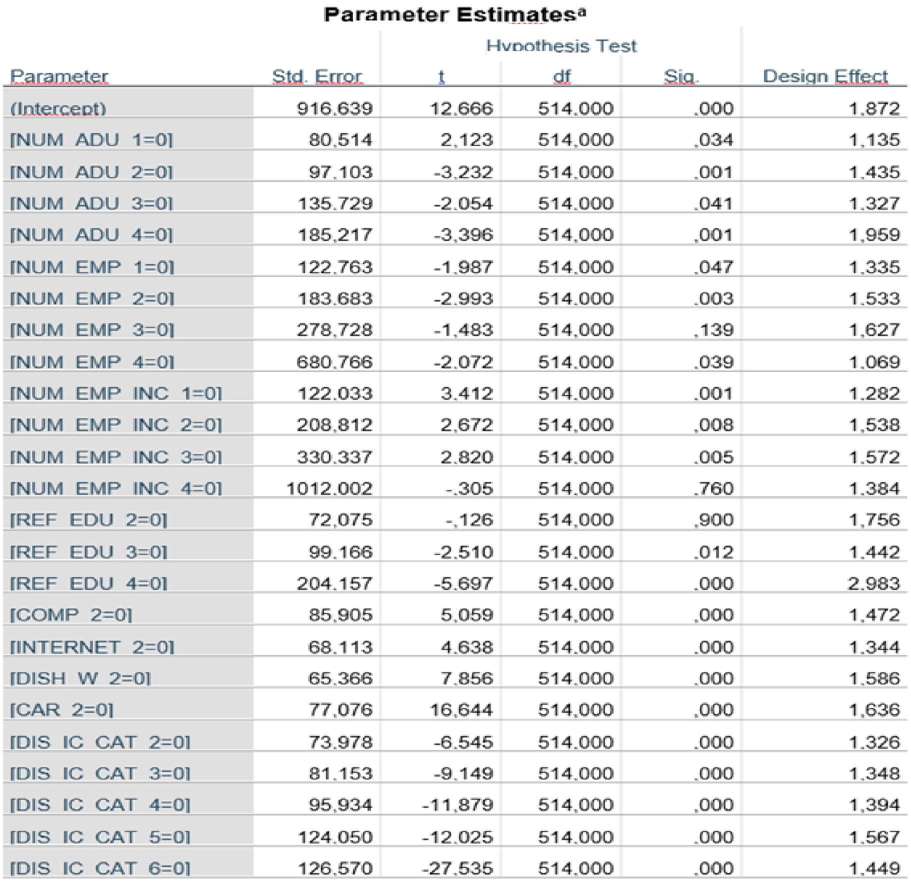
Plan File: D:\TuikUser\24890187290\Desktop\HACETTEPE SAY\TEZ\00A_PROPOSAL VE TİKLER VE TASLAK YAZIMLAR\5_PR_03_01_2022\aaaa.csaplan
Weight Variable: HANE_AGIRLIK
SRS Estimator: Sampling without replacement

Sample Design Information

		N
Unweighted Cases	Valid	11828
	Invalid	0
	Total	11828
Population Size		23599727,487
Stage 1	Strata	26
	Units	540
Sampling Design Degrees of Freedom		514

Variable Information

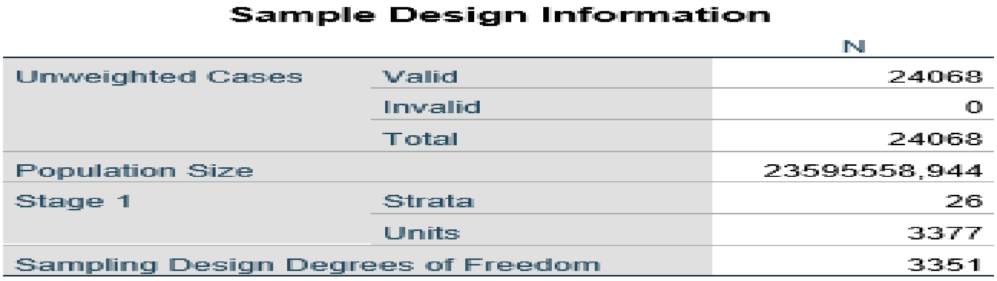
		Mean
Dependent Variable	ZCONSUMPTION	4445,566376663442000



Parameter Estimates[a]

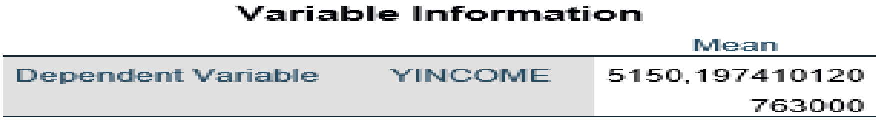
		Hypothesis Test			
Parameter	Std. Error	t	df	Sig.	Design Effect
(Intercept)	916,639	12,666	514,000	,000	1,872
[NUM ADU 1=0]	80,514	2,123	514,000	,034	1,135
[NUM ADU 2=0]	97,103	-3,232	514,000	,001	1,435
[NUM ADU 3=0]	135,729	-2,054	514,000	,041	1,327
[NUM ADU 4=0]	185,217	-3,396	514,000	,001	1,959
[NUM EMP 1=0]	122,763	-1,987	514,000	,047	1,335
[NUM EMP 2=0]	183,683	-2,993	514,000	,003	1,533
[NUM EMP 3=0]	278,728	-1,483	514,000	,139	1,627
[NUM EMP 4=0]	680,766	-2,072	514,000	,039	1,069
[NUM EMP INC 1=0]	122,033	3,412	514,000	,001	1,282
[NUM EMP INC 2=0]	208,812	2,672	514,000	,008	1,538
[NUM EMP INC 3=0]	330,337	2,820	514,000	,005	1,572
[NUM EMP INC 4=0]	1012,002	-,305	514,000	,760	1,384
[REF EDU 2=0]	72,075	-,126	514,000	,900	1,756
[REF EDU 3=0]	99,166	-2,510	514,000	,012	1,442
[REF EDU 4=0]	204,157	-5,697	514,000	,000	2,983
[COMP 2=0]	85,905	5,059	514,000	,000	1,472
[INTERNET 2=0]	68,113	4,638	514,000	,000	1,344
[DISH W 2=0]	65,366	7,856	514,000	,000	1,586
[CAR 2=0]	77,076	16,644	514,000	,000	1,636
[DIS IC CAT 2=0]	73,978	-6,545	514,000	,000	1,326
[DIS IC CAT 3=0]	81,153	-9,149	514,000	,000	1,348
[DIS IC CAT 4=0]	95,934	-11,879	514,000	,000	1,394
[DIS IC CAT 5=0]	124,050	-12,025	514,000	,000	1,567
[DIS IC CAT 6=0]	126,570	-27,535	514,000	,000	1,449

B.10. SILC Regression Results with survey design variables



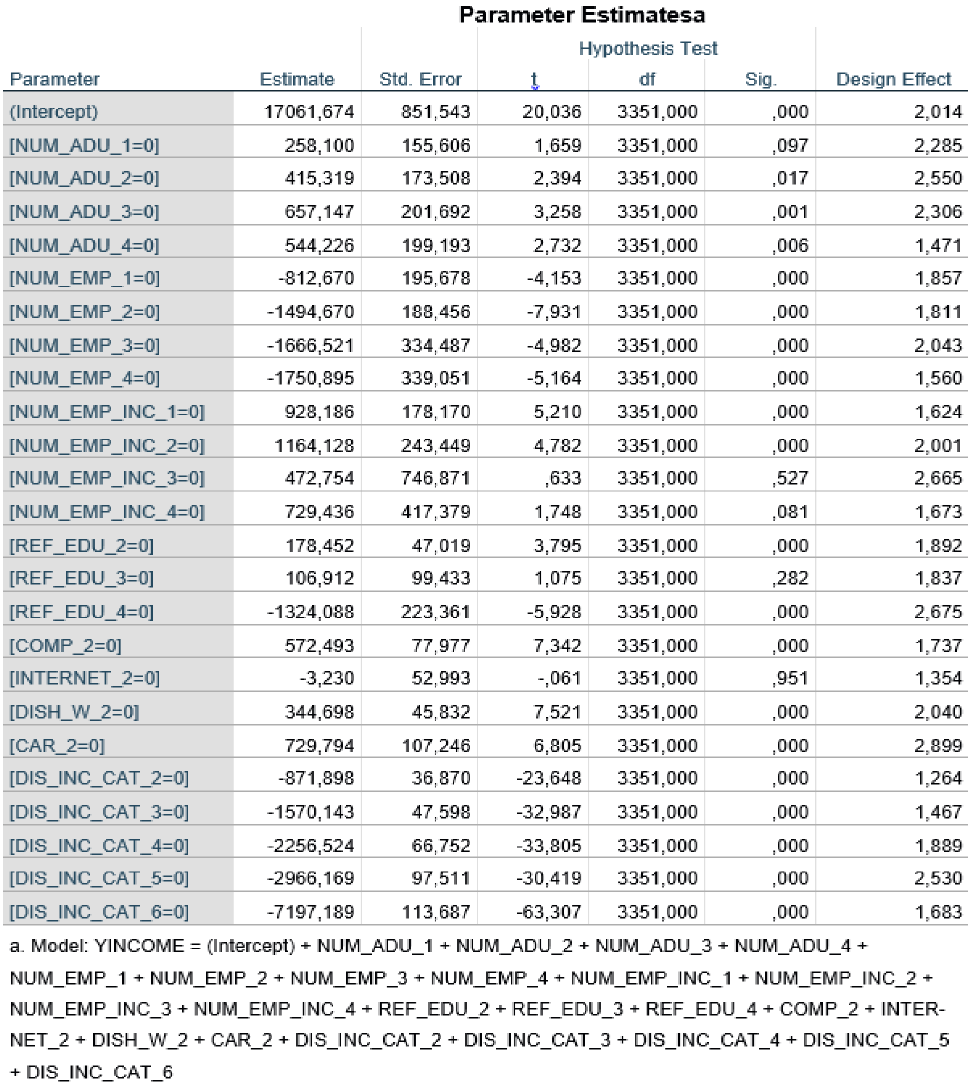
Sample Design Information

		N
Unweighted Cases	Valid	24068
	Invalid	0
	Total	24068
Population Size		23595558,944
Stage 1	Strata	26
	Units	3377
Sampling Design Degrees of Freedom		3351



Variable Information

		Mean
Dependent Variable	YINCOME	5150,197410120763000

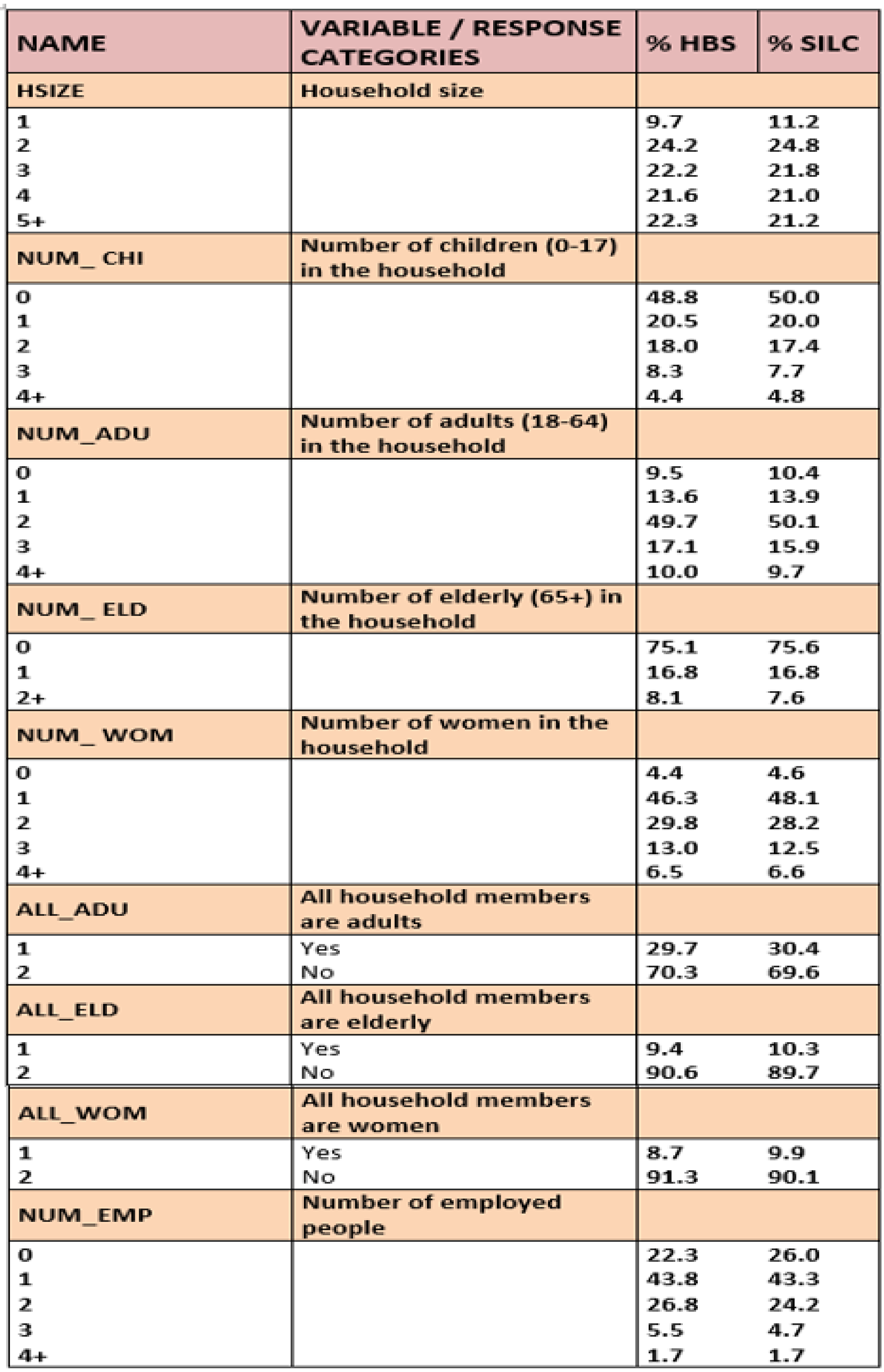


Parameter Estimatesa

Parameter	Estimate	Std. Error	Hypothesis Test			Design Effect
			t	df	Sig.	
(Intercept)	17061,674	851,543	20,036	3351,000	,000	2,014
[NUM_ADU_1=0]	258,100	155,606	1,659	3351,000	,097	2,285
[NUM_ADU_2=0]	415,319	173,508	2,394	3351,000	,017	2,550
[NUM_ADU_3=0]	657,147	201,692	3,258	3351,000	,001	2,306
[NUM_ADU_4=0]	544,226	199,193	2,732	3351,000	,006	1,471
[NUM_EMP_1=0]	-812,670	195,678	-4,153	3351,000	,000	1,857
[NUM_EMP_2=0]	-1494,670	188,456	-7,931	3351,000	,000	1,811
[NUM_EMP_3=0]	-1666,521	334,487	-4,982	3351,000	,000	2,043
[NUM_EMP_4=0]	-1750,895	339,051	-5,164	3351,000	,000	1,560
[NUM_EMP_INC_1=0]	928,186	178,170	5,210	3351,000	,000	1,624
[NUM_EMP_INC_2=0]	1164,128	243,449	4,782	3351,000	,000	2,001
[NUM_EMP_INC_3=0]	472,754	746,871	,633	3351,000	,527	2,665
[NUM_EMP_INC_4=0]	729,436	417,379	1,748	3351,000	,081	1,673
[REF_EDU_2=0]	178,452	47,019	3,795	3351,000	,000	1,892
[REF_EDU_3=0]	106,912	99,433	1,075	3351,000	,282	1,837
[REF_EDU_4=0]	-1324,088	223,361	-5,928	3351,000	,000	2,675
[COMP_2=0]	572,493	77,977	7,342	3351,000	,000	1,737
[INTERNET_2=0]	-3,230	52,993	-,061	3351,000	,951	1,354
[DISH_W_2=0]	344,698	45,832	7,521	3351,000	,000	2,040
[CAR_2=0]	729,794	107,246	6,805	3351,000	,000	2,899
[DIS_INC_CAT_2=0]	-871,898	36,870	-23,648	3351,000	,000	1,264
[DIS_INC_CAT_3=0]	-1570,143	47,598	-32,987	3351,000	,000	1,467
[DIS_INC_CAT_4=0]	-2256,524	66,752	-33,805	3351,000	,000	1,889
[DIS_INC_CAT_5=0]	-2966,169	97,511	-30,419	3351,000	,000	2,530
[DIS_INC_CAT_6=0]	-7197,189	113,687	-63,307	3351,000	,000	1,683

a. Model: YINCOME = (Intercept) + NUM_ADU_1 + NUM_ADU_2 + NUM_ADU_3 + NUM_ADU_4 + NUM_EMP_1 + NUM_EMP_2 + NUM_EMP_3 + NUM_EMP_4 + NUM_EMP_INC_1 + NUM_EMP_INC_2 + NUM_EMP_INC_3 + NUM_EMP_INC_4 + REF_EDU_2 + REF_EDU_3 + REF_EDU_4 + COMP_2 + INTERNET_2 + DISH_W_2 + CAR_2 + DIS_INC_CAT_2 + DIS_INC_CAT_3 + DIS_INC_CAT_4 + DIS_INC_CAT_5 + DIS_INC_CAT_6

Appendix C. Distributions of the Variables



NAME	VARIABLE / RESPONSE CATEGORIES	% HBS	% SILC
HSIZE	Household size		
1		9.7	11.2
2		24.2	24.8
3		22.2	21.8
4		21.6	21.0
5+		22.3	21.2
NUM_ CHI	Number of children (0-17) in the household		
0		48.8	50.0
1		20.5	20.0
2		18.0	17.4
3		8.3	7.7
4+		4.4	4.8
NUM_ADU	Number of adults (18-64) in the household		
0		9.5	10.4
1		13.6	13.9
2		49.7	50.1
3		17.1	15.9
4+		10.0	9.7
NUM_ ELD	Number of elderly (65+) in the household		
0		75.1	75.6
1		16.8	16.8
2+		8.1	7.6
NUM_ WOM	Number of women in the household		
0		4.4	4.6
1		46.3	48.1
2		29.8	28.2
3		13.0	12.5
4+		6.5	6.6
ALL_ADU	All household members are adults		
1	Yes	29.7	30.4
2	No	70.3	69.6
ALL_ELD	All household members are elderly		
1	Yes	9.4	10.3
2	No	90.6	89.7
ALL_WOM	All household members are women		
1	Yes	8.7	9.9
2	No	91.3	90.1
NUM_EMP	Number of employed people		
0		22.3	26.0
1		43.8	43.3
2		26.8	24.2
3		5.5	4.7
4+		1.7	1.7

NAME	VARIABLE / RESPONSE CATEGORIES	% HBS	% SILC
NUM_EMP_INC	**Number of individuals with employee income**		
0		41.7	42.7
1		40.3	40.1
2		15.4	14.8
3		2.1	2.0
4+		0.5	0.3
NUM_SELF_EMP_INC	**Number of individuals with self-employed income**		
0		76.5	81.1
1		22.6	18.2
2+		0.9	0.7
NUM_RET_INC	**Number of individuals with retired income**		
0		78.6	70.1
1		18.8	25.2
2+		2.6	4.7
REF_SEX	**Reference person's sex**		
1	Male	84.6	78.7
2	Female	15.4	21.3
REF_AGE	**Reference person's age group**		
<25		1.0	4.7
>=25 and <35		13.3	18.7
>=35 and <45		23.4	24.2
>=45 and <65		43.7	36.0
>=65		18.5	16.5
REF_MAR	**Reference person's marital status**		
1	Married	81.7	76.3
2	Never married	4.0	9.8
3	Widow	9.8	9.8
4	Divorced	4.4	4.2
REF_EDU	**Reference person's education**		
1	No formal education	11.1	11.6
2	Less than high school	56.6	51.0
3	High school	17.3	18.0
4	Higher education	15.0	19.4
REF_PRO	**Reference person's professional status**		
0	Doesn't work	33.1	30.9
1	Regular employee	39.1	44.8
2	Casual employee	4.5	4.6
3	Employer	4.0	4.0
4	Own account worker	19.3	15.3
5	Unpaid family worker	0.1	0.4

NAME	VARIABLE / RESPONSE CATEGORIES	% HBS	% SILC
REF_OCC	Reference person's occupation		
0	Doesn't work	33.1	30.9
1	Legislators, senior, officials and managers	4.7	4.9
2	Professionals	5.7	8.1
3	Technicians and associate professionals	3.6	4.4
4	Clerks	3.2	3.4
5	Service workers and shop and market sales workers	11.3	12.8
6	Skilled agricultural, and fishery workers	12.5	9.2
7	Craft and related trades workers	10.2	10.8
8	Plant and machine operators and assemblers	8.1	8.3
9	Elemantary occupations	7.5	7.3
REF_ECO	Reference person's economic activity of work		
0	Doesn't work	33.1	30.9
1	Agriculture, Forestry and Fishing	13.8	9.9
2	Mining and Quarrying	0.6	0.7
3	Manufacturing	11.7	12.6
4	Electricity, Gas, Steam, Water Supply, Sewerage etc.	0.6	1.0
5	Construction	6.2	6.6
6	Whole-sale and Retail Trade	8.6	9.3
7	Transportation and Storage	3.9	3.7
8	Accommodation and Food Service Activities	3.2	3.5
9	Information and Communication	0.5	0.5
10	Financial and Insurance Activities	0.8	0.8
11	Real Estate Activities	0.8	0.8
12	Professional, Scientific and Technical Activities	1.2	1.5
13	Administrative and Support Service Activities	2.0	2.5
14	Public Administration and Defence	6.4	6.0
15	Education	2.8	4.1
16	Human Health and Social Work Activities	1.7	3.1
17	Arts, Entertainment and Recreation	0.3	0.4
18	Other Activities	2.0	2.2
REF_WHRS	Reference person's number of weekly working hours		
0		33.1	30.9
<20		1.9	0.8
>=20 and <40		7.2	5.4
>=40 and <60		40.0	42.2
>=60		17.8	20.8
DWE	Dwelling type		
1	Detached / semidetached	45.4	40.2
2	Apartment	54.6	59.8
TENURE	Tenure status		
1	Owner	60.7	59.4
2	Tenant	23.4	23.9
3	Lodging	1.6	1.4
4	Not owner but accommodation is provided free	14.3	15.3

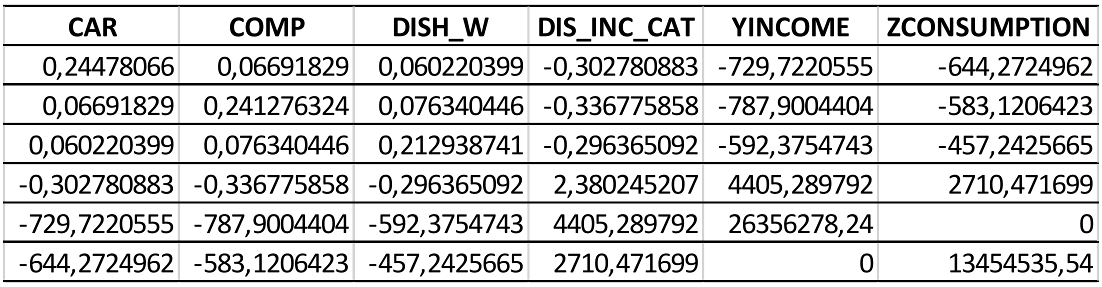
NAME	VARIABLE / RESPONSE CATEGORIES	%	%
		HBS	SILC
RENT_CAT	Current rent related to occupied dwelling (including imputed rent)		
<500		38.0	37.3
>=500 and <1000		47.7	50.5
>=1000 and <1500		9.1	9.0
>=1500		5.3	3.3
ROOM_NUM	Number of rooms (except for kitchen, bathroom and toilet) available to the household		
1		0.7	1.1
2		6.7	7.9
3		39.1	37.7
4+		53.5	53.3
TOT_AR	Total space available to the household (m2)		
<=60		5.8	6.0
>60 and <=80		13.8	13.2
>80 and <=100		32.2	31.5
>100 and <=120		23.9	21.7
>120		24.4	27.6
HEAT_SYS	Heating system of the dwelling		
1	Stove (Coal, gas, natural gas, electricity, etc.)	48.6	45.6
2	Radiator (Joint or central heating)	10.0	11.5
3	Radiator (Heating system for only a flat/combi boiler)	38.4	39.2
4	Air conditioner	2.8	3.6
5	Other	0.2	0.1
BATH	Bath or shower in dwelling		
1	Yes	98.7	98.1
2	No	1.3	1.9
TOILET	Indoor flushing toilet for sole use of household		
1	Yes	93.0	94.8
2	No	7.0	5.2
PIPED_WAT	Piped water		
1	Yes	99.8	99.3
2	No	0.2	0.7
HOT_WAT	Hot water		
1	Yes	94.1	91.8
2	No	5.9	8.2
MOBILE	Mobile		
1	Yes	98.3	97.3
2	No	1.7	2.7
COMP	Computer		
1	Yes	40.1	40.9
2	No	59.9	59.1

NAME	VARIABLE / RESPONSE CATEGORIES	% HBS	% SILC
INTERNET	**Internet**		
1	Yes	63.2	65.6
2	No	36.8	34.4
WASH_M	**Washing machine**		
1	Yes	97.9	97.8
2	No	2.1	2.2
REFRIG	**Refrigerator**		
1	Yes	99.2	99.0
2	No	0.8	1.0
DISH_W	**Dishwasher**		
1	Yes	68.1	69.8
2	No	31.9	30.2
AIR_CON	**Air conditioner**		
1	Yes	19.9	21.6
2	No	80.1	78.4
CAR	**Car**		
1	Yes	43.2	42.6
2	No	56.8	57.4
DIS_INC_CAT	**Total disposable household income**		
<=1000		2.3	5.0
>1000 and <=2000		12.0	22.5
>2000 and <=3000		21.9	23.0
>3000 and <=4000		18.3	16.8
>4000 and <=5000		13.8	11.0
>5000		31.6	21.8

Appendix D. R Codes

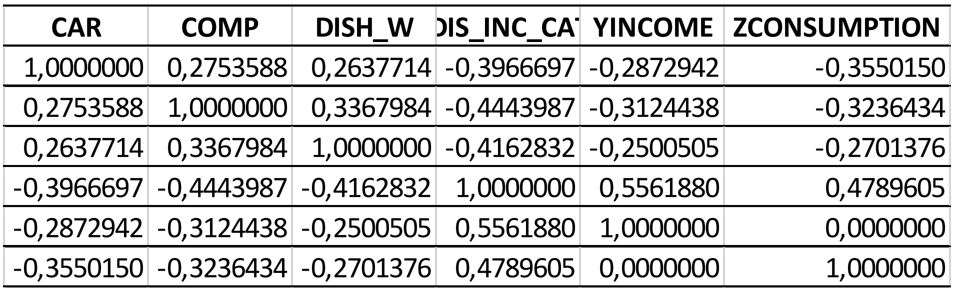
```
“#Parametric Micro #
AA$COMP=as.numeric(AA$COMP)
AA$DISH_W=as.numeric(AA$DISH_W)
AA$CAR=as.numeric(AA$CAR)
AA$DIS_INC_CAT=as.numeric(AA$DIS_INC_CAT)

BB$COMP=as.numeric(BB$COMP)
BB$DISH_W=as.numeric(BB$DISH_W)
BB$CAR=as.numeric(BB$CAR)
BB$DIS_INC_CAT=as.numeric(BB$DIS_INC_CAT)

#regression Y vs. X in AA#
>reg.yx <- lm(YINCOME~CAR,data=AA)
>coefficients(reg.yx)
(Intercept)        CAR
  9428.579   -2984.882
>out <- summary(reg.yx)
>out$sigma #residual sd s_Y|X “4917.23”
>pred.y.B <- predict(reg.yx, newdata=BB) #predicted values
>imp.y.B <- pred.y.B + rnorm(nrow(BB), mean=0, sd=out$sigma)
>BB$YINCOME <- imp.y.B “fill in Y in BB

#regression Y vs. X in AA#
>reg.zx <- lm(ZCONSUMPTION~CAR, data=BB)
>coefficients(reg.zx)
(Intercept)    CAR
  5697.353   -2625.395
>out <- summary(reg.zx)
>out$sigma #residual sd s_Z|X   “3429.878”
>pred.z.A <- predict(reg.zx, newdata=AA) #predicted values
>imp.z.A <- pred.z.A + rnorm(nrow(AA), mean=0, sd=out$sigma)”fill Z in AA”
>AA$ZCONSUMPTION <- imp.z.A

**
AUB <- rbind(AA,BB) “concatenation AA U BB”
head(AUB)
cor(AUB) “estimated var-cov”
#Parametric Macro#
#Preliminary procedures on the donor and recipient data#
AA$COMP=as.factor(AA$COMP)
AA$DISH_W=as.factor(AA$DISH_W)
AA$CAR=as.factor(AA$CAR)
AA$DIS_INC_CAT=as.factor(AA$DIS_INC_CAT)
AA$HANE_AGIRLIK=as.numeric(AA$HANE_AGIRLIK)
AA=as.data.frame(AA)
str(AA)
BB$COMP=as.factor(BB$COMP)
BB$DISH_W=as.factor(BB$DISH_W)
BB$CAR=as.factor(BB$CAR)
BB$DIS_INC_CAT=as.factor(BB$DIS_INC_CAT)
BB$HANE_AGIRLIK=as.numeric(BB$HANE_AGIRLIK)
BB=as.data.frame(BB)
```

```
#Parametric Macro MS #
>x.mtc.MS<- c("CAR","COMP","DISH_W","DIS_INC_CAT")

>mix.MS <- mixed.mtc(data.rec=AA, data.don=BB,
match.vars=x.mtc.MS,y.rec="YINCOME", z.don="ZCONSUMPTION",
method="MS",rho.yz=0, micro=FALSE, constr.alg="lpSolve")
       input value for rho.yz is 0
       low(rho.yz)= -0.4172
       up(rho.yz)= 0.9894
       The input value for rho.yz is admissible
>names(mix.MS)
"rho.yz"  "mu"      "vc"      "cor"     "phi"     "res.var" "call"
>mix.MS$rho.yz
  start low.lim  up.lim    used
 0.0000 -0.4172  0.9894  0.0000
>mix.MS$mu "estimated means"
  CAR          COMP        DISH_W     DIS_INC_CAT      YINCOME  ZCONSUMPTION
1.572292     1.593437    1.307472    4.171607       4729.138906  4206.196097
>mix.MS$vc "estimated var-cov matrix"
```



CAR	COMP	DISH_W	DIS_INC_CAT	YINCOME	ZCONSUMPTION
0,24478066	0,06691829	0,060220399	-0,302780883	-729,7220555	-644,2724962
0,06691829	0,241276324	0,076340446	-0,336775858	-787,9004404	-583,1206423
0,060220399	0,076340446	0,212938741	-0,296365092	-592,3754743	-457,2425665
-0,302780883	-0,336775858	-0,296365092	2,380245207	4405,289792	2710,471699
-729,7220555	-787,9004404	-592,3754743	4405,289792	26356278,24	0
-644,2724962	-583,1206423	-457,2425665	2710,471699	0	13454535,54

```
>mix.MS$cor "estimated corelation matrix"
```



CAR	COMP	DISH_W	DIS_INC_CA	YINCOME	ZCONSUMPTION
1,0000000	0,2753588	0,2637714	-0,3966697	-0,2872942	-0,3550150
0,2753588	1,0000000	0,3367984	-0,4443987	-0,3124438	-0,3236434
0,2637714	0,3367984	1,0000000	-0,4162832	-0,2500505	-0,2701376
-0,3966697	-0,4443987	-0,4162832	1,0000000	0,5561880	0,4789605
-0,2872942	-0,3124438	-0,2500505	0,5561880	1,0000000	0,0000000
-0,3550150	-0,3236434	-0,2701376	0,4789605	0,0000000	1,0000000

```
#Parametric Macro ML #
>x.mtc.ML<- c("CAR","COMP","DISH_W","DIS_INC_CAT")
>mix.ML <- mixed.mtc(data.rec=AA, data.don=BB,
match.vars=x.mtc.MS,y.rec="YINCOME", z.don="ZCONSUMPTION",
method="ML",rho.yz=0, micro=FALSE, constr.alg="lpSolve")
>names(mix.ML)
"start.prho.yz" "mu"   "vc"    "cor"      "res.var"       "call"
>mix.ML$rho.yz
  start low.lim  up.lim    used
 0.0000 -0.4172  0.9894  0.0000
>mix.ML$mu "estimated means"
  CAR          COMP        DISH_W   DIS_INC_CAT      YINCOME     ZCONSUMPTION
1.572292     1.593437     1.307472      4.171607  4784.531225  4149.378999
```

```
>mix.ML$vc “estimated var-cov matrix”
```

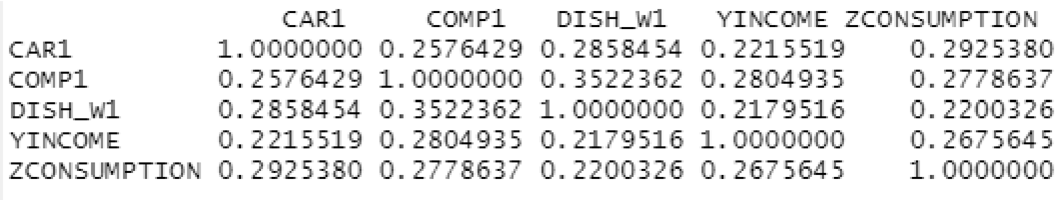
CAR	COMP	DISH_W	DIS_INC_CAT	YINCOME	ZCONSUMPTION
0,244773841	0,066916426	0,060218722	-0,302772449	-717,5699705	-662,0679521
0,066916426	0,241269603	0,076338319	-0,336766476	-770,0380204	-608,1201799
0,060218722	0,076338319	0,212932809	-0,296356835	-564,3334358	-491,5549238
-0,302772449	-0,336766476	-0,296356835	2,380178898	4298,447515	2868,505438
-717,5699705	-770,0380204	-564,3334358	4298,447515	26162724,4	5545247,392
-662,0679521	-608,1201799	-491,5549238	2868,505438	5545247,392	13649749,03

```
>mix.ML$cor “estimated corelation matrix”
```

CAR	COMP	DISH_W	DIS_INC_CAT	YINCOME	ZCONSUMPTION
1,000000000	0,275358783	0,263771421	-0,396669721	-0,283556933	-0,362207744
0,275358783	1,000000000	0,336798382	-0,444398667	-0,306492168	-0,335101041
0,263771421	0,336798382	1,000000000	-0,416283173	-0,239096384	-0,288329066
-0,396669721	-0,444398667	-0,416283173	1,000000000	0,544709996	0,503255516
-0,283556933	-0,306492168	-0,239096384	0,544709996	1,000000000	0,293438511
-0,362207744	-0,335101041	-0,288329066	0,503255516	0,293438511	1,000000000

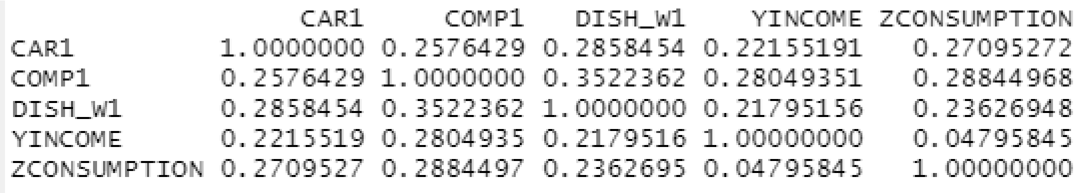
```
# Preliminary procedures on the donor and recipient data for mixed matc.#
AA_MIX$COMP=as.factor(AA_MIX$COMP)
AA_MIX$DISH_W=as.factor(AA_MIX$DISH_W)
AA_MIX$CAR=as.factor(AA_MIX$CAR)
AA_MIX$DIS_INC_CAT=as.factor(AA_MIX$DIS_INC_CAT)
AA_MIX$HANE_AGIRLIK=as.numeric(AA_MIX$HANE_AGIRLIK)
AA_MIX=as.data.frame(AA_MIX)
str(AA_MIX)
BB_MIX$COMP=as.factor(BB_MIX$COMP)
BB_MIX$DISH_W=as.factor(BB_MIX$DISH_W)
BB_MIX$CAR=as.factor(BB_MIX$CAR)
BB_MIX$DIS_INC_CAT=as.factor(BB_MIX$DIS_INC_CAT)
BB_MIX$HANE_AGIRLIK=as.numeric(BB_MIX$HANE_AGIRLIK)
BB_MIX=as.data.frame(BB_MIX)
str(BB_MIX)
```

```
#Mixed matching ML#
> ML Method
> X.mtc <- c("CAR","COMP","DISH_W",“DIS_INC_CAT”)
> mix.ML <- mixed.mtc(data.rec=AA_MIX, data.don=BB_MIX,
            match.vars=X.mtc,y.rec="YINCOME", z.don="ZCONSUMPTION",
            method="ML",rho.yz=0, micro=TRUE, constr.alg="lpSolve")
> fill.ML <- create.fused(data.rec=AA_MIX,
            data.don=BB_MIX,mtc.ids=mix.ML$mtc.ids, z.vars="ZCONSUMPTION")

> cor(mix.ML$filled.rec)
```



```
                  CAR1     COMP1    DISH_W1   YINCOME ZCONSUMPTION
CAR1         1.0000000 0.2576429 0.2858454 0.2215519    0.2925380
COMP1        0.2576429 1.0000000 0.3522362 0.2804935    0.2778637
DISH_W1      0.2858454 0.3522362 1.0000000 0.2179516    0.2200326
YINCOME      0.2215519 0.2804935 0.2179516 1.0000000    0.2675645
ZCONSUMPTION 0.2925380 0.2778637 0.2200326 0.2675645    1.0000000
```

```
#Mixed matching MS#
> MS Method
> X.mtc <- c("CAR","COMP","DISH_W",“DIS_INC_CAT”)
> mix.MS <- mixed.mtc(data.rec=AA_MIX, data.don=BB_MIX,
             match.vars=X.mtc,y.rec="YINCOME", z.don="ZCONSUMPTION",
             method="MS",rho.yz=0, micro=TRUE, constr.alg="lpSolve")

> fill.MS <- create.fused(data.rec=AA_MIX,
             data.don=BB_MIX,mtc.ids=mix.MS$mtc.ids, z.vars="ZCONSUMPTION")
> cor(mix.MS$filled.rec)
```



```
                  CAR1     COMP1   DISH_W1    YINCOME ZCONSUMPTION
CAR1         1.0000000 0.2576429 0.2858454 0.22155191   0.27095272
COMP1        0.2576429 1.0000000 0.3522362 0.28049351   0.28844968
DISH_W1      0.2858454 0.3522362 1.0000000 0.21795156   0.23626948
YINCOME      0.2215519 0.2804935 0.2179516 1.00000000   0.04795845
ZCONSUMPTION 0.2709527 0.2884497 0.2362695 0.04795845   1.00000000
```

```
#Random Hot Deck / nonparametric micro / traditional unweighted with donor
classes#
> group_1 <- c("CAR")
> group_2 <- c("CAR", "COMP")
> group_3 <- c("CAR", "COMP", "DISH_W")
> group_4 <- c("CAR", "COMP", "DISH_W", "DIS_INC_CAT") # donation
             classes #

> out.rnd_1<- RANDwNND.hotdeck(data.rec = AA, data.don = BB,
             don.class= group_1)
>fA.rnd_1 <- create.fused(data.rec=AA,
             data.don=BB,mtc.ids=out.rnd_1$mtc.ids,z.vars="ZCONSUMPTION")
> out.rnd_2<- RANDwNND.hotdeck(data.rec = AA, data.don = BB,
             don.class = group_2)
>fA.rnd_2<- create.fused(data.rec=AA,
             data.don=BB,mtc.ids=out.rnd_2$mtc.ids,z.vars="ZCONSUMPTION")
> out.rnd_3<- RANDwNND.hotdeck(data.rec = AA, data.don = BB,
             don.class = group_3)
> fA.rnd_3<- create.fused(data.rec=AA,
             data.don=BB,mtc.ids=out.rnd_3$mtc.ids,z.vars="ZCONSUMPTION")
> out.rnd_4<- RANDwNND.hotdeck(data.rec = AA, data.don = BB,
             don.class = group_4)
> fA.rnd_4<- create.fused(data.rec=AA,
             data.don=BB,mtc.ids=out.rnd_4$mtc.ids,z.vars="ZCONSUMPTION")
#Random Hot Deck / nonparametric micro / weighted with donor classes#
> out.rnd_1w <-RANDwNND.hotdeck(data.rec=AA, data.don=BB,
             match.vars=NULL,don.class=group_1,weight.don="HANE_AGIRLIK")
> fA.rnd_1w<- create.fused(data.rec=AA, data.don=BB,
             mtc.ids= out.rnd_1w$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_2w<-RANDwNND.hotdeck(data.rec=AA, data.don=BB,
             match.vars=NULL,don.class=group_2,weight.don="HANE_AGIRLIK")
> fA.rnd_2w<- create.fused(data.rec=AA, data.don=BB,
             mtc.ids= out.rnd_2w$mtc.ids, z.vars="ZCONSUMPTION"
> out.rnd_3w<-RANDwNND.hotdeck(data.rec=AA, data.don=BB,
             match.vars=NULL,don.class=group_3,weight.don="HANE_AGIRLIK")
```

```
> fA.rnd_3w<- create.fused(data.rec=AA, data.don=BB,
             mtc.ids= out.rnd_3w$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_4w<-RANDwNND.hotdeck(data.rec=AA, data.don=BB,
             match.vars=NULL,don.class=group_4,weight.don="HANE_AGIRLIK")
> fA.rnd_4w<- create.fused(data.rec=AA, data.don=BB,
             mtc.ids= out.rnd_4w$mtc.ids, z.vars="ZCONSUMPTION")
#Random Hot Deck / nonparametric micro / “rot”, “min”, “exact”,
“constrained”, “unconstrained” options#
> group_5 <-c("DISH_W","DIS_INC_CAT")
> X.mtc <- c("CAR","COMP") # matching variables #
> “exact”

> rnd_opt_1 <- RANDwNND.hotdeck(data.rec=AA, data.don=BB,
           match.vars=X.mtc,don.class=group_5,dist.fun="gower",
           cut.don="exact", k=10)
> fA_opt_1 <- create.fused(data.rec=AA, data.don=BB,
             mtc.ids= rnd_opt_1$mtc.ids, z.vars="ZCONSUMPTION")
> “rot”
> rnd_opt_2 <- RANDwNND.hotdeck(data.rec=AA, data.don=BB,
           match.vars=X.mtc,don.class=group_5,dist.fun="gower",
           cut.don="rot")
> fA_opt_2 <- create.fused(data.rec=AA, data.don=BB,
             mtc.ids= rnd_opt_2$mtc.ids, z.vars="ZCONSUMPTION")
> “min”
> rnd_opt_3 <- RANDwNND.hotdeck(data.rec=AA, data.don=BB,
           match.vars=X.mtc,don.class=group_5,dist.fun="gower",
           cut.don="rot")
> fA_opt_3 <- create.fused(data.rec=AA, data.don=BB,
             mtc.ids= rnd_opt_3$mtc.ids, z.vars="ZCONSUMPTION")

> “constrained” “variables must be numeric for un/constrained”
AA$COMP=as.numeric(AA$COMP)
AA$DISH_W=as.numeric(AA$DISH_W)
AA$CAR=as.numeric(AA$CAR)
AA$DIS_INC_CAT=as.numeric(AA$DIS_INC_CAT)

BB$COMP=as.numeric(BB$COMP)
BB$DISH_W=as.numeric(BB$DISH_W)
BB$CAR=as.numeric(BB$CAR)
BB$DIS_INC_CAT=as.numeric(BB$DIS_INC_CAT)
> rnd_opt_4 <- RANDwNND.hotdeck(data.rec=AA, data.don=BB,
            match.vars=X.mtc,don.class=group_5, constrained=TRUE,
            constr.alg="Hungarian")
> fA_opt_4 <- create.fused(data.rec=AA, data.don=BB,
             mtc.ids= rnd_opt_4$mtc.ids, z.vars="ZCONSUMPTION")
> “unconstrained”
```

#Random Hot Deck “rot”, “min”, “exact”, “un/constrained” with 4 donor classes#

```
X.mtc_a <- c("CAR")
X.mtc_b <- c("CAR", "COMP")
X.mtc_c <- c("CAR", "COMP", "DISH_W")
X.mtc_d <- c("CAR", "COMP", "DISH_W", "DIS_INC_CAT")
group_6 <-c("CAR", "COMP", "DISH_W", "DIS_INC_CAT")
```

“exact”

```
> rnd_opt_a <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_a,don.class=group_6,dist.fun="exact
      matching")
> fA_opt_a <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_a$mtc.ids, z.vars="ZCONSUMPTION")


> rnd_opt_b <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_b,don.class=group_6,dist.fun="exact
      matching")
> fA_opt_b <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_b$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_c <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_c,don.class=group_6,dist.fun="exact
      matching")
> fA_opt_c <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_c$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_d <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_d,don.class=group_6,dist.fun="exact
      matching")
> fA_opt_d <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_d$mtc.ids, z.vars="ZCONSUMPTION")
```

“rot”

```
> rnd_opt_rot_a <- RANDwNND.hotdeck(data.rec=AA,
             data.don=BB,match.vars=X.mtc_a,don.class=group_6,dist.fun="gow
             er",cut.don="rot")
> fA_opt_rot_a <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_rot_a$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_rot_b <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_b,don.class=group_6,dist.fun="gower",cut
      .don="rot")
> fA_opt_rot_b <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_rot_b$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_rot_c <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_c,don.class=group_6,dist.fun="gower",cut
      .don="rot")
> fA_opt_rot_c <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_rot_c$mtc.ids, z.vars="ZCONSUMPTION")
```

```
> rnd_opt_rot_d <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_d,don.class=group_6,dist.fun="gower",cut
      .don="rot")
> fA_opt_rot_d <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_rot_d$mtc.ids, z.vars="ZCONSUMPTION")
```

"**min**"

```
> rnd_opt_min_a <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_a,don.class=group_6,dist.fun="gower",cut
      .don="min")
> fA_opt_min_a <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_rot_a$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_min_b <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_b,don.class=group_6,dist.fun="gower",cut
      .don="min")
> fA_opt_min_b <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_rot_b$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_min_c <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_c,don.class=group_6,dist.fun="gower",cut
      .don="min")
> fA_opt_min_c <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_rot_c$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_min_d <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_d,don.class=group_6,dist.fun="gower",cut
      .don="min")
> fA_opt_min_d <- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      rnd_opt_rot_d$mtc.ids, z.vars="ZCONSUMPTION")
```

“**constrained**”

```
> rnd_opt_c_a <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_a,don.class=group_6, constrained=TRUE,
      constr.alg="Hungarian")
> fA_opt_c_a <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_c_a$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_c_b <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_b,don.class=group_6, constrained=TRUE,
      constr.alg="Hungarian")
> fA_opt_c_b <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_c_b$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_c_c <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_c,don.class=group_6, constrained=TRUE,
      constr.alg="Hungarian")
> fA_opt_c_c <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_c_c$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_c_d <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_d,don.class=group_6, constrained=TRUE,
      constr.alg="Hungarian")
> fA_opt_c_d <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_c_d$mtc.ids, z.vars="ZCONSUMPTION"
“unconstrained”
> rnd_opt_Uc_a <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_a,don.class=group_6, constrained=FALSE,
      constr.alg="Hungarian")
> fA_opt_Uc_a <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_c_a$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_Uc_b <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_b,don.class=group_6, constrained=FALSE,
      constr.alg="Hungarian")
> fA_opt_Uc_b <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_c_b$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_Uc_c <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_c,don.class=group_6, constrained=FALSE,
      constr.alg="Hungarian")
> fA_opt_Uc_c <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_c_c$mtc.ids, z.vars="ZCONSUMPTION")

> rnd_opt_Uc_d <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_d,don.class=group_6, constrained=FALSE,
      constr.alg="Hungarian")
> fA_opt_Uc_d <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_c_d$mtc.ids, z.vars="ZCONSUMPTION"
"exact weighted rnd"
> out.rnd_w_a <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_a,don.class
      =group_6,weight.don="HANE_AGIRLIK",weight.rec="HANE_AGIRLIK",dist.fun
      ="exact matching")
> fA.rnd_w_a<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_a$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_w_b <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_b,
      don.class=group_6,weight.don="HANE_AGIRLIK",weight.rec="HANE_AGIRLIK"
      ,dist.fun="exact matching")
> fA.rnd_w_b<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_b$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_w_c <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_c,
      don.class=group_6,weight.don="HANE_AGIRLIK",weight.rec="HANE_AGIRLIK"
      ,dist.fun="exact matching")

> fA.rnd_w_c<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_c$mtc.ids, z.vars="ZCONSUMPTION")

> out.rnd_w_d <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_d,
      don.class=group_6,weight.don="HANE_AGIRLIK",weight.rec="HANE_AGIRLIK"
      ,dist.fun="exact matching")
```

```
> fA.rnd_w_d<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_d$mtc.ids, z.vars="ZCONSUMPTION")

"rot weighted rnd"
> out.rnd_w_ra <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_a,don.class
      =group_6,weight.don="HANE_AGIRLIK",dist.fun="gower",
      cut.don="rot")
> fA.rnd_w_ra<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_ra$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_w_rb <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_b,
      don.class=group_6,weight.don="HANE_AGIRLIK",dist.fun="gower",
      cut.don="rot")
> fA.rnd_w_rb<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_rb$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_w_rc <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_c,
      don.class=group_6,weight.don="HANE_AGIRLIK",dist.fun="gower",
      cut.don="rot")
> fA.rnd_w_rc<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_rc$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_w_rd <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_d,
      don.class=group_6,weight.don="HANE_AGIRLIK",dist.fun="gower",
      cut.don="rot")
> fA.rnd_w_rd<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_rd$mtc.ids, z.vars="ZCONSUMPTION")
"rot weighted rnd"
> out.rnd_w_ma <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_a,don.class
      =group_6,weight.don="HANE_AGIRLIK",dist.fun="gower",
      cut.don="min")
> fA.rnd_w_ma<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_ma$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_w_mb <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_b,don.class
      =group_6,weight.don="HANE_AGIRLIK",dist.fun="gower",
      cut.don="min")
> fA.rnd_w_mb<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_mb$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_w_mc <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_c,don.class
      =group_6,weight.don="HANE_AGIRLIK",dist.fun="gower",
      cut.don="min")
> fA.rnd_w_mc<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_mc$mtc.ids, z.vars="ZCONSUMPTION")
> out.rnd_w_md <-
      RANDwNND.hotdeck(data.rec=AA,data.don=BB,match.vars=X.mtc_d,don.class
      =group_6,weight.don="HANE_AGIRLIK",dist.fun="gower",
      cut.don="min")
```

```
> fA.rnd_w_md<- create.fused(data.rec=AA, data.don=BB,mtc.ids=
      out.rnd_w_md$mtc.ids, z.vars="ZCONSUMPTION")
#constrained weighted rnd#
> rnd_opt_cw_a <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_a,don.class=group_6,
      weight.don="HANE_AGIRLIK", constrained=TRUE, constr.alg="Hungarian")
> fA_opt_cw_a <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_cw_a$mtc.ids, z.vars="ZCONSUMPTION")
> rnd_opt_cw_b <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_b,don.class=group_6,
      weight.don="HANE_AGIRLIK", constrained=TRUE, constr.alg="Hungarian")
> fA_opt_cw_b <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_cw_b$mtc.ids, z.vars="ZCONSUMPTION")
> rnd_opt_cw_c <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_c,don.class=group_6,
      weight.don="HANE_AGIRLIK", constrained=TRUE, constr.alg="Hungarian")
> fA_opt_cw_c <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_cw_c$mtc.ids, z.vars="ZCONSUMPTION")
> rnd_opt_cw_d <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_d,don.class=group_6,
      weight.don="HANE_AGIRLIK", constrained=TRUE, constr.alg="Hungarian")
> fA_opt_cw_d <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_cw_d$mtc.ids, z.vars="ZCONSUMPTION")
#unconstrained weighted rnd#
> rnd_opt_Ucw_a <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_a,don.class=group_6,
      weight.don="HANE_AGIRLIK", constrained=FALSE, constr.alg="Hungarian")
> fA_opt_Ucw_a <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_Ucw_a$mtc.ids, z.vars="ZCONSUMPTION")
> rnd_opt_Ucw_b <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_b,don.class=group_6,
      weight.don="HANE_AGIRLIK", constrained=FALSE, constr.alg="Hungarian")
> fA_opt_Ucw_b <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_Ucw_b$mtc.ids, z.vars="ZCONSUMPTION")
> rnd_opt_Ucw_c <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_c,don.class=group_6,weight.
      don="HANE_AGIRLIK", constrained=FALSE, constr.alg="Hungarian")
> fA_opt_Ucw_c <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_Ucw_c$mtc.ids, z.vars="ZCONSUMPTION")
> rnd_opt_Ucw_d <- RANDwNND.hotdeck(data.rec=AA,
      data.don=BB,match.vars=X.mtc_d,don.class=group_6,weight.
      don="HANE_AGIRLIK",constrained=FALSE, constr.alg="Hungarian")
> fA_opt_Ucw_d <- create.fused(data.rec=AA, data.don=BB, mtc.ids=
      rnd_opt_Ucw_d$mtc.ids, z.vars="ZCONSUMPTION")
#rank data prep.#
X.mtc_a <- c("CAR")
X.mtc_b <- c("CAR", "COMP")
X.mtc_c <- c("CAR", "COMP", "DISH_W")
X.mtc_d <- c("CAR", "COMP", "DISH_W", "DIS_INC_CAT")
group.rnk <- c("CAR","COMP","DISH_W", "DIS_INC_CAT")
AA$CAR_N<-as.numeric(AA$CAR)
```

```
AA$COMP_N<-as.numeric(AA$COMP)
AA$DISH_W_N<-as.numeric(AA$DISH_W)
AA$DIS_INC_CAT_N<-as.numeric(AA$DIS_INC_CAT)
BB$CAR_N<-as.numeric(BB$CAR)
BB$COMP_N<-as.numeric(BB$COMP)
BB$DISH_W_N<-as.numeric(BB$DISH_W)
BB$DIS_INC_CAT_N<-as.numeric(BB$DIS_INC_CAT)
X.mtc.a_n=c("CAR_N")
X.mtc.b_n=c("CAR_N","COMP_N")
X.mtc.c_n=c("CAR_N","COMP_N","DISH_W_N")
X.mtc.d_n=c("CAR_N","COMP_N","DISH_W_N", "DIS_INC_CAT_N")


AA$X.mtc.a_n=paste0(AA$CAR_N)
AA$X.mtc.b_n=paste0(AA$CAR_N,AA$COMP_N)
AA$X.mtc.c_n=paste0(AA$CAR_N,AA$COMP_N,AA$DISH_W_N)
AA$X.mtc.d_n=paste0(AA$CAR_N,AA$COMP_N,AA$DISH_W_N,AA$DIS_INC_CAT_N)
BB$X.mtc.a_n=paste0(BB$CAR_N)
BB$X.mtc.b_n=paste0(BB$CAR_N,BB$COMP_N)
BB$X.mtc.c_n=paste0(BB$CAR_N,BB$COMP_N,BB$DISH_W_N)
BB$X.mtc.d_n=paste0(BB$CAR_N,BB$COMP_N,BB$DISH_W_N,BB$DIS_INC_CAT_N)
#rank unweighted exact#
> rnk.a <- rankNND.hotdeck(data.rec=AA, data.don=BB,var.rec="X.mtc.a_n",
      var.don="X.mtc.a_n",don.class = group.rnk,dist.fun="exact matching")
> fA.rnk.a <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.a$mtc.ids,z.vars="ZCONSUMPTION", dup.x=TRUE,
      match.vars=X.mtc_a)
> rnk.b <- rankNND.hotdeck(data.rec=AA, data.don=BB,var.rec="X.mtc.b_n",
      var.don="X.mtc.b_n",don.class = group.rnk,dist.fun="exact matching")
> fA.rnk.b <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.b$mtc.ids,z.vars="ZCONSUMPTION", dup.x=TRUE,
      match.vars=X.mtc_b)
> rnk.c <- rankNND.hotdeck(data.rec=AA, data.don=BB,var.rec="X.mtc.c_n",
      var.don="X.mtc.c_n",don.class = group.rnk,dist.fun="exact matching")
> fA.rnk.c <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.c$mtc.ids,z.vars="ZCONSUMPTION", dup.x=TRUE,
      match.vars=X.mtc_c)
> rnk.d <- rankNND.hotdeck(data.rec=AA, data.don=BB,var.rec="X.mtc.d_n",
      var.don="X.mtc.d_n",don.class = group.rnk,dist.fun="exact matching")
> fA.rnk.d <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.d$mtc.ids,z.vars="ZCONSUMPTION", dup.x=TRUE,
      match.vars=X.mtc_d)
#rank unweighted unconstrained#
> rnk.ra <- rankNND.hotdeck(data.rec=AA, data.don=BB,var.rec="X.mtc.a_n",
      var.don="X.mtc.a_n",don.class =
      group.rnk,constrained=TRUE,constr.alg="Hungarian")
> fA.rnk.ra <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.ra$mtc.ids,z.vars="ZCONSUMPTION", dup.x=TRUE,
      match.vars=X.mtc_a)
```

```
 rnk.rb <- rankNND.hotdeck(data.rec=AA, data.don=BB,var.rec="X.mtc.b_n",
      var.don="X.mtc.b_n",don.class =
      group.rnk,constrained=FALSE,constr.alg="Hungarian")
> fA.rnk.rb <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.rb$mtc.ids,z.vars="ZCONSUMPTION", dup.x=TRUE,
      match.vars=X.mtc_b)
> rnk.rc <- rankNND.hotdeck(data.rec=AA, data.don=BB,var.rec="X.mtc.c_n",
      var.don="X.mtc.c_n",don.class =
      group.rnk,constrained=FALSE,constr.alg="Hungarian")
> fA.rnk.rc <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.rc$mtc.ids,z.vars="ZCONSUMPTION", dup.x=TRUE,
      match.vars=X.mtc_c)
> rnk.rd <- rankNND.hotdeck(data.rec=AA, data.don=BB,var.rec="X.mtc.d_n",
      var.don="X.mtc.d_n",don.class =
      group.rnk,constrained=FALSE,constr.alg="Hungarian")

> fA.rnk.rd <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=rnk.rd$mtc.ids,z.vars="ZCONSUMPTION", dup.x=TRUE,
      match.vars=X.mtc_d)
#nnd unweighted#
X.mtc_a <- c("CAR")
X.mtc_b <- c("CAR", "COMP")
X.mtc_c <- c("CAR", "COMP", "DISH_W")
X.mtc_d <- c("CAR", "COMP", "DISH_W", "DIS_INC_CAT")
group.nnd <- c("CAR","COMP","DISH_W", "DIS_INC_CAT")
> out.nnd_a <- NND.hotdeck(data.rec=AA, data.don=BB,match.vars=X.mtc_a,
      don.class=group.nnd, dist.fun="Gower")
> out.nnd_b <- NND.hotdeck(data.rec=AA, data.don=BB,match.vars=X.mtc_b,
      don.class=group.nnd, dist.fun="Gower")
> out.nnd_c <- NND.hotdeck(data.rec=AA, data.don=BB,match.vars=X.mtc_c,
      don.class=group.nnd, dist.fun="Gower")
> out.nnd_d <- NND.hotdeck(data.rec=AA, data.don=BB,match.vars=X.mtc_d,
      don.class=group.nnd, dist.fun="Gower")
> fA.nnd_a <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=out.nnd_a$mtc.ids,match.vars = group.nnd,
      z.vars="ZCONSUMPTION")
> fA.nnd_b <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=out.nnd_b$mtc.ids,match.vars = group.nnd,
      z.vars="ZCONSUMPTION")
> fA.nnd_c <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=out.nnd_c$mtc.ids,match.vars = group.nnd,
      z.vars="ZCONSUMPTION")
> fA.nnd_d <- create.fused(data.rec=AA,
      data.don=BB,mtc.ids=out.nnd_d$mtc.ids,match.vars = group.nnd,
      z.vars="ZCONSUMPTION")
write_xlsx(fA.nnd_a,"NND_UNWEIGHTED.X.mtc_a.xlsx")
write_xlsx(fA.nnd_b,"NND_UNWEIGHTED.X.mtc_b.xlsx")
write_xlsx(fA.nnd_c,"NND_UNWEIGHTED.X.mtc_c.xlsx")
write_xlsx(fA.nnd_d,"NND_UNWEIGHTED.X.mtc_d.xlsx")
```

```
#weighted NND hot deck#
> out.nnd.wa <- NND.hotdeck(data.rec=AA, data.don=BB,match.vars=X.mtc_a,
       don.class=group.nnd,
       dist.fun="Gower",weight.rec="HANE_AGIRLIK",weight.don="HANE_AGIRLIK")
> out.nnd.wb <- NND.hotdeck(data.rec=AA, data.don=BB,match.vars=X.mtc_b,
       don.class=group.nnd,
       dist.fun="Gower",weight.rec="HANE_AGIRLIK",weight.don="HANE_AGIRLIK")
> out.nnd.wc <- NND.hotdeck(data.rec=AA, data.don=BB,match.vars=X.mtc_c,
       don.class=group.nnd,
       dist.fun="Gower",weight.rec="HANE_AGIRLIK",weight.don="HANE_AGIRLIK")
> out.nnd.wd <- NND.hotdeck(data.rec=AA, data.don=BB,match.vars=X.mtc_d,
       don.class=group.nnd,
       dist.fun="Gower",weight.rec="HANE_AGIRLIK",weight.don="HANE_AGIRLIK")
> fA.nnd.wa <- create.fused(data.rec=AA,
       data.don=BB,mtc.ids=out.nnd.wa$mtc.ids,match.vars = group.nnd,
       z.vars="ZCONSUMPTION")
> fA.nnd.wb <- create.fused(data.rec=AA,
       data.don=BB,mtc.ids=out.nnd.wb$mtc.ids,match.vars = group.nnd,
       z.vars="ZCONSUMPTION")
> fA.nnd.wc <- create.fused(data.rec=AA,
       data.don=BB,mtc.ids=out.nnd.wc$mtc.ids,match.vars = group.nnd,
       z.vars="ZCONSUMPTION")
> fA.nnd.wd <- create.fused(data.rec=AA,
       data.don=BB,mtc.ids=out.nnd.wd$mtc.ids,match.vars = group.nnd,
       z.vars="ZCONSUMPTION")
#Renssen calibration#
> tt.AA <- xtabs(HANE_AGIRLIK~CAR+COMP+DISH_W+DIS_INC_CAT, data=AA)
> tt.BB <- xtabs(HANE_AGIRLIK~CAR+COMP+DISH_W+DIS_INC_CAT, data=BB)
> (prop.table(tt.AA)-prop.table(tt.BB))*100
> comp.prop(p1=tt.AA, p2=tt.BB, n1=nrow(AA),n2=nrow(BB), ref=FALSE)
> $meas
       tvd    overlap      Bhatt       Hell
0.05860870 0.94139130 0.99674776 0.05702843
> $chi.sq
   Pearson         df      q0.05   delta.h0
204.485089  47.000000  64.001112   3.195024
#Creating and attaching survey design objects#
> svy.A <- svydesign(~1, weights=~HANE_AGIRLIK, data=AA)
> svy.B <- svydesign(~1, weights=~HANE_AGIRLIK, data=BB)
$p.exp
, , DISH_W = 1, DIS_INC_CAT = 1
   COMP
CAR            1            2
  1 0.0003979189 0.0005312219
  2 0.0007919607 0.0031372133
, , DISH_W = 2, DIS_INC_CAT = 1
   COMP
CAR            1            2
  1 0.0000931184 0.0008285302
  2 0.0009921041 0.0198193299
, , DISH_W = 1, DIS_INC_CAT = 2
```

```
   COMP
CAR            1            2
  1 0.0021898416 0.0061537724
  2 0.0067532972 0.0300455062
, , DISH_W = 2, DIS_INC_CAT = 2
   COMP
CAR            1            2
  1 0.0011493955 0.0061164584
  2 0.0052249433 0.0615110452
, , DISH_W = 1, DIS_INC_CAT = 3
   COMP
CAR            1            2
  1 0.0144559212 0.0237475472
  2 0.0228636517 0.0627742931
, , DISH_W = 2, DIS_INC_CAT = 3
   COMP
CAR            1            2
  1 0.0036929340 0.0132418442
  2 0.0076011858 0.0512493560
, , DISH_W = 1, DIS_INC_CAT = 4
   COMP
CAR            1            2
  1 0.0262953447 0.0269645204
  2 0.0284672745 0.0433635364
, , DISH_W = 2, DIS_INC_CAT = 4
   COMP
CAR            1            2
  1 0.0038008531 0.0090840738
  2 0.0084459845 0.0226842650
, , DISH_W = 1, DIS_INC_CAT = 5
   COMP
CAR            1            2
  1 0.0335772605 0.0229215867
  2 0.0255050832 0.0263902921
, , DISH_W = 2, DIS_INC_CAT = 5
   COMP
CAR            1            2
  1 0.0040749626 0.0060404218
  2 0.0061771858 0.0110567458
, , DISH_W = 1, DIS_INC_CAT = 6
   COMP
CAR            1            2
  1 0.1611964067 0.0472796801
  2 0.0701151603 0.0338568807
, , DISH_W = 2, DIS_INC_CAT = 6
   COMP
CAR            1            2
  1 0.0069660140 0.0104460599
  2 0.0082104147 0.0117176020
#harmonizing#
```

```
hz.org <- harmonize.x (svy.A=svy.A, svy.B=svy.B,
form.x=~CAR:COMP:DISH_W:DIS_INC_CAT-1 ,cal.method="linear")
> options "linear", "raking", "poststratification"
> new calibrated weights for A and B
summary(hz.org$weights.A)
summary(hz.org$weights.B)
> tt.1 <- xtabs(hz.org$weights.A~CAR+COMP+DISH_W+DIS_INC_CAT, data=AA)
> tt.2 <- xtabs(hz.org $weights.B~CAR+COMP+DISH_W+DIS_INC_CAT, data=BB)
> c1 <- comp.prop(p1=tt.1, p2=tt.2, n1=nrow(AA),n2=nrow(BB), ref=FALSE)
> c1$meas
       tvd    overlap      Bhatt       Hell
0.05365261 0.94634739 0.99708869 0.05395653
> comb.samples(svy.A=out.hz$cal.A,svy.B=out.hz$cal.B,svy.C=NULL,
y.lab="YINCOME",z.lab="ZCONSUMPTION",form.x=~CAR:COMP:DISH_W:
DIS_INC_CAT-1,estimation="STWS",micro="TRUE")”
```

Appendix E. Original Article

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***, s**ince 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 $n\mathrm{R}i$), 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)

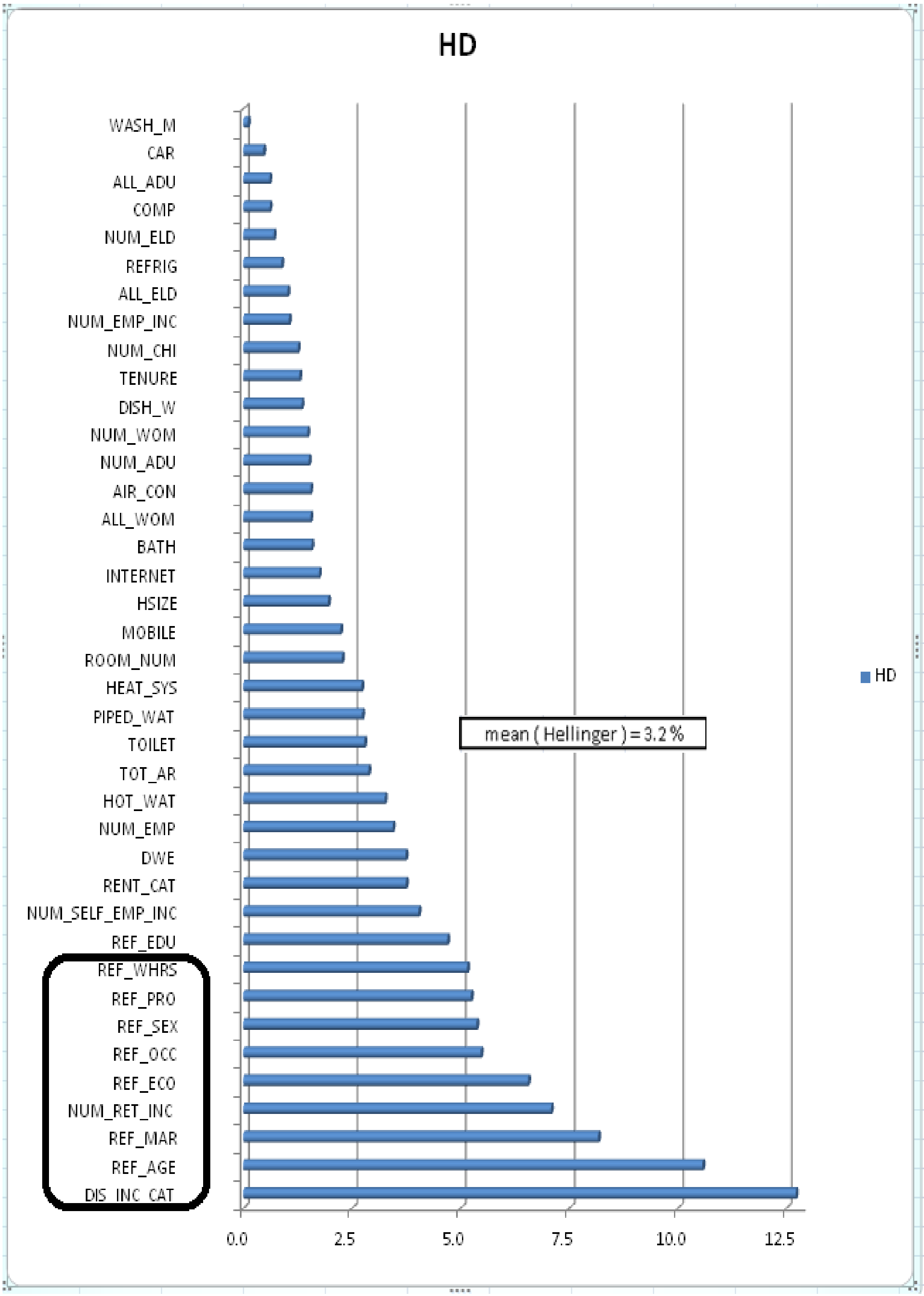
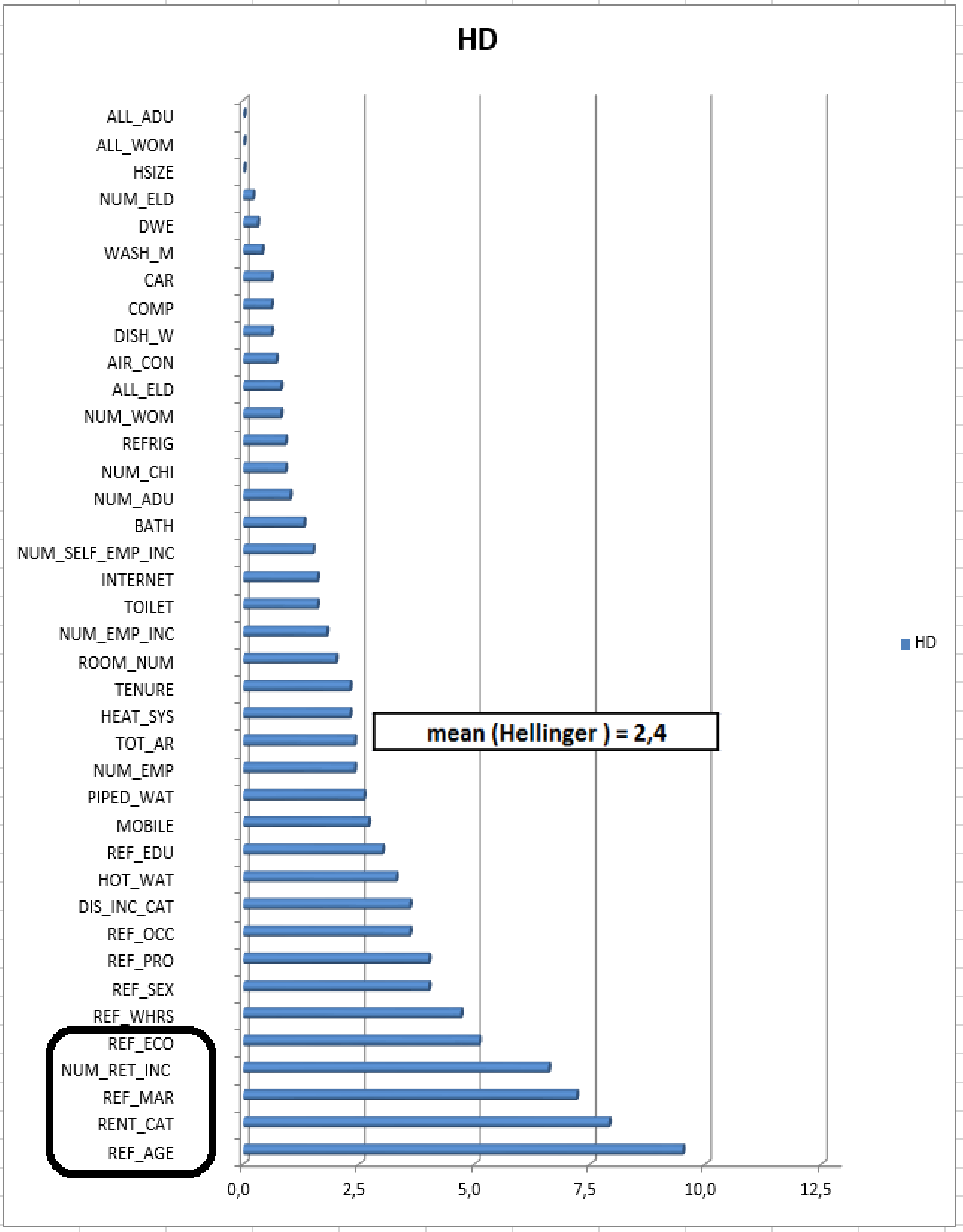


Figure 2. Hellinger distance results of the common variables (weighted)



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.

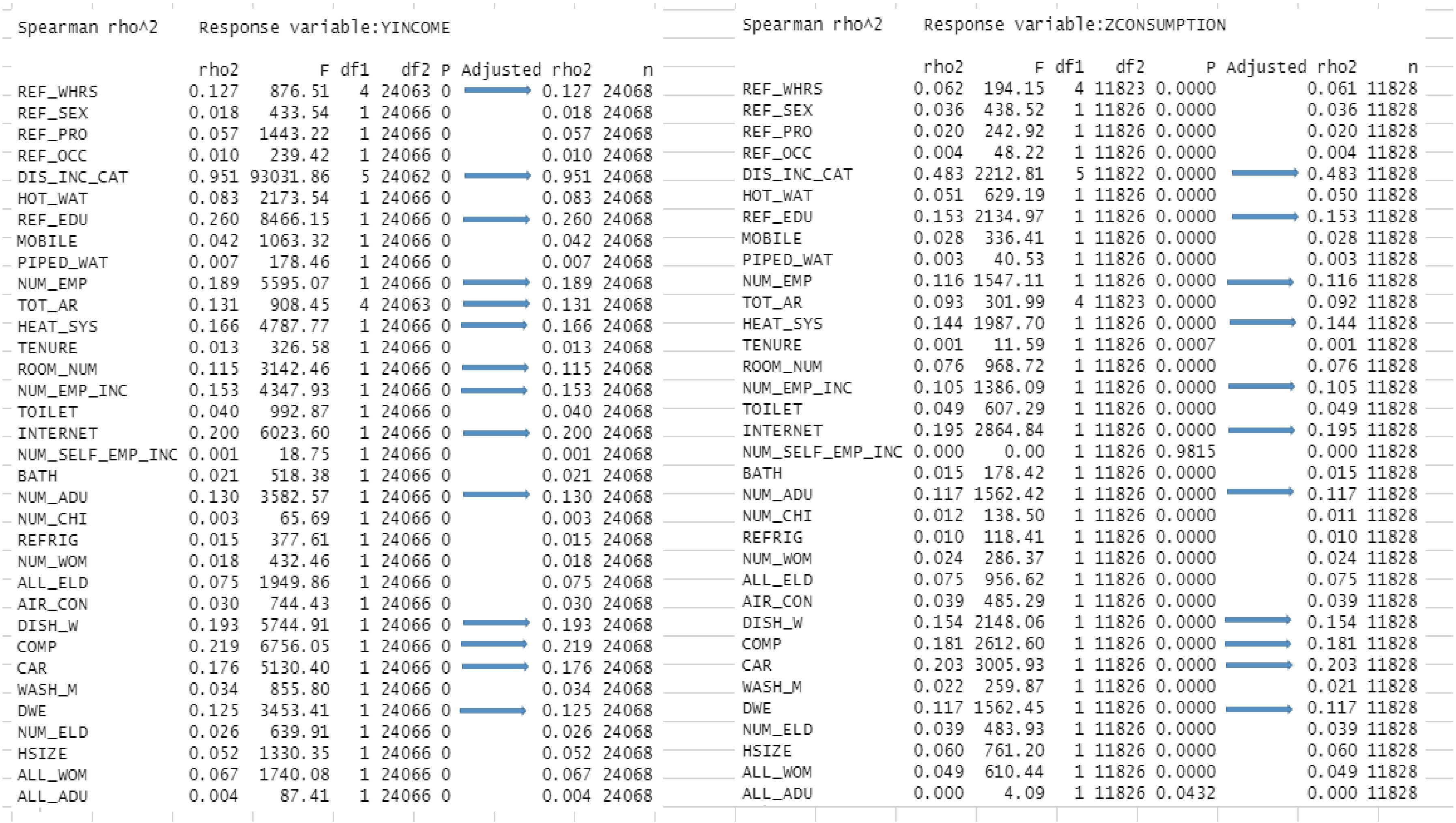
Table 3. Weighted and unweighted scores of the 4 variables

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