

HACETTEPE UNIVERSITY
INSTITUTE OF POPULATION STUDIES

**NONPARAMETRIC STATISTICAL MATCHING
METHODS: AN APPLICATION ON HOUSEHOLD
SURVEYS IN TURKEY**

Cansu ÖZTÜRK

Department of Social Research Methodology
Master's Thesis

Ankara
November 2019

HACETTEPE UNIVERSITY
INSTITUTE OF POPULATION STUDIES

**NONPARAMETRIC STATISTICAL MATCHING
METHODS: AN APPLICATION ON HOUSEHOLD
SURVEYS IN TURKEY**

Cansu ÖZTÜRK

Supervisor

Assist. Prof. Dr. Tuğba ADALI

Department of Social Research Methodology

Master's Thesis

Ankara

November 2019

NONPARAMETRIC STATISTICAL MATCHING METHODS: AN APPLICATION ON HOUSEHOLD SURVEYS IN TURKEY

by Cansu Öztürk

Submission date: 15-Nov-2019 10:36AM (UTC+0300)

Submission ID: 1169075879

File name: CHING_METHODS_AN_APPLICATION_ON_HOUSEHOLD_SURVEYS_IN_TURKEY.docx (100.02K)

Word count: 14030

Character count: 78262

NONPARAMETRIC STATISTICAL MATCHING METHODS: AN APPLICATION ON HOUSEHOLD SURVEYS IN TURKEY

ORIGINALITY REPORT

7%	5%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ec.europa.eu Internet Source	2%
2	cran.um.ac.ir Internet Source	1%
3	www.istat.it Internet Source	<1%
4	Nathaniel Schenker. "Combining information from multiple surveys to enhance estimation of measures of health", Statistics in Medicine, 04/15/2007 Publication	<1%
5	indiastat.com Internet Source	<1%
6	krishikosh.egranth.ac.in Internet Source	<1%
7	Submitted to Harrisburg University of Science and Technology Student Paper	<1%

ACKNOWLEDGEMENTS

I am thankful to my supervisor, Assist. Prof. Dr. Tuğba ADALI, for the advice and support she has provided me with throughout the preparation of this thesis.

I would like to thank Prof. Dr. Yaprak Arzu ÖZDEMİR and Prof. Dr. A. Sinan TÜRKYILMAZ for their valuable and constructive critics at the thesis defense.

I thank my mother, Sultan ÖZTÜRK, my father, Saadettin ÖZTÜRK and my brother, Can ÖZTÜRK, for their daily moral support.

ABSTRACT

The use of administrative registers and sample surveys together in estimation process has a significant effect on increasing the accuracy of statistical information and reducing the response burden, time, cost and labour. However, this requires integration of data sources.

Record linkage and statistical matching are two techniques improved for data integration. Record linkage is a method used when there is a perfect agreement between indicators. If there is no such variable in the data set but there are some common variables and samples of the surveys refer to same target population then statistical matching is used. Common variables between surveys are used as matching variables and fused data sets are obtained using different matching approaches such as parametric, non-parametric and mixed.

Aim of this dissertation was to apply different non-parametric statistical matching methods on household based sample surveys and compare their results regarding similarity of variables' distributions. 2014-2015 Time Use Survey of Turkey and 2014 Life Satisfaction Survey of Turkey were used in the implementation. Three non-parametric hot deck methods named as nearest neighbour distance, random and rank hot deck were used in the matching with their constrained and unconstrained versions. For each method, fused data sets were obtained using four different combinations of matching variables and four target variables. Comparisons were made both between methods, and within methods; through changes in applications for each specific method.

It was found that results vary according to both matching variables and target variables. Moreover, contrary to expectations, constrained hot deck did not provide better result.

Key words: Statistical matching, non-parametric hot deck, constrained and unconstrained hot deck

ÖZET

İdari kayıtların ve örneklem arařtırmalarının tahmin sürecinde birlikte kullanılması, istatistiki bilgilerin doęruluęunu arttırmakla beraber cevaplayıcı yükünü, zamanı, maliyeti ve işgücünü azaltmada önemli bir etkiye sahiptir. Ancak, bu yöntem farklı veri kaynaklarının birleřtirilmesini gerektirmektedir.

Kayıt baęlama ve istatistiksel eřleřtirme, veri birleřtirme için geliřtirilmiř iki tekniktir. Kayıt baęlama, veride birebir eřleřtirme saęlayan eřsiz bir deęiřken olduęu zaman kullanılan yöntemdir. Veri setinde böyle bir deęiřken yoksa fakat bazı ortak deęiřkenler varsa ve arařtırmaların hedef kitleleri aynı ise, istatistiksel eřleřtirme yöntemi kullanılmaktadır. Arařtırmalar arasındaki ortak deęiřkenler eřleřtirme deęiřkenleri olarak kullanılmakta ve birleřik veri setleri parametrik, parametrik olmayan ve karma gibi farklı eřleřtirme yaklařımları kullanılarak elde edilmektedir.

Tezde, hanehalkı bazlı örneklem arařtırmalarında farklı parametrik olmayan istatistiksel eřleřtirme yöntemlerinin kullanılması ve deęiřkenlerin daęılımlarının benzerlięi kullanılarak sonuçların karřılařtırılması amaçlanmıřtır. Uygulamada, 2014-2015 Türkiye Zaman Kullanımı Arařtırması ve 2014 Türkiye Yařam Memnuniyeti Arařtırması kullanılmıřtır. En yakın komřu, rastgele ve sıralıhot deck yöntemleri olmak üzere üç tane parametrik olmayan istatistiksel eřleřtirme yöntemi, kısıtlıve kısıtlı olmayan seçenekleri ile kullanılmıřtır. Kullanılan tüm yöntemlerde, birleřtirilmiř veri setleri eřleřtirme deęiřkenlerin 4 farklı kombinasyonu ve dört hedef deęiřken kullanılarak elde edilmiřtir. Her yöntem için farklı uygulamalar yapıldıęından hem yöntemlerin kendi içinde, hem de yöntemler arasında karřılařtırmalar yapılmıřtır.

Sonuçların hem eřleřtirme deęiřkenlerine hem de hedef deęiřkenlere göre deęiřtięi tespit edilmiřtir. Ayrıca, beklentilerin aksine, kısıtlı seçeneęi kullanılarak gerçeęleştirilen eřleřtiriminin daha iyi bir sonuç vermedięi gözlemlenmiřtir.

Anahtar Kelimeler: İstatistiksel eřleřtirme, parametric olmayan hot deck, kısıtlı ve kısıtlı olmayan hot deck

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	i
ABSTRACT	ii
ÖZET	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	vi
LIST OF FIGURES	vii
ABBREVIATIONS	viii
CHAPTER 1. INTRODUCTION	1
CHAPTER 2. LITERATURE REVIEW	5
CHAPTER 3. METHOD AND DATA SOURCES	12
3.1. Methodology of the Statistical Matching	12
3.1.1. Statistical Matching Approaches under the CIA	15
3.1.1.1. <i>Parametric Approach</i>	16
3.1.1.2. <i>Nonparametric Approach</i>	16
3.1.1.3. <i>Mixed Approach</i>	17
3.2. Statistical Matching for Complex Sample Surveys	18
3.2.1. Statistical Matching Approaches for Complex Sample Surveys	20
3.3. Preparation Steps Before Applying Statistical Matching	21
3.3.1. Reconciliation of the Data Sources and Harmonization of Variables	21
3.3.2. Choosing the Matching Variables	22
3.4. DATA SOURCES	24
3.4.1. Life Satisfaction Survey	24
3.4.1.1. <i>Survey Objectives</i>	24
3.4.1.2. <i>History of the Life Satisfaction Survey</i>	24
3.4.1.3. <i>Questionnaire Design and Variables of Turkey Life Satisfaction Survey 2014</i>	25
3.4.1.4. <i>Survey Design and Weighting Procedure</i>	28
3.4.1.5. <i>Data Collection</i>	30
3.4.2. Time Use Survey	31
3.4.2.1. <i>Survey Objectives</i>	31
3.4.2.2. <i>History of the Survey Time Use Survey</i>	31

3.4.2.3. <i>Questionnaire Design and Variables Turkey</i>	
<i>Time Use Survey, 2014-2015</i>	32
3.4.2.4. <i>Survey Design and Weighting Procedure</i>	37
3.4.2.5. <i>Data Collection</i>	38
3.4.3. Harmonization of Common Variables in LSS and TUS	39
CHAPTER 4. RESULTS	41
4.1. CHOOSING THE MATCHING VARIABLES	41
4.2. NON-PARAMETRIC MICRO STATISTICAL MATCHING	45
4.2.1. Nearest Neighbour Distance Hot Deck	46
4.2.2. Random Hot Deck	48
4.2.3. Rank Hot Deck	51
CHAPTER 5. DISCUSSION	53
CHAPTER 6. CONCLUSION	56
REFERENCES	57
APPENDIX	60
Appendix A. NUTS Code List	60
Appendix B. ISCED-97 Code List	63
Appendix C. ICSE-93 Code List	64
Appendix D. HETUS Activity Coding List	65
Appendix E. NACE Rev. 2 Code List	69
Appendix F. ISCO-08 Code List	70
Appendix G. Harmonized Common Variables	71
Appendix H. Logistic Regression Results of Cinema	72
Appendix I. Logistic Regression Results of Theatre	73
Appendix J. Logistic Regression Results of Television	74
Appendix K. Logistic Regression Results of Social Media	75
Appendix L. Logistic Regression Results of Daily Tours or Nature Walks	76
Appendix M. Logistic Regression Results of Solving a Puzzle	77
Appendix N. Logistic Regression Results of Going to Cafe or Bar	78
Appendix O. Logistic Regression Results of Going to Mall	79

LIST OF TABLES

Table 4.1.1. Hellinger Distances of Common Variables	41
Table 4.1.2. Significant Variables for Target Variables in TUS	43
Table 4.1.3. Ordinal Logistic Regression Results on Happiness	44
Table 4.2.1.1. Nearest Neighbour Distance Hot Deck Results	46
Table 4.2.1.2. Nearest Neighbour Distance Constrained Hot Deck Results	47
Table 4.2.2.1. Results of Random Hot Deck Using "rot"	49
Table 4.2.2.2. Results of Random Hot Deck Using "min"	50
Table 4.2.3.1. Rank Hot Deck Results	51
Table 4.2.3.2. Rank Constrained Hot Deck Results	52

LIST OF FIGURES

Figure 3.1.1. Structure of overall sample under first way integration	13
Figure 3.1.2. Structure of data sources under second way integration	13
Figure 3.2.1. Structure of the data sources in SM for complex surveys	19

ABBREVIATIONS

CIA	Conditional Independence Assumption
Eurostat	European Statistical Office
HD	Hellinger Distance
HETUS-ACL	Harmonised European Time Use Surveys - Activity Coding List
ICSE-93	International Classification of Status in Employment
ISCED-97	International Standard Classification of Education
ISCO-08	International Standard Classification of Occupations
i.i.d.	Independently and identically distributed
LSS	Life Satisfaction Survey
NACE Rev.2	Statistical Classification of Economic Activities in European Community
NND	Nearest Neighbour Distance
NUTS	Nomenclature of Territorial Units for Statistics
SM	Statistical Matching
TurkStat	Turkish Statistical Institute
TUS	Time Use Survey

CHAPTER 1. INTRODUCTION

Surveys have been conducted for a long time to obtain information on issues, which are unknown or wondered. For the purpose of estimating population attributes and characteristics, sample based surveys are used since such surveys have a big role on enriching data for statistical purposes. There are many sample surveys conducted by different institutes, offices or organizations on various issues in Turkey. According to their scopes and aims, some of these surveys are applied to households and some to enterprises. *Innovation, Life Satisfaction and Household Labour Force Surveys* conducted by Turkish Statistical Institute, *Demographic and Health Survey* conducted by Hacettepe University Institute of Population Studies, *Family Structure Survey* conducted by Ministry of Family, Labour and Social Services can be given as examples of these surveys. However, conducting a survey is not an easy process. When household surveys are considered, there are about 82 million people and 23 million household addresses in Turkey, from which collecting data require very high budget, time and labour.

In addition to sample surveys, administrative sources might also be used to obtain information. There are various administrative registers collected by Turkish Government Offices and Ministries. Nevertheless, most of them are not collected for statistical purposes. Therefore, researchers come across many problems when they try to use registers for their studies. Definition, duplication and coverage problems can be given as an example to these difficulties. On the other hand, even if registers are collected taking care of the statistical purposes, sometimes it cannot be used due to completeness and up-to-date problems. Thus, using administrative data sources instead of sample surveys are not useful most of the time. Beside that, some types of data or information such as individual's emotions or thoughts cannot be obtained from administrative sources, therefore; surveys still need to be conducted. However, it is not always possible to conduct a new sample surveys due to the cost and labour considerations.

For many years, researchers have been working on data integration methods to benefit from existing data sources. These methods improve the complementary use of existing data sources (Leulescu and Agafitei, 2013). In this way, the quality and efficiency increases while cost decreases. These are the potential advantages of integrating existing data sources.

There are two techniques improved for integrating data. One of them is “record linkage” and the other one is “statistical matching”. Record linkage is a method used when there is a perfect agreement between indicators. Personal Identification Number is most known and useful one of these indicators since it is a unique for each person. However; if there is no such variable in the data set but there are some common variables and samples of the surveys refer to same target population then statistical matching is used.

In recent years, statistical matching has used to obtain joint information based on different surveys by European Union Statistical Office (Eurostat) especially on issues related with economic indicators. However, this subject is not a widely studied one in Turkey, especially on variables related with social life. This thesis will focus on statistical matching of household surveys that covers social topics. Generally, in studies on statistical matching, results of different approaches namely, non-parametric, parametric and mixed are compared. Usually, one non-parametric approach is compared against the parametric approach, yet the choice of non-parametric method in these studies seems somehow arbitrary. Nevertheless, various non-parametric approaches can be used in statistical matching and there is almost no study discussing the results of different non-parametric statistical matching methods. For this reason, the main objective of this study will be to compare non-parametric statistical matching methods in terms of their validity, in other words; preserving the marginal distribution of the variables by using two household surveys conducted in Turkey. Evaluating important issues that should be considered while combining different data sets and presenting statistical matching techniques are the additional objectives of this study.

As mentioned before, researchers have focused on data integration methods to increase quality of the survey results and decrease the disadvantages of conducting a survey. Therefore, it should be applied as much as possible on various issues. Different non-parametric statistical matching methods will be applied and their results will be compared in this study in order to investigate whether the statistical matching can be used on social surveys. Results of this study might be beneficial for the researchers or statisticians while determining the statistical matching method to be used in their study. Moreover, it might give new ideas on what kind of surveys or variables can be used for data integration.

There are 6 chapters in this dissertation and these are “Introduction”, “Literature Review”, “Method and Data Sources”, “Results”, “Discussion” and “Conclusion”, respectively.

Introduction chapter represents data integration concept briefly with reasons and potential advantages. Objectives and the importance of the study are also explained in this chapter. The history of statistical matching method and its evaluation are covered in the Literature Review chapter. Moreover, various studies conducted by using statistical matching method and their results are given in this chapter.

In Chapter 3, Method part reviews the statistical matching method. Important details of matching procedure such as comparison of the data sets, harmonization of the variables, how to choose matching variables and matching methods are presented in this chapter separately. Life Satisfaction and Time Use Surveys, data sources used in the study, are covered in Data Sources part. First, their background information are given briefly. Then, their history in Turkey and detailed information on micro data sets are explained.

In Results chapter, results of the statistical matching procedure according to used nonparametric approaches are presented. After that, results of different statistical matching methods are compared and discussed in Discussion chapter.

Conclusion chapter concludes the result of the discussion as a final chapter and presents some advices to researchers for further studies.

CHAPTER 2. LITERATURE REVIEW

In this chapter, brief information about history of statistical matching (SM) and some studies conducted on statistical matching in various countries are explained.

Due to difficulties of conducting a new survey, statistical offices and researchers have tried to develop methodology on integrating different sources to benefit more from existing data sets. Various methods have been considered to combine different surveys. Kish (1999) conceptualized the issue under several headings such as multidomain designs, rolling samples and combining experiments in his paper. Under combining experiments, he also discussed combining surveys, for which he mentioned methods were newly emerging. Although both methods have similar aims in order to investigate useful relations between two distinct topics, they differ from each other in terms of three main points. First, combining experiments can be applied on the reports of experiments already performed while combining surveys need advance preparation, planning and coordination. Second, combining experiments includes randomization of variables over subjects while combining surveys is based on probability sampling with randomized selections of subjects not variables. Third, combining experiments can be used at the end of the analyses. However, combining surveys provides a full statistical analysis for joint analysis. It is also indicated that designs for multipopulation (multinational) and periodic surveys have become more important and common since they provide spatial and temporal comparisons, respectively. Besides that, it is stated that the variations between the populations are commonly confirmed as obstacles to combinations for multinational samples or on cumulations of periodic surveys, thus they are used for restricting the sample estimates to single populations. In this thesis, combination of sample surveys, especially the household surveys, coming from single population will be the interest.

There are various applications of combining surveys or combining information from different surveys in the literature. Schenker et al. (2002) combined National Health Interview Survey and National Nursing Home Survey to obtain prevalence estimates for several chronic health conditions and demonstrate benefits of combining complementary surveys. Both are combined and then the combined prevalence rate estimated by treating the two target populations as sampling strata of a single overall population. However, it should be noted that these two surveys cover different populations and have different definitions and reference periods for variables. Then, Schenker and Raghunathan (2007) studied on four different cases and combined information coming from multiple surveys to address non-sampling error due to non-coverage, missing data, measurement error and response error. Surveys used in these cases cover different populations and have different interview modes. They emphasized that since information was obtained from different surveys, incomparability of information can be the most important problem. In the paper, five potential sources of incomparability of information were identified as follows: i) differences in the types of respondents and/or the sources of the respondents' information, ii) differences in the modes of interviewing, iii) differences in the survey contexts, iv) differences in the sample designs, v) differences in survey questions.

In this thesis, focus was on combining surveys covering same population, having same interview mode, sample design and similar context by taking considerations on incomparability of information into account. Therefore, literature was limited to relevant studies, which are on statistical matching - the method used for this purpose.

Statistical matching is one of the methods used in integration on different data files for single populations. By this way, statistically matched file can be used by economists for various purposes as an extensive micro data set (Yoshizoe and Araki, 1999). According to Moriarity and Scheuren (2003), statistical matching is widely used by economists for policy microsimulation modelling in government and also plays a role in many business settings as well. Nevertheless, all of these purposes

cannot be expected when the file is created and this is called “statistical file matching” problem (Rubin, 1986). Therefore, different methods have been studied by different researchers (Okner 1972; Ruggles and Ruggles 1974; Wolff 1977; Rubin 1986; Rodgers 1984; Singh et al. 1993). Rodgers (1984) gives the review of practical issues in his paper.

Most of the studies given above belong to USA. Besides that different countries around the world have applied different methods independently since 1960’s (Rässler, 2002). Therefore, it is hard to give the literature on the development of the statistical matching techniques since there are various terminology used for statistical matching. Rässler (2002) gives a comprehensible summary on the history of statistical matching techniques in different countries and indicates that “Often our information is based on unpublished reports supplied by experts practicing statistical matching. Thus we hope to fill a gap in the literature and explain what often is left to the reader's imagination”.

As mentioned before, statistical matching has been studied for years. There are a lot of studies conducted on different statistical matching methods using different data sets. However, according to Rässler (2002), some of the early studies were not successful studies due to technical restrictions. Now, thanks to improvement in technology, applying matching on big data sets by using computer programmes make the process easier for researchers. Therefore, attention on statistical matching has increased lately. Summaries of the several studies conducted on statistical matching in different countries and different issues are given below.

Gavin (1985) presented a study on statistical matching of health services research for creating databases for health care issues. Survey of Income and Education (SIE) and Health Interview Survey (HIS), both representative samples of the U.S. population, were statistically matched. SIE was used as the recipient data file and HIS as the donor file. Least-squares regression equations were used and matching was applied on individual level. Partially constrained matching was applied. This means that there would be a maximum number set for records to be

used, however, there could be exceptions to this constraint if used records were still better matches than other records. According to results, it is indicated that SM worked well with no significant bias.

Yoshizoe and Araki (1999) proposed at the time of their study that they had virtually no experience of SM with Japanese household surveys. For this purpose, they applied different statistical matching methods on two household based surveys namely, Family Income and Expenditure Survey (FIES) and Family Savings Survey (FSS). Monthly income and consumption is included only in FIES while assets and liabilities are included in FSS. Annual income and other household attributes are the common variables between surveys. These surveys share a portion of households, in other words; they have a common part. However, two files are matched with no common part and the common part is used to make comparison between the results of different statistical matching methods. Model of the consumption function is estimated using disposable income and financial assets variables. Three different combinations of the common variables are used as matching variables of statistical matching and unconstrained matching is implemented on household level using different distance functions. As a result of this study, they remarked that for successful statistical matching it is important to have external information. Unconstrained matching provides unsuccessful results for some of them. Statistical matching gave some bias in most of the cases since conditional independence assumption (CIA) did not hold. Mahalanobis distance using all matching variables did not provide good matching results. They remarked that if the estimation of statistical models is the interest, SM is not an efficient method of extracting information.

Examples on statistical matching are not limited to household surveys. Two surveys on agricultural enterprises, Farm Structure Survey and Farm Accountancy Data Network Survey, were matched using statistical matching techniques in Italy (Ballin et al. 2009). Most of the large enterprises are covered in both surveys but the portion of their total number in whole data is small. Matching was applied under three approaches in the study. First approach was matching under CIA. Second was

applying the Expectation-Maximization (EM) algorithm and the third one is the applying the EM algorithm by splitting the concatenated file in two non-overlapping part such as small and large farms. According to results, it is observed that first and second approaches provide similar results while third method provides quite different results. This was an expected result when association between file is concerned.

Leulescu and Agafitei (2013) conducted two studies on statistical matching. First, they matched European Union Statistics on Income and Living Conditions (EU-SILC) and 2007 European Quality of Life Survey (EQLS) in order provide integrated statistics on economic and subjective well-being aspects of people's life. Results were presented for Finland and Spain. EU-SILC was used as recipient data file whereas EQLS used as donor. Distance unconstrained hot deck (nearest neighbour) and model based methods were applied. Since all of the variables were binary variables, distance defined on the similarity of coefficient of Dice was used. Logistic regression and predictive mean matching were used for the model based methods. Results based on four target variables analysed and they showed that both methods preserved well the variables' marginal distributions. Yet, results tended to differ when it comes to joint distributions of variables not observed together since capturing the dependence relationship between variables became difficult because of the limitation on the number of variables used in the model. Moreover, detailed analysis showed that optimal solutions at national level require tailored approaches.

Then, Leulescu and Agafitei (2013) statistically matched two surveys, European Union Statistics on Income and Living Conditions (EU-SILC) and Labour Force Survey (LFS), since there was a need to analyse labour market information and employment related income together. These two surveys share a large common set of variables and their definitions, classifications, marginal and joint distributions were consistent with each other. EU-SILC was used as donor and LFS used as recipient. The Hellinger distance was applied to compare similarity of common variables' marginal distributions. To select the matching variables, multivariate analysis and modelling techniques were applied. Then, hot deck, regression based methods, predictive mean matching method and probabilistic decision algorithms were tested.

The results were presented based on the data of seven countries. According to results, it was said that matching provided good results for marginal and joint distributions when pre-requisites of coherence are met. For the variables not observed together more complex quality checks were needed such as the uncertainty analysis. Nevertheless, SM provided good inferences for specific estimates when model assumptions held. Moreover, it was emphasised that, due to differences between countries in terms of wage and population definitions, and based on how good the assumptions regarding SM are met; there exists a need for tailored applications and “fine-tuning” for different countries.

Webber and Tonkin (2013) aimed to compare people’s exposure to poverty using income, expenditure and material deprivation measures. Since there is no single data set that includes income, expenditure and material deprivation variables together, they statistically matched Household Budget Survey (HBS) with EU-SILC for the 2005 United Kingdom data. Hot-deck, parametric and mixed methods were applied assuming conditional independence. Validity of the matching procedures was tested comparing the distributions of the matched and observed variables in four ways. Results indicated that SM is broadly effective. Parametric approach did not provide good results on similarity of joint distributions. Suitable matching variables were limited since 2005 data set includes information on only household level. Therefore, they proposed that when there is a small number of variables, the risk of model misspecification increases, resulting in decreased reliability of parametric matching.

Roszka (2015) matched HBS and EU-SILC conducted in Poland. Aim was the extension of the scope of the estimates and target variables were household expenditure from EU-SILC and head of household income from HBS. Multiple imputation and mixed approaches were used in the application and matching was applied on household level. Imputations were made based on a linear regression models. According to analysis, both approaches preserved well the essential characteristics of the distribution. Moreover, results showed that when imputations done from smaller to larger dataset, multiple imputation approach provide better

result while mixed seems better otherwise. Using conditional independence assumption can help constructing high-quality estimators without using additional information. It was also indicated that for the selection of the appropriate matching variables and integration model, each target variable should be analysed separately.

2010 wave of HBS and EU-SILC of six European Union countries were matched by Serafino and Tonkin (2017) using non-parametric, parametric and mixed statistical matching methods and assuming conditional independence. This study was built on preliminary work carried out by Webber and Tonkin (2013) and comparisons were made in a similar way. All three methods provide good results. Hot-deck and mixed methods performed marginally better overall. Beside these, it was indicated that the number of potential matching variables was less in 2010 than 2005. Therefore, the quality of matching can reduce because of this reduction.

2011 HBS and 2012 SILC of Turkey were matched by Ahi (2015) in same manner with Webber and Tonkin (2013). Results were the same, in other words; non-parametric hot deck and mixed approaches performed well overall.

Time Use Survey and Consumer Expenditure Survey of Burkina Faso were matched by Anil et al. (2017) using three different statistical matching procedures. According to results, they proposed that SM can be effectively used while combining time use survey with consumers' expenditures.

CHAPTER 3. METHOD AND DATA SOURCES

Method of statistical matching and data sources are covered in this chapter. First, methodological framework of statistical matching then the important issues to be considered in the implementation of matching techniques are presented. Afterwards, micro data sets of Life Satisfaction and Time Use Surveys, used household surveys in the dissertation, are stated.

3.1. Methodology of the Statistical Matching

Statistical matching (SM), also called data fusion or synthetic matching, is a method used for integration of two or more data sources. This provides the joint information on the not jointly observed variables. Matching can be applied on micro or macro level. These two approaches are explained in the literature (D'Orazio, et al., 2006; Netherlands, 2015) as follows.

Micro level matching aims to estimate the target variable in one data source by using information coming from the other source. Thereby, complete synthetic data set, including information on all variables, are constructed using records of individuals in different sources. Constructed new data set called as complete since it includes whole variables of interest and called as synthetic since it is not obtained in a direct way but obtained by using some appropriate matching methods. The synthetic data set can be constructed in two ways. First, two data sources can be concatenated and then missing values that come from both data sources, can be filled using the information of each other. In this situation, sample size of this constructed new data set equals to total sample size of two surveys. Secondly, integration can be done considering only one data source and missing values in this data source can be imputed using the available information in the other source (D'Orazio, 2013). In this situation, these two data sources are called as the *recipient* and the *donor*, respectively. Structure of the data sources under these two approaches is given in

Figure 3.1.1 and Figure 3.1.2, respectively. Empty cells represent the unobserved variables in samples.

Figure 3.1.1. Structure of overall sample under first way integration

Sample	X_1	...	X_P	Z_1	...	Z_Q	Y_1	...	Y_R
<i>Sample₁</i>	x_{11}^1	...	x_{1P}^1	z_{11}^1	...	z_{1Q}^1			
			
	$x_{n_1 1}^1$...	$x_{n_1 P}^1$	$z_{n_1 1}^1$...	$z_{n_1 Q}^1$			
<i>Sample₂</i>				z_{11}^2	...	z_{1Q}^2	y_{11}^2	...	y_{1R}^2
			
				$z_{n_2 1}^2$...	$z_{n_2 Q}^2$	$y_{n_2 1}^2$...	$y_{n_2 R}^2$

Figure 3.1.2. Structure of data sources under second way integration

<i>Sample₁</i> (Donor)						<i>Sample₂</i> (Recipient)					
X_1	...	X_P	Z_1	...	Z_Q	Z_1	...	Z_Q	Y_1	...	Y_R
x_{11}^1	...	x_{1P}^1	z_{11}^1	...	z_{1Q}^1	z_{11}^2	...	z_{1Q}^2	y_{11}^2	...	y_{1R}^2
...
$x_{n_1 1}^1$...	$x_{n_1 P}^1$	$z_{n_1 1}^1$...	$z_{n_1 Q}^1$	$z_{n_2 1}^2$...	$z_{n_2 Q}^2$	$y_{n_2 1}^2$...	$y_{n_2 R}^2$



<i>Synthetic data set</i>									
\hat{X}_1	...	\hat{X}_P	Z_1	...	Z_Q	Y_1	...	Y_R	
\hat{x}_{11}^1	...	\hat{x}_{1P}^1	z_{11}^2	...	z_{1Q}^2	y_{11}^2	...	y_{1R}^2	
...	
$\hat{x}_{n_2 1}^1$...	$\hat{x}_{n_2 P}^1$	$z_{n_2 1}^2$...	$z_{n_2 Q}^2$	$y_{n_2 1}^2$...	$y_{n_2 R}^2$	

According to D'Orazio et al. (2006), the decision of which source will be the recipient and which the donor differs according to many factors. Nevertheless, accuracy of information in the data sources is the most important issue. Therefore, role of the donor or recipient assign to equally reliable sources in general since they come from the same distribution. Another important criterion to be considered while

assigning role to the data sources is their sample sizes. Common practise is choosing the smaller data file as the recipient. Otherwise, some records would have to be imputed several times and by this way the variability of the distribution of the imputed variable would be artificially modified. On the other hand, as an alternative view, D'Orazio (2013) is stated that using the larger data set as recipient provides more accurate results since the further statistical analyses will be depend on it. Yet, sample sizes of the sources should not be very different from each other, otherwise; statistical matching might provide inaccurate results.

Macro level matching aims to construct parametric models for whole data. The parameters of these models are estimated for the purpose of estimating population parameters of interest. In other words, data sources are used for direct estimation of the joint distribution function of the variables of interest that are not observed in common.

In this dissertation statistical matching procedure will be explained on two different data sets (samples). For the integration of more than two data sets, two of them can be matched first, and then the other sources can be matched with this new data set one by one. Let's assume two independent sample surveys, *Sample*₁ and *Sample*₂, coming from the same infinite population and having sample sizes n_1 and n_2 , respectively. Samples are assumed to have independent and identically distributed (i.i.d.) observations. *Sample*₁ (donor) has information on variables X and Z, whereas *Sample*₂ (recipient) has information on variables Y and Z, matched variable would be denoted by \hat{X} . Since they come from the same infinite population, X, Y and Z values are assumed to be a set of random variables and their joint distribution follows a given model. This model might be known or unknown. In this way, inference is said to be model based (D'Orazio, 2013). Common variables contained in both data sources are used to impute the missing items/records in the recipient data set using the donor data set (D'Orazio et al., 2006; De Waal, 2015). Here, Z includes the common variables from both data sources; therefore, Z variables will be used as matching variables. However, matching variables do not have to

contain all of the common variables. A subset of the common variables is generally used as matching variables for several reasons. How to choose these matching variables will be explained later in this chapter.

Relationship between X and Y is another important issue that should be considered in SM procedure. Unfortunately, there is a limitation on measuring the association between X and Y conditional on Z. It cannot be estimated and because of that it is assumed to be zero in general. This assumption is named as *Conditional Independence Assumption (CIA)*. This limitation leads to an important concern on the results since it affects validity. Matching procedure provides almost perfect match (linkage) and produces accurate estimates from integration of multiple sources if the assumption holds. On the other hand; if the assumption does not hold, in other words if the conditional independence cannot be provided and there is no additional/auxiliary information, some identification problems may occur and produced synthetic data set may lead to incorrect inferences (Leulescu and Agafitei, 2013). Uncertainty analysis and use of auxiliary information are two main approaches used to tackle with the conditional independence assumption. Here, uncertainty analysis refers to the sensitivity of the results according to different assumptions.

Up to now, brief information about methodology of statistical matching procedure is given. Statistical matching approaches applied in the matching procedure will be presented as the next section.

3.1.1. Statistical Matching Approaches under the CIA

There are various methods used in statistical matching procedure. These methods can be grouped under three main approaches as parametric, nonparametric and mixed, respectively. Information about them is given below.

3.1.1.1. Parametric Approach

Parametric approach is one of the approaches used in SM. This approach can be applied on both micro and macro level matching.

A regression model needs to be specified for the parametric approach. Then, predicted values used in imputation are obtained from this regression model. D’Orazio (2013) stated that this method might become too burdensome since it deals with many and mixed type of variables. Transformation of variables might be needed in some cases and such operation might create undesired noise. Besides that, regression towards the mean might be another problem in this approach (Serafino and Tonkin, 2017). Here, the main important point is that a wrong specification provides unreliable results.

3.1.1.2. Nonparametric Approach

Nonparametric micro matching procedures are also known as hot deck imputation methods. These methods use one of the data set as recipient and the other one as donor then fill the non-observed variables in the recipient with values exist in the donor according to some distance functions. Once the distance function is determined, it is calculated for the common variables and each record in the recipient is associated with the nearest record that shows a smallest distance in the donor file. When more than one records having equal distance, one of them is selected randomly. Distance functions can be defined in many ways and a weighted distance can also be adopted (Leulescu and Agafiței, 2013).

In the dissertation, three mostly used hot decks methods namely nearest neighbour, random and rank hot deck methods were used in the statistical matching. These methods were applied using ‘StatMatch’ package in R software.

- Nearest Neighbour Distance (NND) Hot Deck: selects the donor according to minimum distance. Many distance functions can be used. Default is the

‘Manhattan’ distance. Manhattan distance converts all non-numeric variables to numeric. On the contrary, all variables are converted to character in ‘Exact’ distance function. Exact distance was used in the application since matching variables are categorical. It converts variables to character and then calculates ‘Gower’ distance. Its formula is given below (Gower, 1971). NND allows the definition of donation classes and this reduces the effort in computing distances. Moreover, it is possible to put a constraint for avoiding the selection of a donor more than once (constrained matching). Here, the important point is that NND turns to random hot deck when matching variables are categorical (Chen and Shao, 2000).

$$\text{Gower distance} = \frac{\sum_k (s_{ijk} * w_k)}{\sum_k (d_{ijk} * w_k)} \quad (3.1.)$$

Here,

- s_{ijk} equals to 1 for a pair of TRUE logicals or matching factor levels, and the absolute difference for metric variables.
- w_k is weight of the variable.
- d_{ijk} equals to 0 for missings or a pair of FALSE logicals, and 1 else.

- Random Hot Deck: selects the donor randomly from a suitable subset of all the available donors. Donation classes can be fixed or “moving” classes. When donation class is defined a donor is picked up completely at random within the same donation class. The selection of the donor can be done with probability proportional to a weighting variable. Different distance functions can be used like in NND hot deck. ‘Exact’ distance function used in application. Moreover, identification of the subset of the closest donor records can be formed in many ways. Default is the ‘rot’. ‘rot’ and ‘min’ options were used in the application. For the ‘rot’, the number of the closest donors to retain is given by $\sqrt{\text{Total number of available donors} + 1}$. In ‘min’, donors at the minimum distance from the recipient are retained.

- Rank Hot Deck: selects the donor closer based on the distance between the percentiles of the empirical cumulative distribution function of the continuous common variable. It allows the definition of donation classes in this case the empirical cumulative distribution is estimated separately class by class. Weights can be used in matching. Constrained matching is possible like in NND.

3.1.1.3. Mixed Approach

Mixed approach contains two steps. First, parametric approach is applied, in other words; a regression model is fitted and all parameters are estimated. Then, nonparametric approach is applied to create a synthetic data set. According to D’Orazio (2017), two steps procedure has some advantages. It provides protection against model misspecification and reduces the risk of bias in the marginal distribution of the imputed variable.

So far methodology of statistical matching procedure has been presented in general concept. Now, statistical matching for complex sample surveys will be covered.

3.2. Statistical Matching for Complex Sample Surveys

In the previous section, the general concept of statistical matching is presented assuming i.i.d. samples coming from an infinite population. In this situation, randomness or variation is induced by the model which generates the data. However, this is not the case in real life. Most of the time, data is compiled from surveys coming from finite populations and consisting of N units ($N < \infty$). In this case X , Y and Z values are viewed as fixed and randomness comes from the probability criterion. In other words, sampling design is the source of variation. Inference is based on inclusion probabilities, non-null probability of being included in the sample. For simple sample surveys, observed values in samples can be assumed i.i.d. since the sample is selected by simple random sampling.

Unfortunately, for the complex sampling designs, having multiple stage sample selection with stratification and clustering, it is difficult to hold the i.i.d. assumption. In this situation, taking sampling design and sample weights into account becomes an important issue in order to provide reliable matching results. To do that design variables should be known and if they are not related with variables of interest then the sampling design can be ignored.

According to D'Orazio, 2013, there are some other issues that might create problems when integrating two complex sample surveys coming from the same target population. These issues are given below.

- Unit nonresponse caused by non-contacts, refusals, etc.
- Discarding of ineligible units
- Missing values or values identified as erroneous
- Values affected by measurement errors
- The final weights adjusted according to unit nonresponse, under coverage and population totals
- Partially available design variables due to risk of disclosure

Data sources that will be used in SM should be formed by initial sampling design and final weights (w) set. Sampling design might be fully or partially known and the final weights include modified/corrected initial weights. Structure of the data sources is presented Figure 3.2.1.

Figure 3.2.1. Structure of the data sources in SM for complex surveys

<i>Sample₁</i> (Donor)						<i>Sample₂</i> (Recipient)					
X_1	...	W^1	Z_1	...	Z_Q	Z_1	...	Z_Q	Y_1	...	W^2
x_{11}^1	...	w_1^1	z_{11}^1	...	z_{1Q}^1	z_{11}^2	...	z_{1Q}^2	y_{11}^2	...	w_1^2
...
$x_{n_1 1}^1$...	$w_{n_1}^1$	$z_{n_1 1}^1$...	$z_{n_1 Q}^1$	$z_{n_2 1}^2$...	$z_{n_2 Q}^2$	$y_{n_2 1}^2$...	$w_{n_2}^2$

For matching of complex sample surveys, synthetic data sets can be created under the micro approach or finite population parameters can be estimated regarding the relationship between X and Y under the macro approach, like in the general SM procedure.

3.2.1. Statistical Matching Approaches for Complex Sample Surveys

As mentioned in the previous section, the sampling design and the weights are important issues that should be taken into account for complex sample surveys. Different approaches are developed considering the design and the weights in different stages of statistical matching in complex survey. These approaches can be grouped mainly under two headings; “approaches not considering the sampling weights” and “approaches considering the sampling weights”, respectively.

Approaches not considering the sampling weights in SM for complex sample survey are also called as “*Naive Micro Approach*” (D’Orazio, 2013). This approach contains nonparametric micro methods and use hot deck methods such as rank, random and nearest neighbour distance hot deck. They are applied without taking the sampling design and the sample weights into account. While obtaining the synthetic data set, the sampling design and the weights are not the issue; however, after the constructing of the synthetic file both sampling design of the recipient file and weights of the units are taken into account while the analyses are carried out. See the D’Orazio, 2013, pp.35-37 for the examples.

Alternatively, design variables and sample weights can be used to form donation classes. By this way, sample weights are used in the matching procedure then fused data set will be created.

3.3. Preparation Steps Before Applying Statistical Matching

So far, the definition of statistical matching, its aim and different techniques developed according to purpose of the matching have been discussed. It is underlined that the selection of the appropriate matching method regarding the properties of data sources and the aim of the matching are very important issues, since they affect the results and might create some undesired situations. However, selection of the method is just one of the important stages in the statistical matching procedure. There are other even more essential issues that should be considered in the matching stages.

Leulescu and Agafiței (2013) with Serafino and Tonkin(2017) stated that the statistical matching has certain pre-requisites about harmonization and coherence of data sources that will be matched. In the previous sections, samples are assumed to be integrable, homogenous in terms of their definitions and concepts. Nevertheless, data sources generally have so many different forms. Two sample surveys may be incompatible in many ways even if they are conducted by same institution. For this reason, D’Orazio et al. (2006) proposed the stages of statistical matching as follows:

First of all, reconciliation process should be implemented on data sources to enable the joint analysis of multiple sources. Then, matching variables should be selected according to multivariate analysis. Modelling techniques have to be implemented for the selection. After that, matching techniques can be applied. Here, monitoring each stage carefully is very important to produce accurate results.

3.3.1. Reconciliation of the Data Sources and Harmonization of Variables

Data sources should be integrable to achieve successful matching results. In general, this rarely happens. Most of the time data sources are not compatible and making these sources compatible requires a reconciliation of the data sources in terms of their concepts and definitions. This process is applied on micro level.

Van der Laan (2000) presents nine steps to be performed for the harmonization of the different sources. These are:

- i. harmonization of units
- ii. harmonization of reference periods
- iii. completion of populations (coverage)
- iv. harmonization of variables
- v. harmonization of classifications
- vi. adjusting for measurement errors (accuracy)
- vii. adjusting for missing data (item non-response)
- viii. derivation of variables

Harmonization of units refers to the checking whether the statistical units are defined uniformly in all of the sources. Reference periods of the sources should be checked whether they refer to the same period or the same point in time. Same target population should be covered in all sources. Corresponding variables should be defined and classified in the same way. After harmonizing definitions, corresponding variables should be checked whether they have the same value or not. Then missing data should be adjusted in a way that all variables possess a value.

Common variables included in both datasets needed to be harmonized between sources in order to be used in matching. This is achieved recoding the variables in a way that they have the same degree of detail.

3.3.2. Choosing the Matching Variables

Before using common variables in the matching procedure, all of them should be checked in terms of two criteria. First, the distribution of the variables must be similar between two sources. Second, the variables must be significant in explaining variations in the target variables.

To compare similarity of variables' distributions, weighted frequency distributions or measure such as Hellinger Distance (HD) can be used. Webber and Tonkin (2013) with Serafino and Tonkin (2017) states that HD is convenient for comparison of the similarity in distribution of two variables since it provides a single number as a measure. Therefore, interpretation of it is easy and it allows comparisons across variables and surveys. There is no certain rule about what degree of similarity is suitable for SM purposes. Webber and Tonkin (2013) proposed that a HD of over 5% should raise concerns about the similarities in distributions. HD calculation formula is given in Formula 3.2.

$$HD (V, V') = \sqrt{\frac{1}{2} \sum_{i=1}^K \left(\sqrt{\frac{n_{Di}}{N_D}} - \sqrt{\frac{n_{Ri}}{N_R}} \right)^2} \quad (3.2.)$$

Here,

- V: donor data set
- V': recipient data set
- K: total number of cells in the contingency table
- n_{Di} : the frequency of cell i in the donor data set
- n_{Ri} : the frequency of cell i in the recipient data set
- N: total size of the specific contingency table.

Matching variables have direct effect on accuracy of the matching results. Therefore, the choice of matching variable is a very important point in SM. Power of the matching variable depends on its power on the explaining variations in the target variable. It should be a good predictor of the target variable. Various methods can be applied to find the optimum set of predictors. Regression, factor analysis and deriving new common variables with highest possible explanatory power are some of these methods. In the dissertation, HD was used to compare similarity of the common variables' distributions. Then regression was applied on target variables in order to determine the matching variables. After determining the matching variables, statistical matching can be applied.

3.4. DATA SOURCES

Micro data sets of two household surveys, Life Satisfaction and Time Use Surveys, were used in the thesis. First, aims and scopes of the surveys and general information on their history will be presented briefly. Then, detailed information about micro data sets will be given in this chapter.

3.4.1. Life Satisfaction Survey

Life Satisfaction Survey will be covered in this session. First, its objectives and history will be explained briefly. Then, detailed information on 2014 Turkey Life Satisfaction Survey will be given since its micro data set was used in thesis.

3.4.1.1. Survey Objectives

Life Satisfaction Survey (LSS) is a survey conducted to measure people's personal perception of their quality of life and to monitor the changes in their perception. It aims to examine how they feel about their lives in terms of issues such as education, health, social security, employment and income, work-life balance, personal development, personal security and justice services, etc.

3.4.1.2. History of the Life Satisfaction Survey

Life satisfaction and happiness is an issue that has been discussed and tried to be explained in various dimensions since years. Obtaining data from the field on happiness level was began late 1940's over the world. "World Data Base of Happiness" has data bases containing rich information on theoretical and practical studies conducted in this field. Moreover, there is a published journal about life satisfaction named as "Journal of Happiness Studies".

Life Satisfaction Survey was conducted by TurkStat in 2003 for the first time and it was the first research on happiness produced as an official statistics in Turkey. Besides that, it is also a first survey of TurkStat including subjective items and also social elements.

2003-LSS was implemented in the Household Budget Survey as an additional module in 2003. After that, it has been carried out regularly every year. LSS gives estimation on country level. However, only in 2013, it was conducted to give estimation on provincial (NUTS-3) level to monitor the differences between regions and provinces. After the first implementation of survey, it was recognized that sufficient information to give estimation could not be obtained for some questions. Therefore, these questions were excluded from the questionnaire. At the same time, some questions about views on municipal services and Turkey's potential European Union membership were added to questionnaire. There have been some other changes, questions were excluded or added, in the LSS questionnaire over the years according to national and international needs.

3.4.1.3. Questionnaire Design and Variables of Turkey Life Satisfaction Survey 2014

Two questionnaires, "Household" and "Individual", were used to compile data in 2014 Turkey-LSS. Household Questionnaire was used to obtain information on household. Similarly, Individual Questionnaire was used to obtain information on individuals. Variables compiled in the survey are given under two main categories below in detail.

i. Household Variables

Variables related with housing conditions, education, income and the safety of the house are compiled in the household questionnaire.

Household size which is the number of the household members, the ownership status of accommodation, number of the rooms and area of the dwelling

unit, the presence of properties in accommodation such as municipal water, piped water system, toilet and bathroom, having a problem with issues such as leaky roof, dark (insufficient daylight), sewage overflows, noise from the neighbours and street, power and water outage, warming are the housing conditions variables.

Questions related with education such as taken courses for entrance examination in 2014, having a problem with registration to the school, quality of the education and tools at the school, satisfaction on attitude of the school administration and teachers, satisfaction on conditions at the school are asked after the housing conditions.

Number of individuals who bring income to the household, monthly net income group of the household and whether this income meets the needs of the household are asked as income questions. Last questions in the household questionnaire are about safety of the household. Being exposed to burglary or experienced crime victimization at home/workplace/fields/garden/car/motor, whether applying to police and the reason for not applying are asked in this section. Beside that, it is asked whether they are experienced any other crime victimization in the household. Reference date is year 2014 for these questions.

ii. Individual Variables

Variables complied via the individual questionnaire can be given under 7 categories as background information, happiness and satisfaction from individual situation, utilization and satisfaction from the public services, environmental safety, hope, self-evaluation and expectations by 5-year periods, values and view to European Union. Questions included in these sections are explained below in detail.

Sex, completed age, marital status, education status, working status and sector (public or private), situation at work and problems in the workplace are variables under “Background information”.

Level of happiness and life satisfaction, the person who makes happiest in life, the value which makes happiest in life, individuals' satisfaction level from own health/education, etc., satisfaction level of individuals in social circles, problems about work and welfare perception are asked in the “Happiness and satisfaction from individual situation” section.

Satisfaction level from the health/public order/etc. services, the beneficiary social security organization, by whom benefiting the social security and level of satisfaction, channel meets costs of the treatment or medication during the illness, institution or person first referred during the illness and the reason for choosing that institution or person, problems with health/public safety/judicial/educational services, problematic health institution, the institution received public order service, satisfaction from education received, satisfaction from public services' obtaining information process and satisfaction with municipal/provincial special administration/transportation services are asked to individuals under “Utilization and satisfaction from the public services” section.

“Environmental safety” includes questions on the level of feelings safe in the home and in the living environment, person from which individual get help when needed and events related with the public order.

Level of hope, the development level according to 5 years ago, estimation of the development after 5 years, personal and national (for Turkey) expectations for the next year are asked in “Hope, self-evaluation and expectations by 5-year periods”.

Important values to be honourable in society, people's given importance to the situation in the environment, perception of the social pressure, changes in the lives in the last 1 year and interest level in social issues are variables under the “Values” section.

Individuals' thoughts on which direction Turkey's membership of the European Union affects their life if Turkey becomes a member of the European

Union and the vote of the individual if there is a referendum on Turkey's membership of the European Union are asked in “View to European Union” section.

Three classifications, NUTS, ISCED-97, and ICSE-93, were used in Turkey Life Satisfaction Survey, 2014. “*NUTS*” stands for “Nomenclature of Territorial Units for Statistics” and it is used for dividing the territories into hierarchical levels. This classification provides cross-border statistical comparisons. “*ISCED-97*” stands for “International Standard Classification of Education”. It is used to classify individuals’ education. “*ICSE-93*” stands for “International Classification of Status in Employment”. Their codes are given in Appendix A-C.

3.4.1.4. Survey Design and Weighting Procedure

Survey design and weighting procedure of Turkey Life Satisfaction Survey-2014 will be presented in this section regarding coverage and sampling frame, sample design and selection, weighting procedures and variance estimation, respectively.

i. Coverage and Sampling Frame

Geographical coverage of Turkey Life Satisfaction Survey (TLSS), 2014 is whole settlement areas within the territory of Turkey. TLSS covers all household members who are 18 years old and over and living in the territory of Turkey including Turkish citizens and foreign people, except for institutionalized population. In other words, nomadic population and population who live at rest and elderly homes, dormitories, military barracks, recreation quarters for officers, correctional facilities, special hospitals, etc. were excluded from the survey coverage. Small settlement areas that have smaller than 20 households were also excluded from the coverage. Population in these excluded settlements are under 1% of the total population.

Sampling frame of TLSS consists of all household addresses in Turkey. A household address is included in sampling frame only if there is at least one individual who lives in that address. This sample frame was obtained via Address Based Population Registration System and National Address Database.

ii. Sample Design and Selection Procedure

Estimation level of Turkey Life Satisfaction Survey-2014 is whole Turkey. For this purpose, 4560 households were selected using two-stage stratified cluster sampling. At the first stage, clusters including nearly 100 households were selected using probability proportional to size (PPS) method. Then at the second stage, households were selected systematically from these sample clusters.

New administrative division in 2014 has changed the definition of urban and rural. Therefore previous and current estimations on urban and rural areas cannot be comparable. For this reason, urban and rural were not used as an estimation level. Nonetheless, they were used as design domain and sample selection was done taking care of this. As a result, 58 from rural and 398 from urban totally 456 clusters were selected. Then, 10 households were selected from each sample cluster. 3908 household were interviewed.

Over coverage and household level unit response rate were 9.16% and 94.35%, respectively for TLSS, 2014. Since loss rate was taken to consideration in sample size calculation, substitution was not used in the survey.

iii. Weighting Procedure and Variance Estimation

Five steps were followed in the weighting procedure of TLLS. These are Design weight calculation, Non-response adjustment, Integrative calibration, Trimming for outliers and Overall inflation factor calculation.

“Design weights” are calculated as inverse of overall selection probabilities, in other words; inverse of multiplication of first stage and second stage selection

probabilities. Households within the same cluster have equal design weights. “*Non-response adjustment factor*” is calculated as inverse of response rate at household level. Households within the same cluster have equal non-response adjustment factor. Besides that, individual non-response adjustment factor is calculated in each age group-gender considering design domain at individual level.

“*Integrative calibration*” uses iterative proportional fitting. It was applied to equalize the weights of individuals in the household and household weight. Projected age-gender, NUTS1-rural/urban and household size population distributions were used in calibration as auxiliary variables. “Trimming” checks the outliers/extreme values and large variation in weights after the calculation of non-response adjustment factor and calibration weights. These values are recoded to boundary of the limits if they are outside the limits. Until all weights fall between the boundaries, calibration and trimming follow each other.

“*Overall Inflation Factor*” is calculated by dividing mid dated total projected population of the fieldwork to the sampled population. As a result, final weights are calculated as product of all weights of individuals and overall inflation factor.

Variance estimations are calculated using Taylor Linearization approach. This approach uses convergence method under particular assumptions for complex surveys.

3.4.1.5. Data Collection

Turkish Life Satisfaction Survey was designed to produce annually estimates by Turkey level. Field application of TLSS is held in November for each year. Hence, fieldwork of TLSS-2014 was conducted between 3rd November and 1stDecember, 2014. Data are compiled with Computer Assisted Personal Interview.

3.4.2. Time Use Survey

Time Use Survey will be covered in this session. First, its objectives and history will be explained briefly. Then, detailed information on 2014-2015 Turkey Time Use Survey will be given since its micro data set was used in thesis.

3.4.2.1. Survey Objectives

Time Use Survey (TUS) is a survey conducted in many countries to examine how people spend their time on daily life activities. There are several objectives of TUS and these are:

- To measure how people divide their time among various daily life activities such as housework, childcare, work, etc.,
- To observe time usage differences between various population groups in terms of different characteristics such as gender, age, work status, etc.,
- To provide data for estimations of the gross domestic product,
- To obtain internationally comparable data on time usage issue.

3.4.2.2. History of the Survey Time Use Survey

Time Use Survey has been conducted in many countries, especially in Europe since 1960s. Although many countries obtain data on time usage, they were not internationally comparable. Within this scope, Statistical Office of the European Union (Eurostat) has started to work on obtaining internationally comparable Time Use Survey data at the beginning of the 1990s. The pilot studies were carried out between the years 1996 and 1997 in 18 countries, nine of which were European Member countries and nine of which were transitional countries. ‘Guidelines on Harmonised European Time Use Surveys’ were published in 2000 and the first results of pilot studies were disseminated in 2004 via Eurostat Press Release. After

that, between 2006 and 2007 Eurostat performed the second step of harmonizing works then published a new Guide Book in 2008.

Within the scope of internationally comparable Time Use Survey, Turkish Statistical Institute (TurkStat) conducted a pilot survey in 1996. 117 households were interviewed and evaluation report was prepared. For 2006 Turkey-TUS, another pilot survey conducted as part of the preparation studies. 78 households with 3 households in each Regional Directorate were interviewed between 25th July and 7th August 2005. After that, 2006 TUS of Turkey was conducted by Labour Force and Living Conditions Department of TurkStat between 1st January and 31th December 2006 on 5070 households. Second TUS of Turkey was conducted by Demographic Statistics Department of TurkStat between 1st August 2014 and 31th July 2015 on 11440 households. There are differences between two surveys in terms of estimation level, target population and questionnaire.

3.4.2.3. Questionnaire Design and Variables Turkey Time Use Survey, 2014-2015

Three questionnaires were used to compile data in 2014-2015 Turkey-TUS and these were “Household Questionnaire”, “Individual Questionnaire” and “Diaries”. Household Questionnaire was used to obtain information on household and it was filled by interviewing with a household individual who has information on household and who is 18 and over. Other questionnaires were filled by individuals who are 10 and over in the household. Variables compiled in the survey are given under four main categories below in detail.

i. Household Variables

Variables related with housing and living conditions, growing plants and keeping or breeding animals, household income, and received help are located under this category.

“Housing and living conditions” variable includes information on the dwelling type and the ownership status of accommodation, the presence of properties in accommodation such as balcony, garage and garden, numbers of rooms in accommodation, belongings used by the household, the status of building and accommodation to live in and having extensive repair in the accommodation. *“Growing plants and keeping or breeding animal”* variables includes information on status of keeping a pet, breeding domestic animal, growing herbal product and making money from the herbal or domestic products. *“Household income”* variable contains information related with all income sources and the main income source that household earn and net monthly average household income. *“Received help”* variable includes information on persons who received help except household in the form of service.

ii. Household Composition Variables

Information on household composition such as individuals constituting the household, their gender and age and relationship of the household individuals to head of the household is located under this category. Moreover, information on childcare is also given in this category. It includes the status of care of the children who are smaller than 10, information about the persons taking care of them and the frequency of care.

iii. Individual Variables

Variables related with individuals’ background information, education, health, owned technological products, social participation, volunteer work, help and services to others, employment and unemployment, time use and eldercare are compiled under this category. Detailed information about these 11 parts is given below.

- “Background information” includes questions about individuals composing the household, their sex, completed age, place of birth (born in Turkey or not) and marital status and relationship to head of the household.

- “Formal and mass education” part includes information on individual’s educational status, the school attending and the taken courses.

- “Health” part includes information on individual’s general state of health and chronic illnesses or disability status.

- “Owned technological products” part includes information on whether the individual has technological products such as mobile phone, laptop, Ipod or mp3 player, VCD or DVD, Photo camera, camera, game console or other products apart from these.

- “Social participation” part includes information on individual’s present membership status of a non-profit civil society organizations, cooperative and professional association, union, political party, sports club, foundation and association.

- “Volunteer work” part includes information about voluntary activities done by individual in last four weeks for welfare groups (elderly, disabled, children charity groups), sports clubs and associations, help to a place of worship (build a mosque, cleaning, repair etc.), political groups or clubs, the groups formed by young people (youth groups, scouts, guidelines etc.), security/first aid groups (The Red Crescent etc.), environmentalist groups, justice/human rights groups, fellow countrymen associations etc. regional solidarity groups, art and hobby groups, professional solidarity associations, parent-teacher association, adult education groups and any other organization apart from these.

- “Help and services to others” part includes information on helps done to persons who are outside the household in last four weeks for food preparation, house cleaning and tidying, clothes washing, ironing and maintenance, gardening (watering flowers, etc.), domestic animals

maintaining and circulating, payment of bills, care of children, children's home, school, day-care centre transport, care of disabled, patient or elderly people (bath, haircut, etc.), transportation of other adults (to market, doctor etc.), shopping, repairing and maintenance of the house tools and equipment, furniture montage or repair, car wash, education given to the family members who are outside the household, health-related services, other help and services apart from these.

-For “Employment” part, respondents are individuals who are 15 years and older. It includes information on individual’s working status in the last week, main activity of work place, reasons for not working in last week, occupation and employment in work place, weekly normal working hours, net monthly average and annual individual income earned from the job, working status and weekly normal working hours in the second job, net monthly average individual income earned from the second job.

-Respondents are again 15 years and older individuals for “Unemployment” part. This part includes information on individuals’ job searching status in the last four weeks, job searching channels and reasons for not looking a job or starting a work.

-“Time use” part includes information on whether the individual go to entertaining and cultural activities such as cinema, theatre, concert, ballet and opera, art exhibition or museum, library, sports activities, cafe or bar, internet cafe, shopping mall, visiting relatives and friends, kermis or fair, picnic, daily tours or nature walks in the last four weeks and their frequencies. It also includes information about whether individuals read a book, newspaper or magazines, watch a television, listen a radio, solve a puzzle, and spend time on the social media in the last four weeks. Sports activities done regularly such as walking or jogging, cycling, skiing, swimming, football, basketball, volleyball, judo or karate, sailing or surfing, instrumental sports activities and their frequency are also questioned. Other

issues covered in this part are frequency of the daily jobs' intensity, whether the individual think frequently that s/he cannot do activities that s/he want on weekdays because of insufficient time and which activity the individual prefers if s/he has enough time.

-“Eldercare” part includes information on status and frequency of providing unpaid assistance or care for an old person in the last four months, to whom and how long this unpaid eldercare was provided.

iv. Diary Variables

Diary variables include information on individual's sex and age, dairy day, activity done (sleeping, eating, working, household and family care, etc.) and its duration with starting and ending time, location of the activity and person/s near the individual during activity time. Weighting factor of individuals are also given as a last variable in dairy data set.

“HETUS-ACL”, “NACE Rev.2”, “ISCO-08”, “ISCED-97” and “ICSE-93” were used as classifications in TTUS, 2014-2015. “*HETUS-ACL*” stands for “Harmonised European Time Use Surveys - Activity Coding List”. It is used to classify daily activities of individuals. “*NACE Rev.2*” is a "Statistical Classification of Economic Activities in European Community". It is used to classify economic activities of individuals who are employed. “*ISCO-08*” stands for "International Standard Classification of Occupations". It is used to classify employed individuals' occupations and profession groups. “*ISCED-97*” stands for “International Standard Classification of Education”. It is used to classify individuals' education. “*ICSE-93*” stands for “International Classification of Status in Employment”. Their codes are given in Appendix B-F.

3.4.2.4. Survey Design and Weighting Procedure

Survey design and weighting procedure of Turkey Time Use Survey, 2014-2015 will be presented in this section regarding coverage and sampling frame, sample design and selection, weighting procedures and variance estimation, respectively.

i. Coverage and Sampling Frame

Geographical coverage of Turkey Time Use Survey (TTUS), 2014-2015 is whole settlement areas within the territory of Turkey. TTUS covers all household members who are 18 years old and over and living in the territory of Turkey including Turkish citizens and also foreign people, except for institutionalized population. This means that population who live at rest and elderly homes, dormitories, military barracks, recreation quarters for officers, correctional facilities, special hospitals, etc. were excluded from the survey coverage. Nomadic population was also excluded from the coverage.

Sampling frame of TTUS consists of all household addresses in Turkey. A household address is included in sampling frame only if there is at least one individual who lives in that address. This sample frame was obtained via Address Based Population Registration System and National Address Database.

ii. Sample Design and Selection Procedure

Estimation level of TTUS is whole Turkey. For this purpose, 11.440 households were selected using two-stage stratified cluster sampling. At the first stage, clusters including nearly 100 households were selected using probability proportional to size (PPS) method. Then at the second stage, households were selected systematically from these sample clusters.

Because of the same reason explained in LSS, Urban and rural were not used as an estimation level but used as design domain and sample selection was done

taking care of this. As a result, 156 from rural and 988 from urban totally 1144 clusters were selected. Then, 10 households were selected from each sample cluster.

Over coverage, household level unit response rate and individual level unit response rates were 10.03%, 88.15% and 92.5%, respectively for TTUS, 2014-2015.

iii. Weighting Procedure and Variance Estimation

Five steps same with Life Satisfaction Survey were followed in the weighting procedure of TTUS. “*Design weights*” and “*Non-response adjustment factor*” in TTUS are calculated in the same manner as LSS. “*Integrative calibration*” uses iterative proportional fitting, like in LSS. However, while NUTS1 level projected population distributions used in LSS, NUTS2 level ones are used in TTUS as auxiliary variables. “*Trimming*”, “*Overall inflation factor*” and “*Variance estimations*” are also calculated in the same manner as LSS.

3.4.2.5. Data Collection

Fieldwork of TTUS, 2014-2015 was conducted between 1st August 2014 and 31st July 2015. Working month numbers from 1 to 13 were assigned to sample clusters since there are 13 working months in a year. Besides that, Day number from 1 to 5 for weekdays and 1 to 2 for weekend was assigned to sample households. Thus, number of sample clusters and households were equalized per each working month and representation of each day was provided equally.

Data collection of TTUS consists of two parts. First, questionnaires except from diaries were filled by Computer Assisted Personal Interview. Then, diaries in which persons saved their daily activities were filled in the web environment.

3.4.3. Harmonization of Common Variables in LSS and TUS

Happiness is the target variable in Life Satisfaction Survey. Variables related with social and daily activities such as going to cinema, theatre, daily tours or nature walk, mall and cafe, watching television and solving puzzle are the variables of interest in Time Use Survey. There are eight common variables between Life Satisfaction and Time Use Surveys. Six of them belong to individual variables and two of them belong to household variables. Common variables between LSS and TUS:

- Individual Variables:
 - i. Sex
 - ii. Completed Age
 - iii. Marital Status
 - iv. Completed Level of Education
 - v. Activity Status
 - vi. Employment Status at Work

- Household Variables:
 - i. Number of Room
 - ii. Ownership Status of the House

Age, education level, activity status and employment status at work were needed to be recoded to provide harmonization between surveys. Life Satisfaction Survey covers population aged 18 and older while Time Use Survey covers individuals 10 years old and older. Their population totals differ because of that. To solve this problem, individuals who are 18 years old and older in TUS were filtered. Moreover, completed age variable were given as single age in LSS. It starts from 18 and goes to 98. However, in TUS, the situation is different. While ages between 10 and 19 were given as single age, ages older than 19 were given as five years age groups until the 80+. Therefore, age variable in LSS were recoded as five years age

groups in order to provide harmonization between them. Similarly; education level, activity status and employment status were recoded to harmonize variables. Categories and codes of harmonized common variables are given in Appendix G.

CHAPTER 4. RESULTS

4.1. CHOOSING THE MATCHING VARIABLES

Hellinger Distance was used to compare similarity between distributions of common variables. Results are given Table 4.1.1. below.

Table 4.1.1. Hellinger Distances of Common Variables

Common Variables	Hellinger Distance (%)
Sex	1.6
Age Group	1.8
Marital Status	2.5
Completed Education Level	2.1
Activity Status	5.3
Employment Status at Work	5.9
Number of Room	2.4
Ownership Status of the House	2.6

For the variables having Hellinger Distance less than 5%, it can be said that distributions of these variables are similar and so they might be used as matching variables between data sources. According to tables above, it can be seen that marginal distributions of sex, age, education level, marital status, number of room and ownership status of the house are similar since their Hellinger Distance values are less than 5%. However, activity status and employment status at work cannot be used as matching variables since their distributions differ between data sources according to Hellinger Distance values. This difference may arise because of the difference between reference periods of the surveys.

By comparing similarity of variables' distributions, common variables that will be used and not be used in matching are determined as explained above. After that matching variables should be determined according to their effect on target variable. Regression was applied on target variables using the common variables.

Target variables in TUS are going to cinema, theatre, daily tours or nature walk, mall and cafe, watching television, spending time on social media and solving puzzle. Cinema, theatre, watching television and spending time on social media have two categories which are "yes" and "no". There are two kinds of variables for daily tours or nature walk, mall, cafe and puzzle variables. First one includes just two categories "yes" and "no". Second one is a numeric variable including the number of the activity. At first, it was considered to use numeric variables but half of the records have zero value. Moreover, the ranges are wide therefore the frequency is low.

As a result, binary variables were used as target variables and logistic regression was done. Besides that, marital status was also recoded since there are small frequencies in "divorced" and "widow" categories. It was recoded as single or married according to current situation. Results of the regressions are given in Appendix H.-O.. According to these results, significant variables for response variables are given in Table 4.1.2.

Table 4.1.2. Significant Variables for Target Variables in TUS

	Sex	Age Group	Marital Status	Education Level	Number of Room	Ownership Status of the House
Cinema	✓	✓	✓	✓	✓	
Theatre	✓		✓	✓	✓	
Watching a Television	✓		✓	✓	✓	✓
Social Media	✓	✓	✓	✓	✓	✓
Daily Tours or Nature Walks				✓		
Solving a Puzzle	✓	✓		✓	✓	
Going to Cafe or Bar	✓	✓	✓	✓	✓	
Going to Mall	✓	✓	✓	✓	✓	

Sex, age, marital status, education level and number of room are the common variables found significant in most of the models. These variables can be used as matching variables. On the other hand, happiness is the target variable in LSS. It has likert scale, therefore; ordinal logistic regression applied for this variable. Result is given Table 4.1.3.

Table 4.1.3. Ordinal Logistic Regression Results on Happiness

Model							
Dependent Variable		Happiness (reference group=lowest)					
Independent Variables		Sex (ref. group= Male) Age Group (ref. group= 20-24) Marital Status (ref. group= Currently Single) Education Level (ref. group= No School Completed) Number of Room The Ownership Status of House (ref. group= Owner)					
Frequencies of Responses		1	2	3	4	5	
		645	3700	2456	698	227	
		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs 7726 Max deriv 6e-11		LR chi2	429.53	R2	0.059	C	0.606
		d.f.	22	g	0.503	Dxy	0.211
		Pr(> chi2)	<0.0001	gr	1.654	gamma	0.212
				gp	0.119	tau-a	0.138
				Brier	0.236		
		Coef	S. E.	Wald Z	Pr (> z)		
y>= 2		3.2374	0.1477	21.92	<0.0001		
y>= 3		0.5037	0.1418	3.55	0.0004		
y>= 4		-1.3075	0.1432	-9.13	<0.0001		
y>= 5		-2.8379	0.1541	-18.42	<0.0001		
Sex=2		-0.3031	0.0453	-6.70	<0.0001		
Age Group=25-29		0.3195	0.1038	3.08	0.0021		
Age Group=30-34		0.6467	0.1049	6.16	<0.0001		
Age Group=35-39		0.8992	0.1062	8.46	<0.0001		
Age Group=40-44		0.8888	0.1094	8.12	<0.0001		
Age Group=45-49		1.0641	0.1123	9.47	<0.0001		
Age Group=50-54		0.9588	0.1141	8.40	<0.0001		
Age Group=55-59		0.8984	0.1175	7.65	<0.0001		
Age Group=60-64		0.9222	0.1241	7.43	<0.0001		
Age Group=65-69		0.4944	0.1347	3.67	0.0002		
Age Group=70-74		0.3999	0.1530	2.61	0.0090		
Age Group=75-79		0.4213	0.1700	2.48	0.0132		
Age Group=80+		0.2428	0.1566	1.55	0.1212		
Marital Status=2		-0.8380	0.0587	-14.27	<0.0001		
Education Level=1		-0.1052	0.0701	-1.50	0.1337		
Education Level=2		-0.0916	0.0901	-1.02	0.3095		
Education Level=3		-0.1979	0.0868	-2.28	0.0225		
Education Level=4		-0.4508	0.0891	-5.06	<0.0001		
Number of Room		-0.1660	0.0285	-5.83	<0.0001		
The Ownership Status of House=2		0.3426	0.0533	6.43	<0.0001		
The Ownership Status of House=3		-0.0900	0.2376	-0.38	0.7049		
The Ownership Status of House=4		0.2232	0.0729	3.06	0.0022		

Sex, age, marital status and number of room are found significant in explaining variation in happiness. Therefore, these variables can be used in matching. Half of the education levels are significant so education level might also be used.

To sum up, matching variables are determined as sex, age, marital status and number of room. Target variables that are affected from these variables used in matching, others were not used. Then, since education level is found partially significant, it was added to matching variables in order to see the difference and results were compared. Besides that, effect of the number of room was wondered since it is just variable related with household not the individual. For this purpose, different matching variables were used in the matching and these are given in next section.

4.2. NON-PARAMETRIC MICRO STATISTICAL MATCHING

Nearest neighbour, random and rank hot deck methods were applied by not using sample weights and using sample weights. Their results were compared using Hellinger Distance. Four combinations of matching variables were used in implementation.

- i. X.mtc.1 : sex, age group and marital status
- ii. X.mtc.2 : sex, age group, marital status and education level
- iii. X.mtc.3 : sex, age group, marital status and number of room
- iv. X.mtc.4 : sex, age group, marital status, education level and number of room

4.2.1. Nearest Neighbour Distance Hot Deck

Unconstrained and constrained Nearest Neighbour Distance was applied using sex, age group and marital status as donation class with using and not using sample weights. Defining a donation class reduces the effort in computing distances. In unconstrained matching each record in donor data can be used as a donor more than once. On the contrary, when constrained option is used each record in donor data can be used as a donor only once. Results are given in Table 4.2.1.1 and Table 4.2.1.2..

Table 4.2.1.1. Nearest Neighbour Distance Hot Deck Results

Variables	Hellinger Distance (%)			
	Matching without using sample weights		Matching using sample weights	
	Compared not using sample weights	Compared using sample weights	Compared not using sample weights	Compared using sample weights
CAFE_BAR				
X.mtc.1	0.8	0.2	0.7	0.6
X.mtc.2	0.1	0.5	0.3	0.3
X.mtc.3	0.9	0.3	1.6	1.0
X.mtc.4	0.4	0.1	0.5	0.0
CINEMA				
X.mtc.1	0.4	0.1	0.1	0.7
X.mtc.2	0.2	0.5	0.2	0.5
X.mtc.3	0.6	0.3	0.8	0.4
X.mtc.4	0.6	0.8	0.1	0.2
SOCIAL MEDIA				
X.mtc.1	0.9	0.1	0.4	0.6
X.mtc.2	0.4	0.3	0.3	0.3
X.mtc.3	0.3	0.5	0.2	0.9
X.mtc.4	0.1	0.1	0.1	0.0
MALL				
X.mtc.1	0.2	0.2	0.0	0.5
X.mtc.2	0.4	0.1	0.1	0.4
X.mtc.3	0.1	0.1	0.3	0.3
X.mtc.4	0.4	0.3	0.2	0.2

According to Table 4.2.1.1., Hellinger Distances are less than 2% for all situations also most of them are less than 1%. Therefore, it can be said that distributions of the variables are similar between synthetic and donor data sets. There is no much difference between using and not using sample weights. When performances of matching variables are compared, it is hard to say that this one is better than the others since result changes according to variable.

Table 4.2.1.2. Nearest Neighbour Distance Constrained Hot Deck Results

Variables	Hellinger Distance (%)			
	Matching without using sample weights		Matching using sample weights	
	Compared not using sample weights	Compared using sample weights	Compared not using sample weights	Compared using sample weights
CAFE_BAR				
X.mtc.1	5.1	6.5	5.1	6.5
X.mtc.2	2.5	3.4	2.5	3.4
X.mtc.3	4.6	7.0	4.6	7.0
X.mtc.4	2.1	4.2	2.1	4.2
CINEMA				
X.mtc.1	4.1	4.9	4.1	4.9
X.mtc.2	3.0	2.8	3.0	2.8
X.mtc.3	4.0	7.0	4.0	7.0
X.mtc.4	3.1	4.2	3.1	4.2
SOCIAL MEDIA				
X.mtc.1	6.6	7.5	6.6	7.5
X.mtc.2	3.9	4.0	3.9	4.0
X.mtc.3	6.3	7.0	6.3	7.0
X.mtc.4	3.7	4.2	3.7	4.2
MALL				
X.mtc.1	6.2	6.7	6.2	6.7
X.mtc.2	3.4	3.5	3.4	3.5
X.mtc.3	5.7	6.1	5.7	6.1
X.mtc.4	3.2	3.3	3.2	3.3

According to Table 4.2.1.2., most of the Hellinger Distances are close or greater than 5%. Therefore, it can be said that distributions of the variables are not similar between synthetic and donor data sets. This means that constrained NND provides worse result according to unconstrained one. When performances of matching variables are compared, it can be said that 2nd and 4th matching variables provide better result than 1st and 3rd. This shows that using education level as matching variable provides more similarity on variables' distributions between data sources.

4.2.2. Random Hot Deck

Random Hot Deck was applied using sex, age group and marital status as donation class with using and not using sample weights and also using 'rot' and 'min' options. For the 'rot', the number of the closest donors to retain is given by $\sqrt{(Total\ number\ of\ available\ donors + 1)}$. In 'min', donors at the minimum distance from the recipient are retained. Results are given in Table 4.2.2.1 and Table 4.2.2.2..

Table 4.2.2.1. Results of Random Hot Deck Using "rot"

		Hellinger Distance (%)			
		Matching without using sample weights		Matching using sample weights	
Variables		Compared not using sample weights	Compared using sample weights	Compared not using sample weights	Compared using sample weights
CAFE_BAR					
X.mtc.1		2.9	4.2	2.6	3.7
X.mtc.2		0.9	1.4	0.7	1.5
X.mtc.3		1.5	2.8	1.5	2.6
X.mtc.4		1.1	1.7	1.5	2.2
CINEMA					
X.mtc.1		6.2	7.0	5.3	6.5
X.mtc.2		3.7	3.7	3.3	3.4
X.mtc.3		6.2	7.0	5.5	6.3
X.mtc.4		3.2	3.8	2.3	2.7
SOCIAL MEDIA					
X.mtc.1		4.9	6.0	4.7	5.7
X.mtc.2		3.7	4.0	2.8	3.2
X.mtc.3		4.9	5.8	3.7	4.8
X.mtc.4		3.5	4.1	2.9	3.1
MALL					
X.mtc.1		0.4	1.0	2.4	3.2
X.mtc.2		0.5	0.5	0.4	0.2
X.mtc.3		1.0	1.6	1.4	2.0
X.mtc.4		1.4	1.8	1.7	1.8

According to Table 4.2.2.1., Hellinger Distances of Cafe-Bar and Mall variables are less than 5% while Cinema and Social Media variables have HD values greater than or close to 5%. In this situation it can be said that Random Hot Deck with 'rot' option provide similar distribution for Cafe-Bar and Mall variables. Performances of matching variables differ according to variables, therefore; generalization cannot be done.

Table 4.2.2.2. Results of Random Hot Deck Using "min"

Variables	Hellinger Distance (%)			
	Matching without using sample weights		Matching using sample weights	
	Compared not using sample weights	Compared using sample weights	Compared not using sample weights	Compared using sample weights
CAFE_BAR				
X.mtc.1	0.7	0.5	1.4	0.8
X.mtc.2	0.9	0.6	0.1	0.2
X.mtc.3	0.3	0.6	0.9	0.4
X.mtc.4	0.7	0.3	0.1	0.3
CINEMA				
X.mtc.1	0.5	0.9	0.5	0.0
X.mtc.2	0.1	0.1	0.4	0.6
X.mtc.3	1.1	0.5	0.7	0.5
X.mtc.4	1.0	0.6	0.3	0.5
SOCIAL MEDIA				
X.mtc.1	0.3	0.7	1.6	0.9
X.mtc.2	0.4	0.4	0.6	0.8
X.mtc.3	0.4	0.5	0.8	0.3
X.mtc.4	0.2	0.1	0.1	0.2
MALL				
X.mtc.1	0.3	0.0	0.7	0.1
X.mtc.2	0.2	0.2	0.6	0.6
X.mtc.3	0.5	0.2	0.2	0.5
X.mtc.4	0.5	0.1	0.3	0.1

‘min’ option provides better matching result than ‘rot’ option since all HD values are less than 2% according to Table 4.2.2.2.. Similarly, performances of matching variables differ according to variables, therefore; generalization cannot be done.

4.2.3. Rank Hot Deck

Unconstrained and constrained Rank Hot Deck was applied using sex, age group and marital status as donation class with using and not using sample weights. Results are given in Table 4.2.3.1 and Table 4.2.3.2..

Table 4.2.3.1. Rank Hot Deck Results

Variables	Hellinger Distance (%)			
	Matching without using sample weights		Matching using sample weights	
	Compared not using sample weights	Compared using sample weights	Compared not using sample weights	Compared using sample weights
CAFE_BAR				
X.mtc.1	0.7	0.2	0.6	0.5
X.mtc.2	0.2	0.0	0.2	0.3
X.mtc.3	1.1	0.0	1.1	0.2
X.mtc.4	0.1	0.4	0.4	0.7
CINEMA				
X.mtc.1	0.4	0.1	0.8	0.0
X.mtc.2	0.5	0.4	0.5	0.7
X.mtc.3	0.7	0.3	0.2	0.9
X.mtc.4	0.2	0.2	0.2	0.1
SOCIAL MEDIA				
X.mtc.1	1.1	0.1	0.8	0.0
X.mtc.2	0.2	0.1	0.6	0.4
X.mtc.3	0.6	0.0	0.1	0.6
X.mtc.4	0.6	0.5	0.3	0.4
MALL				
X.mtc.1	0.1	0.0	0.6	0.5
X.mtc.2	0.8	0.5	1.0	0.8
X.mtc.3	0.3	0.2	0.4	0.8
X.mtc.4	0.1	0.4	0.9	1.2

It can be said that Rank Hot Deck provides a good result in matching since all HD values are less than 2% according to Table 4.2.3.1.. Generalization about the matching variables and using weights in matching cannot be done since it differs.

Table 4.2.3.2. Rank Constrained Hot Deck Results

	Hellinger Distance (%)			
Variables	Matching without using sample weights		Matching using sample weights	
	Compared not using sample weights	Compared using sample weights	Compared not using sample weights	Compared using sample weights
CAFE_BAR				
X.mtc.1	5.1	6.5	5.1	6.5
X.mtc.2	2.6	3.4	3.0	3.7
X.mtc.3	4.2	5.5	4.2	5.5
X.mtc.4	1.8	2.2	2.2	2.5
CINEMA				
X.mtc.1	4.1	4.9	4.1	4.9
X.mtc.2	3.0	2.7	3.1	2.9
X.mtc.3	3.8	4.5	3.8	4.6
X.mtc.4	2.4	2.5	2.8	2.7
SOCIAL MEDIA				
X.mtc.1	6.6	7.5	6.6	7.5
X.mtc.2	4.1	4.2	4.4	4.6
X.mtc.3	5.9	6.7	5.8	6.7
X.mtc.4	3.1	3.4	3.3	3.5
MALL				
X.mtc.1	6.2	6.7	6.2	6.7
X.mtc.2	3.6	3.7	4.0	4.0
X.mtc.3	5.2	5.5	5.1	5.3
X.mtc.4	2.2	2.4	2.6	2.7

Like in constrained NND, constrained Rank Hot Deck provides worse result according to unconstrained one since most of the Hellinger Distances are greater than or close to 5%. Therefore, it can be said that distributions of the variables are not similar between synthetic and donor data sets. When performances of matching variables are compared, it can be said that 2nd and 4th matching variables provide better result than 1st and 3rd. This shows that using education level as matching variable provides more similarity on variables' distributions between data sources.

CHAPTER 5. DISCUSSION

The main objective of this study is to compare different nonparametric statistical matching methods used in household based sample surveys. For this purpose, three hot deck method applied to create a synthetic data set includes happiness variable coming from 2014 Life Satisfaction Survey of Turkey and social activity variables coming from 2014-2015 Time Use Survey of Turkey.

First, common variables between harmonized between surveys and then their similarities compared using Hellinger Distance. It was found that sex, age group, marital status, completed education level, number of room and the ownership status of house are the variables having a similar distribution between two sources. On the other hand, activity status and employment status at work cannot be used for matching since their distributions differ between surveys according to HD. This result was expected according to previous studies conducted on these common variables (Webber and Tonkin, 2013; Serafino and Tonkin, 2017).

Second, matching variables was determined according to their power on explaining variations in target variables. According to regression results, it was found that the ownership status of house is not significant for almost all of the target variables. Education level found partially significant in explaining happiness. As a result, four different combination of matching variables were used in matching.

Then, nonparametric statistical matching methods namely nearest neighbour distance, random and rank hot deck were applied with using and not using sample weights. Nearest neighbour distance and rank hot deck methods applied in two ways such as unconstrained and constrained. Constrained hot deck is expected to give better result according to previous statistical matching studies. According to D'Orazio et al. (2019), when sex and age used as a donation class, constrained NND performs worse than others. However, the result differs when large geographic area is added to donation class (D'Orazio, 2017). In other words, constrained NND

performs better when geographic area included. In the study, sex age and marital status was used a donation class and it was found that constrained NND and constrained rank hot deck performs worse than unconstrained once. Unfortunately, effect of geographical area could not be examined since it is not included in the data sets. This variable should be asked to include in the data sets and examined for further researches.

In the literature, nearest neighbour is preferred instead of random hot deck since its results does not change. However, this situation changes when matching variables are categorical and nearest neighbour turns to random hot deck (Chen and Shao, 2000). When results of NND and random hot deck was compared, it was found that when 'min' option of random hot deck that provides retaining the donor having minimum distance is used, they provide similar result.

When results of three methods compared it can be said that all of them perform quite well in terms of preservation of the variables' distributions since most of the variables have Hellinger Distance less than 5%, except for constrained ones and random hot deck with 'rot' option. According to results, generalization on the basis of target variables and matching variables cannot be done. There are just two things can be said. First, using education as matching variable gives better result in constrained hot deck methods. Second, random hot deck with 'rot' option provides worse result on Cinema and Social Media variables.

Several naive procedures was compared by D'Orazio et al. (2012) and it was found that when weights used in rank and random hot deck they tend to perform well in terms of preservation of the distributions. On the other hand, according to previous study conducted by Linskens (2015), it was found that the use of survey weights in matching does not have considerable effect even when matching applied not using sample weights provides more accurate result both for distance and random hot deck. In this study, when usage of the weights compared using distance , random and rank hot deck it was found that using weights in matching does not have a considerable

effect on the results. It provides better result for some variables and worse for some but differences are small.

CHAPTER 6. CONCLUSION

For many years, researchers have been working on data integration methods to benefit from existing data sources. Using different data sources in estimation has a significant effect on increasing the accuracy of statistical information and reducing the response burden, time, cost and labour.

Statistical matching is a method used in data integration especially when there is no unique variable in data sources for matching records. Common variables between data sources are used as matching variables and fused data sets are obtained using different matching approaches such as parametric, non-parametric and mixed.

Life Satisfaction and Time Use Surveys, including variables related with social activity, were used in the study. Three non-parametric hot deck methods, named as nearest neighbour distance, random and rank hot deck, were used with their unconstrained and constrained version in the matching.

It was found that results vary according to both matching variables and target variables. In general, all of the methods performed well in terms of preservation of the distribution of the variables. Constrained nearest neighbour distance and rank hot deck did not provide better result. Random hot deck performed better when ‘min’ option used and provide similar result with nearest neighbour hot deck.

To conclude, this thesis showed that the application of SM for variables of social nature was challenging, yet further applications on different variables and surveys could enhance its practice. It was seen that every variable had its own specific challenges. As suggested by Ballin et al. (2009), “statistical matching, such as any small piece of applied statistics, is more than a collection of tools and technical solutions to be applied following specific guidelines: it is a practice which requires a deep understanding of the data, of the way they are collected, of subject related issues. More an art than a science...”.

REFERENCES

- Ahi, L. (2015). Veri Madenciliği Yöntemleri İle Ana Harcama Gruplarının Paylarının Tahmini.
- Alpman, A., Gardes, F., & Thiombiano, N. (2017). Statistical Matching for Combining Time-Use Surveys with Consumer Expenditure Surveys: An Evaluation on Real Data.
- 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.
- Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of official statistics*, 16(2), 113.
- Conti, P., L., Marella, D. & Neri, A. (2015). *Statistical matching and uncertainty analysis in combining household income and expenditure data*, Banca D'Italia, TD No. 1018
- De Waal, T. (2015). *Statistical matching: experimental results and future research questions*. Statistics Netherlands.
- Denk, M. and Hackl, P., *Data Integration: Techniques and Evaluation*, Austrian Journal of Statistics, 1&2:135-152, 2004.
- D'Orazio, M., *Introduction to Statistical Matching*, Statistical Matching: Methodological issues and practice with R-StatMatch, 21-22, 2013.
- D'Orazio, M. (2017). Statistical Matching and Imputation of Survey Data with StatMatch.
- D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons.
- D'Orazio, M., Di Zio, M., & Scanu, M. (2012, May). Statistical matching of data from complex sample surveys. In *Proceedings of the European Conference on Quality in Official Statistics-Q2012* (Vol. 29).
- 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*.
- Gavin, N. I. (1985). An application of statistical matching with the survey of income and education and the 1976 Health Interview Survey. *Health services research*, 20(2), 183.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.
- Kish, L. (1999). Cumulating/combining population surveys. *Survey Methodology*, 25(2), 129-138.
- Leulescu, A. and Agafiței, M., (2013) *Statistical matching: a model based approach for data integration*, EUROSTAT, European Union.
- Linskens, S.J. (2015), *Statistical Matching: A Comparison of Random and Distance Hot Deck*. Report, Tilburg University, The Netherlands.

- Marella, D. and Conti, P. L., (2016), *ESTP Course on Statistical Matching and Record Linkage*. Retrieved from https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2016%20ESTP%20PROGRAMME/36.%20Statistical%20matching%20and%20record%20linkage%2c%2019%20E2%80%93%2021%20September%202016%20-%20Organiser_%20DEVSTAT/1_record_linkage.pdf
- Moriarity, C., & Scheuren, F. (2003). A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 21(1), 65-73.
- Okner, B. (1972). Constructing a new data base from existing microdata sets: the 1966 merge file. In *Annals of Economic and Social Measurement, Volume 1, number 3* (pp. 325-362). NBER.
- Rasler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer.
- Rodgers, W. L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2(1), 91-102.
- Roszka W., *Some Practical Issues Related to the Integration of Data from Sample Surveys*, Statistika, 95:60-75, 2015.
- Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1), 87-94.
- Ruggles, N., & Ruggles, R. (1974). A strategy for merging and matching microdata sets. In *Annals of Economic and Social Measurement, Volume 3, number 2* (pp. 353-371). NBER.
- Schenker, N., Gentleman, J. F., Rose, D., Hing, E., & Shimizu, I. M. (2002). Combining estimates from complementary surveys: a case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*.
- Schenker, N., & Raghunathan, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in medicine*, 26(8), 1802-1811.
- Serafino, P. and Tonkin, R. (2017), *Statistical matching of European Union statistics on income and living conditions (EU-SILC) and household budget survey*, Statistical working papers, Eurostat, European Union, Luxembourg.
- 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(1), 59-79.
- Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Retrieved from https://www.economie.gouv.fr/files/finances/presse/dossiers_de_presse/090914mesure_perf_eco_progres_social/synthese_ang.pdf
- Turkish Statistical Institute, (n.d.-a), Life Satisfaction Survey Micro Data Set, 2014. Retrieved from http://www.tuik.gov.tr/MicroVeri/YMA_2014/index.html

- Turkish Statistical Institute, (n.d.-b), Time Use Survey Micro Data Set, 2014-2015. Retrieved from http://www.tuik.gov.tr/MicroVeri/ZKA_2014/english/index.html
- Van der Laan, P. (2000). Integrating administrative registers and household surveys. *Netherlands Official Statistics*, 15(2), 7-15.
- Van der Putten, P., Kok, J. N., & Gupta, A. (2002). Data fusion through statistical matching.
- Webber, D. and 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.
- Wolff, E. N. (1980). Estimates of the 1969 size distribution of household wealth in the US from a synthetic data base. In *Modeling the distribution and intergenerational transmission of wealth* (pp. 223-272). University of Chicago Press.
- Yoshizoe, Y. and M. Araki (1999). *Statistical Matching of Household Survey Files*.

APPENDIX

Appendix A. NUTS Code List

CITY	CITY CODE	NUTS1	NUTS2	NUTS3
İSTANBUL	34	TR1	TR10	TR100
EDİRNE	22	TR2	TR21	TR211
KIRKLARELİ	39	TR2	TR21	TR212
TEKİRDAĞ	59	TR2	TR21	TR213
BALIKESİR	10	TR2	TR22	TR221
ÇANAKKALE	17	TR2	TR22	TR222
İZMİR	35	TR3	TR31	TR310
AYDIN	9	TR3	TR32	TR321
DENİZLİ	20	TR3	TR32	TR322
MUĞLA	48	TR3	TR32	TR323
AFYON	3	TR3	TR33	TR331
KÜTAHYA	43	TR3	TR33	TR332
MANİSA	45	TR3	TR33	TR333
UŞAK	64	TR3	TR33	TR334
BİLECİK	11	TR4	TR41	TR411
BURSA	16	TR4	TR41	TR412
ESKİŞEHİR	26	TR4	TR41	TR413
BOLU	14	TR4	TR42	TR421
KOCAELİ	41	TR4	TR42	TR422
SAKARYA	54	TR4	TR42	TR423
YALOVA	77	TR4	TR42	TR424
DÜZCE	81	TR4	TR42	TR425
ANKARA	6	TR5	TR51	TR510
KONYA	42	TR5	TR52	TR521
KARAMAN	70	TR5	TR52	TR522
ANTALYA	7	TR6	TR61	TR611
BURDUR	15	TR6	TR61	TR612
ISPARTA	32	TR6	TR61	TR613
ADANA	1	TR6	TR62	TR621
İÇEL	33	TR6	TR62	TR622
HATAY	31	TR6	TR63	TR631
KAHRAMANMARAŞ	46	TR6	TR63	TR632
OSMANİYE	80	TR6	TR63	TR633
KIRŞEHİR	40	TR7	TR71	TR711
NEVŞEHİR	50	TR7	TR71	TR712
NİĞDE	51	TR7	TR71	TR713
AKSARAY	68	TR7	TR71	TR714

Appendix A. (continued)

KIRIKKALE	71	TR7	TR71	TR715
KAYSERİ	38	TR7	TR72	TR721
SİVAS	58	TR7	TR72	TR722
YOZGAT	66	TR7	TR72	TR723
ZONGULDAK	67	TR8	TR81	TR811
BARTIN	74	TR8	TR81	TR812
KARABÜK	78	TR8	TR81	TR813
ÇANKIRI	18	TR8	TR82	TR821
KASTAMONU	37	TR8	TR82	TR822
SİNOP	57	TR8	TR82	TR823
AMASYA	5	TR8	TR83	TR831
ÇORUM	19	TR8	TR83	TR832
SAMSUN	55	TR8	TR83	TR833
TOKAT	60	TR8	TR83	TR834
ARTVİN	8	TR9	TR90	TR901
GİRESUN	28	TR9	TR90	TR902
GÜMÜŞHANE	29	TR9	TR90	TR903
ORDU	52	TR9	TR90	TR904
RİZE	53	TR9	TR90	TR905
TRABZON	61	TR9	TR90	TR906
ERZİNCAN	24	TRA	TRA1	TRA11
ERZURUM	25	TRA	TRA1	TRA12
BAYBURT	69	TRA	TRA1	TRA13
AĞRI	4	TRA	TRA2	TRA21
KARS	36	TRA	TRA2	TRA22
ARDAHAN	75	TRA	TRA2	TRA23
İĞDIR	76	TRA	TRA2	TRA24
BİNGÖL	12	TRB	TRB1	TRB11
ELAZIĞ	23	TRB	TRB1	TRB12
MALATYA	44	TRB	TRB1	TRB13
TUNCELİ	62	TRB	TRB1	TRB14
BİTLİS	13	TRB	TRB2	TRB21
HAKKARİ	30	TRB	TRB2	TRB22
MUŞ	49	TRB	TRB2	TRB23
VAN	65	TRB	TRB2	TRB24
ADIYAMAN	2	TRC	TRC1	TRC11
GAZİANTEP	27	TRC	TRC1	TRC12
KİLİS	79	TRC	TRC1	TRC13
DİYARBAKIR	21	TRC	TRC2	TRC21
ŞANLIURFA	63	TRC	TRC2	TRC22

Appendix A. (continued)

MARDİN	47	TRC	TRC3	TRC31
SİİRT	56	TRC	TRC3	TRC32
BATMAN	72	TRC	TRC3	TRC33
ŞIRNAK	73	TRC	TRC3	TRC34

Appendix B. ISCED-97 Code List

Code	Explanation
0	Pre-primary level of education
1	Primary level of education
2	Lower secondary level of education
3	Upper secondary level of education
4	Post-secondary, non-tertiary education
5	First stage of tertiary education
6	Second stage of tertiary education

Appendix C. ICSE-93 Code List

Code	Explanation
1	Employees
2	Employers
3	Own-account workers
4	Members of producers' cooperatives
5	Contributing family workers
6	Workers not classifiable by status

Appendix D. HETUS Activity Coding List

Main and secondary activities	
0	PERSONAL CARE
01	SLEEP
011	Sleep
012	Sick in bed
02	EATING
021	Eating
03	OTHER PERSONAL CARE
031	Washing and dressing
039	Other or unspecified personal care
1	EMPLOYMENT
11	MAIN JOB AND SECOND JOB
111	Working time in main and second job (including coffee breaks and travel at work)
12	ACTIVITIES RELATED TO EMPLOYMENT
121	Lunch break
129	Other or unspecified activities related to employment
2	STUDY
20	UNSPECIFIED STUDY
200	Unspecified study
21	SCHOOL OR UNIVERSITY
211	Classes and lectures
212	Homework
22	FREE TIME STUDY
221	Free time study
3	HOUSEHOLD AND FAMILY CARE
30	UNSPECIFIED HOUSEHOLD AND FAMILY CARE
300	Unspecified household and family care
31	FOOD MANAGEMENT
311	Food preparation, baking and preserving
312	Dish washing
32	HOUSEHOLD UPKEEP
321	Cleaning dwelling
322	Cleaning garden
323	Heating and water
324	Arranging household goods and materials
329	Other or unspecified household upkeep
33	MAKING AND CARE FOR TEXTILES
331	Laundry
332	Ironing
333	Handicraft and producing textiles

Appendix D. (continued)

339	Other or unspecified making of and care for textiles
34	GARDENING AND PET CARE
341	Gardening
342	Tending domestic animals
343	Caring for pets
344	Walking the dog
349	Other or unspecified gardening and pet care
35	CONSTRUCTION AND REPAIRS
351	House construction and renovation
352	Repairs to dwelling
353	Making, repairing and maintaining equipment
354	Vehicle maintenance
359	Other or unspecified construction and repairs
36	SHOPPING AND SERVICES
361	Shopping
362	Commercial and administrative services
363	Personal services
369	Other or unspecified shopping and services
37	HOUSEHOLD MANAGEMENT
371	Household Management
38	CHILDCARE
381	Physical care and supervision
382	Teaching the child
383	Reading, playing and talking with child
384	Accompanying child
389	Other or unspecified childcare
39	"HELP TO AN ADULT FAMILY MEMBER
(Codes at three digit level, 391, 392 and 399, are voluntary)"	
391	Physical care of a dependent adult household member
392	Other help of a dependent adult household member
399	Help to a non dependent adult household member
4	VOLUNTARY WORK AND MEETINGS
41	ORGANISATIONAL WORK
411	Organisational work (work for or through an organisation)
42	INFORMAL HELP TO OTHER HOUSEHOLDS
421	Construction and repairs as help
422	Help in employment and farming
423	Care of own children living in another household
424	Other childcare as help to another household
425	Help to an adult of another household

Appendix D. (continued)

429	Other or unspecified informal help to another household
43	PARTICIPATORY ACTIVITIES
431	Meetings
432	Religious activities
439	Other or unspecified participatory activities
5	SOCIAL LIFE AND ENTERTAINMENT
51	SOCIAL LIFE
511	Socialising with family
512	Visiting and receiving visitors
513	Celebrations
514	Telephone conversation
519	Other or unspecified social life
52	ENTERTAINMENT AND CULTURE
521	Cinema
522	Theatre and concerts
523	Art exhibitions and museums
524	Libraries
525	Sports events
529	Other or unspecified entertainment and culture
53	RESTING — TIME OUT
531	Resting — Time out
6	SPORTS AND OUTDOOR ACTIVITIES
61	PHYSICAL EXERCISE
611	Walking and hiking
612	Jogging and running
613	Cycling, skiing and skating
614	Ball games
615	Gymnastics and fitness
616	Water sports
619	Other or unspecified sports or outdoor activities
62	PRODUCTIVE EXERCISE
621	Productive exercise (e.g. hunting, fishing, picking berries, mushrooms or herbs)
63	SPORTS RELATED ACTIVITIES
631	Sports related activities
7	HOBBIES AND COMPUTING
71	ARTS AND HOBBIES
711	Arts (visual, performing, literary)
712	Collecting
713	Correspondence
719	Other or unspecified hobbies

Appendix D. (continued)

72	COMPUTING
721	Computing - programming
722	Information by computing
723	Communication by computing
729	Other or unspecified computing
73	GAMES
731	Solo games and play, gambling
732	Parlour games and play
733	Computer games
739	Other or unspecified games
8	MASS MEDIA
81	READING
811	Reading periodicals
812	Reading books
819	Other or unspecified reading
82	TV, VIDEO AND DVD
821	Watching TV, video or DVD
83	RADIO AND RECORDINGS
831	Listening to radio or recordings
9	TRAVEL AND UNSPECIFIED TIME USE
	TRAVEL BY PURPOSE
910	Travel to/from work
920	Travel related to study
936	Travel related to shopping and services
938	Travel related to childcare
939	Travel related to other household care
940	Travel related to voluntary work and meetings
950	Travel related to social life
960	Travel related to their leisure
980	Travel related to changing locality
900	Other or unspecified travel purpose
	AUXILIARY CODES
995	Filling in the time use diary
998	Unspecified leisure time
999	Other or unspecified time use

Appendix E. NACE Rev. 2 Code List

Code	Explanation	Divisions
A	Agriculture, forestry and fishing	01 – 03
B	Mining and quarrying	05 – 09
C	Manufacturing	10 – 33
D	Electricity, gas, steam and air conditioning supply	35
E	Water supply; sewerage, waste management and remediation activities	36 – 39
F	Construction	41 – 43
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	45 – 47
H	Transportation and storage	49 – 53
I	Accommodation and food service activities	55 – 56
J	Information and communication	58 – 63
K	Financial and insurance activities	64 – 66
L	Real estate activities	68
M	Professional, scientific and technical activities	69 – 75
N	Administrative and support service activities	77 – 82
O	Public administration and defence; compulsory social security	84
P	Education	85
Q	Human health and social work activities	86 – 88
R	Arts, entertainment and recreation	90 – 93
S	Other service activities	94 – 96
T	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	97 – 98
U	Activities of extraterritorial organisations and bodies	99

Appendix F. ISCO-08 Code List

Codes	Explanations
1	Managers
2	Professionals
3	Technicians and Associate Professionals
4	Clerical Support Workers
5	Services and Sales Workers
6	Skilled Agricultural, Forestry and Fishery Workers
7	Craft and Related Trades Workers
8	Plant and Machine Operators and Assemblers
9	Elementary Occupations
0	Armed Forces Occupations

Appendix G. Harmonized Common Variables

Common Variables	Codes:
Sex	1: Male 2: Female
Age Group	20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79 80+
Marital Status	1: Currently Single 2: Currently Married
Completed Education Level	0: No school completed 1: Primary school 2: Primary education /General lower secondary school/Vocational and technical junior secondary school 3: General high school/Vocational and technical high school 4: Post secondary of 2 or 3 years/Faculty of 4 years/ Master / Doctorate
Activity Status	1: Worked 2: Did not worked but interest continues with work 3: Did not worked
Employment Status at Work	0: Missing 1: Waged or salaried/Casual (people working at seasonal or daily work) 2: Employer 3: Self employed 4: Unpaid family workers
Number of Room	1, 2, ...
The Ownership Status of House	1: Owner 2: Renter 3: Housing 4: Not owner but not paying rent

Appendix H. Logistic Regression Results of Cinema

Model					
Dependent Variable	Cinema				
Independent Variables	Sex Age Group Marital Status Education Level Number of Room The Ownership Status of House				
Deviance Residuals	Min	1Q	Median	3Q	Max
		-3.6662	0.0866	0.1896	0.4346
Coefficients	Estimate	Std. Error	Z value	Pr (> z)	
(Intercept)	6.09067	0.47468	12.831	< 2e-16	***
Sex 2	-0.14454	0.05696	-2.537	0.011169	*
Age Group 25-29	-0.03005	0.08922	-0.337	0.736238	
Age Group 30-34	0.40596	0.10358	3.919	8.89e-05	***
Age Group 35-39	0.35522	0.11117	3.195	0.001397	**
Age Group 40-44	0.40215	0.12072	3.331	0.000865	***
Age Group 45-49	0.57865	0.13313	4.347	1.38e-05	***
Age Group 50-54	0.75061	0.14581	5.148	2.63e-07	***
Age Group 55-59	1.05698	0.17706	5.970	2.38e-09	***
Age Group 60-64	1.48162	0.23894	6.201	5.62e-10	***
Age Group 65-69	1.35148	0.28786	4.695	2.67e-06	***
Age Group 70-74	0.96750	0.32583	2.969	0.002984	**
Age Group 75-79	14.02858	162.36786	0.086	0.931148	
Age Group 80+	2.26398	0.72219	3.135	0.001719	**
Marital Status 2	0.80791	0.06927	11.663	< 2e-16	***
Education Level 1	-2.30223	0.45893	-5.017	5.26e-07	***
Education Level 2	-3.11890	0.45913	-6.793	1.10e-11	***
Education Level 3	-4.19864	0.45370	-9.254	< 2e-16	***
Education Level 4	-4.67039	0.45341	-10.301	< 2e-16	***
Number of Room	-0.26580	0.03597	-7.390	1.47e-13	***
The Ownership Status of House 2	-0.07566	0.06538	-1.157	0.247140	
The Ownership Status of House 3	-0.13491	0.22641	-0.596	0.551269	
The Ownership Status of House 4	0.06104	0.10171	0.600	0.548387	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 11280 on 20157 degrees of freedom					
Residual deviance: 8599 on 20135 degrees of freedom					
AIC: 8645					
Number of Fisher Scoring iterations: 16					

Appendix I. Logistic Regression Results of Theatre

Model					
Dependent Variable	Theatre				
Independent Variables	Sex Age Group Marital Status Education Level Number of Room The Ownership Status of House				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-3.6560	0.0674	0.0992	0.2009	0.5745
Coefficients	Estimate	Std. Error	Z value	Pr (> z)	
(Intercept)	7.17079	0.66621	10.763	< 2e-16	***
Sex 2	-0.31200	0.11961	-2.609	0.00909	**
Age Group 25-29	-0.16411	0.20797	-0.789	0.43006	
Age Group 30-34	-0.20478	0.22452	-0.912	0.36172	
Age Group 35-39	-0.31340	0.23608	-1.327	0.18435	
Age Group 40-44	0.10352	0.27853	0.372	0.71015	
Age Group 45-49	-0.12649	0.27637	-0.458	0.64718	
Age Group 50-54	-0.18903	0.28638	-0.660	0.50920	
Age Group 55-59	-0.31583	0.30005	-1.053	0.29253	
Age Group 60-64	0.52260	0.45208	1.156	0.24768	
Age Group 65-69	1.83605	1.02041	1.799	0.07197	
Age Group 70-74	0.55072	0.73852	0.746	0.45584	
Age Group 75-79	13.93399	462.13439	0.030	0.97595	
Age Group 80+	13.96396	441.02597	0.032	0.97474	
Marital Status 2	0.47439	0.14510	3.269	0.00108	**
Education Level 1	-1.17698	0.61186	-1.924	0.05440	.
Education Level 2	-1.84148	0.62611	-2.941	0.00327	**
Education Level 3	-2.85360	0.59667	-4.783	1.73e-06	***
Education Level 4	-3.68196	0.59260	-6.213	5.19e-10	***
Number of Room	-0.16185	0.07487	-2.162	0.03064	*
The Ownership Status of House 2	-0.04305	0.13957	-0.308	0.75774	
The Ownership Status of House 3	0.19367	0.51735	0.374	0.70814	
The Ownership Status of House 4	-0.22126	0.19986	-1.107	0.26826	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 3103.3 on 20157 degrees of freedom					
Residual deviance: 2675.8 on 20135 degrees of freedom					
AIC: 2721.8					
Number of Fisher Scoring iterations: 18					

Appendix J. Logistic Regression Results of Television

Model					
Dependent Variable	Television				
Independent Variables	Sex Age Group Marital Status Education Level Number of Room The Ownership Status of House				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-0.9559	-0.3305	-0.2685	-0.2280	2.9732
Coefficients	Estimate	Std. Error	Z value	Pr (> z)	
(Intercept)	-1.55873	0.19636	-7.938	2.05e-15	***
Sex 2	0.29192	0.07185	4.063	4.85e-05	***
Age Group 25-29	0.12841	0.14223	0.903	0.366602	
Age Group 30-34	-0.08615	0.15239	-0.565	0.571860	
Age Group 35-39	-0.14544	0.15881	-0.916	0.359756	
Age Group 40-44	-0.09030	0.16013	-0.564	0.572787	
Age Group 45-49	-0.41598	0.17725	-2.347	0.018937	*
Age Group 50-54	-0.34921	0.17328	-2.015	0.043875	*
Age Group 55-59	-0.64140	0.19468	-3.295	0.000985	***
Age Group 60-64	-0.62590	0.20274	-3.087	0.002020	**
Age Group 65-69	-0.10288	0.18684	-0.551	0.581878	
Age Group 70-74	0.40759	0.17934	2.273	0.023045	*
Age Group 75-79	0.47165	0.18817	2.506	0.012194	*
Age Group 80+	0.86037	0.16811	5.118	3.09e-07	***
Marital Status 2	-0.41726	0.07916	-5.271	1.36e-07	***
Education Level 1	-0.78105	0.09600	-8.136	4.10e-16	***
Education Level 2	-0.68213	0.13267	-5.142	2.72e-07	***
Education Level 3	-0.70194	0.12165	-5.770	7.92e-09	***
Education Level 4	-0.68489	0.12799	-5.351	8.74e-08	***
Number of Room	-0.17687	0.04114	-4.299	1.71e-05	***
The Ownership Status of House 2	-0.17029	0.08969	-1.899	0.057625	.
The Ownership Status of House 3	0.56782	0.28512	1.992	0.046422	*
The Ownership Status of House 4	-0.23959	0.11690	-2.050	0.040411	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 7850.1 on 20157 degrees of freedom					
Residual deviance: 7384.2 on 20135 degrees of freedom					
AIC: 7430.2					
Number of Fisher Scoring iterations: 6					

Appendix K. Logistic Regression Results of Social Media

Model					
Dependent Variable	Social Media				
Independent Variables	Sex Age Group Marital Status Education Level Number of Room The Ownership Status of House				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-3.2060	-0.6944	0.2552	0.6234	2.2629
Coefficients	Estimate	Std. Error	Z value	Pr (> z)	
(Intercept)	3.3675	0.19940	16.889	< 2e-16 ***	
Sex 2	0.55299	0.03963	13.955	< 2e-16 ***	
Age Group 25-29	-0.00786	0.07662	-0.103	0.91829	
Age Group 30-34	0.01588	0.07878	0.202	0.84022	
Age Group 35-39	0.25001	0.08096	3.088	0.00201 **	
Age Group 40-44	0.49748	0.08413	5.913	3.36e-09 ***	
Age Group 45-49	0.83181	0.08947	9.297	< 2e-16 ***	
Age Group 50-54	1.36536	0.09661	14.132	< 2e-16 ***	
Age Group 55-59	1.77887	0.11253	15.808	< 2e-16 ***	
Age Group 60-64	2.03011	0.13407	15.142	< 2e-16 ***	
Age Group 65-69	2.73775	0.20292	13.492	< 2e-16 ***	
Age Group 70-74	2.81656	0.28042	10.044	< 2e-16 ***	
Age Group 75-79	3.38380	0.46711	7.244	4.35e-13 ***	
Age Group 80+	4.31321	0.71831	6.005	1.92e-09 ***	
Marital Status 2	0.40242	0.05230	7.695	1.42e-14 ***	
Education Level 1	-2.04479	0.17389	-11.759	< 2e-16 ***	
Education Level 2	-2.90812	0.17566	-16.555	< 2e-16 ***	
Education Level 3	-3.71374	0.17359	-21.393	< 2e-16 ***	
Education Level 4	-4.33839	0.17517	-24.767	< 2e-16 ***	
Number of Room	-0.18865	0.02557	-7.377	1.62e-13 ***	
The Ownership Status of House 2	-0.38750	0.04585	-8.451	< 2e-16 ***	
The Ownership Status of House 3	-0.03263	0.17133	-0.190	0.84893	
The Ownership Status of House 4	-0.27173	0.06203	-4.381	1.18e-05 ***	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 24381 on 20157 degrees of freedom					
Residual deviance: 16298 on 20135 degrees of freedom					
AIC: 16344					
Number of Fisher Scoring iterations: 8					

Appendix L. Logistic Regression Results of Daily Tours or Nature Walks

Model					
Dependent Variable	Daily Tours or Nature Walks				
Independent Variables	Sex Age Group Marital Status Education Level Number of Room The Ownership Status of House				
Deviance Residuals	Min	1Q	Median	3Q	Max
		-0.5515	-0.2390	-0.1870	-0.1508
Coefficients	Estimate	Std. Error	Z value	Pr (> z)	
(Intercept)	-5.53034	0.37012	-14.942	< 2e-16	***
Sex 2	0.01784	0.09535	0.187	0.851574	
Age Group 25-29	0.43406	0.20974	2.070	0.038494	*
Age Group 30-34	0.38236	0.22040	1.735	0.082771	.
Age Group 35-39	0.48476	0.22610	2.144	0.032033	*
Age Group 40-44	0.44191	0.23681	1.866	0.062030	.
Age Group 45-49	0.31510	0.25253	1.248	0.212120	
Age Group 50-54	0.77306	0.23722	3.259	0.001119	**
Age Group 55-59	0.95185	0.24193	3.934	8.34e-05	***
Age Group 60-64	0.94066	0.25932	3.627	0.000286	***
Age Group 65-69	1.15230	0.27736	4.155	3.26e-05	***
Age Group 70-74	0.31552	0.42715	0.739	0.460122	
Age Group 75-79	-0.02469	0.61316	-0.040	0.967878	
Age Group 80+	0.45699	0.49192	0.929	0.352891	
Marital Status 2	-0.19189	0.11971	-1.603	0.108935	
Education Level 1	1.19591	0.26134	4.576	4.74e-06	***
Education Level 2	1.33728	0.29732	4.498	6.87e-06	***
Education Level 3	1.96892	0.27115	7.261	3.83e-13	***
Education Level 4	2.61310	0.26630	9.813	< 2e-16	***
Number of Room	-0.06275	0.06173	-1.017	0.309390	
The Ownership Status of House 2	0.19189	0.10984	1.747	0.080624	.
The Ownership Status of House 3	0.06605	0.39498	0.167	0.867197	
The Ownership Status of House 4	-0.29749	0.18592	-1.600	0.109584	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 4536.5 on 20157 degrees of freedom					
Residual deviance: 4272.5 on 20135 degrees of freedom					
AIC: 4318.5					
Number of Fisher Scoring iterations: 7					

Appendix M. Logistic Regression Results of Solving a Puzzle

Model					
Dependent Variable	Solving a Puzzle				
Independent Variables	Sex Age Group Marital Status Education Level Number of Room The Ownership Status of House				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-1.1611	-0.5997	-0.4240	-0.1026	3.4912
Coefficients	Estimate	Std. Error	Z value	Pr (> z)	
(Intercept)	-6.26678	0.28989	-21.618	< 2e-16	***
Sex 2	-0.23797	0.04427	-5.376	7.62e-08	***
Age Group 25-29	0.22388	0.09841	2.275	0.02291	*
Age Group 30-34	0.41720	0.10017	4.165	3.11e-05	***
Age Group 35-39	0.60413	0.10196	5.925	3.12e-09	***
Age Group 40-44	0.81621	0.10365	7.875	3.42e-15	***
Age Group 45-49	0.79949	0.10770	7.423	1.14e-13	***
Age Group 50-54	0.75711	0.11041	6.857	7.01e-12	***
Age Group 55-59	0.90308	0.11550	7.819	5.32e-15	***
Age Group 60-64	1.13395	0.12189	9.303	< 2e-16	***
Age Group 65-69	1.22775	0.13885	8.842	< 2e-16	***
Age Group 70-74	1.00756	0.17609	5.722	1.05e-08	***
Age Group 75-79	0.60259	0.25253	2.386	0.01702	*
Age Group 80+	0.27761	0.28665	0.968	0.33281	
Marital Status 2	-0.01415	0.05925	-0.239	0.81122	
Education Level 1	2.97064	0.26376	11.263	< 2e-16	***
Education Level 2	3.85671	0.26808	14.386	< 2e-16	***
Education Level 3	4.36864	0.26529	16.467	< 2e-16	***
Education Level 4	4.56123	0.26571	17.166	< 2e-16	***
Number of Room	0.06840	0.02797	2.446	0.01445	*
The Ownership Status of House 2	0.15339	0.05217	2.940	0.00328	**
The Ownership Status of House 3	-0.03006	0.19326	-0.156	0.87639	
The Ownership Status of House 4	0.03201	0.07383	0.434	0.66456	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 16037 on 20157 degrees of freedom					
Residual deviance: 14067 on 20135 degrees of freedom					
AIC: 14113					
Number of Fisher Scoring iterations: 7					

Appendix N. Logistic Regression Results of Going to Cafe or Bar

Model					
Dependent Variable	Cafe_Bar				
Independent Variables	Sex Age Group Marital Status Education Level Number of Room The Ownership Status of House				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-1.7884	-0.8236	-0.5214	0.9418	2.6195
Coefficients	Estimate	Std. Error	Z value	Pr (> z)	
(Intercept)	-1.79669	0.12248	-14.670	< 2e-16	***
Sex 2	-0.99711	0.03543	-28.143	< 2e-16	***
Age Group 25-29	-0.10687	0.07330	-1.458	0.144854	
Age Group 30-34	-0.17501	0.07527	-2.325	0.020075	*
Age Group 35-39	-0.30175	0.07793	-3.872	0.000108	***
Age Group 40-44	-0.32716	0.08040	-4.069	4.72e-05	***
Age Group 45-49	-0.22343	0.08278	-2.699	0.006956	**
Age Group 50-54	-0.20290	0.08381	-2.421	0.015475	*
Age Group 55-59	-0.16853	0.08846	-1.905	0.056767	.
Age Group 60-64	-0.18554	0.09546	-1.944	0.051942	.
Age Group 65-69	-0.14935	0.10797	-1.383	0.166600	
Age Group 70-74	-0.51632	0.13536	-3.814	0.000136	***
Age Group 75-79	-0.54817	0.16220	-3.380	0.000726	***
Age Group 80+	-0.89537	0.17361	-5.157	2.50e-07	***
Marital Status 2	-0.45009	0.04573	-9.843	< 2e-16	***
Education Level 1	1.25669	0.08299	15.143	< 2e-16	***
Education Level 2	1.44586	0.09255	15.622	< 2e-16	***
Education Level 3	2.08187	0.08760	23.767	< 2e-16	***
Education Level 4	2.68000	0.08907	30.089	< 2e-16	***
Number of Room	0.09052	0.02186	4.141	3.46e-05	***
The Ownership Status of House 2	0.12816	0.04200	3.051	0.002277	**
The Ownership Status of House 3	-0.19015	0.16785	-1.133	0.257286	
The Ownership Status of House 4	0.02645	0.05717	0.463	0.643631	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 24519 on 20157 degrees of freedom					
Residual deviance: 20635 on 20135 degrees of freedom					
AIC: 20681					
Number of Fisher Scoring iterations: 5					

Appendix O. Logistic Regression Results of Going to Mall

Model					
Dependent Variable	Mall				
Independent Variables	Sex Age Group Marital Status Education Level Number of Room The Ownership Status of House				
Deviance Residuals	Min	1Q	Median	3Q	Max
	-2.0307	-0.9134	-0.4680	0.9593	2.8068
Coefficients	Estimate	Std. Error	Z value	Pr (> z)	
(Intercept)	-3.253155	0.112359	-28.953	< 2e-16	***
Sex 2	0.144040	0.033012	4.363	1.28e-05	***
Age Group 25-29	0.093325	0.070803	1.318	0.187473	
Age Group 30-34	0.278824	0.072203	3.862	0.000113	***
Age Group 35-39	0.320005	0.073804	4.336	1.45e-05	***
Age Group 40-44	0.163662	0.075282	2.174	0.029706	*
Age Group 45-49	0.083513	0.078046	1.070	0.284597	
Age Group 50-54	0.006223	0.078735	0.079	0.936999	
Age Group 55-59	-0.101532	0.083488	-1.216	0.223937	
Age Group 60-64	-0.182453	0.090798	-2.009	0.044490	*
Age Group 65-69	-0.297562	0.104513	-2.847	0.004412	**
Age Group 70-74	-0.619833	0.130586	-4.747	2.07e-06	***
Age Group 75-79	-0.782128	0.165255	-4.733	2.21e-06	***
Age Group 80+	-1.396816	0.197042	-7.089	1.35e-12	***
Marital Status 2	0.147690	0.044483	3.320	0.000900	***
Education Level 1	1.176578	0.066213	17.770	< 2e-16	***
Education Level 2	1.713769	0.075819	22.603	< 2e-16	***
Education Level 3	2.407249	0.072642	33.139	< 2e-16	***
Education Level 4	2.894906	0.076467	37.858	< 2e-16	***
Number of Room	0.293262	0.020789	14.107	< 2e-16	***
The Ownership Status of House 2	0.070735	0.039826	1.776	0.075712	.
The Ownership Status of House 3	0.279490	0.164114	1.703	0.088563	.
The Ownership Status of House 4	-0.021231	0.052712	-0.403	0.687113	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 27302 on 20157 degrees of freedom					
Residual deviance: 22880 on 20135 degrees of freedom					
AIC: 22926					
Number of Fisher Scoring iterations: 5					