Børge Strand

Digital archives in historical research

Gender effects on labour market participation in post-industrial Norway – a quantitative approach

Thesis for the degree of Doctor Philosophiae

Trondheim, December 2014

Norwegian University of Science and Technology
Faculty of Humanities
Department of Historical Studies

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Doctor Philosophiae

Faculty of Humanities
Department of Historical Studies

ISBN 978-82-326-0604-7 (printed ver.)
ISBN 978-82-326-0605-4 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2014:346

Printed by NTNU-trykk

It's Not Where You're Going – It's How You Get There!

Charles W. Shirriff

CONTENTS

INDEX OF FIGURES 3

INDEX OF TABLES 4

INDEX OF SAS DATA STEPS 6

PREFACE 7

PREFACE BY THE NATIONAL ARCHIVIST OF NORWAY 9

AIMS OF THE THESIS 11

DEFINITIONS AND CONCEPTS 16

- DIGITAL ARCHIVES VERSUS COMPUTERISED ARCHIVES 16
- INFORMATIONAL VALUE VERSUS EVIDENTIAL VALUE 25
- NATIONAL IDENTIFIERS AND CODED INFORMATION 27
  - The national identity number 28
  - The D-number 29
  - Numeric address code 30
  - Organisation number 31
  - Coded information 32

A CASE STUDY: SUBJECT, METHODS AND POSSIBLE DIGITAL SOURCES 34

- IMPORTANCE AND RELEVANCE OF THE CASE STUDY SUBJECT 34
- QUANTITATIVE ANALYSIS 36
- THE POST-INDUSTRIAL LABOUR MARKET 40
- HYPOTHESES ABOUT THE CASE STUDY SUBJECT 46
- REPRESENTATION OF THE CONCEPTS FROM THE THEORETICAL MODEL 54
- A FRAMEWORK FOR THE PROCESS OF IDENTIFYING POSSIBLE DATA SOURCES 58
- HEURISTIC CONSIDERATIONS – WHERE AND HOW TO IDENTIFY DATA SOURCES? 61
  - Possible data from the Central Population Register? 62
  - Possible records creators and data sources – a summary 66
- THE TECHNICAL METADATA: 'INFORMATION ABOUT INFORMATION' 68

BUILDING THE DATA MATRIX 70

- REQUIREMENTS FOR PANEL POPULATION AND VARIABLES 70
- HOW TO USE RAW DATA? 72
- THE PANEL POPULATION 74
- GENDER AND YEAR OF BIRTH 74
- THE DEPENDENT VARIABLE: LABOUR MARKET PARTICIPATION 75
- MATCHING PANEL POPULATION AND RATE OF LABOUR MARKET PARTICIPATION 82
- GEOGRAPHIC CONNECTION 83
- RESIDENTIAL CHARACTERISTICS EXPRESSED BY 'MUNICIPALITY CLASSIFICATION'? 84
- GEOGRAPHIC MOBILITY 88
- FAMILY RESPONSIBILITIES 89

Finalising the data matrix 92
HISTORICAL CRITICISM 95
Authenticity and reliability 99
consistency and match results 102
validity and coverage of the variables 105
The ID number 106
Gender 107
Birth cohort 108
Rate of labour market participation 109
Geographic variables 112
Family responsibilities 115
testing the model fit 117
THE THEORETICAL MODEL APPLIED ON PHYSICAL DATA 122
Correlation analysis 122
Regression analysis 124
Bivariate effect of gender at county level: Evident regional differences 128
Municipalities and bivariate effect of gender: Considerable local differences 132
The association between post-industrialism and female labour market participation 140
Research findings in other comparative studies 144
DIGITAL ARCHIVES IN HISTORICAL RESEARCH – RESULTS AND EXPERIENCES 149
Electronic registers, administrative structures and empirical findings 154
APPENDIX 1: SAS DATA STEPS 159
Data steps for establishing the panel population 160
Data steps for the dependent variable 166
Data steps for geographic variables 171
Data steps for adding family obligations – number of children 174
Data steps for finalising the data matrix 179
example data steps for the historical criticism process 181
APPENDIX 2: EXTRACT OF TECHNICAL METADATA FOR THE TAX REGISTER FOR PERSONAL TAXPAYERS 185
APPENDIX 3:'BASIC AMOUNT' (G) - AND AVERAGE PENSIONABLE INCOME 1967 – 2007 187
APPENDIX 4: NORWAY – COUNTIES AND MUNICIPALITIES. 189
APPENDIX 5: TECHNICAL METADATA FOR 1990 TAX REGISTER FOR PERSONAL TAXPAYERS, PAGE 1. 191
SOURCES 193
Digital sources 193
Paper-based sources 194
Legislation and regulations 194
BIBLIOGRAPHY 196
INDEX 203

INDEX OF FIGURES

Figure 1. Computerized archives: Scanned image from church register..... 17
Figure 2. Extract from system with numeric and coded data. 19
Figure 3. Model for explaining variation in the rate of labour market participation. 48
Figure 4. DSF table structure..... 68
Figure 5. Average labour income for men and women 1967 – 2007 and two G 1967 – 2007. NOK. Current value. 81
Figure 6. Match between two data sets with partly a joint population and partly separate populations. 104
Figure 7. Panel population and relative share of each birth cohort with maximum employment by gender. Per cent. 111
Figure 8. Panel population and relative share of each birth cohort without labour market participation. Male and female. Per cent. 112
Figure 9. Adjusted model for explaining variation in labour market participation. 121
Figure 10. Labour market participation and bivariate effect of gender. Standardized Beta coefficients. County. .. 130
Figure 11. Labour market participation and bivariate effect of gender. Standardized Beta coefficients. Municipality. 137
Figure 12. Vardø and Iveland. Males and females by number of years of employment. Per cent. 138
Figure 13. Persons 16 years and older with tertiary education by county. Men and women. Per cent. 1970. 140

INDEX OF TABLES

TABLE 1. HYPOTHESES ABOUT VARIATION IN LABOUR MARKET PARTICIPATION IN POST-INDUSTRIAL NORWAY. 47

TABLE 2. A SKETCH OF THE DATA MATRIX. 53

TABLE 3. SELECTED AMOUNTS OF BASIC AMOUNT (G) AND AVERAGE LABOUR INCOME 1967 AND 2007. NOK - CURRENT VALUE. 80

TABLE 4. NUMBER OF WAGE EARNERS AND SELF EMPLOYED, 1970, 1980, 1990, AND 2000 – IN 1000. 82

TABLE 5. GROSS AND NET PANEL POPULATION. 84

TABLE 6. CLASSIFICATION OF MUNICIPALITIES – MAIN CLASSES. 87

TABLE 7. EXTRACT FROM THE FINAL DATA MATRIX WITH A FEW RANDOM OBSERVATIONS AND SELECTED VARIABLES. 94

TABLE 8. OVERVIEW OF OBSERVATIONS AND VALUE RANGE OF VARIABLES IN THE DATA MATRIX. 106

TABLE 9. PANEL POPULATION BY CODE FOR GENDER. 107

TABLE 10. NET PANEL POPULATION AND POPULATION IN OFFICIAL STATISTICS BY BIRTH COHORT GROUPS. 108

TABLE 11. PANEL POPULATION AND LABOUR MARKET PARTICIPATION: MAXIMUM, MINIMUM AND AVERAGE. TOTAL, MEN AND WOMEN. 110

TABLE 12. PANEL POPULATION BY NUMBER OF YEARS OF LABOUR MARKET PARTICIPATION 1967 – 2007. TOTAL, MALE AND FEMALE. 110

TABLE 13. PANEL POPULATION BY MUNICIPALITY CLASSIFICATION. 113

TABLE 14. PANEL POPULATION BY NUMBER OF MUNICIPALITY CHANGES 1967 – 2007. TOTAL, MALE AND FEMALE. 114

TABLE 15. AVERAGE NUMBER OF CHILDREN FOR SELECTED COHORTS BY GENDER. OFFICIAL STATISTICS AND TRP. 115

TABLE 16. ANALYSIS OF VARIANCE (ANOVA) TEST. 118

TABLE 17. COEFFICIENTS. DEPENDENT VARIABLE: EMPLOYMENT HISTORY. 119

TABLE 18. TEST FOR MULTICOLLINEARITY. DEPENDENT VARIABLE: EMPLOYMENT HISTORY 119

TABLE 19. HYPOTHESIS ABOUT LABOUR MARKET PARTICIPATION IN POST-INDUSTRIAL NORWAY. ADJUSTED. 120

TABLE 20. MULTIVARIATE CORRELATION. NATION. 122

TABLE 21 A. COEFFICIENTS. LABOUR MARKET PARTICIPATION AND BIVARIATE EFFECT OF GENDER. 125

TABLE 21 B. COEFFICIENTS. LABOUR MARKET PARTICIPATION AND BIVARIATE EFFECT OF BIRTH COHORT. 125

TABLE 21 C. COEFFICIENTS. LABOUR MARKET PARTICIPATION AND BIVARIATE EFFECT OF GEOGRAPHIC MOBILITY. 125

TABLE 22. MODEL SUMMARY - EXTRACT. LABOUR MARKET PARTICIPATION AND BIVARIATE EFFECT OF GENDER, BIRTH COHORT AND GEOGRAPHIC MOBILITY, RESPECTIVELY. 126

TABLE 23. LABOUR MARKET PARTICIPATION AND TRIVARIATE EFFECT OF GENDER AND BIRTH COHORT 126

TABLE 24. MODEL SUMMARY. TRIVARIATE REGRESSION. NATION. 127

TABLE 25. BIVARIATE EFFECT OF GENDER BY BIRTH COHORT. NATION. 128

TABLE 26. BIVARIATE EFFECT OF GENDER BY COUNTY. SORTED BY STANDARDIZED BETA COEFFICIENTS. 129

TABLE 27. BIVARIATE EFFECT OF GENDER BY MUNICIPALITY CLASS. CLASSIFICATION AS OF YEAR 20. 131

TABLE 28. TEN MUNICIPALITIES WITH THE STRONGEST GENDER EFFECTS. 134

TABLE 29. TEN MUNICIPALITIES WITH THE WEAKEST GENDER EFFECTS. 136

TABLE 30. DESCRIPTION DSF – DEMOGRAPHIC TABLE. 160
TABLE 31. COMPONENTS OF THE NATIONAL ID NUMBER. 162
TABLE 32. STRUCTURE OF EACH COHORT TABLE. 165
TABLE 33. DESCRIPTION DSF – INCOME TABLES. 166
TABLE 34. STRUCTURE OF NUMERIC ADDRESS. 171
TABLE 35. COHORT GROUPS AND EDITION OF MUNICIPALITY CLASSIFICATION FOR MATCH. 174
TABLE 36. DESCRIPTION FOR SPOUSE TABLE ASSEMBLED FROM TRP. 177
TABLE 37. DESCRIPTION – FINAL DATA MATRIX. 180

INDEX OF SAS DATA STEPS

SAS DATA STEP 1. FIRST SELECTION OF POPULATION. 161
SAS DATA STEP 2. VALIDITY CHECK OF NATIONAL ID NUMBER. 161
SAS DATA STEP 3. DUPLICATE CHECK. 162
SAS DATA STEP 4. DERIVE GENDER AND CALCULATE FOUR-DIGIT YEAR OF BIRTH. 163
SAS DATA STEP 5. SELECTING POPULATION COHORTS. 164
SAS DATA STEP 6. SELECT INCOME BY TYPE AND BY INCOME YEAR. 167
SAS DATA STEP 7. MATCH LABOUR INCOME BY TYPE AND SUM UP INCOME FIELDS. 167
SAS DATA STEP 8. TEST INCOME AMOUNT AND REPLACE WITH CODE FOR EMPLOYMENT. 168
SAS DATA STEP 9. MATCHING THE 1937 COHORT WITH EMPLOYMENT CODE FOR 1967. 169
SAS DATA STEP 10. MATCHING THE 1937 COHORT WITH EMPLOYMENT CODE 1968. 169
SAS DATA STEP 11. SUMMING UP CODED FIELDS FOR EMPLOYMENT HISTORY. 170
SAS DATA STEP 12. READ MUNICIPALITY IDENTIFIER AND SELECT VALID IDENTIFIERS. 172
SAS DATA STEP 13. MATCH BETWEEN GROSS AND NET POPULATION. 173
SAS DATA STEP 14. READING 1980 GENERATION OF TRP. 175
SAS DATA STEP 15. DATA STEP FOR SPLITTING BY PERSONAL CODE. 176
SAS DATA STEP 16. CREATING SPOUSE DATA SET FROM TRP. 178
SAS DATA STEP 17. CREATE ANONYMOUS DATA SET. 179
SAS DATA STEP 18. REDIRECT POPULATION FROM EXPIRED MUNICIPALITIES. 181
SAS DATA STEP 19. MATCH BETWEEN DSF INCOME TABLE 1990 AND TRP 1990 WITH SAS LOG. 182
SAS DATA STEP 20. CALCULATE DIFFERENCE BETWEEN PENSIONABLE INCOME IN DSF AND IN TRP. 183

PREFACE

This is a revised edition of the book *Digital archives in historical research.Gender effects on labour market participation in post-industrial Norway – a quantitative approach* published 2013 in Riksarkivarens skriftserie 38.

Hamar, August 2014
Børge Strand

PREFACE BY THE NATIONAL ARCHIVIST OF NORWAY[1]

This book is the result of a research project conducted by Børge Strand, as an archivist at the Regional State Archives of Hamar. The National Archives of Norway has stored digital archives since the 1980s, but has to date offered no general service for the use of these archives. This is partly due to the fact that access to digitally created archives is restricted for privacy reasons, and partly because the technical threshold for use of such archives is quite high.

Digital archives in the custody of the National Archives mainly contain personal information which is strictly protected by legislation. The release date for free distribution of personal information is 60 years or more. Though access to digital archives is generally restricted, there are exceptions. Access may be granted for research purposes based on exceptions laid down in the Norwegian Public Administration Act, but the technical barriers still have to be overcome. Strand's research project aims at demonstrating a method for applying digital archives in historical research exemplified by a case study.

Digital archives are usually categorized as either 'subject-specific systems' or 'systems for internal administration'. The first category was gradually implemented in central government administration from the 1960s. A common term for such systems is 'registers' which is often reflected in the system names: the Central Population Register, the Tax Register for Personal Taxpayers, the Central Vehicle Register and the Register of Business Enterprises, to mention a few. Electronic registers constitute an increasingly important part of our digital cultural heritage. For instance, the Central Population Register is described as the jewel in the crown of this cultural heritage.

The sources for Strand's case study are basically 'registers' as defined above; hence other categories of digital archives are not discussed in any detail. A simple description of a register is that it stores data in a structured format, in tables, columns, rows and cells. In an archival repository, register data must be stored in technology-independent format, also known as 'raw data'. Technology-independent storage means that the original functionality for queries, retrieval and processing is lost. Some of the technological barriers are related to the raw data format, but at the same time raw data offers vast opportunities for creating 'tailor-made' research data sets.

[1] For the 2013 edition

It is a great pleasure to be able to present this study in *Riksarkivarens skriftserie*. This is the first time I can present a publication in English for international researchers and readers. I hope this book reaches a large group of readers and extend my thanks to Børge Strand for his research and for his willingness to publish his text.

The Central Office of the National Archives of Norway, May 2013

Ivar Fonnes
National Archivist of Norway

AIMS OF THE THESIS

The main aim of this thesis is to explore some of the potential for historical research based on digitally created archives and at the same time demonstrate methods and techniques for this purpose. The best way to achieve this aim is to apply real, digitally created data in an example enquiry: a case study. Thus the main results of the case study are the experience that can be drawn from and the lessons to be learned about the use of digital archives in the context of historical research. The possible results of the case study subject itself are subordinate, though this study may be read as an independent historical enquiry.

A fundamental condition for this enquiry is that possible data sources had to be found in the collection of digital archives in the custody of the National Archives of Norway. Preserving digitally created archives in a long-term perspective presents numerous challenges. Different strategies for this purpose exist, among these is one strategy known as the 'migration strategy' where the basic principle is 'technology-independent storage'. Data stored in technology-independent format is also known as 'raw data'. Raw data is not prepared and readymade for any kind of utilisation, but must undergo a re-creation process before any utilisation can commence. An important aim for this thesis is therefore to identify some of the technical barriers that must be overcome when data sources appear as raw data, and hence the technical part of preparing data for the case study will be carefully documented and discussed.

For the purpose of the case study, a research data set must be constructed. This data set will be based on digitally created archives and designed as a longitudinal study. A basic condition for the planned data set is therefore that the entity has to be the individual person. Each individual in this data set will be observed repeatedly over a given period of time, and each individual will be supplied with the same set of variables for the observation period. Through this approach, limitations, weaknesses and strengths as well as possibilities for historical research based on digitally created archives will be closely investigated.

As the construction of this specific research data set requires the processing of personal information, this raises the issue of privacy and confidentiality. Access to personal information subject to the 60-year release date in the custody of the National Archives of Norway is generally restricted. My access to the majority of the data sources is granted by the National Archivist of Norway in accordance with legislation and with the National Archivist's practice and policy in matters of access to restricted material. Access to one of the data sources is, however, also subject to the consent of the records creator: the Norwegian Labour and Welfare Service authorities. A condition set by this records creator for access to this particular data source is that the research data set must be deleted when the analyses have been completed.

In any event, the research data set will be anonymised before any analysis is carried out, no identifiable personal information will be published, and when all analyses are finished, the research data set will be deleted. Nonetheless, the documentation included in this thesis (source code, technical metadata etc.) will be sufficient to reproduce the research data set for control or other purposes, if desirable, and provided that the same access rights apply.

When the data set for the case study has been completed, the data will be analysed quantitatively, mainly by regression analysis. Results from the analyses will be presented as tables, graphics and in plain text. Identifiable individual information will not be published.

The National Archives of Norway has stored digital archives since the mid-1980s, but as of 2013 has not offered any general application service for these archives. Access to digitally created archives is restricted for privacy reasons, besides which, since these are stored as raw data, the technical threshold for use is quite high, both in terms of understanding the technical metadata as part of the heuristic process and in terms of data retrieval. Raw data is not intuitively understandable or accessible; see e.g. Figure 2 and Appendix 2. Consequently, use of digitally created archives from raw data format as stored by the National Archives of Norway has been very limited so far. This also implies that the usability of the migration strategy has not really been tested. At present, there is no 'recipe' for utilising such records for historical research. Digitally created archives generally appear to be much more inaccessible than computerised archives, as discussed below.

Data computerised from paper-based originals are frequently used in research, in one way or another, provided the necessary preparations have been made. Privacy considerations are not an issue in the case of sources from paper-based originals, as personal information in computerised archives normally is older than any stipulated release date. Personal information originating from digitally born archives, but in aggregated shape, such as statistics, is also frequently used in research as such data are not subject to privacy restrictions.

The research data set to be constructed for the upcoming case study will be based on a full scale population and not a sample. Thus the data set will cover approximately 1.2 million persons, or observations, i.e. the number of observations that meet a set of requirements which will be presented later.

An important admission at this point is that the case study subject has been chosen because it can be examined using certain methods and techniques, and requires utilisation of digitally created archives from raw data format. However, this is probably not the first time research methodology and source material have played a part in defining a research subject. Time limits, as well as other limitations for the case study subject, are imposed by the data sources that are available. This will be discussed later, but as a general statement, I have simply accepted the limitations imposed by the data sources and focused on the possibilities existing within the given constraints. The time limits for the

case study set by accessible and relevant data sources will be outlined later. To a large extent, given this background, the following case study must necessarily be experimental.

As mentioned earlier, the research unit or observational unit in the panel is the individual person, mainly referred to as an observation. Observations for the case study will be constituted by selected birth cohorts. Together, all observations constitutes the research population. Each observation will be followed for a period of 20 years of his or her life – by my decision from the age of 30 to the age of 50, which means that the 20-year period will be different for each birth cohort. The 1937 birth cohort will be 30 years old in 1967, so the observation period for this cohort will be 1967 – 1986. On the other hand, the observation period for the 1958 cohort will be 1988 - 2007. Thus the 1937 and the 1958 birth cohorts, respectively, represent the outer limits of the population to be included in the research data set. This population will later be referred to as the 'panel population'. For the suggested case study, these are some basic limits for what can actually be investigated at present when the conditions in question are a combination of 'digitally created archives' in raw data format and 'historical research'.

This presentation basically consists of sections where the chapters are ordered in accordance with the natural sequence of stages of the research project, beginning with general definitions, prerequisites and constraints related to digitally created data sources. An exception is the data steps required for the step-by-step construction of the research data set. These data steps are taken out of their native chapters and instead collected and presented in Appendix 1.

The presentation begins with technical and conceptual definitions. Distinctions must be drawn between digitally created archives on the one hand and archives computerised from paper-based originals on the other hand. However, it is also essential to point out distinctions and differences within the category digitally created archives in order to understand the conditions, challenges and methods required for use of such archives in a specific research project.

In the next section of this thesis, the historical background and framework for the enquiry in question – possible gender effects on participation in the post-industrial labour market - and theories and hypotheses for the case study subject are outlined. Against this background, the case study is gradually materialised through questions and by means of a theoretical model which illustrates and specifies the subject, and suggests preliminary answers to the questions. This section also clarifies requirements for the research population and variables for the theoretical model.

The next step is to identify possible data sources from which the population and the specific variables must be operationalized in order to create the research data set, i.e. a data matrix. The operationalization of variables and how the research data set is constructed will be thoroughly presented in this section.

Source criticism is not just a theoretical exercise, but to a much greater degree an actual examination and test of the physical data assembled in the data matrix. Hence, once the research data

set has been physically created, source criticism must necessarily follow. During the source criticism process, the coherence and consistency within the data matrix and accordance with the origin in electronic registers must be enquired and tested. The risk of writing erroneous source code during creation of the data matrix is obviously present. Consistency controls are important to detect any such mistakes. The adaptability of the physical data on the theoretical model for regression analysis has to be tested when the data matrix has been completed, but before the analysis takes place. Logically, therefore, this test is part of the source criticism section. Testing the suitability of the theoretical model on physical data may result in adjustments of the theoretical model.

In the following section, the final theoretical model will be subjected to regression analysis based on the physical data matrix, and the results will be presented by graphics and tables. This is where the preliminary answers from the theoretical model are tested on empirical data. The regional and local perspective is essential during the analysis.

The final part of the thesis deals with the experience and knowledge that can be drawn from the case study, where results relating to the main aim of the thesis are summarised and evaluated.

The technical solutions for creating the data matrix are described in a separate appendix. This appendix contains selected source code written in SAS (cf. below) and comments documenting the process of transforming raw data into the completed data matrix.

A few remarks about the concepts and notation applied in this thesis are now appropriate. The terms ‘system’, ‘data set’, ‘table’ and ‘file’ will be used frequently. The distinction between a ‘system’ and a ‘data set’ is important. ‘System’ is normally a short form of the concept ‘information system’. An information system includes one or more databases, as well as procedures and software for registering, updating and retrieving data, creating reports, etc. Thus an information system comprises both the stored information itself and its specific software and hardware environment. In other words, this is technology-dependent storage which is incompatible with requirements for long-term storage. To be more specific, long-term storage means storage over at least a period of centuries.

Information from a system can be broken down into units which we may call ‘data sets’. Defining a data set within a given system is often a method for creating distinct physical units for the purpose, for instance, of extracting data for transfer to an archival institution. Various criteria can be applied to define and generate a data set, but chronology is often a determinant, as well as regional criteria.

Furthermore, a data set can be composed by one or more ‘files’. A file is the set of all occurrences of a given table structure. A ‘table’ is the relational database equivalent of a file. The term ‘file’ is used both in the sense of a data set and in the sense of a table, but in this book the ‘file’ concept is mainly applied in a technical context.

The term ‘record’ is also used in the technical sense: one row in a table is one record. Inside a record – or a row - there will be a number of ‘fields’ or ‘variables’. When speaking in a technical context I will use the term ‘field’, and in an analytic context I will use the term ‘variable’.

The term ‘register’ will be used about both a system and a file, but the context will reveal in which sense the concept is applied. ‘Register’ is explained in greater detail in the next chapter. The term ‘database’ is not frequently used, because this does not quite cover the oldest history of electronic administrative systems as some of these were developed before database technology was common. Furthermore, the database concept does not clearly distinguish between ‘system’ and ‘file’. The Tax Register for Personal Taxpayers (TRP) (*Ligningsregisteret*) – established by the Central Taxation Authorities in Norway may be regarded as a system. This system has existed since 1967 and produces annual volumes of data, where each volume constitutes one data set.

Software for establishing my research data set will be SAS, cf. the chapter ‘How to use raw data’ as well as Appendix 1. As far as the construction of the research data set is concerned, selected code snippets written in SAS are included in Appendix 1. Inside each code example, field names, data type, etc. are presented in standard SAS notation: as a basic rule, SAS key words are written in uppercase – e.g. DATA _NULL_. File names are written in lowercase italics like *data.population*, *population.txt* or *matrix.dat*. The first notation is always used for data sets in SAS format, while the last notation with the file extension *.dat* or *.txt* is used for data sets in text format. When variable names appear outside SAS code examples, the notation is in lowercase italics, e.g. *gender*. Finally, acronyms are normally typed in UPPER CASE.

In the following text, terms like ‘record linkage’, ‘merging’ and ‘matching’ will occur, but basically in the same sense: assembling and combining variables from different data sets and data tables with each observation.

All data processing in this project has been carried out behind the secured zones of the National Archives of Norway. All temporary files created during the process have been deleted when they were no longer needed. I have considered the risk of losing temporary data to be immaterial, since such data could quickly be recreated by means of the source code and the archived primary data.

DEFINITIONS AND CONCEPTS

DIGITAL ARCHIVES VERSUS COMPUTERISED ARCHIVES

Unfortunately, there is a certain risk of confusing the concepts 'digital' archives and 'computerised' archives, which is why both must be mentioned in a definition context. To the question 'digital archives – what do we mean?' - the short answer is that we mean archives that are created digitally. The concepts 'digital archives' and 'electronic archives' are used synonymously, and cover all types of archives that are created digitally, as opposed to 'computerised' archives which have a paper-based origin, but which have been computerised in the last few decades. A few examples may help to clarify this distinction: the Norwegian Central Population Register (CPR) was born digitally in 1964 and is an example of digital archives, while the 1801 Population Census was created on paper questionnaires in February 1801, and computerised in the early 1970s, and is an example of 'computerised archives'. Computerised archives generally do not appear as raw data, and consequently do not require any re-creation process before utilisation.

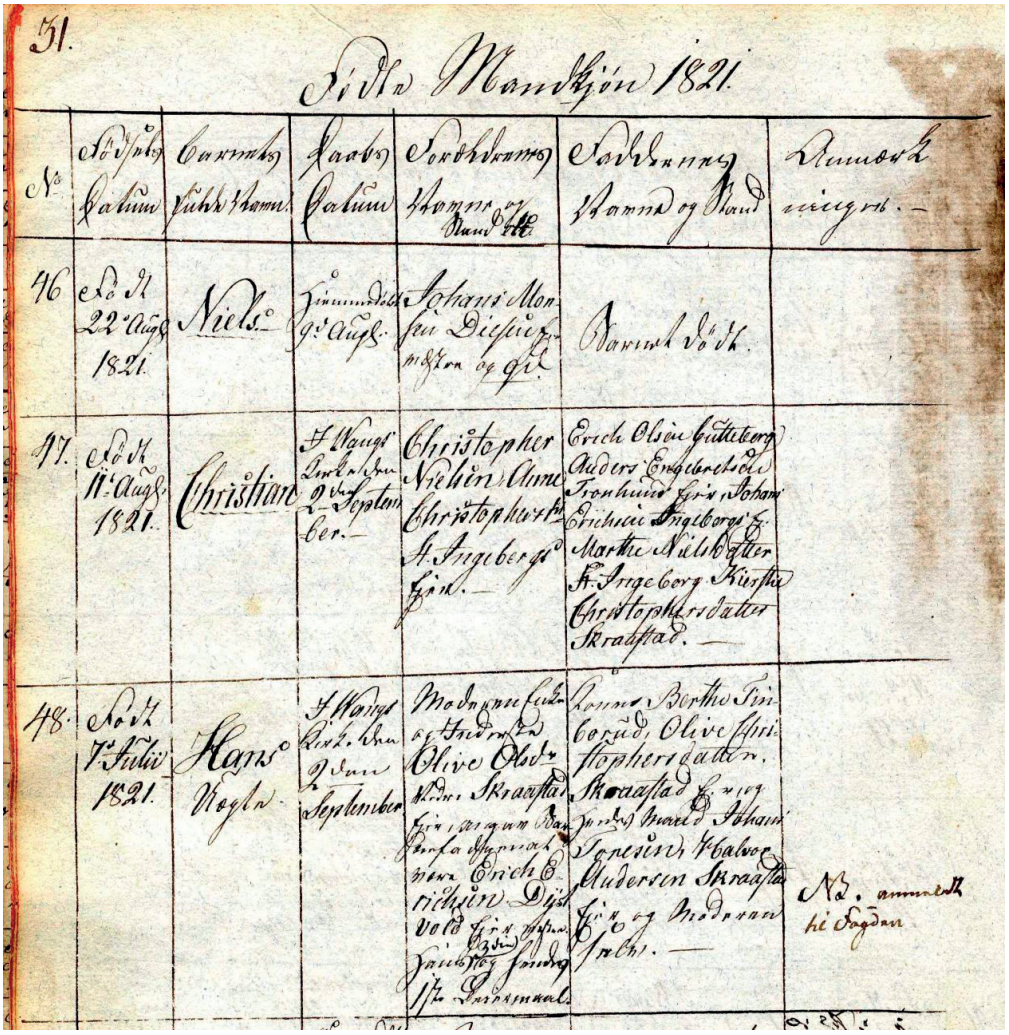
Any electronic system in active administrative use will have a user interface which makes the information accessible and understandable. But a system with its given user interface, functionality and hardware environment cannot be preserved over time. The price to pay for long-term storage of digitally created information is that the accessibility provided by a living system is lost. For digitally created archives there is no paper-based version that can serve as a 'backup', which is why technology-independent storage is so critical for preservation.

On the other hand, computerised data are always entered into and presented within an active and living information system, with a user interface which can be compared to any active, contemporary system. While computerisation of paper-based archives serves several purposes, it will primarily enhance accessibility and distribution, and reduce attrition on paper originals. So far, computerisation has never been carried out in order to destroy the paper-based originals. The possible loss of a computerised version of part of an archive represents no loss of information, as the paper-based original will always be accessible. For this reason, conversion of computerised data to technology-independent format is not required to secure the information itself, but as computerisation in addition involves an economic investment, the question of using technology-independent format for computerised data as well is gradually becoming more apparent. If or when that happens, computerised data will have to be converted to and re-created from raw data format, but that is not yet the normal situation.

All groups and subgroups of digital archives as well as computerised archives have their specific characteristics and their own specific nature, requiring qualitatively different methods of access and utilisation. Obviously, an audio file must be accessed and used differently from an image,

and plain text information must be used differently from coded information. Understanding these differences is fundamental to understanding the methods for accessing, re-creating and using raw data extracts from electronic registers as they appear in the archives, which will be elaborated on through this thesis.

Figure 1. Computerized archives: Scanned image from church register[2].



By no means can archives with the tag 'digital' or 'electronic' be perceived and treated as if they were a homogeneous group. Some basic concepts must therefore be defined, and important

[2] Church register for Vang Parish in Hedmark County (1809 – 1827), p. 31.

distinctions need to be emphasized. A major aim of this clarification is to prevent confusion between subgroups of archives within the category 'digital archives', but also between digitally created archives and paper-based, subsequently computerised archives. Both digital and computerised archives comprise groups and subgroups that are fundamentally different.

Paper-based archives have been computerised to enhance accessibility in general, and for research purposes in particular. Experience from research based on computerised archives cannot automatically be transferred to research based on digital archives, or vice versa: from a research perspective, it is not indifferent whether the origin of the research data is paper-based or digital. Use of data in a research context differs fundamentally for the data exemplified in Figure 1 and in Figure 2, respectively. Within the category computerised archives, there are also significant differences between scanned images of protocol pages as in Figure 1, and transcribed data. Furthermore, in the case of transcribed data, there are different ways of utilisation for transcribed, plain text information and encoded information, respectively. Information in picture scanned formats (i.e. graphics) can only be used visually. If such information is planned to be processed electronically, a transcription is required and, ultimately also an encoding process must precede any computer based analysis.

The image in Figure 1 has been scanned from a church register and shows a list of baptisms dating from September 1821. The image is digitized, but the text in the image is obviously not machine-readable. In this image, the name "Hans" appears e.g., but it is not possible to make a computer-based query for this name, or any other name, text string or letter, in the image. Altogether computerised archives retain their paper-based nature and characteristics also after computerisation. Compared to digitally created information, there are significant differences as regards the technical utilisation of computerised archives.

In contrast, the example in Figure 2 is an extract from a digitally created electronic register containing numeric and coded information. Any digit or group of digits in this file is machine-readable. This kind of data cannot be utilised visually, but must be accessed and processed by a computer.

Routines and procedures for computerising data from paper-based originals have been dealt with and debated by Gunnar Thorvaldsen[3]. Thorvaldsen discusses difficulties connected with interpreting and standardising information from paper-based originals which display a diversity in spelling of names, profession, place of living etc. Normally computerisation from paper-based originals follows two steps: first, transcription as free text, letter by letter, followed by an encoding process, which means that information passes through two interpretative steps before any analysis may commence. This is e.g. the sequence followed by the 'Ullensaker project'[4], cf. below. Each step of interpretation, data entry and encoding obviously represents a possible source of error which

[3] Thorvaldsen, Gunnar (1999): *Databehandling for historikere.* Oslo: Tano Aschehoug, pp. 107 and pp. 204.
[4] Langholm, Sivert (1974): *Historie på individnivå.* Oslo: Historisk Tidsskrift, 3/1974,

necessarily requires attention and critical assessment. Consequently, an encoded version of a set of computerised archives is a tertiary version.

Opposed to this - digitally created archives, in particular electronic registers, are designed for encoding, and as these are encoded by the records creator during the creation process, they appear encoded in their primary version. Source criticism is still required, but a major point is that digitally created data sources are not influenced by the researcher‘s interpretation prior to preparation for encoding. Machine-readable codes in the primary version of digital archives are ready to be applied in computer-based research.

Figure 2 shows a table extract from a public system (the Tax Register for Personal Taxpayers) in raw data format where basically all the information is represented by numerals. The format in this data table is comma separated, i.e. record and field length is variable. In any event, each row of this table represents information about one taxpayer – in the example there are ten taxpayers – and each field represents specific information about the taxpayer. The content of one field is either one single digit or a group of digits. The meaning of each field is explained by technical metadata which must always accompany a raw data table extract. Without technical metadata it is impossible to understand the meaning of the content of each field. An extract of technical metadata for this data set is shown in Appendix 2[5]. For privacy reasons the personal ID numbers have been excluded from this extract.

Figure 2. Extract from system with numeric and coded data.

;960603;261485;5;7;10;0;5;0;95;;12;1;0;515;0;17640;0;0;0;0;
;960603;351451;12;7;10;0;5;0;95;;12;1;55;765;0;17640;0;0;0;0;
;960913;532280;56;28;10;0;5;0;95;;12;1;0;219;0;8820;0;0;0;0;
;960530;365316;11;4;10;0;5;0;95;;12;1;443;1146;0;17640;0;0;0;0;
;960826;385302;39;15;10;0;5;0;95;;12;1;111;682;0;14700;0;0;0;0;
;960828;115574;45;0;10;0;-;0;95;;12;;0;0;0;0;0;0;0;0;
;960912;532327;56;0;10;0;-;0;94;;12;;0;0;0;0;0;0;0;0;
;960603;351500;12;7;10;0;5;0;94;;12;1;7;619;0;17640;0;0;0;0;
;960830;370546;37;18;10;0;5;0;94;;12;1;0;0;0;2940;0;0;0;0;
;960905;61185;39;22;10;0;5;0;94;;12;1;0;211;0;8820;0;0;0;0;

The table example in Figure 2 is an extract from the main table of the 1995 Tax Register for Personal Taxpayers, showing a few selected fields of a total of nearly 200 fields. Some of the selected fields contain coded information, and some fields contain various types of income amounts. This

[5] The technical metadata in Appendix 2 only specifies the first field in the table extract: ‘Production date’ (...) >*Date for last change of correction - format YYMMDD*<. The rest of the information is context information.

volume of the register comprises about 3.5 million taxpayers. This example should suffice to illustrate why information represented by codes and numerals cannot be used visually, and a printout of such material would be both meaningless and useless. This kind of information can only be utilized electronically; on the other hand, this offers vast opportunities for research.

Furthermore, the concept 'digital archives' needs to be defined precisely in terms of groups and subgroups. What is characteristic of one group of digital archives is not necessarily characteristic of a second, or a third group. Last, but not least, different categories of digital archives vary considerably in terms of utilisation.

By definition, digital archives have a short history, starting with the digital era from the 1950s and onwards, depending on which country we have in mind. In Norway, digital archives have emerged since the first half of the 1960s, as fundamental functions in public administration began to be computerised. Since the 1960s, the Norwegian public administration has undergone increasing computerisation. Today, there are considerable numbers of administrative registers, databases and information systems in all public institutions in Norway. Together, these systems constitute the informational infrastructure of the modern welfare state.

A fundamental condition from the very beginning was that digital information should be exchangeable, between systems and between institutions and organisations. Some basic information relating to people, real estate, employers, etc. could be shared and reused thanks to what we might call a set of 'national identifiers', as outlined in the chapter 'National identifiers and coded information'.

Systems developed and implemented during the first decades of the computer era were mainly a continuation of old, manual register systems, card files, etc. where the entities were persons, real estate, business organisations or other legal entities. A common denominator for the information in these manual systems was that information had a structured nature – organised in columns, rows and cells – though the information appeared in plain text. In other words, it was just the kind of information that would be profitable to computerise, and hence structured information was first in line for computerisation and encoding. Thanks to computerisation, formerly time-consuming processes could be automated - e.g. calculation of direct taxes has been computer-based since 1967.

All these systems were created primarily to serve administrative purposes. Whether the purpose was to enhance civil registry, fiscal or taxation functions or some other administrative objective, the advantages of computerisation for administrative purposes are easy to spot. Though administrative registers are designed to serve specific administrative functions, they represent valuable sources for research as well. But this is not a new aspect; paper-based church registers served ecclesiastical purposes, and provide information about how the clergy performed their official duties, but such archives are obviously also highly applicable in e.g. demographic research due to their detailed demographic information.

In terms of cultural evidence and research potential, many of the electronic administrative registers can be compared to censuses. An important difference is that administrative sources are continuously updated, while censuses are periodical and are never updated once they are completed. Continuously updated registers with national identifiers open up for large-scale, longitudinally designed studies.

However, problems raised by the computerisation of administrative functions are also easy to spot. It is sufficient to mention problems concerning confidentiality, protection of personal integrity, risk of data corruption or loss of information. Such problems will be dealt with in the chapter 'Historical criticism' and in other contexts as they become relevant.

Whether we like it or not, a side effect of an increasingly computerised public administration is of course that a growing share of the Norwegian cultural heritage is born digitally, and consequently must remain in digital form also in the archives. Preserving digital information for the future implies huge methodological and technological challenges. Since the mid-1980s, the National Archives of Norway has dealt with these issues and received and stored digital archives from the central government administration.

For a long time The National Archives of Norway has practised a distinction between the 'transfer' and the 'deposit' of archives. This practice has implications for access to the archives. A transfer is defined as *'conveyance of physical records and the ownership of these records*', while a deposit is a *'conveyance of physical records without transfer of ownership'*[6]. The main rule has been a transfer, and for paper-based archives this is still mainly the case, but for digital archives deposits are more and more frequently used. The formal shift from a deposit to a transfer normally takes place 25 years after the deposit or alternatively by individual agreement. From the utilisation point of view this distinction is decisive – transferred archives are accessible for research purposes with the consent of the National Archivist of Norway, while access to deposits is normally not possible, although exceptions may be granted by the records creator.

Whatever the formal status, the technical aspects of a transfer or a deposit are the same: from the records creator's active system, an archival extract has to be exported from the production format in the native system, to archival format for long-term storage, according to the principles of technology-independent storage. Future retrieval and use of digitally created archives will necessarily be based on technology that will be very different from the technology used when the information was created. The positive side effect is that digital archives represent a huge potential for historical, as well as other, research.

[6] Forskrift om utfyllende tekniske og arkivfaglige bestemmelser om behandling av offentlige arkiver. FOR-1999-12-01-1566: VIII Bestemmelser om elektronisk arkivmateriale som avleveres eller overføres som depositum til Arkivverket. § 8-2 (Definitions).

In the mainframe era, the landscape of digital archives was quite uniform, and dominated by registers and databases storing structured information with numeric and coded information. Since the 1980s, technological development has led to a greater variety of digital archives. In particular, systems emerged for processing and handling information in unstructured format– or 'free text' – of the type that can be found in correspondence. At this stage of the thesis, a categorisation of digital archives is therefore appropriate. The following groups and subgroups are commonly applied:

1. Systems for internal administration
    1A. Electronic record-keeping systems/case-handling systems
    1B. Other systems for internal administration
2. Subject-specific systems (databases/registers)
    2A. Systems for subject-specific administrative purposes (administrative registers)
    2B. Systems for non-administrative purposes (statistical and/or research registers)
3. Other systems.

Systems in Subgroup 1A are generally record-keeping systems according to the concept used in archival theory today[7]. Subgroup 1B is a residual category for a variety of systems for other internal administration functions: accounts systems, payroll systems, management information systems and office information systems, to mention a few.

The National Archives of Norway has been occupied with all aspects of digital archives since the early 1980s, both in terms of legislation, standardisation matters, records capture, appraisal and preservation[8]. One result of these efforts is the development of a standard for systems requirements for electronic record-keeping systems: the NOARK standard[9], which was introduced in 1984. Based on this standard, systems developers have designed and developed systems which have been implemented in both public and private sector. The NOARK standard has been revised and modernised from time to time. The latest generation of this standard, NOARK-5, was launched in 2008.

While earlier generations of the NOARK standard merely were designed as electronic registry and retrieval tools for paper-based case documents, the latest generations (NOARK- 4 and NOARK-5) also implemented solutions for the electronic storage of case documents, during the case handling process as well as in the archives. In other words, the standard now offers a solution for the electronic handling of documents at every stage of the life cycle of the archives.

Main group 2 'Subject-specific systems', is a generic term for most other informational systems, provided they are systems entailing a minimum of permanent data storage. Administrative

[7] Cf. ISO15489 and MoReq e.g.
[8] Sirevåg, Trond (2002): *De elektroniske arkivene - Hva har vært Arkivverkes strategi? Resultater og utfordringer framover.* Oslo: Riksarkivarens skriftserie 13,p. 303.
[9] '**No**rsk **ark**ivsystem' – original acronym, later "**No**rsk **ark**ivstandard'

registers comprise systems for both central and local public administration. Administrative registers are developed to serve specific purposes, often initiated and authorised by legislation, such as systems for civil registry, for calculating taxes, for real estate administration, for administering pension rights and pension payments, for medical information etc., and thus represent a huge diversity of design, technology and content, according to their purpose and function. Actually, legislation is often a good place to start in order to obtain information about a specific system in this subgroup.

A common term for systems in Main group 2 is simply 'registers', the common term for Subgroup 2A is 'administrative registers' while systems in Subgroup 2B are commonly referred to as 'research registers'. These terms will be used frequently in this presentation.

The final category, 'Other systems', is a residual category for other digitally created archives, including audio and multimedia files and systems. Police questionings may e.g. exist purely as audio files.

As a rule, central administrative registers share the characteristic of covering total populations. To serve their purpose, many systems have been developed to collect information about the total Norwegian population. For other central administrative systems a basic condition is that they must cover the total population for whom specific legislation and a corresponding information system is developed, such as *all* legal entities, *all* vehicles, *all* employees – in general *all* members included by some administrative regulation. This quality always counts in favour of preservation when digital archives are appraised, as discussed and concluded by 'Bevaringsutvalget'[10], a committee initiated by the National Archivist, cf. below.

There is no reason to suspect that registration in electronic administrative registers is arbitrary in a sense that might cause any bias in terms of gender, age groups, place of residence etc. Once a unit fulfils the legal requirements for entry into an administrative register, the registration takes place. Registration in the Central Population Register e.g. is not optional; every legal resident in the country is entered into this system. The fact that the kind of information to be found about an individual in a given register might be erroneous is a totally different issue. In a historical research context, identifying and addressing this issue is an objective for historical criticism.

Electronic registers designed and constructed for research or statistical purposes - often a combination of the two – may comprise total populations, but more frequently we find sample surveys in this category, where the input is based on interviews or questionnaires, or a combination of input partly from questionnaires and partly from public administrative registers. The records creators for such systems are typically research institutions, among which Statistics Norway is a major contributor. In Norway, non-administrative processing of electronic registers with identifiable personal

[10] Riksarkivaren (2002): *Rapport fra Bevaringsutvalget.* Oslo: Riksarkivaren. Rapporter og retningslinjer nr. 10. Pp. 65.

information from raw data format (i.e. also from the primary version) is primarily carried out within the walls of Statistics Norway and other research institutions.

Electronic registers containing personal information have been used for decades for research purposes, but of course always in compliance with the regulatory framework in force. Statistics Norway distributes research data in electronic form on a large scale, but never as identifiable personal information.

It must also be mentioned that collections of register-based data sets for social science research are available through the website of the Norwegian Social Science Data Services (NSD). NSD has distributed survey data for research through their website for several years, but this information is never published in a way that makes personal individual information traceable. A common feature of electronic research data distributed by either NSD or Statistics Norway is that such information is always anonymised if the entity is the individual. This also includes data at individual level with encrypted identification keys, but encryption means that data is no longer authentic and is not an option for the National Archives of Norway; see the chapter 'Authenticity and reliability'. Data sets with encrypted identifiers can be matched provided the same encryption key is used in the data sets. Otherwise it is not possible to combine variables from different sources by means of matching at individual level.

Population and housing censuses have, more or less, been carried out every ten years in Norway since 1801. Since 1960 the censuses have been born digitally, though much of the input still came from questionnaires. But since the 1970 census, a growing portion of the input has been collected from electronic administrative registers, and the portion derived from questionnaires has gradually decreased. For the 2001 Population and Housing Census mainly housing information was based on questionnaires. With a few exceptions, any future Norwegian population and housing census will be entirely based on digital input from administrative and statistical registers. This involves combinations of information from hundreds of electronic administrative registers, and is an expression of the dynamics in digitally created administrative information.

There are also other essential differences between the main groups of digitally created archives, differences that are reflected in both archival solutions as well as in technical solutions for future retrieval and utilisation. Electronic record keeping systems and case handling systems based on the NOARK standard are constructed with integrated solutions for future retrieval and use, which makes archival versions of such systems quite user-friendly. That is not the situation for administrative and research registers. In technical terms, archival versions of registers are always table exports from the original system in technology-independent format, i.e. raw data.

The informational content of a document, such as a letter, from a case handling system must basically be accessed and understood visually, as a retrieved document to be read on a computer screen, or alternatively, as a paper print-out. On the other hand, a table export from a register cannot

be understood visually (see the example in Figure 2), but such information must be processed and made understandable electronically, as the following case study will demonstrate.

A final, but quite obvious, distinction between the main groups of digitally created archives is related to their age: we find the oldest systems, with the oldest historical information, among administrative and research registers. Record keeping systems have a shorter history, and accordingly information stored in this category of systems is still fairly young.

Finally, the distinction between digital archives and published digital material (newspapers, periodicals, books, photos, films, music, web pages and social media) must be mentioned. More and more published material is created digitally. A very important distinction is that digitally created and published material is not subject to the same restrictions as digitally created archives as far as privacy matters are concerned. With a very few exceptions, digitally created archives as defined above cannot be published freely. In administrative terms, long-term storage of digital archives is the responsibility of the National Archives of Norway and other archival institutions, while long-term storage of published digital material is the responsibility of the National Library of Norway.

INFORMATIONAL VALUE VERSUS EVIDENTIAL VALUE

The distinction between the main groups of digital archives also reflects different aspects of utilisation. One aspect is investigating and utilising the 'evidential value' of a system, while the second aspect is investigating and utilising the 'informational value' of a system. In short, the evidential value means the information a set of archives might document about the records creator's own history and own functions and transactions. Informational value on the other hand is what a set of archives might document about its subject, e.g. the society and social conditions at a given point in time. A population and housing census is a source of information about both the records creator (e.g. about Statistics Norway), i.e. the evidential value, and the population, housing and social conditions in a given society on one particular date, i.e. the informational value.

This difference was pointed out by Theodore Schellenberg in 1956: 'There is a basic dualism between evidential value and informational value'[11]. This distinction obviously applies to any archival material, but in the context of digital archives this has given rise to some debate internationally about whether systems mainly representing a research potential through their informational value should be preserved at all.

In the Norwegian Archives Act this dualism is reflected in section 1, where the two aspects are considered equal. The dualism reflected in section 1 of the Archives Act was also discussed by

[11] Schellenberg, Theodore (1956): *Modern archives: Principles and Techniques.* Chicago: The University of Chicago Press, p. 52 and pp. 148.

'Bevaringsutvalget'[12]. The report from this committee, in which a systematic procedure for the appraisal process is proposed, claims that both evidential value and informational value must be considered in the appraisal process.

There is a close connection between which of the main groups a system belongs to, and which of these aspects it mainly reflects. In somewhat simplified terms, the evidential value is primarily reflected in record keeping systems, and informational value is mainly reflected in administrative and research registers. But case handling documents also have their independent informational value, although mainly in plain text and unstructured format which to a large extent precludes computer-based analyses.

There seems to be little focus on digital archives of the electronic administrative register type in present international archival theory, perhaps because they are not 'record keeping' or 'records management' systems. It is true that in this sense databases and registers may only to a minor degree, if at all, be said to reflect the case handling transactions – the evidential value. Instead they must be seen as the continuation of classic archival series of the type found in population registers, census material, parish registers, emigration protocols, passport registers, ship passenger lists, files of fiscal listings and many, many others, and thus are carriers of information about population, society and social conditions. These are all examples of archives where the informational value has been frequently demanded and investigated.

This issue is also discussed by Ivar Fonnes in an archival manual for public administration[13]. As Fonnes points out, in the Norwegian discussion the demarcation between evidential and informational value has been disputed, and still is, but on the other hand there is a long tradition in Norway of preserving material where the decisive consideration has been the possible informational value of the archives. For example, all population censuses, since the very first from 1769, have been preserved. Each and every one of them will of course have potential evidential value. Their provenance would give valuable information about the records creator, in this case about Statistics Norway and its predecessors, if that was a subject of interest to us. However, their main asset is their informational value, i.e. whatever they might be able to reveal about the Norwegian population and society at a given point in time or over a period of time when a series of censuses are compared.

Furthermore, there is much evidence to show that there is considerable demand for informational value in the archives. To meet this demand, substantial effort and resources have been spent by the National Archives of Norway and others on computerising information from 19th-century archives. Examples include the nominative population censuses from 1801, 1865, 1900 and 1910, which have all been completely transcribed, as well as various lists from parish registers, passenger

[12] Op. cit., pp. 53.
[13] Fonnes, Ivar (2000): *Arkivhåndboken for offentlig forvaltning.* Oslo: Kommuneforlaget, pp. 129.

lists from emigration ships, etc. Resources have also been spent on establishing a coded version of the 1801 census in order to analyse the information electronically.

Computerised archives are distributed through the website 'Digitalarkivet' (www.digitalarkivet.no), which is a public service provided by the National Archives of Norway, and the computerised records are available for free use. More recently, scanned images of parish registers have also been published through this site; see Figure 1. Some figures from 'Digitalarkivet' may exemplify the demand: in 2010 almost 65 million database queries were counted, and nearly 90 million pages from parish registers and the like were downloaded. Though we may never know whether the actual utilisation of this information reflects the informational or the evidential aspect, it is hardly debatable to assume that the informational value is predominant. Given the subject of my case study, I mainly need to look for possible sources among administrative registers and investigate their informational value.

NATIONAL IDENTIFIERS AND CODED INFORMATION

From the very start of the computerisation of the Norwegian public sector, it was a fundamental idea that, whenever possible, information was to be shared and reused. To avoid multiple entry of the same information, data was to be prepared for exchange between different systems, across institutions, across administrative levels and across private and public organisations and institutions so that it could be reused in a new context.

However, an initial problem arose with regard to how information about an entity in one system could be linked to the same entity in another system. The public administrative registers had to be prepared for data exchange and reuse of information through record linkage (matching), and some standardised, national identifiers were needed for this purpose.

Unique national identifiers were defined for basic entities such as persons, legal entities and real estate. As a result, automated linkage of individual entities between different registers is possible as long as they have one of these identifiers in common. That is why such identifiers can be called 'national identifiers'.

The three main identifiers are:

- A national identity number for persons (including a subgroup called the D-number)
- A numeric address code for real estate
- An organisational number for legal entities

These identifiers serve a dual purpose as they are usually internal keys in an actual information system, as well as keys for external linkage, i.e. linkage between separate systems. From a researcher's point of view, this opens up the possibility of a wide range of enquiries and research design. In the following enquiry the national identity number for persons and, to some extent, the numeric address code are important.

THE NATIONAL IDENTITY NUMBER

For the following case study, the national identity number is the most important of the identifiers mentioned above. The history of the national identity number dates back to the late 1950s and early 1960s, and was a political response to a desire from the Norwegian banking industry. The 1960 Population and Housing Census was decided to become the population base for entry in the Central Population Register as of 10 October 1964 and the allocation of national identity numbers. All persons registered as residents in this census were granted a national identity number, as well as all persons born in Norway and all residents who had immigrated since the census date. The oldest person ever entered in the Central Population Register and granted a national ID number was born in 1855[14].

The legal basis for the national identity number is provided by the National Population Register Act of 16 January 1970[15] and the National Population Register Regulations of 25 January 1971[16]. The detailed construction of the identity number is set out in section 2-2 in the Regulations. According to the Regulations, the national identity number shall consist of eleven digits, and have the following components:

- Date of birth in the format DDMMYY
- 'Personal number' - five digits composed of:
  - 'individual number' – three digits
  - 'control digits' – the last two digits are a product of the nine preceding digits according to a defined algorithm.

Some useful qualities are linked to the identity number. The person's gender may be derived from digit number nine: even numbers are used for females and odd numbers for males. A second quality of the ID number is of course that its construction makes it possible to derive the year of birth as a four digit number, century included. The structure of the ID number is illustrated in Table 31.

---

[14] Skatteetaten (2010): *Folkeregistrering.* Virtual document about civil registry without pagination at www.skatteetaten.no

[15] Folkeregisterloven. LOV 1970-01-16 nr 01

[16] Forskrift om folkeregistrering. FOR-2007-11-09-1268

The national ID number also serves as identification for the nuclear family. All members of the same family according to their official marital status are identified by the national identity number of the reference person in the family. For married couples the reference person (normally the male or the oldest person) has his own ID number as family ID, while all other members of the same family have the ID number of the reference person as their family ID.

To distinguish between parents and children in the nuclear family a one-digit 'personal code' is added. Information about family number and personal code will be put to use later on when establishing the variable *family responsibilities* expressed by *number of children*.

An important comment is that the 'family' concept in the formal sense does not reflect the 'household', e.g. in cases where people live together without being formally married. In order to identify a household, a household identifier is required, cf. 'numeric address code'.

Finally, due to the construction with the control digits, the validity of the national identity number can be tested in an automated validity check, which is useful to bear in mind when it comes to historical criticism and questions of reliability and data quality in general.

Data on individual persons are found both in computerised and digital source material, but one of many fundamental differences between the two is that computerised individual data of course do not contain a national ID number, or any other kind of unique identifiers, and automated linkage between entities is thus impossible.

THE D-NUMBER

The D-number is an ID number for non-resident persons in Norway. On certain conditions, non-resident persons are registered in the Central Population Register as well as in other public systems. It was necessary to develop an identifier for this purpose, but at the same time an identifier which differs from the national ID-number.

The Directorate for Seamen invented a solution for such an identifier in 1978. Their solution was an ID number called the 'D-number'. The D-number is composed in the same way as the national ID number, with one difference: the number '4' is added to the first digit of the date. The date of birth in a national ID number for a person born on 30 May 1960 looks like this: '300560', while the D-number for a non-resident person born on the same date would appear as '700560'. The D-number is a valid format for the date field in the ID number, and is accepted by the validity check mentioned above. (D-numbers may also occur in systems older than 1978, as they have been entered in posterity).

Originally, most of the 'D-number population' were foreign citizens employed on ships owned by Norwegian shipping companies. As the Norwegian oil industry expanded from the 1970s and onwards, many foreign citizens were hired for the activity on the Norwegian Continental Shelf, and accordingly were given a D-number in the Central Population Register. In recent years, there has been

a vast increase in the D-number population due to the more open labour market and increased demand for labour in Norway.

In general, the D-number is issued to persons who do not intend to stay permanently in Norway (i.e. less than six months), but who have a connection to the country which requires their formal registration in the Central Population Register. Regardless of the duration of the period of employment, there is payment involved, and income from labour is always taxable income under national legislation. From a legal standpoint, it makes no difference whether such income is earned by a resident or a non-resident. The employer must in any case report the amount to the tax authorities, which means that D-numbers appear in several public databases and registers, as was proved in the study 'The D-number population'[17].

A general issue in a research context is whether to include or exclude this population in a research population. For the present case study, this issue is further elaborated below.

NUMERIC ADDRESS CODE

The numeric address code is a 25-digit identifier for geographic entities and real estate, but it can also serve as an identifier for a household. This identifier is hierarchically composed where the components are as follows: code for county, code for municipality, code for street, house number in the street, code for entrance in the house and a possible sub-code for other units within each entrance number. The last component is the code for the dwelling unit[18]. Table 34 in Appendix 1 shows the structure of the numeric address in technical terms.

Each component of the address code is dependent on all preceding components. The municipality identifier is a two-component number where the two first digits identify which county the municipality belongs to: thus the municipality code '0403' identifies a municipality within county '04'. For codes with values below 10 a leading zero is required. The street code is only unique in combination with the municipality code, etc. In Norway there are 19 counties, but numbers range from '01' to '20', as for a period there were 20 counties[19]. The last 5 digits of the numeric address code identify a dwelling unit (an apartment), and is only used for buildings with multiple dwellings.

The numeric address code comprises a 'household identifier', based on the idea that people who share the same dwelling unit constitute a household. In buildings with multiple dwellings, all 25 digits are needed to identify a household; otherwise it is sufficient to include only the house number.

---

[17] Strand, Børge (1996): *D-nummerpopulasjonen.* Oslo: Statistisk sentralbyrå, Notater 96/39. P. 14

[18] Central Mapping Authorities(2010): *Matrikkel/Adresse/Bolignummer*. Virtual document without pagination at www.statkart.no.

[19] County number '13' no longer exists - the city of Bergen was once county '13', but was later merged with county '12',

For my case study, only the county and the municipality identifiers are needed. The valid value range for county identifiers is from '01' to '20' and for municipalities from '0101' to '2030'. Identifiers with values outside this scope exist, but any such value is only a technical identifier in the sense that it does not identify a municipality. In some public registers a few administrative units with municipality identifier substitutes will occur. For administrative purposes it is convenient, and sometimes mandatory, to connect individuals to a geographic unit, even if this geographic unit is not a municipality in the real sense. The Norwegian Continental Shelf is an example of an administrative unit which is not a municipality, but which has nevertheless been given a numeric identifier in accordance with the structure of the municipality identifier. The Continental Shelf south of 62 ° N has the identifier '2311', and '2312' north of 62 ° N. The island of Spitsbergen also has a similar identifier, '2111', but none of these units are municipalities in the real sense. Such extraordinary 'municipalities' will occur in public information systems, but they are not classified in the 'Standard Classification of Municipalities' by Statistics Norway which is essential for my analysis, cf. the chapter 'Residential characteristics expressed by municipality classification?'.

Anyway the Continental Shelf and Spitsbergen are real geographic units, as opposed to units of a virtual nature. For technical purposes, records creators sometimes issue identifiers like '5000', '6000' or some other value needed only for technical purposes, but which are impossible to confuse with a genuine municipality identifier. Such substitute municipality identifiers are also found in a main source for the present case study, - 'The Central System for National Social Security' (*Det Sentrale Folketrygdsystem)*. There is no corresponding municipality, nor any municipality classification for such identifiers. These considerations are included in the test to deselect such substitutes in SAS data step number 12.

ORGANISATION NUMBER

Although it is not much used in the following case study, the organisation number deserves to be mentioned in this connection. The nine-digit organisation number identifies a legal entity. This is the standard identifier for legal entities to be found in various administrative and research registers, including the Employer/Employee register, cf. below. The native system for legal entities is the Central Coordinating Register for Legal Entities (*Enhetsregisteret*), which was established in 1995. The legal basis for this identity number is the Act on the Central Coordinating Register for Legal Entities[20]**.** The construction of the organisation number makes it subject to automated validity control, like the national identity number.

[20] Enhetsregisterloven. LOV-1994-06-03-15

CODED INFORMATION

As mentioned above, electronic administrative registers are characterised by an extensive use of coded information, meaning that variables are represented by codes whenever possible, and to a large extent by numeric codes. This principle has, of course continued in new generations of these systems. This is also the case for research registers designed for computer-based analyses.

The code system for given variables may vary from the simplest type comprising only a few different codes, to very complex code systems with hundreds or thousands of codes. 'Gender' is very commonly described by the codes '0' for male and '1' for female, and 'marital status' is expressed by 9 different codes. Other variables may be represented by hundreds or thousands of codes, as in the case of education[21], industrial classification, etc. Code systems often have a hierarchic structure similar to the numeric address code.

Codes are used to adapt variables for computer-based procedures, for exact categorization in research and for exact retrieval in administrative use. A query for a specific individual by national ID number is exact, as opposed to a query for the same individual by name.

Coded information also represents a major advantage for future use in research as the codes make computer-based operations and analyses possible. Some of these qualities and functionalities will be explored and demonstrated in the following enquiry.

With reference to the issue of digitally created versus computerised archives, the need for coded information in social science is obviously also present for computerised archives. After computerisation by means of transcription of paper-based sources, some efforts have been spent on converting information from transcribed to coded information: the 1801 population and housing census exists in both a transcribed version, well suited for genealogists' single-person queries, and a coded version, which was developed for more advanced, computer-based research purposes.

Computerised data were introduced and used on a larger scale in historical research in Norway during the 1970s in an extensive programme for master students in history at the University of Oslo[22]. This project is known as the 'Ullensaker project'. At that time, electronic data processing made its way into social sciences, and among history students was applied mainly to 19th-century material. This project and its methods are presented in an essay on history at individual level by Sivert Langholm[23]. For this project, nominative censuses, parish registers, emigration lists and several other sources were transcribed and then coded for computer-based analysis. A fundamental idea for the project was to combine variables from different sources with the individual as the combination unit, i.e. the entity –

---

[21] Statistics Norway (2003), *Norwegian Standard Classification of Education C751.*

[22] E.g.: Pryser, Tore (1974): *Thranittene i Ullensaker: en sosialhistorisk analyse.* Oslo, and Sande, Per (1978): *Småbrukerne i Ullensaker 1835–1865: en sosialhistorisk analyse.* Oslo.

[23] Langholm, op. cit.

the observational unit – was the individual person. By Langholm's definition this is 'micro history', and in this sense the following case study is also micro history.

A second feature of many of the data sets for the Ullensaker project was their longitudinal design. Individuals were to be followed throughout their life cycle, or parts of their life cycle, and for each individual a set of variables had to be added from various types of sources. The data sources for the Ullensaker project comprised data of a structured nature, but not standardized in terms of spelling, contents and concepts, and apparently not coded. In any event, the Ullensaker project required data sets, machine-readable to a degree and prepared for computer-based analysis.

The Ullensaker project revealed some substantial obstacles to large-scale use of computerised material. Even if the material was transcribed and to some extent machine- readable, some serious weaknesses were apparent: the lack of unique individual identifiers was obvious, and plain text information had to be coded, and hence interpreted, prior to the encoding process. As the longitudinal aspect was very much desired and intended in the project, the difficulties with identifying individuals through a variety of sources and over time turned out to be a cumbersome manual process. And even then, the identification for a number of observations was uncertain. Moreover, the risk of lacunae is generally high in longitudinal studies when based on traditional, paper-based sources. Langholm emphasises that the linkage and identification process is both time consuming and expensive[24]. This is a common obstacle for all paper-based sources, as well as an impediment to more widespread use of computerised source material. For digitally created archives with unique national identifiers this is no problem, as we will see.

Due to the resource-intensive procedures summarised above, computerised data have been 'recycled' to a certain degree in historical research. Once computerised, the same data have been utilized in several research projects over the years. One effect is that a few local communities have been very thoroughly investigated, such as Ullensaker and Rendalen (cf. below), while a majority of communities and local societies are not investigated at all. Inevitably, the frequent use and reuse of existing, computerised and encoded material exposes the data to attrition and exhaustion.

With this - hopefully – clarifying review of definitions and concepts, it is time to go one step further and introduce the case study.

[24] Langholm, op. cit., p. 260.

A CASE STUDY: SUBJECT, METHODS AND POSSIBLE DIGITAL SOURCES

The subject of the case study is 'variation in labour market participation in post-industrial Norway - in a gender, generational and spatial perspective'. The intention is to investigate this subject based on data collected from digital archives stored by the National Archives of Norway, and to analyse the subject by applying quantitative techniques, mainly linear regression analyses.

The main purpose of the case study is to serve as a demonstration and documentation of how digitally created archives can be applied in a historical research project, i.e. from an idea about an enquiry, via the heuristic phase to the establishment of a specified research data set, including historical criticism and analysis.

All electronic administrative registers are created to serve administrative purposes and functions, and use of their possible informational value in research represents an alternative and secondary use of information. Some register information may be transmitted directly into a research context, such as in demographic studies where information about year of birth, gender, migration, etc. is relevant, while other types of information need to be deduced to fit with a desired research subject, like the dependent variable for this particular enquiry which is explained below.

IMPORTANCE AND RELEVANCE OF THE CASE STUDY SUBJECT

The transition from the industrial to the post-industrial society is one of the most momentous events in most of the developed countries. This shift involved significant and permanent changes, among these women's emancipation and gender equality issues. One of the characteristics of the era is women's increasing participation in the labour market. Thus the single-income family was replaced by the dual-income family, but not at the same time and not to the same extent when compared across time and space. This transition is crucial also because it concerns the difficult combination of family life, parenthood and labour market participation on the one hand and political and institutional measures for enhancing this combination on the other hand.

For decades an important objective for Norwegian family and equalisation politics have been to facilitate equal opportunities for men and women to participate in the labour market in combination with family life, e.g. by regulating parental leave, child care facilities etc. According to the 2000 Lisbon Strategy the *'EU-member states have committed themselves to increase the labour market participation among women'*, as referenced by Tomas P. Boje[25]. The Lisbon Strategy is also followed up by the EFTA states, among them Norway.

[25] Boje, Thomas P. (2007): 'Welfare and work. The gendered organisation of work and care in different European countries'. *United Kingdom: European Review,* Vol. 15, No. 3, p. 378.

However, the transition from industrial to post-industrial societies has occupied researchers nationally and internationally – in historical, economic and social science enquiries. In a report from the Agder counties by Magnussen et. al.[26] the authors emphasize in the preface that there are large, regional issues that remain to be investigated in terms of gender equality. This view is shared by Mari Teigen[27], who identifies the regional and geographic dimension of the subject 'gender integration' as a subject which still remains to be examined more closely.

Female labour market participation is generally considered to be one of the most important indicators over time on gender equality. A common characteristic for more recent studies is that also the spatial dimension is included, either as comparative studies within countries or between countries (see the chapter 'Research findings in other comparative studies'). Objectives for the local and regional perspective is the comparative aspect - to identify and localize difference. Local and regional differences must necessarily be identified prior to more qualitative explanations.

Some common features for the more recent national and international studies of the post-industrial labour market are found in terms of data sources, in terms of method and in terms of variables: data sources are commonly digital, micro data or aggregated data sets, which are analysed quantitatively. Studies of the subject 'labour market participation' or 'labour force participation' in post-industrial societies apply largely the same variables – economic, demographic (i.e. gender, family constellation, with or without children, number and age of children etc.) and also spatial variables in comparative studies across nations and regions.

Social history as a separate historical discipline developed from the 1950s and 1960s is recognized by methods and subjects of study common with the social sciences (to some extent also physical sciences), like quantitative and computer based analyses where data sources are digitalized or digitally born. A common subject for social history that emerged from the 1970s was the specific women's history, later gender history in general and in particular gender equality issues. Demographic history, social and geographic mobility, labour market, regional and local history are all among the classic subjects for social history. These are also prominent subjects for several other social sciences like economy, sociology and statistics. It is a challenge for contemporary history that the demarcation between the sciences is washed out as the data sources as well as methods often are common. Knut Kjeldstadli characterizes social history as 'total history', including working life, the history about 'ordinary people', religious beliefs, social structures and collective phenomena[28].

---

[26] Magnussen, May-Linda, Trond Stalsberg Mydland og Gro Kvåle(2005): *Arbeid ute og hjemme: Sørlandske mødres valg og vurderinger.* Rapport fra prosjektet Likestilling og arbeidsliv på Agder. FoU-rapport nr. 5/2005. Kristiansand

[27] Teigen, Mari (2006): *Det kjønnsdelte arbeidslivet: en kunnskapsoversikt.* Oslo: Institutt for samfunnsforskning

[28] Kjeldstadli, Knut (1992): *Fortida er ikke hva den en gang var*. Oslo: Universitetsforlaget. p. 66

These characteristics are appropriate also for the case study, which is both contemporary history and social history, as it deals with typical social historical subjects and applies social historical methods. The case study, however, is strictly 'purist' in terms of digital data sources to match the main objective for the thesis, the analyses are quantitative and computer based. Incorporated in the subject for the case study are some classic Norwegian dichotomies – central versus peripheral, urban versus rural, local versus regional, and the interaction between them. Appliance of this type of data sources almost necessitates a social historical and a quantitative approach, distanced from the hermeneutic conception of history.

Data sources for the case study are definitely micro data. That does not necessarily make it 'micro history'. According to Sivert Langholm's[29] definition it is, but micro history is a wide concept, and applied on certain subjects more than on data sources and method. Arnfinn Kjelland[30] points out some of the varieties of the concept internationally, and in particular that German and Italian understanding of micro history as different from the Norwegian. A Danish definition from 1999 referenced by Kjelland is close to Langholm's definition. According to the Danish definition the objective for micro history is to uncover historical patterns and structures, as the case study also is an example of.

QUANTITATIVE ANALYSIS

A quantitative approach for the following enquiry is quite obvious, given the characteristics and qualities of electronic administrative registers described above. Regression analysis is preferable for several reasons. One reason is that through the model specification this type of analysis introduces a steering element for the research data, and thus points out a clear direction for the heuristic stage very early in the research process (see Table 2).

As regards the quantitative aspects of this case study, the implementation is mainly based on the following textbooks: *'Making History Count – a Primer in Quantitative Methods for Historians'* by Charles Feinstein and Mark Thomas (2002), *'Regresjonsanalyse for samfunnsvitere'* (*'Regression analysis for social scientists')* by Tor Midtbø (2007), and *'Å forklare sosiale fenomener'* (*'Explaining social phenomena')* by Ole-Jørgen Skog (2004).

Midtbø gives an introduction to regression analysis in social science, devoting considerable attention to the fundamental stages of the research process prior to the analysis, as well as the analysis itself. This is a tutorial for the total research process, from developing a theoretical, explanatory model, through the operationalization of variables, testing the suitability of the model when all variables have been operationalized, to finally analysing the data and interpreting results. The

[29] Op. cit.
[30] Kjelland, Arnfinn (2009): 'Norsk lokalhistorie og 'nyare' mikrohistorie'. *Heimen,* 46/2009

sequence of activities outlined by Midtbø accords very well with the main aim of this thesis: to examine how data collected from electronic registers created for administrative purposes can be operationalized and applied in historical research and analysed quantitatively. The case study basically follows the patterns outlined by Midtbø. The data applied by Midtbø for the textbook examples have been collected from contemporary and not historical sources, but the method is applicable in social science in general.

'Making History Count' by Feinstein and Thomas also deals with quantitative analytical methods, mainly regression analysis, but contrary to Midtbø, their primer is directed especially towards historians and historical research. The authors deal with fundamental concepts and quantitative method techniques. Feinstein and Thomas are more concerned with understanding and interpreting regression analysis results and focus less attention on the modelling and operationalization of variables than Midtbø.

'Making History Count' refers to a few British case studies based on historical data. The data sets for the case studies referred to by Feinstein and Thomas are data which have been computerised from a paper-based origin. For instance, the source for one of these data sets is the 1831 Census of Population, while the source for a second data set is a questionnaire-based enquiry from rural parishes in 1832[31]. In other words, the data sets for these cases are interpreted and coded versions of the primary data, i.e. they have been through a manual process to convert them from paper-based to computer-readable data. Difficulties and judgements in building the data sets for quantitative historical research are commented on only briefly by the authors[32]. A feature common to both Midtbø and Feinstein and Thomas is that they are not concerned with preparing a research data set from raw data. They apply data that have already been collected and prepared for research, i.e. in technology-dependent format, and that have previously been applied in completed studies.

Both textbooks focus extensively on quantitative methods as such, and mainly on correlation and regression analysis. Neither of them adopts a strict mathematical approach, but simply assert that this aspect of quantitative analysis is very well taken care of through the use of statistical software packages, such as SPSS, SAS (analytical tool modules, e.g. Enterprise Guide) or STATA, etc.

Ole Jørgen Skog is more heavily oriented towards causal explanations in social science. Skog also discusses various research designs – cross-sectional, time-series and longitudinal design respectively, and the advantages and disadvantages related to each of them. Both Skog and Midtbø emphasize the advantages of longitudinal design, i.e. panel design, compared to time-series and cross-sectional design. They also seem to agree that panel design is generally applied too seldom in social science and historical research, but recognize that such design is more time consuming and resource-

[31] Feinstein, Charles and Mark Thomas (2002): *Making History Count. A Primer in Quantitative Methods for Historians.* Cambridge: Cambridge University Press, p. 496.
[32] Op. cit., pp. 502.

intensive than cross-sectional and time-series design. The opportunity to make a choice between different designs, however, is not always present, as this depends on the nature of the data sources actually available. A clear conclusion based on their assessments is that a panel design should be aimed at whenever data sources allow. For the present case study the design issue is discussed below.

These three textbooks deal with analysis based on samples from populations. The three agree that the larger the number of observations, the less testing is needed; for instance, significance tests are considered unnecessary in large population samples. Total populations are probably regarded as almost unattainable in electronic form when the origin is paper-based, and experience indicates that this is likely the case. Exceptions are computerised (but not coded) versions of the 19th-century nominative Norwegian population censuses and now also the 1910 Population Census. On the other hand, when electronic administrative registers become accessible for research, access to data on total populations will be the general rule.

A distinction between prospective longitudinal design and retrospective longitudinal design is specified by Ole Jørgen Skog. A design is prospective when the observations are followed from a given starting point and forwards in time, while a retrospective design collects information about events prior to the starting point[33]. A combination of the two is quite possible, by both collecting historical information about the observations and then following the same observations a period ahead. This is quite common in social science when some information is collected from administrative registers and then combined with information from questionnaires. For my case study the design is retrospective: all information is historic and already collected in administrative registers.

Linear regression analysis requires the determination of a dependent variable (Y), and an independent variable, or a set of independent variables (X). The dependent variable is the result, the empirical situation that can be observed, and the independent variables are variables that influence the result, also called explanatory variables. With one explanatory variable the analysis is bivariate; with two or more explanatory variables the analysis is multivariate, or multiple.

Some questions that can be answered by regression analyses and that are of special relevance for the present enquiry, are[34]:

- To describe the strength and the direction of the association between the dependent variable (Y) and the explanatory variable(s) – (X).
- To decide the relative importance of different explanatory variables, i.e. decide which explanatory variables that have a strong effect on Y, and which explanatory variables that have a weak effect on Y.

[33] Skog, Ole-Jørgen (2004): *Å forklare sosiale fenomener: en regresjonsbasert tilnærming.* Oslo: Gyldendal akademisk, p. 75.
[34] Op. cit., p. 214.

- To examine whether the association between the dependent variable and a set of explanatory variables is the same in different social groups, or in this enquiry between groups locally, as far as persons belonging to one local labour market constitute one group.
- To examine whether there is interaction between the explanatory variables, i.e. whether the effect of one explanatory variable depends on which value an observation has on other explanatory variables.

Variables to be included in regression analyses must be measurable or quantitative, which in technical terms means that all variables must be numeric. Metric variables are the most precise type of quantitative variables. Each value in a metric variable is a number on which mathematic operations may be performed. For instance, it is meaningful to calculate an average for a given metric variable and the distance between each value can be measured. Any metric variable can be included in the regression analyses.

A variable type with less precision is the ordinal. Ordinals can rank classes within a variable and distinguish one class from another, but calculating an average or summarising values for ordinal variables does not make sense. Nevertheless, variables of the ordinal type can be included in regression analyses, and treated either as dummy variables or as metric variables if the number of classes is high enough.

The least precise type of variable is the nominal type. Gender is an example of a nominal variable with only two values: male and female, e.g. represented by the codes '0' and '1'. However, this variable can be redefined as a dichotome variable (i.e. converted to a binary field in the data matrix), and treated as a dummy variable in the regression analysis.

The observational unit in the following case study is the single person, which requires that all variables should be individual by nature or at least be possible to relate to an individual. Furthermore, as the longitudinal perspective is prominent in this enquiry, variables should possess diachronic qualities. The type of variables that it is actually possible to extract from the sources in question will be summarized after the operationalization process.

An example from recent years where regression analysis is applied in Norwegian historical research is Hans Henrik Bull's dissertation *'Marriage decisions in a peasant society'*[35], which is a study of the parish of Rendalen, Norway, during the period 1750 - 1900. How 'decision to marry' (Y) is influenced by 'death/retirement of parents', 'size of farms', 'working capacity of parents' (X variables) and other explanatory variables is analysed by logistic regression. Despite the difference in era, this dissertation bears some similarities to my case study as it is planned: the design is a panel

[35] Bull, Hans Henrik (2006): *Marriage decisions in a peasant society.* Oslo: Faculty of Humanities, University of Oslo Unipub.

study and the analyses are performed electronically, which requires machine-readable and quantifiable data. The observational unit is the individual person, and the observations are followed in a longitudinal perspective. Variables are added to these observations from different sources through record linkage, but by means of a manual and not a computer-based linkage, which is a major difference. Data for the Rendalen case were entered into a data base from a paper-based origin, which means that the electronic data exist only in technology-dependent format without any intervening existence as raw data. As opposed to technology-independently stored data, there is no need for a recreation process, and technical metadata are included in the software.

The Rendalen study also illustrates a major difference between computerised, paper-based sources and modern electronic administrative registers: the lack of continuous registration and incomplete registration of events as there is no nationwide collection of data where individuals may be traced across municipality borders or over time. On the whole, such data require methods and techniques to compensate for incomplete registration known as survival analyses[36].

While logistic regression is applied by Hans Henrik Bull in his study, my intention is to apply linear regression. Logistic regression is appropriate when the dependent variable is dichotome. But in my case study, the dependent variable will be a metric type variable and thus suited for linear regression, besides which the possible associations between my intended variables are expected to be approximately linear.

THE POST-INDUSTRIAL LABOUR MARKET

The terms 'labour market' and 'post-industrial' are key concepts for the following case study. In short, the 'labour market' is the marketplace for labour demand and labour supply within a given geographic area. Actors in this marketplace are those who belong to the labour force, i.e. either employed persons or non-employed persons seeking work. According to recommended definitions by the International Labour Organisation (ILO), a condition for being employed is that there is payment involved [37] – either payment earned as an employee or payment earned through self-employment. This criterion is also included in the definition applied by Statistics Norway in their Labour Force Sample Surveys. Comparable definitions are important for consistency tests and source criticism later on.

A consequence of this definition of employment is that various types of unpaid work are excluded, such as unpaid work in private households, unpaid voluntary work, education, etc. Unpaid work obviously represents an important contribution in the macro-economic picture, but almost by definition this size is difficult to quantify even at macro level. As the ILO definition will be followed

[36] E.g. Allison, Paul D. (1995): *Survival Analysis Using the SAS System. A practical Guide.* SAS Institute Inc, NC, USA.

[37] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå, p. 233.

in this case study, unpaid work is not part of the labour market concept applied in the following enquiry. Women's transition into 'paid work' is also essential in Ragnhild Steen Jensen's study[38] which is elaborated below.

The geographic aspect in the 'labour market' definition allows a focus on national, regional and local labour markets.

Labour market and labour force participation in the last few decades are frequently investigated in international studies, in particular in studies with an economic perspective. The terminology shifts between *labour force participation* and *labour market participation.* The main difference between the concepts is that people *seeking work* are included in *the labour force* concept, which is not necessarily identical to empirical labour market participation. When seeking work has resulted in actual occupation or employment, the empirical labour market participation is a fact.

In general, a 'post-industrial' society is recognised when employment in knowledge and technology-based service industries exceeds employment in manufacturing and commodity-producing industries. A side remark is that digital archives are themselves a product of such technology-based service industries. Norwegian history from the end of the 1960s and onwards is often denoted as 'the post-industrial era' - an era which implied noticeable structural shifts. Some basic development trends for the years 1962 - 1991 according to statistical figures are:

> (…) a major redistribution of employment from commodity-producing industries to service industries. (…) There has been a substantial employment increase in wholesale and retail trade, hotels and restaurants, with 110 000 new jobs, and in financial and business services with about 105 000 new jobs. The strongest increase has taken place in local government, which grew 330 000 from 1962 to 1991[39].

Employment in manufacturing, mining and quarrying and electricity industries reached its peak in the mid-1970s with 408 000 employees at the most[40]. This redistribution obviously also implied redistribution of people in terms of geographic mobility from rural to central areas. 'The drift from the countryside' was a demographic change observed from the 1950s. By and large, this was a desired development and, in part at least, an effect of politics[41].

In 1993 Dennis Fredriksen published a macro-level analysis of structural changes in the labour market in the period 1950 -1990[42]. Among the trends pointed out by Fredriksen are changes in

---

[38] Jensen, Ragnhild Steen (2004): *Sted, kjønn og politikk: Kvinners vei inn i lønnsarbeid.* Oslo: Unipax Institutt for samfunnsforskning

[39] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå, p.234.

[40] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå, table 9.6

[41] Furre, Berge (2000): *Norsk historie 1914 – 2000.* Oslo: Det Norske Samlaget, pp. 159.

[42] Fredriksen, Dennis Finn (1993): *Strukturelle endringer på arbeidsmarkedet 1950 – 1990*. Oslo: Statistisk sentralbyrå. Sosialt utsyn 1993, ch. 5.3

industrial structure and employment, from primary to secondary industries, and from the 1970s substantial growth in tertiary industries. Fredriksen points out a strong growth in the total labour force and employment from 1973 to 1981, and explains this mainly by the strong increase in female employment.

Legislation was passed and major social reforms were planned and implemented during the 1960s, and the effects became evident in the following decades. Two major reforms, the design of the new welfare state and the educational reform respectively, led to comprehensive and permanent changes in the labour market as well as in several other areas of society[43].

A milestone in the development of the Norwegian welfare state was reached when the social security solutions were joined in the form of the National Social Insurance System (*Folketrygden*) in 1967. The educational system was reformed in the same period, with a main objective of securing equal rights to higher education independently of gender, as well as of social and economic background.

Both reforms demanded new hands and new skills and generated an expanding public sector. New labour force was demanded e.g. in health care, in the educational system, and in local and central administration. There was, however, an unused labour force capacity available among the female population, and especially among 'married women', which is an often quoted postulate. According to official statistics, 'females accounted for about three-fourths of the increase in employment which occurred during the 1970s and 1980s up until 1987'[44].

General characteristics of post-industrial Norwegian society are depicted by Berge Furre[45]. According to Furre, it was assumed that married women themselves wanted paid work; they were not forced into paid work by the need to meet living expenses and the like[46]. Anyway women could meet the increased demand for labour in tertiary industries, in both the private and public sector, but particularly in the public sector. These trends have been observed through macro-level studies. A gradually improving parental leave and better child care facilities have been regarded as significant components in this development, as asserted, for instance, by Berit Gullikstad[47]. From a geographic perspective, it is important that some social benefits have been introduced simultaneously nationwide, while others have been implemented with clear local differences, e.g. as kindergarten and day-care facilities vary from one municipality to another.

Political measures have also been implemented to maintain population, settlement and employment in specific parts of the country. In Finnmark county and some municipalities in the

---

[43] Furre, Berge (2000): *Norsk historie 1914 – 2000.* Oslo: Det Norske Samlaget

[44] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå, p. 234.

[45] Furre, Berge (2000): *Norsk historie 1914 – 2000.* Oslo: Det Norske Samlaget.

[46] Furre, op.cit., p. 264.

[47] Gullikstad, Berit (2002): *Kvinnelig livsoppgave - mannlig lønnsarbeid? Kjønn og arbeid under velferdsstatens oppbygging ca. 1945 – 1970.* Trondheim: Historisk institutt, Senter for kvinne- og kjønnsforskning, NTNU

northern part of Troms county, also called the 'action zone', special benefits like extra tax deductions and extra child care benefits have existed for decades[48].

The following summary of the period from a statistical point of view, and of special interest for my focus, is taken from Statistics Norway '*Theme pages on Labour market'* dating from 2008[49]:

> Compared with other countries, a high percentage of the adult population in Norway is in employment. This is mainly due to the majority of Norwegian women being in employment. 7 out of 10 women and almost 8 out of 10 men are currently in employment. Thirty years ago, less than half of all Norwegian women were employed or actively seeking work. There is roughly the same number of employed men today as there was in the mid-1970s.

This conclusion points out that as of 2008 women had increased their participation in the labour market, almost to the same extent as men. While the participation rate for men has been fairly stable for 30 years or more, there is an increased participation among women which should be traceable by generation. It should thus be possible to quantify and measure the effect on labour market participation both by gender and by generation by means of regression analysis.

It is one thing to cite the statistical figures, but this is a description and not an explanation. Another very important consideration is that this is a description at national level. By utilising the digital archives as intended, the situation at lower geographic levels can be investigated. The statement by Statistics Norway quoted above covers the last 30 years or so, which briefly coincides with the time limitations for this case study: 1967 – 2007.

Internationally, a major finding is that female labour force participation has increased strongly in most OECD countries during the same period, but also that:

> (…) the timing of the increase has varied across countries, with some countries starting earlier (e.g. the Nordics and the United States), and in the last two decades the largest increases have been observed in lower income countries (…). However, large cross-country differences in the levels of female participation persist[50].

---

[48] Stortingsmelding nr. 8 (2003-2004): *Rikt mangfold i nord. Om tiltakssonen i Finnmark og Nord-Troms.* Kommunal- og Regionaldepartemenet. Ch. 2. Sammendrag

[49] Statistics Norway (2008): *Theme pages on the labour market.* Virtual document without pagination at www.ssb.no

[50] Jaumotte, Florence (2004): *Labour Force Participation of Women: Empirical Evidence on The Role of Policy and Other Determinants in OECD Countries. OECD Economic Studies*, Vol. 2003/2, p. 52.

A study of labour force participation in the Euro area by Balleer et al., finds that age and cohort effects account for a substantial part of the recent increase in labour force participation[51]. The cohort effects are particularly relevant for women and indicate in particular that cohorts of the 1960s and early 1970s are more likely to participate over the lifecycle. This study also shows substantial variation across countries.

Despite the increasing female labour market participation, the Norwegian labour market is still very gender-segregated: men and women work to a large extent in different branches and sectors. Systematic differences between men and women are proved by Inger Håland and Gunnlaug Daugstad in an article on the gender-segregated labour market from 2003[52]. Their conclusion is that women to a higher extent than men work in the public sector, particularly in health care and in education, and to a less extent in management. Håland and Daugstad's figures actually show that 8 out of 10 employees in health care and social services are women. This clearly indicates that the industrial structure of a local labour market must be expected to have a gender impact on labour market participation. The rate of female labour market participation might be very dependent on labour market opportunities in terms of the industries and sectors that exist locally.

The report *'Det kjønnsdelte arbeidslivet'* by Mari Teigen[53] is a review of both national and international research on gender segregation in working life and in education. Comparative studies show that the Scandinavian countries and Norway in particular are characterised by a high degree of gender-segregated labour markets. Mari Teigen claims that in recent years there has been a change towards more gender integration in Norway, a trend she explains mainly by 'women entering male-dominated areas'[54].

To a large extent, data sources for comparative studies on this subject have been labour market statistics, which limits flexibility in terms of age groups, part-time versus full-time employment, etc. Teigen's report suggests that comparative studies would benefit from individually based data sources in addition. The main focus of Teigen's report, however, is to identify special fields of research within this subject which still remain to be examined more closely. One such special field is the regional and geographic dimension of the subject[55].

---

[51] Balleer, Almut, Ramón Gómez-Salvador and Jarkko Turunen (2009): *Labour force participation in the Euro area. A cohort-based analysis.* Germany: European Central Bank - Working Paper Series, no. 1049/May 2009

[52] Håland, Inger og Gunnlaug Daugstad (2003): *Den kjønnsdelte arbeidsmarknaden.* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 6/2003

[53] Teigen, Mari (2006): *Det kjønnsdelte arbeidslivet : en kunnskapsoversikt.* Oslo: Institutt for samfunnsforskning

[54] Op. cit., Abstract, p 44.

[55] Op. cit., p 23.

In the dissertation on space, gender and politics, by Ragnhild Steen Jensen[56],women's transition into paid work over the last 30 – 40 years is investigated from a spatial point of view. Steen Jensen points out the general lack of spatial dimension in analyses of this subject, and argues strongly for including this dimension. Her solution is a micro study of four selected municipalities: Årdal, Nord-Odal, Rana and Elverum. Each of her selected municipalities has its own labour market characteristics in terms of industrial structure, i.e. whether primary, secondary or tertiary industries are predominant. Among these four municipalities, Rana and Årdal represent the most single-industry labour markets. Steen Jensen argues for including the spatial dimension because previous studies of the era and the subject deal with the concept 'post-industrial' at national level, which is very general and aggregated. The fundamental questions discussed by Steen-Jensen are whether a national labour market really exists, and whether local labour markets are just miniature reflections of the national aggregate. A central issue for the spatial perspective of my case study is that underlying the national trends there must be a number of local labour markets with different development trends, possibly differentiated by local labour market characteristics, as well as by gender and generation.

Post-war policies in Norway, in the 1950s and 1960s, prioritised rural labour markets, and the result was often a state-initiated effort to establish 'cornerstone' industries. The consequence was often single-industry labour markets, enhancing traditional, male employment. In the long-term perspective, this seemed to make modernisation more difficult, as illustrated by the Årdal community which is one of the municipalities closely examined by Steen Jensen.

Special attention has been paid to promoting settlement, industry and commerce, and employment in the 'action zone' established in 1990, and was intended to counteract negative trends in settlement, population and employment in Finnmark and the northern parts of Troms. All of the municipalities in Finnmark are included in the action zone as well as seven municipalities in the northern part of Troms. (Kvæfjord municipality in Troms county which appears in Table 29 is not among them, however). Some of the most important measures in the action zone are exemption from employers' national insurance contributions, reduction in personal taxation and an increase in family allowance – the so-called 'Finnmark supplement', according to the Ministry of Local Government and Regional Development[57]. Many of these measures are not gender-specific, or at least are not intended to be, but the consequences may very well appear to be gender-differentiating.

The quotation from Statistics Norway above manifests some basic observations about the post-industrial labour market. One basic observation is that men's labour market participation differed from that of women in terms of both a higher rate of participation and more continuity throughout the

[56] Jensen, Ragnhild Steen (2004): *Sted, kjønn og politikk: Kvinners vei inn i lønnsarbeid.* Oslo: Unipax Institutt for samfunnsforskning

[57] Stortingsmelding nr. 8 (2003-2004): *Rikt mangfold i nord. Om tiltakssonen i Finnmark og Nord-Troms.* Kommunal- og Regionaldepartementet. Ch. 1.2.

period, which also implies that a generation difference might be weak or even untraceable for the male part of the population. The Statistics Norway quotation also claims that male and female labour market participation has become more similar in the period, indicating that there has been a generation difference between women in terms of labour market participation. The labour market behaviour of younger generations of women differed in the form of a higher rate of participation and continuity in the labour market, while older generations of women had a lower rate of participation and less continuity. However, there is reason to investigate whether this development was uniform across the country.

HYPOTHESES ABOUT THE CASE STUDY SUBJECT

A summary of the most prominent characteristics of the period in question must be made quantifiable to be suitable for regression analysis. Although the national picture seems to be a steady development towards a post-industrial labour market and increased female employment, there must be regional and local variations.

An introductory key word for this enquiry is 'measurement'. More specifically, this subject implies 'variation in rate of labour market participation' during the observation period, which might be associated with gender, generational and regional qualities. My intention is to develop valid measurements for these variables at the individual level. The question to be answered by the regression analyses may be summarised as follows: what explains variation in labour market participation in post-industrial Norway at national, regional and local level?

An operationalized concept for the result – the dependent variable – is the individual *rate of labour market participation* – for which the short term *employment history* will also be applied. Furthermore, the subject obviously requires *gender* and *generation*, or *birth cohort*, as explanatory variables. By specifying birth cohorts and an observation period for each birth cohort, a dynamic and changing historical background will be captured by the panel population. Obviously, the 1937 generation had poorer welfare services during their observation period than the 1958 generation, in terms of paternal leave, child care, education opportunities, etc. In other words, since opportunities were different in the 1970s, the 1980s and the 1990s, different generations would benefit differently from the development of the welfare state. Various child care services in both a generational and a local perspective also list *family obligations* as a desirable, explanatory variable. *Geographic mobility* is another desirable explanatory variable. Finally, an explanatory variable that can express characteristics of a residential nature, so far classified as *residential qualities,* should be examined.

The regression analysis produces compact measures in terms of coefficients. A major objective for my case study is to compare regression coefficients between local labour markets. The

regression analysis applied at municipality level is an instrument to identify and localize the most dissimilar labour markets expressed by such coefficients. For this purpose, a population and a set of variables must be materialised.

Some of my intended variables are straightforward and possible to establish directly, like *gender* and *year of birth*, while others must be deduced and established indirectly, like *employment history*. Variables of the latter type are very experimental, and their usability will be uncovered and discussed during tests and controls.

In brief, post-industrial Norway refers to the period from around 1970 onwards. In any event, variation in the individual *rate of labour market participation* is the result, and hence the variable to be explained: the Y variable, or the dependent variable, in the regression model. Variables which influence and possibly explain the result are the X variables in the regression model.

Table 1. Hypotheses about variation in labour market participation in post-industrial Norway.

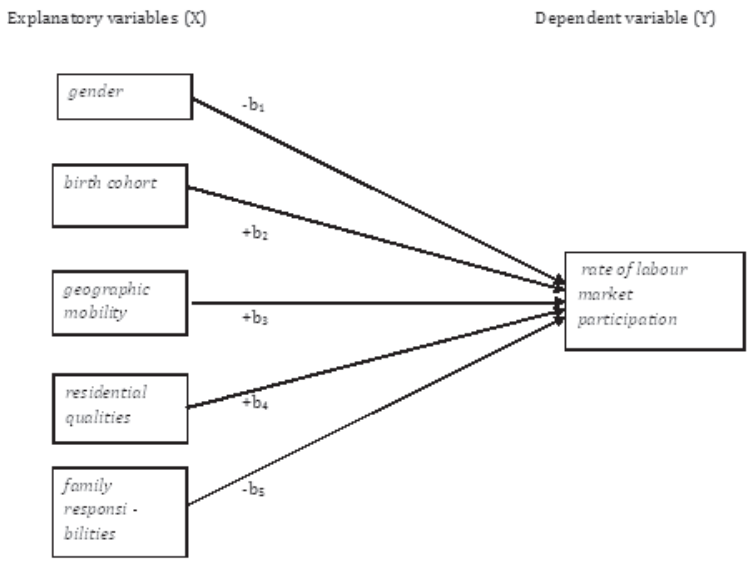
The null hypothesis	**Alternative or research hypothesis**
Gender had no influence on the rate of labour market participation	*Gender made a difference: women and men had different rates of participation in the labour market – male accounted for a higher rate of participation than female*
Generation had no impact on the rate of labour market participation	*Generation made a difference: younger generations had a higher rate of participation than older generations*
Geographic mobility did not affect the rate of labour market participation	*Geographic mobility made a difference: a high rate of mobility accounted for a high rate of labour market participation*
Residential/spatial qualities and characteristics had no effect on labour market participation	*Residential/spatial qualities and characteristics made a difference: residence in central, urban areas led to higher participation than residence in rural areas.*
Family responsibilities did not affect the rate of labour market participation	*Family responsibilities made a difference: heavy family responsibilities led to a lower rate of employment*

At this point, the subject may be summarised as a systematic set of hypotheses about labour market participation in the post-industrial period. The hypotheses are preliminary answers to the question 'What explains variation in labour market participation in post-industrial Norway, at the national, regional and local level?' The analyses will later show whether these preliminary answers are

sustainable when tested on actual data. The null hypothesis and the alternative hypothesis are stated in the past tense, all with reference to the period of measurement 1967 – 2007**,** as presented in Table 1.

Based on what has been outlined above as regards women's increasing employment, the null hypothesis may perhaps seem a bit improbable. However, it is necessary to bear in mind that these trends have been observed at the national and aggregated level, apart from a few local studies. My assumption and focus is that the trend cannot have been similar in all local communities; the pace of development must have been different and the extent and rate of female employment must have differed during the years in question, 1967 - 2007. Such differences will be revealed through analyses at the local level.

Figure 3. Model for explaining variation in the rate of labour market participation.



For the next step, a model is required; see Figure 3. The 'b1', 'b2', etc. in the model are the coefficients, to be replaced with actual numbers through the analysis. The signs + or – indicate whether the effect is expected to be positive or negative. A positive effect means that the X and the Y variables move in the same direction, while a negative effect means that the X and Y variables move in opposite directions. The arrows in the model indicate the anticipated direction of the influence of each of the X variables. The result, the dependent variable, is the individual rate of labour market participation which is assumed to be influenced by factors such as gender, which birth cohort a person

belonged to, qualities specific to the municipality where a person lived, i.e. *residential qualities*, by geographic mobility and the kind of family responsibilities a person had.

For some of these hypotheses, opposite effects are conceivable for men and women. For instance, generation and family responsibilities may have opposite effects for women and men, but the total effect is still expected to be negative for the two genders jointly.

One condition for linear regression is that there should be no explanatory variable that causes the dependent variable (the Y variable) and at the same time is correlated with any other explanatory variable (X variables), i.e. confounding variables. Anyway, the model is less vulnerable to confounding variables when it is expanded with more explanatory variables to create a multivariate model.

The model is constructed for computer software analyses, but when defining the model there is a clear distinction between the responsibility of the human mind, and what can be left to the software. The direction of causality depicted in the model, visualized by the arrows, is based on logical expectations and empirical observations; no software is able to detect causality. *Gender* may very well cause variation in *labour market participation*, but the reverse is unthinkable. Variation in *labour market participation* cannot cause *gender*. Furthermore, the effect of *gender* on the rate of *labour market participation* is assumed to be negative because women's participation in the labour market has been different – lower and less continuous – from men's participation in the period in question. Only towards the end of the period can an equalisation be traced, but this is assumed to be too late for the effect to be noticed as not measurable or positive in the model. The arrow in the model indicates a one-way effect by *gender*, and the sign indicates that this effect is negative. The intention is to measure this effect at the national, regional and local level.

Generation may cause variation in *labour market participation,* but the opposite causality is not conceivable. My expectation with regard to the effect of *generation* is the younger the generation, the higher its rate of participation in the labour market. This effect is expected to be stronger for women than for men, but men and women must be measured jointly, and the effect is assumed to be strong enough to be measurable.

Only a one-way direction of causality between *family responsibilities* and *variation in labour market participation* is reasonable: *family responsibilities* operationalized by *number of children* can cause *variation in labour market participation*, but the opposite causality is not possible. A negative effect is expected for *family responsibilities*. The anticipated effect is based on the concept that the 30-50 year age groups are particularly squeezed in the difficult combination of family obligations and employment. It should also be borne in mind that the observation period covers the years from 1967 to 2007 where child care benefits were more strongly upgraded only towards the end of the period. Thus a high rate of *family responsibilities* is expected to be associated with a low rate of *labour market participation.*

Sometimes the direction of causality can be quite difficult to determine. A dubious causality must be considered in the model when it comes to the direction of the effect of *geographic mobility* and of *residential qualities,* respectively, on variation in *labour market participation*. An example of a dubious causality is cited by Feinstein and Thomas[58] in one of their case study examples: this example was collected from a study by Benjamin and Kochin[59] on effects of unemployment benefits on unemployment in inter-war Britain. Their model assumes that 'generous payments' of benefits to the unemployed caused the high level of unemployment in inter-war Britain. But an opposite direction of causality is actually conceivable, i.e. that a high level of unemployment would be the cause of high unemployment benefits. Their discussion includes empirical data and tests by alternative models which ultimately lead to the initial assumption about causality being maintained.

A two-way causality is conceivable between *geographic mobility* and variation in *labour market participation.* High mobility is assumed to facilitate labour market participation due to the general redistribution of the population from rural to central areas in the period, which is also assumed to be a movement towards improved labour market opportunities. This issue is investigated by Kjetil Sørlie in a study on the recruitment of women to coastal and rural Norway[60]. This study concentrates on coastal and rural areas of Norway, and on female migration in particular, but a main question is to what extent migration is motivated by job options and the labour market. Sørlie's conclusion is that an important motivation for migration is labour market opportunities, although in combination with other factors. The sequence of events empirically is that migration comes prior to labour market participation, i.e. mobility is the cause of labour market participation in the empirical sense and not vice versa, and the direction of causality is one way. In the model, *geographic mobility* is expected to have a positive effect on *labour market participation*.

Sørlie's study is a panel-designed study where selected birth cohorts are followed from the age of 15 to the age of 30. It is well documented that younger people have a higher mobility than older people. As my panel population is followed from the age of 30 to the age of 50, an implication might of course be that migration frequency is past its peak for this population, which might make variation difficult to measure.

Between *labour market participation* and *residential qualities* there is obviously interaction and mutual influence. A two-way causality between general labour market properties, including labour market participation, and *residential qualities* is apparent, but between *residential qualities* and variation in *labour market participation* as an individual variable I assume a one-way directional

---

[58] Feinstein, Charles and Mark Thomas (2002): *Making History Count: a Primer in Quantitative Methods for Historians.* Cambridge: Cambridge University Press, p. 438.

[59] Benjamin, Daniel K. and Levis A. Kochin: *Searching for an explanation for unemployment in interwar Britain*, Journal of Political Economy, 87, 1979, pp. 441 – 478.

[60] Sørlie, Kjetil (2000): *Rekruttering av kvinner til kyst- og bygde-Norge. Sammenhengen mellom innflytting, jobbtilknytning og familieforøkelse.* Oslo: Norsk institutt for by- og regionforskning. Notat 2002:114.

causality. To defend the one-way direction of the causality in my model, chronology is decisive. The possible effect on *labour market participation* has to be measured with a delay in relation to the status of *residential qualities*. This problem has been taken into account during operationalization.

I assume a positive effect of *residential qualities,* i.e. qualities in terms of industrial structure and other characteristics developed over a period of years for a given local labour market. Local labour markets with a high degree of centrality and service industries are, for instance, expected to promote a high rate of labour market participation, while low centrality and a high proportion of primary industries are expected to be associated with a low rate of employment, especially female employment. The general structural changes that took place over the years in question were changes in industrial structure and employment, from primary to secondary industries, and then to tertiary industries, and population redistribution from rural to central areas.

Electronic administrative registers with national coverage of course invite to including the spatial dimension in the analysis. In fact, the data sources I have identified as relevant for my case study allow a study of all local labour markets, not just a selected few. A major advantage of full-scale data sources is that the spatial dimension may be investigated at a very low geographic level, such as municipal level. This is not possible with a sample population, or with paper-based sources, albeit for different reasons.

With the type of data sources available for my subject, there are mainly three ways to design the data for the analysis:

- as cross-sectional data
- as time-series data
- as longitudinal data

Both cross-sectional and time-series designs have been applied in studies of the subject, but studies based on longitudinal data and on a full-scale population are mainly untried. My intention is to build a data set with longitudinal data, with a defined population and a defined set of variables. This type of data set is sometimes called ‘longitudinal data’ and sometimes ‘panel data’. This is actually the same concept. In the present context, the concepts are used synonymously.

Whatever the design that is decided for the enquiry, there are no data already existing and ready for use, but when based on electronic administrative registers, it is in fact possible to assemble and adapt data for all three types of design. Electronic registers make all three types of design possible; for this particular study, there is a genuine choice of design. Based on the design discussion by both Midtbø and by Skog referred to above, my conclusion is simple: the longitudinal design offers advantages that the other design types do not possess. Only a longitudinal design ensures the

diachronic axis – the time dimension of the phenomenon in question. Besides, level fallacies are easier to avoid with panel data.

The foundation for the analysis in the form of a panel population with specified variables must be built brick by brick, from the most basic level available for each type of data source. This is a process that requires a deep plunge into the technical metadata to identify possible sources: records creators, data sets and variables to be operationalized for this particular context. Apart from the technical aspect, however, this part of the process has similarities with a process based on paper-based sources.

A research data set – a data matrix - will form the basis for the analyses, and it has already been determined that the entity in this data set is to be the individual person. Table 2 is a sketch of the data matrix which is materialized later on (see Table 37). The personal identifier (the national identity number) is needed as the linkage key to build the data matrix, and to derive gender and birth cohort, but as soon as the data matrix is finished and no further matching is required, the national identifier will be replaced with a neutral counter as the primary key.

The variables in the data matrix are explained, one by one, below. The data sources referenced in the table, the Central System for National Social Security (DSF) and the Tax Register for Personal Taxpayers (TRP), are also presented in further detail below. In this table sketch, the dependent variable appears in only one alternative. As outlined below, this variable will be deduced from labour income amounts from the Central System for National Social Security income tables. Actually, I have experimented with three alternatives to test which level of income – income limits applied during construction of the variable – would be best suited for the intended measurement. After testing the three alternatives, a conclusion was reached as to which of these alternative limits was ultimately to be applied for the analysis, as explained below.

The value range for each of the variables in Table 2 is partly predictable due to the nature of the variable, and partly unpredictable, which means that the value range can only be observed as actual results in the final data matrix, which is the case for some of the figures inserted in the table. It is also apparent that the value range for some of these variables clearly differs between national and local level. This is hardly conceivable for the dependent variable, but geographic mobility obviously does not range to 16 in each municipality. This is also the case for family responsibilities where the maximum value of 12 children is only counted for a few observations. For birth cohort, the value range is determined by the number of cohorts, but there is a risk of low, or even missing. representation of observations for one or more cohorts in municipalities with a very small number of residents.

Other variables that would have been very desirable to include in the data matrix are education (level and type) and profession. Possible sources for profession would be the Employer/Employee System administered by the Norwegian Labour and Welfare Service, and also the census data bases.

Statistics Norway administers a research register on the highest completed education of the Norwegian population, but access to this register, as well as access to the census data bases, is restricted by the Statistical Act. Access to data collected under the Statistical Act is more restricted than access to data collected for administrative purposes. Under this Act, data may be released 100 years after data entry.

Table 2. A sketch of the data matrix.

Variable	Data source/origin	Note
Primary key: Counter that equals the number of observations	*Generated by a SAS procedure.*	*A neutral counter which makes individual identification impossible replaces the original primary key*
Gender	*Deduced from the national identity number in the DSF population table*	*Defined as a dummy variable with codes 0 for male and 1 for female*
Birth cohort	*Deduced from the national identity number in the DSF population table*	*Coded from year of birth. 1937 = 1, 1938 = 2 etc.* *Values will range from 1 to 22*
Labour market participation/ employment history	*Deduced and coded from the DSF income tables*	*The dependent variable:* *values will range between 0 and 20*
Residence characteristics 'year 10'	*Transferred from Standard of Municipality classification by Statistics Norway*	*Values will range from 1 to 7*
Residence characteristics 'year 20'	*Transferred from Standard of Municipality classification by Statistics Norway*	*Values will range from 1 to 7*
Geographic mobility	*Deduced and calculated from the income tables in the DSF.*	*Actual range of values from 0 to 16*
Family responsibilities	*Deduced from TRP - 'Number of children'*	*Actual range of values from 0 to 12*

A general compensation for all confounding variables is to consider the results of the following analysis in the light of figures from relevant official statistics, e.g. statistics on education and on migration.

Variation in *labour market participation* is obviously also influenced by trade cycles and economic trends that affect offer and demand. These are macro forces which cannot be transformed

into individual variables. However, the construction of pensionable income and benefits in cases of unemployment reduces the impact of unemployment; see below regarding 'basic amount'.

REPRESENTATION OF THE CONCEPTS FROM THE THEORETICAL MODEL

The theoretical model in Figure 3 raises the question of how these concepts can be adequately represented in the data matrix and hence endeavour that the results of the analyses are as close as possible to the intention behind the concepts. In the end only one variable can represent the single concept, but some comments on the operationalised conceptual contents are required. The following review deals with concepts where alternatives exist for representation, or a further clarification is needed.

There is no doubt that somehow family situation and family obligations for men and for women, have influenced individuals', families' and households' attitudes, decisions and behaviour towards labour market and labour market participation in post-industrial societies. Women are commonly characterized as an unused and available labour force resource when the growth in service industries accelerated from the end of the 1960s and early 1970s. Both nationally and internationally historians and others that have been occupied with the post-industrial era have emphasized that the new jobs in the service sector were occupied by women, and to a large scale by 'married women'. It is less well documented whether the same women – i.e. the same individuals – had a career with paid work before they married and possibly re-entered the labour market after a period of non-paid work since marriage.

Berge Furre characterises the 1950s as the 'good housewife era', when the nuclear family had a strong position, and number of divorces were few – around 10 per cent[61], as compared to around 45 per cent in 1990[62]. The single-income family was the norm, and based on the male breadwinner. A female breadwinner on the other hand, would probably not have been sufficient to support a family, due to the generally much lower payment for women. Anyway post-industrialism implied a permanent shift from the single-income household to the dual-income household in most western countries. This is an international picture, frequently investigated and analysed, cf. the chapter 'Research findings in other comparative studies'.

The emphasis on 'marital status' is probably based on an assumption that implicit in 'being married' are commitments, traditions and expectations which affect attitudes and behaviour towards paid work that differ between married and unmarried persons, men and women, but basically to

[61] Furre, Berge (1992): *Norsk historie 1905-1990. Vårt hundreår*. Oslo. Det Norske Samlaget. Pp 280.
[62] Statistisk sentralbyrå (1995): *Historisk statistikk 1994*. Oslo: Statistisk sentralbyrå, p. 57.

capture the obligation of parenthood. In the report from 2005 Magnussen et. al.[63] study the systematic deviance between men's and women's participation in several fields of society, including the labour market, that can be observed in the Agder counties (cf. below). The authors emphasize that their focus on women is primarily on mothers, (i.e. regardless of marital status), based on the idea that the gender related imbalance in division of labour is most distinct for families with children, in particular with small children.

The formal 'marital status' includes 'unmarried' (i.e. pre-marriage), 'married', 'separated', 'divorced' and 'widow/er', and equivalents for partnership since 1993, and in a longitudinal study it is important to remember that marital status is not static. Over a twenty year period marital status will change for many people. If we accept the concept 'unmarried women' from the 1950s an 1960s in the sense of 'pre-marriage and pre-children', unmarried women necessarily had a closer connection to the labour market when they had to be self-supporters. But according to norms of the 1950s and 1960s women left paid work to unpaid work in the home sphere after marriage, and relied on economic support, as the rest of their family, from the male bread-winner in a single-income household. When married women started to enter the labour market on a large scale to paid work in the labour market, this is a persistent mental and cultural shift as well as an economic shift, much more complex than what may be expressed by the 'marital status' at a given point in time.

What is imperative for *family responsibilities* in the case study, is to capture the implications of child births and child care, social expectations, role expectations, tradition and culture for parents - over time – independent of the cross-sectional marital status. Besides the concept must represent the total panel population – men and women – of all birth cohorts throughout the entire observation period. For a number of observations marital status would change during the observation period, also raising the question about which marital status that would be the most representative for the observation period – if marital status were to be selected as representation.

Furthermore 'marital status' would gradually be biased over generations, cf. the growth in cohabitation without formal marriage, see page 90. Younger generations are more inclined to be cohabitants - both with and without children. Given the inclination of participating in the labour market by younger women, (the *generation* effect, cf. below) cohabiting couples constitute de facto dual-income households, but based on formal marital status they would erroneously be classified and treated as single persons in single-income households, and besides not possible to separate from de facto single persons who live in a single-income household. There is also evidence that the practice

[63] Magnussen, May-Linda, Trond Stalsberg Mydland og Gro Kvåle(2005): *Arbeid ute og hjemme: Sørlandske mødres valg og vurderinger.* Kristiansand: Rapport fra prosjektet Likestilling og arbeidsliv på Agder. FoU-rapport nr. 5/2005.

and extent of cohabitation without formal marriage differ between regions[64] , and thus 'marital status' would have a biased impact in the regional analyses.

My conclusion is that 'marital status' is not an optimal representation for *family responsibilities* as intended in the theoretical model. *Number of children* is considered to be a better representation, both because this is a variable with better diachronic qualities than 'marital status', and also because this captures the wider contents of *family responsibilities* – for both men and women.

The choice of *number of children* as the most appropriate representation for *family responsibilities* in a longitudinal study is supported by research nationally and internationally, especially in enquiries from more recent years. 'Marital status' is a formal property and thus found as a separate field in many civil registry systems. Generally fields that are already defined and existing in data sources are convenient to apply as variables in research. It is more time consuming to define and derive more complex, but possibly better, alternatives. Methods, and not least digital data sources have developed over the years and thus enhanced new measurements.

Several international enquiries from the last decades explore the association between women's labour market participation and whether they have dependants, especially number of children. In a comparative study of 14 EU-countries based on the European Labour Force Surveys from 1992 and 2005, Olivier Thévenon[65] investigates women's increased labour force participation and finds that *'Whatever the country, women are more likely to be inactive if they have a child, and this probability increases with the number of children'* . Moreover Thévenon finds that the third child had a stronger effect than child number one or two on the degree of employment. Having a third child would make women more reluctant to re-enter the labour market[66].

Vlasblom et. al.[67] deals with effects of educational level, and family composition in terms of number of children on decisions to participate in the labour force in six EU-countries[68]. They investigate the causes for differences in female labour market participation for the years 1992 – 1999. One of their findings is that there is an effect of children – number of children and age of children - on women's decisions to participate in the labour market. The effect - which is actually called the 'child penalty' – means that women are inclined to reduce participation in the labour market due to presence of children.

---

[64] Ellingsen, Dag og Ulla-Britt Lilleaas (2014): *'Noen vil ha det slik'. Tradisjonelle kjønnsroller og svake levekår på Sørlandet.* Kristiansand: Portal forlag, p 57.

[65] Thévenon, Olivier (2009): 'Increased Women's Labour Force Participation in Europe: Progress in the Work-Life Balance or Polarization of Behaviours?' *Institut national d'études démographiques (INED), Paris*

[66] Op. cit. p. 244.

[67] Vlasblom, Jan Dirk and Joop J. Schippers (2004): 'Increases in Female Labour Force Participation in Europe: Similarities and Differences'. *European Journal of Population 20*: 375–392 and Thévenon, Olivier (2009): 'Increased Women's Labour Force Participation in Europe: Progress in the Work-Life Balance or Polarization of Behaviours?' *Institut national d'études démographiques* (INED), Paris.

[68] France, West Germany, Italy, the Netherlands, Spain and the UK.

In the chapter 'Hypotheses about the case study subject' a possible generation effect on labour market participation is assumed. Apparently generation can be represented by consecutive birth cohorts or by age groups. 'Age groups' are basically aggregated data found in cross-sectional and time-series based enquiries. Even age groups with five year intervals are coarse, compared to consecutive birth cohorts which are continuous as one-year age groups. The generation concept should capture the diachronic axis – e.g. the 1940 cohort is compared at any age between 30 and 50 with e.g. the 1950 cohort at any age between 30 and 50 – like all the rest of the cohorts. The complete series of birth cohorts form the basis for the panel population, and represents as such a continuum which age groups can not represent. Cohorts are step less age groups and in a panel based study these cohorts are observed continuously over years, see also the reasoning about design of the study from page 36. The intention is to capture true generation effects, i.e. whether individuals belonging to one birth cohort act differently from other birth cohorts throughout the period they are observed, rather than life stage effects, i.e. that all cohorts behave the same way when they are in the same life stage and age which is measured when age groups are compared.

Operationalising the *gender* variable as a person's biological sex happens to be a technical routine based on information in the national ID-number in this case, but the gender concept deals with gender inequality in a broad sense. *Gender* incorporates the idea that there is a gap between men and women in terms of power, economy, dependency, opportunities etc. *Gender* is the term used to theorize the issue of sexual difference as defined and discussed by gender theorists, as e.g. put forward by Jule Allyson[69] *'(...) gender is a category where masculine and feminine are understood as behavioural patterns and part of a gender system'*. Historian Joan W. Scott[70] has dealt with the historiographical emergence of ' women's history' and 'gender history' – the latter is generally seen as a continuation of 'women's history' - as separate historical disciplines. She states that 'gender means knowledge about sexual difference'. From her historiographical point of view she separates between a descriptive use and an analytic use of the term. According to Scott 'gender' has been applied synonymously for 'women', e.g. that in stead of 'women' in the title of a book or in an article about women's history, 'women' is simply replaced by 'gender'. Scott regards this as an attempt to make it more neutral and create a distance from 'feminism'. It is noticeable that 'gender' in analytic use is not only 'women' or 'female', but both men and women and the disparity between them in a given context.

In the context of labour market participation *gender* is used to illuminate one of the most prominent features of the history of the last decades: the transition from industrial to post-industrial society and possible implications for the balance between men and women. *Gender* as applied in the

[69] Jule, Allyson(2014): *Gender Theory*. Trinity Western University, Canada
[70] Scott, Joan Wallach (1991): Women's History. In: *Peter Burke (ed.) New perspectives on historical writing.* Polity Press

case study is based on the idea that there are several features for men and women that are different - and above all – unequal and imbalanced in women's disfavour. *Gender* is an explanatory variable in the theoretical model – a variable which also proves to be the most important among other explanatory variables when applied on empirical data. Participation in the labour market is commonly seen as an important way of empowering women. When women entered the labour market to a large and permanent scale, this represented a major change for the balance between genders - in the single household, in the labour market and for the gender equalisation issue on a broad spectre.

Women's participation in the labour market are politically desirable and regarded as promoting gender equality. It has been an important objective for Norwegian family- and gender equality politics, to facilitate combination of family life and professional life for both men and women. Gender related differences must be thoroughly considered during construction of the dependant variable, in particular when setting an income line for pensionable income as detailed in the chapter 'The dependent variable: labour market participation'. The representation for *labour market participation* or *employment history* should reflect whether the 'main source of subsistence' is income from own work or not, as discussed in the same chapter. Self-supportiveness is important in the gender equality perspective, which is why the income line had to exceed a level clearly above 0.

The spatial or geographical perspective is essential for the case study, as elaborated above. One aspect of the municipality emphasis is that the municipalities are themselves employers, and often the major employer in smaller societies. But 'geography' is a comprehensive concept which is not quantifiable and hence not applied as a variable in the analyses. In stead several variables are included in the data matrix to reflect the spatial aspects of the study. These are *geographical mobility* and residential qualities operationalized by *municipality classification codes* as explanatory variables. The municipality classification codes serve a double purpose – also as an auxiliary variable for splitting the data matrix for the class wise analyses. Furthermore the municipality identifiers are also required for several purposes: giving each observation a local, geographic connection which is required for the analysis below national level, it serves as a criterion for cleansing the panel population and functions as auxiliary variables for splitting the data matrix for the municipality wise analyses.

The spatial perspective implies analysis on different geographical levels: the national, the county and the municipality level. The theoretical model may then be applied on all these levels.

A FRAMEWORK FOR THE PROCESS OF IDENTIFYING POSSIBLE DATA SOURCES

The conditions now determined for the case study through choice of regression analysis, panel design, observation period and research population, also forms a framework for possible data sources to apply. Data sources must be available for research in compliance with legislation, they must cover a

population which allows the defined birth cohorts (1937 – 1958) to be extracted as complete cohorts, and the dependant variable and the explanatory variables must cover the total observation period (1967 – 2007). A general requirement for the variables is that they must be comparable and consistent over these 40 years, in other words - what to look for, are data sources where the same kind of variables are found for the total research population throughout the entire observation period.

Based on what was outlined in the chapter 'Definitions and concepts' about groups and subgroups of digital archives the heuristic focus must be concentrated on systems belonging to main group 2, and in particular subgroup 2A. All data sources must be found within the National Archives' collection of digital archives, and preferably from sectors of society that fits with the case study subject, i.e. data sets from other sectors of society such as justice, health, agriculture, transportation, fisheries e.g. are not relevant.

The entity in possible data sources must be the individual person or other entity which is possible to link to individuals, and thus aggregated data sets are not relevant. Any anonymised data is out of the question as record linkage by the national ID-number is a basic condition for assembling the variables with the individual as the combination unit in a data matrix. Only data sets with the national ID-number as well as the municipality identifier intact can be considered. The national identifiers as reviewed above, normally functions as the primary key as well as the personal identifier in administrative registers. A fundamental principle for the National Archives of Norway is that any table extract with individual data must be preserved with the primary key in the original form, and not replaced by e.g. encrypted keys - see also the comments about authenticity and encryption above, and the chapter 'Authenticity and reliability'. Sample populations may have the national ID-number intact, but can not give analyses on low geographic level and are out of the question as possible sources for the panel population.

Another fundamental criterion for the planned case study is how the contents of the fields, i.e. variables, is represented in the data sources. For this particular study only field contents represented in structured and machine readable format is relevant. The analysis in question requires quantitative variables which mean that data sources must contain fields that are numeric, either in their primary form, or fields that may be derived as, or converted to numeric format. Hence fields with unstandardized, free text contents are not relevant.

Given the collection of digital archives of the National Archives of Norway - the basic issue is actually how much and what kind of data sources are there to search among? The general practice by the National Archives of Norway differs for the categories of digital archives which are outlined above. Preservation of systems in main group 1A is mandatory according to regulations[71]. The entities in systems of this category are basically 'case' and 'document' – i.e. case handling documents

[71] Forskrift om offentlege arkiv, FOR-1998-12-11-1193, § 3.20

belonging to each case. Anyway the entity in these systems is not the individual person, and the question of national id-numbers e.g. is irrelevant for this category of digital archives.

When it comes to systems belonging to category 1B, the general practice is that such systems are not preserved according to general regulations by the National Archivist. (However, these regulations are updated and replaced by a new set of regulations from 2014). But so far systems from this category are not prioritised for preservation. Besides there is reason to believe that information relevant for the case study would not be found within this system category.

The issue of preservation of subject-specific systems (main group 2) is always based on individual appraisal and final decision by the National Archivist, cf. also Bevaringsutvalget (above) and its recommended routines for the appraisal process. An important group of systems within category 2A are registers with national, basic data, i.e. data about persons, real estate and legal entities respectively. However, these basic data registers were not established at the same time: The Central Population Register as mentioned, was operative from 1964, the Cadastre from 1980, and the Central Coordinating Register for Legal Entities from 1995. A general problem is therefore that the basic registers as data sources do not cover the entire observation period for the case study. See also the assessment of the Employer/Employee Register in the chapter 'Residential characteristics expressed by municipality classification?'

The basic registers supply numerous other registers, databases and administrative systems both in public and private sector with basic data, cf. data exchange and reuse of data in the chapter 'National identifiers and coded information'. Thus central demographic variables from the CPR, key variables about real estate and addresses from the Cadastre, as well as 'business demography' from the Central Coordinating Register for Legal Entities are found in numerous national and local systems today, but only from the time each of the systems were introduced.

Transferred or deposited systems to the National Archives of Norway in subgroup 2B are very few in numbers. Major deposits in this sub group are the population and housing censuses. Systems from category 3 – 'Other systems' are not considered relevant for the case study as accessions are few, these are from the most recent years, and they are not machine readable in the sense required for the present study.

As methods for preserving digital archives have developed, there has been increasing focus on the importance of technical metadata for the archival packages from subject-specific systems. Technical metadata functions as catalogues for the archived systems, cf. e.g. Appendix 2. The earliest transfers to the National Archives accepted the records creator's own technical metadata which was normally internal, i.e. understandable only for those who were familiar with the 'tribal language' – and not easily accessible and understandable for external users, see example in Appendix 5. Though technical metadata for several of the earliest acquisitions have been standardised (cf. ADDML below) in posterity during the latest years, there are still data sets with insufficient technical metadata among

the oldest acquisitions, cf. the population situation files from Statistics Norway from 1964 – 1972 as elaborated below. A consequence is that some transferred data sets may be difficult to apply, due to insufficient technical metadata.

Among possible data sources for the case study are Statistics Norway's population and housing censuses which exist in digital form since 1960. The National Archives have received deposits of the following census data bases: the 1960, 1970, 1980 and the 1990. Desirable fields for the case study from the census databases would have been 'profession' and 'number of children'. But censuses are only periodical – basically carried out in ten year intervals, and the diachronic qualities of the censuses suffer from this, especially compared to continuously updated and accumulating systems.

Another problem with the censuses is that definitions have changed over years which make comparison from one census to another difficult, see e.g. the various definitions of 'employment' as discussed in the chapter 'The dependent variable: labour market participation'. Besides censuses have a stricter legal protection for secondary use than administrative registers because data is collected pursuant to the Statistical Act, cf. below. All census databases in the custody of the National Archives of Norway have a deposit status. The conclusion is that the census databases are inaccessible for this particular case study for formal reasons, and besides there is reason to question the comparability over years of the desired variables, though this can not actually be tested.

To summarize – a lot of records creators and a lot of system categories can be eliminated very early in the heuristic process, but still there are sufficient records creators and data sources to search among.

HEURISTIC CONSIDERATIONS – WHERE AND HOW TO IDENTIFY DATA SOURCES?

When the subject of a historical research project is undergoing the transition from idea to substance, the historian will have an opinion as to which time span and sector of society possible sources for the subject can be found. The purpose of the meticulous categorisation and description above is linked to the fact that any historical research would start by looking for relevant source material, and a systematic classification of possible sources is always helpful, but even more helpful and essential for digitally created sources.

Obviously, the informational value in electronic administrative registers and electronic research registers represent a large potential for quantitative as well as other types of historical research. For instance, such registers are all suitable for cross-sectional analyses. However, they share this quality with many other digital sources, such as sample survey data or data without identifiers for matching (mainly subcategory 'Research registers'). Suitability for large-scale longitudinal studies is a quality that nearly all administrative registers possess, which is not always the case for research registers and for sample populations in particular. One basic requirement for the design of this

particular case study is that possible data sources must have a potential for computer-based record linkage. This immediately brings to mind administrative registers with their national identifiers. Due to their extensive use of coded information, they are also very well suited for electronic processing and computer-based analyses.

The process of identifying relevant records creators, possible electronic sources and data collection for a history research project normally has to be conducted through an archival institution. The storage of transferred or deposited material is more centralised in the case of digital archives than for paper-based archives. For the present enquiry, all data sources are in the custody of the National Archives of Norway. It is normally not possible to obtain any digitally created source material directly from a given records creator for a given research project, but the consent of a records creator may very well be required (such as for my access to the Central System for National Social Security). Besides, the technical metadata as standardised[72] by the National Archivist of Norway are much easier to understand than any records creator's internal technical metadata.

The transfer and deposit of digital archives from the central government administration to the National Archives of Norway have been carried out for a few decades, but for the purpose of identifying accessible digital sources suitable for longitudinal studies, the number of records creators is still fairly limited. In any event, there is no other alternative than to start looking among the material that has actually been transferred or deposited. Among records creators with electronic registers starting in the 1960s, we find Statistics Norway, the national taxation authorities and the Norwegian labour and welfare authorities. From Statistics Norway, the Housing and Population Censuses have been deposited in digital versions since 1960. However, the Statistical Act and the deposit status of the census archives make them unavailable for any use of individually-based information until they reach the release date.

The Norwegian Directorate of Taxes (*Skattedirektoratet,* or the *SKD)* is an institution with several nation-wide information systems including the Central Population Register and many systems specialised for direct taxation of personal and non-personal taxpayers, as well as for indirect taxation. A major advantage of digital archives from this institution is that these are mainly transfers, and hence access is granted by the National Archivist of Norway. A major exception is, however, data from the Central Population Register.

Possible data from the Central Population Register?

As of 2011 only fragments of data from the Central Population Register was in the custody of the National Archives of Norway. When the Central Population Register was established in 1964, the

[72] ADDML - Archival Data Description Markup Language – is a technical metadata standard for data tables extracted from a native system as flat files. See also Appendix 2.

Central Office for Civil Registry was a unit within Statistics Norway. In 1991 the Central Office for Civil Registry was administratively transferred to the Taxation Authorities.

The CPR has been the main data source for Statistics Norway's population statistics since 1964. For technical and practical convenience tailor-made statistical datasets have been extracted by Statistics Norway from the CPR, e.g. annual files that show the population situation by the end of each year. In the following these are referred to as the 'situation files'. The situation files are pictures of the status for residents and non-residents by 31 December, and contain a few selected fields from the CPR. The situation files must not be confused with the complete CPR, cf. below about the 2012 CPR archival extract.

From the period when the Central Office for Civil Registration administratively belonged to Statistics Norway some situation files have been extracted for the National Archives. This is material which was restored, and to some extent reconstructed due to technical difficulties, from Statistics Norway's historical data repository during the years 2002 – 2005. According to information by Statistics Norway, about 75 000 persons are missing from the 1966 situation file because the storage media for this data set was partly damaged. The archival versions of the situation files exist only for the years 1964 – 1972. The contents of the situation files are cross-sectional and not accumulated population data, as opposed to the accumulated population in the DSF population table, or in the CPR main table from the 2012 extract, cf. below.

As the research population will comprise the birth cohorts 1937 – 1958 and the observation period for the present case study starts in 1967, the archived situation files are relevant, but not optimal data sources. The cross-sectional format means that the majority of individuals will appear in each of the situation files as they are repeatedly included in the annual files, i.e. there will be a high number of duplicates in an appended, and thus accumulated, version of these files[73]. The total number of observations in these files are 4,14 million in 1967 increasing to 4,55 million in 1972. The increase from one year to another is mainly due to new births and immigrants for each year. Selected by residential code there are 3 808 396 residents in the 1967 file and 3 950 146 in the 1972 file. The difference between the gross population in these files and the residents are people who are dead, emigrated, disappeared, unregistered or with expired ID numbers. For comparison - the DSF Population table counted 6,5 million persons (ID numbers and D-numbers). The DSF population table covers the members of the National Social Insurance System for the years 1967 - 2008. For further explanation of the DSF population table and comparison with official statistics, cf. table 10, and the chapter Historical criticism.

[73] When the annual situation files 1967 – 1972 are appended to form one accumulated file, the total number of observations is 26,09 million. A duplicate check on the appended tables shows that only 76 950 observations are not duplicated.

The technical metadata is always the catalogue to the contents of the data sets, and is the starting point for the judgement of whether or not a data set is interesting for an actual research project. The technical metadata for the situation files are very scarce, e.g. in the 1964 table description there are several fields with the additional comment 'No explanation'. The 1964 table also include coded fields for 'type of education' from the 1960 census, but these are not accompanied by code explanations, cf. above about the importance of the quality of technical metadata. The layout for the 1964 – file differs from the files for the following years. Since 1965 the fields in the 125 first positions of the record layout is identical, but some new fields are added at the end of the record from 1966 and onwards.

As a data source for establishing the research population the situation files would have been possible, but more cumbersome to use than an accumulated population file like the DSF population table. To ensure a research population as complete as possible (all birth cohorts for the years 1937 – 1958) all situation files would have to be subject to the same selection procedures, the selected cohorts from each annual file must be merged with one another, and purified from duplicates. It is more secure and actually gives a better result to extract the population through one single query from one table where the population is accumulated and updated, in stead of assembling the population from various versions.

One of the criteria for the panel population was that an observation had to be alive at the age of 50. This requires a data source with updated residential status right to the end of the observation period for the youngest cohort. The DSF archival package was extracted in 2008 which means that the population reflect the status by 2008, all changes and transactions are updated as of 2008, while the situations files are not updated since 1972, i.e. any change since 1972 is not reflected in these files.

For the present case study one desired variable is family responsibilities – operationalized as number of children. One field in the 1964 version is in principle relevant for the research population: 'Number of children from the 1960 census'. But this field is only present in the 1964 situation file. There is some doubt if this field has been updated with births after 1960, or if it only repeats the status by 1960. The technical metadata do not specify this any further. There is no field for number of children in the situation files since 1964. Anyway 1960, and 1964 as well, is outside the scope for the observation period for the case study, and besides this variable would only be possible to link to the oldest cohorts in the research population. The youngest cohorts, e.g. 1948 – 1958 were children themselves in 1964. As an example - the observation period for the 1958 – cohort are the years 1988 – 2007, and the variable number of children must refer to this period for this cohort.

A possible alternative would have been to derive number of children, but for that purpose the fields 'family number' and 'personal code' would have been required, cf. page 29. However none of

these fields are present in any of the situation files. Actually 'family number' as a variable for constituting the nuclear family was developed and introduced from around 1970[74] .

Another explanatory variable for the case study is geographic mobility. History for individual change of address is basically CPR information, but the situation files 1964 – 1972 offer no complete address history. However the fields 'residential municipality' and 'former residential municipality' exist. Technically it would therefore be possible to calculate number of changes similar to the process carried out based on the DSF income tables – see pages 88 – 89. But the situation files only allows this for a very limited part of the observation period – and thus only for the birth cohorts that are observed for this period, namely the 1937 – 1942-cohorts, and for these cohorts only for the years 1967 – 1972. To summarise about the situation files - these are best described as secondary sources and fragments of the CPR with limited value for the present enquiry. The most relevant use of the situation files would probably be as a basis for extracting population samples for historical survey enquiries.

Anyway the situation files contain the national ID-number, which means that 'sex' and 'age' can be derived, and the presence of the national ID-number also make the population linkable to other data sets. There is reason to believe that Statistics Norway checked the quality of the national ID – number in the situation files before transfer to the National Archives: a validity check of the ID-number results in one (!) invalid ID-number. A match between the research population in the data matrix and an accumulated version (1967 – 1972) of the situation files gives a 100 per cent result.

As of June 2012 an archival extract of the CPR was produced for the National Archives by the Central Office for Civil Registration: This package comprises the total CPR by June 2012 with all data contents since 1964. The extract counts 40 tables and the size of the package is 25 GB (flat text file with fixed format). For comparison - the size of the largest of the situation files (1972) with the same format is about 700 megabyte. The main table in the CPR is accumulating – there are no duplicates, and no persons are deleted, but historical status is reflected by codes in separate tables. The total number of persons with national ID-numbers is 8,9 million in the CPR main table, while the D-number table counts 1,5 million persons. The D-number population has increased vastly during the latest years, cf. also the chapter 'National identifiers and coded information'. Other tables in this extract are historical tables for 'marital status',' address', 'migration', 'citizenship', link tables for change of D-number to national ID-number, change of national ID-numbers, several auxiliary tables, code tables etc.

By summer 2014 it is not concluded about the formal status (transfer or deposit) for the complete CPR extract that was produced for the National Archives in 2012. But when this issue is settled, the complete CPR will be the major data source for demographic data in future research, but the status as transfer or deposit will decide from when this data source is accessible. Data for the CPR

[74] Statistisk sentralbyrå (1995): Historisk statistikk 1994. Oslo: Statistisk sentralbyrå, page 61.

are collected pursuant to the National Population Register Act and not the Statistical Act, which influences the formal access rights.

Demographic data from the CPR are among the most distributed and reused basic data. CPR duplicates exist among a few major institutions and among commercial distributors. I.e. demographic data from the CPR is found in numerous systems – nationally and locally, e.g. in taxation systems. The type and number of attributes transferred from the CPR will of course differ between the recipient systems. Though the CPR is the base and the primary source for demographic information, it is therefore not necessary to extract demographic data from the CPR, if the same population and the required variables are found in other data sets. If the CPR is not available, possible alternatives exist in nationwide registers with duplicated CPR information, like the DSF population table. To establish a population base in many research projects only the population and the national ID-number is sufficient. Other variables can then be linked to the given research population from different sources.

POSSIBLE RECORDS CREATORS AND DATA SOURCES – A SUMMARY

A welfare reform in 2006 established the Norwegian Labour and Welfare Administration, known as 'NAV'[75]. However, this reform took place so recently that it does not affect the digital archives created by the original national social welfare authorities which are relevant for my study. I therefore refer to the pre-reform institution, the National Insurance Administration (*Rikstrygdeverket* or the *RTV)*.

All three records creators presented above are institutions with a large number of electronic administrative and research registers dating back to the 1960s. Both the RTV and the SKD have digital archive series with economic and other data starting in 1967, and digital information has been exchanged between these institutions since then. A common feature of systems in both these institutions is that they cover total populations. Moreover, both institutions have had their digital systems recently appraised, and finally, the transfer or deposit of table extracts to the National Archives of Norway has been carried out for some years.

For my purpose, a relevant system from the SKD is the Tax Register for Personal Taxpayers (TRP), which, as mentioned earlier, is produced in annual volumes. The first preservation decision for the TRP was issued by the National Archivist in 1985, but was at that time limited to preservation of every fifth generation. In 2001, this decision was reconsidered and the scope of preservation was extended: as from and including 1990, annual volumes from the TRP were to be preserved[76]. In other words, the continuity in terms of available variables from this data source is broken before 1990. It

[75] Ny Arbeids- og Velferdsforvaltning.
[76] At present the National Archives holds the following generations of the Tax Register for Personal Taxpayers system: 1967, 1970, 1975, 1980, 1985, and annual data sets since 1990.

should be added that this records creator is only required by law to keep each generation of data sets for as long as the taxpayers are entitled to put forward complaints. Upon expiry of the period for submitting complaints, which is normally ten years, the data sets are deleted by the records creator once a raw data extract has been transferred to the National Archives of Norway. Due to complaints, minor updates may take place in the TRP until the closing date for complaints. As a result, the final version of a TRP volume is established many years after the actual income year, and it is always the final version of the annual volumes that is transferred to the National Archives of Norway.

The most complete series relevant for my purpose is created by the RTV. This is a system called 'the Central System for National Social Security' *(Det Sentrale Folketrygdsystem)*. The Norwegian acronym for this system is the DSF which is later used as the reference term in tables, source code and text. The complete history of individual pension entitlements in Norway has been stored in the DSF since 1967. Pension entitlements are the annual amount of income from labour – employment or self-employment – commonly referred to as 'pensionable income'.

Identifying adequate variables is a step-by-step procedure in a hierarchical approach. The first step is to identify records creators, then proceed from systems and data sets by a given records creator down to tables in a data set, and finally to fields in a given table – as found in the technical metadata. Examples can be seen in Table 30 and Table 33. An additional requirement is the format of each field which is also found in the technical metadata. For my purpose, mainly numeric and preferably coded fields are of interest. Fields with a potential for computer-based deduction and encoding are also relevant. Variables extracted for my specific purpose from each system are commented on and explained as they are utilized, and each of them will be recognised in the applied source code (see Appendix 1).

For my particular case study the research population will be extracted from the DSF Population table. The variable *geographic mobility* will be derived from the DSF income tables, while *number of children* will be collected from a number of the annual TRP data sets and merged with the research population. The optimal data source for the variable number of children would have been a CPR extract with contents updated to 2008. But such an extract did not exist in the National Archives before 2012 and as of summer 2014 not yet accessible for research. Anyway the field *number of children* is duplicated and transferred from the CPR annually to the TRP (via 'Skattemanntallet'). The TRP data sets represent the best alternative in the longitudinal perspective for the case study as these exist for the major part of the observation period. The main problem with the TRP files versus the CPR appeared to be the different entities. The entity difference required a time-consuming procedure to move number of children from the reference person to the spouse, cf. the chapter 'Data steps for adding family obligations' and the chapter 'Historical criticism'.

THE TECHNICAL METADATA: 'INFORMATION ABOUT INFORMATION'

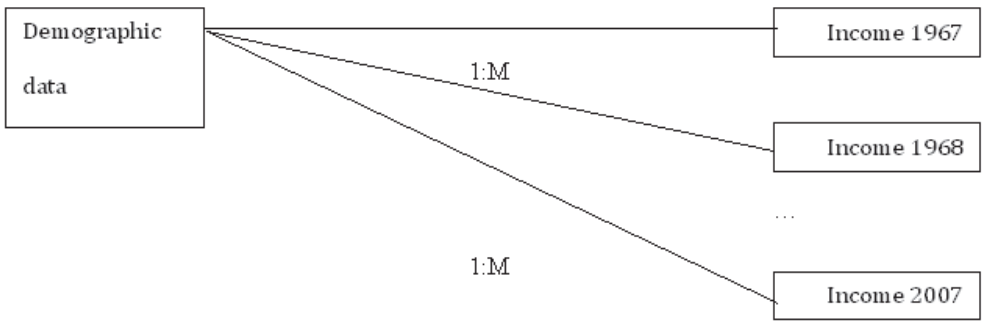
The road from theories and models to a physical, electronic data matrix necessarily always starts with the technical metadata. While access to personal information younger than 60 years, a limit that applies to most digital archives, is restricted, access to 'information about information' is unrestricted. In other words, free distribution of technical metadata for digital archives poses no problem. Technical metadata is the parallel to catalogue information for paper-based archives. However, understanding technical metadata requires a certain minimum of computer skills (see Appendix 2).

For digital archives, technical metadata are the doorway to the contents of the tables exported for the archives. The concept 'technical metadata' encompasses data models and file descriptions at the most detailed level, but also more general descriptions of functionality in a specific system. Examples of technical metadata are shown in Appendix 2 and Figure 4, as well as in the appendix section. The technical metadata concept also includes user manuals and legislation.

The data model in Figure 4 relates to the DSF. As we can see from the figure, the DSF system consists of only two table types, one table type containing demographic data and one table type mainly containing pensionable income data, but also some additional information. Demographic data are accumulated in the demographic table. On the other hand, the pensionable income earnings are collected in one table for each income year, thereby adding one new table to the system year by year.

The data model below is a simplified presentation of the table types in the DSF and the relationship between them. In technical terms, the original system is not a genuine relational data base. Each person occurs only once in the demographic data table, but one person may have different types of pensionable income during an income year, and thus a person may occur several times in the income data table – once for each type of income – for each income year.

Figure 4. DSF table structure



As a further illustration, the technical metadata for the DSF demographic table are shown in Table 30. This data set is the archived extract from the active Central System for National Social Security which was produced for the National Archives of Norway in 2010[77]. The demographic table has a simple structure with very few fields. Anyway – an assessment of each field is required in order to decide whether it is applicable, useful or meaningful in a possible research project, or, as in my case study, in the context of building the panel population.

The national identity number must obviously be selected from the demographic table, both because this is the primary key in the table, as well as the key needed for linkage with other data sets. This field is also needed in order to deduce the variables *gender* and *year of birth*. On the other hand, 'name' in plain text is of no use for the case study, nor is the 'code for type of pension' needed. This code is only activated when people reach retirement and the pension payments start to run, according to RTV's technical metadata. By definition, the present enquiry will only include observations before retirement age as they are followed from age 30 till age 50. But whether or not observations are alive throughout the entire observation period is very important for the analyses, so the field 'possible date of death' must be extracted from this table. These are conclusions that can be drawn based on information from the technical metadata specific for this data table, which is also enough information to write the necessary source code to extract the specified data; see SAS data step 1 in Appendix 1.

There is one income table for each year from 1967 to 2007 and the layout is identical for all income tables, but the number of records in each table varies for each income year. From the income tables most fields appear to be relevant. The field *code for type of income* also contains codes other than those listed in Table 33, but I have chosen to display only codes that must be considered appropriate for my purpose. The DSF income tables otherwise contain a few other fields that are of interest and accordingly must be extracted for further use. The municipality identifier is a very important field, both for regionalization of the data matrix and for adding the variable *residential characteristics*. The municipality code reflects a person's place of residence as of November 1st in the year prior to the income year, as required by law.

This is in principle the method to follow for each set of technical metadata: study the contents and the meaning of each field, and then decide whether or not it is useful in a given research context. The technical solution for extracting the needed variable or variables is explained in the next chapter, and the actual source code is presented in Appendix 1.

[77] Information from this system is preserved pursuant to the decision of 28 October 2005 of the National Archivist of Norway.

BUILDING THE DATA MATRIX

REQUIREMENTS FOR PANEL POPULATION AND VARIABLES

So far the sketch in Table 2 is nothing more than an empty table, which step by step must be filled with observations and variables attached to each observation. The observational unit in the data matrix is the individual person, but more precisely a number of individuals meeting a set of requirements, which together will constitute the research population, i.e. the panel population. The first step is therefore to specify requirements and make decisions about which units should constitute the panel population.

The panel population will consist of a limited, but clearly defined, set of birth cohorts identified by *year of birth*. The first constraint is therefore to set limits for the birth cohorts. These limits are decided by those birth cohorts that most likely would be active in the labour market during the years in question, 1967 – 2007, and also traceable in the data sources for the desired 20 years of their lifetime as described earlier.

The basic data set from which the population is extracted is the DSF's demographic table with its total of 6.5 million observations. This is the total number of individuals registered in the DSF from 1967 to 2007. Observations are never deleted from this system. Inactive observations are identified by codes. By way of comparison, the total number of observations in the Central Population Register is higher, because the Central Population Register started three years earlier. Individuals who died before 1967 were never included in the DSF population. From the DSF demographic table a population that meets the following set of requirements must be derived:

- Only observations with valid ID numbers
- No duplicates
- No D-numbers
- Age span which makes it possible to follow each individual from the age of 30 to the age of 50 within the range of available pensionable income data: 1967 – 2007
- Accordingly the cohort limits will be year of birth = 1937 as the lower limit, and year of birth =1958 as the upper limit.
- To be part of the panel each individual must be alive at the age of 50.
- Valid geographic connection.

The first requirement is there for technical reasons as it is impossible to match invalid ID numbers. A number of explanatory variables will be added from various other data sets at individual

level, and *gende*r and *year of birth* will be derived from the ID number. A valid ID number is actually a condition to fulfil all these requirements. Invalid ID numbers will be excluded through a validity check. The population also has to be checked for possible duplicates as each individual in the panel population must appear only once.

In a longitudinal study only continuity is of interest. The special type of ID number known as the D-number is issued to non-resident persons with a short-term connection to Norway. By definition the D-number population never represents the continuity needed for a longitudinal study. Besides, additional information (e.g. the field *family responsibilities* which is required as an explanatory variable in this case study) will normally not be found in domestic systems at all for non-residents. Hence all D-numbers must be deselected from the panel population.

Each birth cohort in the research data set will be observed continuously from the age of 30 to the age of 50, which implies that each cohort will be observed for a different set of years: the 1937 cohort reaches the age of 30 in 1967. Accordingly, the employment history for this cohort is followed from 1967 up to and including 1986, while the 1958 cohort is observed from 1988 up to and including 2007. Some observations will definitely have an employment history with 20 years of continuity, while some observations, on the other hand, will have no employment history at all in the course of 20 years. In other words, the variable individual *labour market participation* can take any value from 0 to 20.

To ensure comparability within the panel population, the requirement 'being alive at the age of 50' is introduced. The criterion 'being alive at the age of 50' will be based on the field *date of death* in the DSF population table.

The diachronic aspect requires that each cohort is measured while they are in the same cycle of life. The generation perspective – similarities and differences – will be reflected most clearly by the method described above, and the data sources make this possible. A suitable question then is: why 20 years of someone's life – why not 10 or 25 or some other period? And furthermore: why follow each generation from the age of 30 to 50 – why not from 25, or 35 or some other age? The answers lie in the chosen dependent variable, participation in the labour market, which briefly excludes both the youngest and the oldest generations. I assume that people between 30 and 50 are the age groups most likely to be engaged in the labour market, having finished possible higher education, vocational training, etc., and not yet reached retirement. Furthermore, the explanatory variable *family responsibilities* must necessarily be linked to fertile age groups, which fall well within the selected age span 30 to 50 years.

Observation periods other than 20 years would of course have been adequate. The main point is that the observation period must be of some duration, and my decision is a balance between the duration perspective and considerations of population size and birth cohorts selected for the purpose. A shorter observation period would have made a larger population possible, while a longer

observation period would have been at the expense of the size of the population. A smaller population might have put the municipality analysis at stake, due to the expectation of very small numbers of residents in some municipalities. Given these considerations, a 20-year observation period is a compromise.

It is obvious that birth cohorts outside the scope defined above were actors in the labour market during the years 1967 – 2007, and indeed for many, many years. The generations 50 years and above in 1967 would still have 15 – 20 years before reaching retirement age and were active in the labour market into the 1980s. On the other hand, the younger age groups would enter the labour market in their late teens or early twenties. In this sense, there is a larger population from which I have made a selection, and for which the results of the regression analyses can be predicted.

The spatial dimension requires that each member of the panel must have a valid geographic connection expressed by municipality identifiers which is further discussed below.

My decision regarding population, observation period and variables must also be seen against the background of the possibilities offered by and limitations of the data sources. Having decided on this framework, the next step is to write source code and start building the physical data matrix. After making decisions about which variables should be entered into the data matrix, the question of 'how to use raw data' requires a technical solution.

HOW TO USE RAW DATA?

The core of the migration strategy is that information must be separated from its native system and its native technological environment. The software once used to create the information is not preserved, nor is the hardware. For future use of the information, different software and different hardware must be applied to recreate the information.

The files from the Central System for National Social Security described in Table 30 and Table 33 are examples of information extracted for the National Archives of Norway and stored in technology-independent format. Basically technology-independent data must be used from the format in which they are stored, as raw data, but then there are some technical barriers to overcome, e.g. interpreting technical metadata as displayed in Appendix 2 and Table 30.

Raw data are not surrounded by software which can help in accessing the data. Therefore the question of software to access, retrieve and process the data will arise. In general, a Data Base Management System (DBMS) and a query language will always be available. Tables in technology-independent format can be imported to any DBMS. But the question of selecting software is more a question of the intended use of a given number of raw data tables.

It is possible to regenerate any data base from technology-independent format to its original structure on a new technology platform, but regenerating a data base completely can be a formidable job, and normally a 100-per-cent regeneration is not necessary. Besides, in a given research project it will normally be required to assemble and combine data originating from several different data bases. It is quite possible, and usually much more effective, to retrieve only information needed for a specific purpose directly from the raw data format without regenerating the whole data base, and instead match selected rows and fields in raw data format. The creation of the data matrix will demonstrate this.

In terms of storage capacity the size of a data set is not really a problem any longer. It is nevertheless a good idea to reduce the size whenever possible, because in terms of processing and analysing data, size still matters. The size of the final version of the data matrix is only about 40 megabytes, which is not huge, but some of the analytic procedures were actually still quite heavy when measured in terms of processor consumption and elapsed time, not to mention the production of graphics.

When working with raw data it is completely possible, and very advisable, to extract only variables and entities relevant for a specific analysis, while omitting all other data. Thus the only remaining volume should be adequate data, which also reduces the size of data to process. Generally, a research process based on electronic registers will involve two different stages: first, the relevant data must be retrieved from multiple sources, assembled and given a structure and format which make import to analytic software possible. The second stage is the analytic part for which standardised analytic procedures can be applied. For the first stage there are, of course, no standardised procedures available, which necessitates individually constructed source code (see Appendix 1). These two stages will normally also be performed by different software.

As quantitative techniques will be frequently applied when analysing this kind of data, an important requirement is the capability to perform the analyses in question for a given research project, e.g. correlation, regression, clusters, etc. and to depict results by means of graphics. For quantitative analyses in general, specialised business and scientific software packages are available, like SPSS, SAS or STATA, to mention a few. Using SQL is also possible, but requires that the data be imported to a DBMS. With small volumes of data, typically with sample populations, some analyses could even be processed in a spread sheet. Formulas exist for all the types of analysis that I have in mind. However, all relevant formulas for the analyses are included as standard procedures in the analytic software tools. For the researcher, the question is to select the appropriate procedures from the chosen software during the analytic process.

But the standard analytic procedures can only be applied once the research data set has been established. My software choice for establishing the data matrix has been SAS. The preparation of the research data set from raw data might just as well have been accomplished using other programming tools, such as Easytrieve, C++, C # or others.

Software for utilization as defined above must satisfy a few basic requirements. One requirement is the capability to handle, arrange and process raw data, and in this case, also the capability to handle large volumes of data: 5 – 10 – 15 gigabytes or more. The total volume of the DSF in technology-independent format is about 6 gigabytes. Software with any upper limit in terms of number of observations, or with limits in terms of volume to process, should be avoided. Moreover, a very desirable quality is that the software must have the capacity to read different character sets. The earliest acquisitions of digital archives by the National Archives of Norway are stored in their native format with EBCDIC character set and very often compressed numeric fields. SAS data step number 14 is an example of reading data from EBCDIC format with compressed fields.

For the analytic part of this case study, my software choice has been SPSS (abbreviated for Statistical Package for the Social Sciences).

THE PANEL POPULATION

SAS data step number 1 was the first step in selecting the panel population. This code snippet is also put forward as an answer to the question of how to use raw data. There will be a lot more examples later, as the file created in this step will be supplied with additional variables; see the sequence of SAS data steps in Appendix 1.

Only 12 ID numbers failed the validity check, and no duplicates were found. This is interesting information about the records creator and internal quality awareness and quality routines; see also the chapter 'Historical criticism'. The output from the first three data steps is a temporary file checked for duplicates, purged of non-resident persons (D-numbers) and persons with invalid ID numbers.

From the total population in this temporary file the cohorts 1937 – 1958 will be selected to form the panel population. Anyone born outside this scope will be excluded. However, the cohort issue requires a special method described below, so for technical reasons *gender* and *year of birth* are derived first.

GENDER AND YEAR OF BIRTH

Gender is anticipated to be the most important explanatory variable. This variable has to be derived from the national identity number. The national identity number itself is not confidential, yet a modified ID number has been used as an example in Table 31. Digit number nine in the national ID number has even numbers for female, and odd numbers for male. Technically, *gender* may either be

derived by testing on the quotient from integer division, or by testing directly on the value of the ninth digit in the ID number, as in SAS data step 4.

Birth cohort or *generation* is defined by *year of birth,* which also has to be derived from the national identity number. This variable is needed both to separate generations from either being included in, or excluded from the panel population, and secondly as an explanatory variable for possible generation effects within the panel population. The population in the temporary file without duplicates and D-numbers is still composed of persons born in the 1800s and the 1900s, and even includes a few born after 1999.

According to my definition, the time frame 1967 – 2007 is divided into a number of 20-year observation periods. Since each cohort is observed for a different 20-year period, the birth cohort is also an expression of an observation period. Birth cohort 1 equals observation period number 1, i.e. 1967 – 1986, birth cohort 2 is observation period number 2, i.e. 1968 – 1987, etc. to birth cohort 22 which equals observation period number 22, i.e. 1988 – 2007, but *birth cohort* is kept as the variable name.

The total number of observations now constituted by the selected birth cohorts is the maximum number of observations that meet the population requirements listed above. This may not be the final number of observations in the panel, as one important requirement remains: the spatial perspective requires that each observation must have a valid geographic connection which is outlined below. For some observations, a geographic connection may be missing for the total 20-year observation period, or for a large part of this period. Observations without any geographic connection will be deselected later.

THE DEPENDENT VARIABLE: LABOUR MARKET PARTICIPATION

The next step is to identify sources which can meet the requirements for the dependent variable outlined above. It has to be a unit of measurement which is sustainable and comparable over time, more specifically with continuity and comparability for the whole period 1967 – 2007. Furthermore, this measurement has to be available at individual level in an unbroken chain from 1967. That is not a trifling requirement. Obviously such a variable does not exist in any source material – it has to be derived. Some alternatives for deriving it will be discussed, and for a moment disregarding the fact that alternative data sources are not necessarily accessible, either because of privacy or other legal restrictions, or because archival extracts have not yet been produced for the National Archives of Norway.

Statistics Norway has produced statistics based on electronic data sources on occupation, labour market, labour force, employment and unemployment, etc., for a long time, in the population

and housing censuses every ten years, and in the quarterly Labour Force Sample Surveys (LFS)[78] since 1972. Labour market and labour force statistics will also be found in the National Accounts, but only at macro level. Local-level or individual figures cannot be deduced from the National Accounts. Moreover, the National Accounts labour market figures are not a first-hand data source, but consist of data from various sources, among them the LSF.

It is an obvious problem that definitions and measurements have changed over time. Comparability over time is difficult, if at all possible, based on any of these sources. In the population and housing censuses, employment measurements have changed from a 'yes' or 'no' answer to the question: 'Is income from own labour your major source of subsistence?' as in the 1960 census, to be measured by the number of working hours, but with changing reference units: 'Hours worked per week' or 'hours worked per year'. In 1980, Statistics Norway set a limit of at least 1 000 hours worked per year to count a person as employed[79]. Calculations made by Statistics Norway proved that 1 000 hours worked per year coincided very well with 'having a main profession as main source of subsistence'[80]. However, as censuses are periodical, 'working hours' is simply not available as an annual variable at individual level within the 40 years in question for my case study.

The LFS is based on a population sample, and its history only goes back to 1972[81]. Housing and population censuses lack continuity, as they are carried out every ten years, and measurements of employment have changed over time. With reference to the desired variable, the conclusion is that neither the census nor the LFS concepts meet the requirements to fit with the longitudinal and individually-based research data set. Of course, statistics based on the LFS and the censuses will be useful reference for comparison and source evaluation later on.

Measurement of labour market participation by type of income seems possible, however, as 'income', in a broad sense, was very commonly processed electronically from the 1960s and onwards, and thus occurs in several systems from several records creators. There is a variety of income types, both taxable and non-taxable, in electronic administrative registers. Taxable income comprises capital interest and dividends, income from labour, income from pension, gross income, net income, etc., while child care benefits, on the other hand, are non-taxable. The different income types appear in different systems and definitely in separate fields within a given system; technically there is no risk of confusing or mixing any of these income types. The challenge is to identify the most relevant income concept and the primary source system among what is actually preserved and accessible in digital form.

In principle, any taxable income for personal taxpayers is reported annually to the national taxation authorities. For labour income, a possible data source is the Tax Register for Personal

[78] ArbeidsKraftUndersøkelse - AKU
[79] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå, p. 227
[80] Ibid.
[81] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå., p. 228

Taxpayers produced by the national taxation authorities annually since 1967. The content of this register is to a large extent defined and decided by legislation, and by technical solutions required to administer and fulfil these legal requirements. Legislation has changed over the years and such changes are reflected in the system, but basic concepts, like 'labour income', are still comparable over years. The TRP would have been a very desirable data source for my purpose, but the TRP data sets were not preserved in annual volumes until 1990, by the decision of the National Archivist (see above).

An alternative records creator for labour income information is the national social welfare authorities. The National Social Insurance System (*Folketrygden*) was established in 1967, and the Central System for National Social Security (DSF) was operative from the same time as well as other social benefits and pension systems. The national social welfare authorities manage solutions for numerous social security benefits: unemployment, sickness, parental leave, pensions, etc. Different benefits are administered through different systems.

For my subject the Central System for National Social Security is highly relevant. A fundamental principle for retirement benefits is that the size of the future pension is based on individual income from *work* – i.e. 'labour income' or 'pensionable income' – earned by employment or self-employment. To meet these requirements, it is absolutely essential that the 'labour income history' of each member of the national social security system is recorded and preserved. Technically this information is accumulated in the DSF. This system stores data on all labour income, the actual amount in current value, earned per year for each member of the social security system and recorded since 1967. In other words, income from labour is an expression of individual participation in the labour market, and these recordings are continuous since 1967. This labour income concept is coherent with the ILO definitions of employment.

In terms of data exchange, all labour income data are transferred annually from the TRP by the national tax authorities to the social security authorities and stored in the DSF. In other words, the primary source is the TRP, but as not all generations of this system have been preserved, the DSF is the only source for an unbroken labour income history, and becomes the primary source for years when the TRP no longer exists.

According to legislation, labour income earnings must be recorded individually. Hence income from labour is always attributed to the person who earned it, as opposed to other income types which may be shifted between spouses or attributed to a family. In the TRP, all types of taxable income are registered in separate fields for each type. 'Pensionable income' is defined by legislation[82]. Basically the concept 'pensionable income' has remained unchanged since 1967. There has been a certain expansion of its scope, mainly to include the value of fringe benefits. In the 1980s, the value of such benefits was capitalised and added to wages and salaries. However, this hardly affected the number of

[82] LOV 1997-02-28 nr 19: Lov om folketrygd. § 3-15.

people with pensionable income, as the condition of already being employed normally precedes any possible fringe benefit[83]. There is of course no gender-related bias in the definition of pensionable income. According to legislation, pensionable income is entered into registers and reported to authorities independently of whether it is earned by a man or a woman. It is a different issue that the actual pensionable income amounts are generally lower for women than for men, and this has to be accounted for when the labour market participation is operationalized from pensionable income (see below).

In the case of unemployment or sickness, lost labour income is compensated for by the National Social Insurance System. It is important to note that such compensation is defined as pensionable income, either 'wages or salaries' or 'income from self-employment' depending on the type of income that is replaced. Accordingly, a case of unemployment or sickness will not result in immediate interruption of pensionable income. The actual amount of compensation will gradually decline, but discontinuity in employment due to sickness, or unemployment, and in more recent years also parental leave, will only occur after long-term absence, i.e. 12 months' absence or more[84]. This largely excludes unemployment, sickness or parental leave as explanations for discontinuity – over years – in individual employment history. Absence from the labour market for reasons that do not give a right to compensation would result in no labour income or possibly a small amount in a given year, and the intention is to capture this distinction through the requirements for the dependent variable. On the other hand, in the case of a permanent exit from the labour market due to retirement, disability or other reasons, the economic compensation from the Social Insurance System is by definition not labour income.

So far the labour income tables in the Central System for National Social Security seem to be the only possible source for a *labour market participation* measurement. The income amount itself will only be used for auxiliary purposes and is of no interest beyond that. From the labour income amounts it is possible to derive a unit of measurement, an indicator of labour market participation, by setting an income line for each year. This income line is further discussed below.

Whether a person's labour income equals, or is higher or lower than this specific income line, is reflected by a code that is issued to express labour market participation. The value of this code will be either '0' when the amount is below the line, or '1' when the amount equals or is higher than this line. This test must be performed for each member of each birth cohort for a period of 20 years, though a different 20-year period for each cohort, ascending with one year for each cohort. Eventually, each observation will have his or her labour income converted to a labour market participation history, technically as a row of 20 cells, with either '0' or '1' as the value in each cell. The values of these 20

[83] Statistisk sentralbyrå (1995): *Historisk statistikk 1994. Ch. 11*. Oslo: Statistisk sentralbyrå.
[84] Lov om Folketrygd: § 4-15, § 8-12.

cells will then be summarised, thereby giving us our desired variable – individual *labour market participation* - with a value ranging from 0 to 20. The consecutive 20 cells containing '0' or '1' for each observation will be omitted from the data matrix as only the sum cell is needed for the analyses.

The question as to which income line would be the most appropriate for each year requires a detailed explanation. The challenge in defining an income line is to find a balanced and, above all, an unbiased measurement for labour market participation. For reasons not to be further discussed, labour income has been and still is higher on average for men than for women. Hence the most obvious danger of setting an income line is to create a gender-biased measurement. It is beyond doubt that a high income line in general would favour men, and thus create a gender-biased variable. It is also obvious that if the line is set too high, only full-time employees, for instance, would be included, and a large number of part-time employees excluded. Part-time employment is more widespread in typical female sectors[85]. On the other hand, if the line is too low, the risk is that too many will be counted as labour market participants, e.g. people with sporadic employment, with the possible consequence that variation would not be measurable.

By defining a low income line, the effects of possible confounding variables will also be more or less neutralised. A confounding variable which causes income differences is *education*. High education normally generates high income and vice versa.

All pensions and pension rights in the Norwegian welfare system relate to one fundamental concept: the Basic Amount (*Grunnbeløpet*), commonly referred to as 'G'. The amount of G is actually determined each year by parliamentary decision. From time to time it has been adjusted twice a year, sometimes even more frequently. The G is determined in relation to factors such as inflation rate, rise in wages, etc. The lowest limit for earning pension rights is that labour income equals or exceeds the amount of one G. The complete list of the annual amount of G for 1967 - 2007 is shown in Appendix 3. This table shows the annual average of G, as well as two G and four G. For comparison purposes, figures from public statistics on average pensionable income 1967 – 2007[86] are included in Appendix 3.

For this case study, G has been chosen as an instrument for setting an income line. With reference to the sketch of the data matrix in Table 2, three alternatives for this specific income line were tested. Each of the alternatives was based on different requirements for the amount of G:

- Labour market history, alternative I: Amount limit > 1 G
- Labour market history, alternative II: Amount limit > 2 G
- Labour market history, alternative III: Amount limit > 4 G

[85] Håland et. al., op.cit.
[86] Statistics Norway: *Tax statistics 1967 – 2004.*

Compared to the figures from 'Average pensionable (occupational) income' in Appendix 3, the column 'Current value' is quite similar to the amount of 4 G for the whole period 1967 – 2007. Assessments based on experiments using the 4 G limit indicate that this limit is too rigid and hence also gender-biased, while the 1 G limit is too low and would make variation difficult to measure.

A man-labour year is roughly defined as 1 800 – 2 000 hours worked. Average working hours per week have been reduced in the period 1967 – 2007 due to a reduction in standard working time: from 45 to 42.5 hours in 1968, to 40 hours in 1976 and to 37.5 hours per week in 1987[87]. The Statistics Norway definition of 1 000 hours worked per year as of 1980 is slightly more than 50 per cent employment.

Table 3. Selected amounts of basic amount (G) and average labour income 1967 and 2007. NOK - current value.

Year	G	2 G	4 G	Average labour income [88]
1967	5 400	10 800	21 600	21 315
2007	65 505	131 010	262 020	304 200

Finally, the average labour income for men and women, respectively, for the years 1967 – 2007 is compared with the 2 G amount for the same years in Figure 5. In order to avoid a gender bias, the income line for labour market participation should be below the average labour income for women, which it appears to be throughout the entire observation period. The 2 G income line fits quite well with the requirements listed above.

With the 2 G limit any person (man or woman) with a labour income of NOK 10 800 or higher in 1967 will have one year that counts as labour market participation, resulting in the value '1' for the first employment history code in a series of 20 consecutive years. To be even more explicit, a woman with a labour income of NOK 15 000 and a man with a labour income of NOK 150 000 (or vice versa) in 1967 would both fulfil the requirement of being employed this particular year, and both will have the value '1' in the 1967 employment history cell. Thus a high labour income amount does not result in a higher 'value' of employment, and does not favour one gender over the other. It does not make any difference how much an actual income amount exceeds the income line of 2 G.

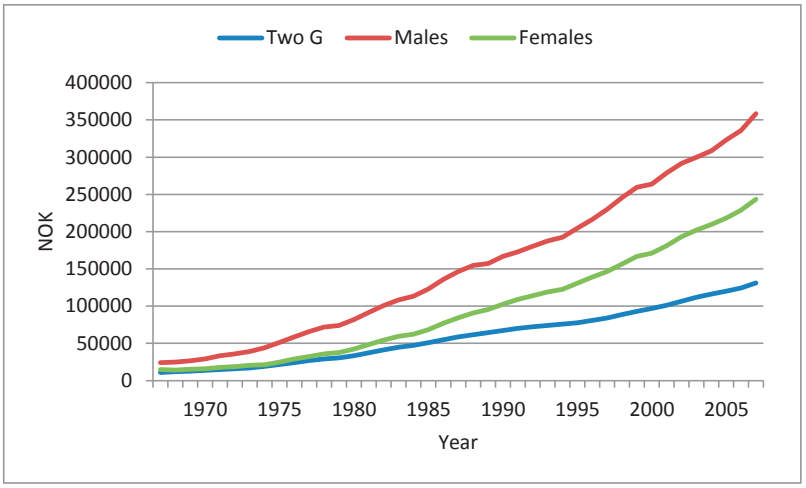
On the other hand, if a person that same year had a labour income amount of NOK 9 100, NOK 2 300 or NOK 0 for that matter, it would in any event be too low for the year to count as a year of employment, so the value of the first cell would be set like '0' for this particular observation. A

[87] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Ch. 9. Oslo: Statistisk sentralbyrå.
[88] Statistics Norway: *Tax statistics 2007.*

similar test has to be performed for each of the 20 succeeding years, for each observation and for each birth cohort, i.e. for a total number of 1.2 million observations. Thanks to the nature and structure of electronic administrative registers this procedure can be performed electronically.

This is the principle that was followed for each of the DSF income tables for the years 1967 – 2007: set amount limits according to the actual 2 G for each year and generate a code by means of this test. An example is displayed in SAS data step number 8 in Appendix 1, based on the 1967 income year, where the amount limit is NOK 10 800 (see also Appendix 3).

Figure 5. Average labour income for men and women 1967 – 2007[89] and two G 1967 – 2007. NOK. Current value.



Labour income or pensionable income is a general term for all income from labour. This could be either an employer's wages, income from self-employment, or both. Anyway, the distinction between self-employment and employment is actually reflected in the DSF income tables, as may be seen from the field 'Code for type of income' in Table 33.

All income tables in the DSF have the same structure for each of the years 1967 – 2007, while the number of observations of course is individual for each year. The code values for labour income type reflect whether a person is a wage earner, i.e. in an 'employer-employee' situation, or is self - employed with no employer. A complete labour market participation history must comprise both employees and self-employed persons, as both categories are participants in the labour market. The technical solutions to implement this requirement are found in SAS data steps number 6 and number 7.

[89] Statistics Norway: *Tax statistics 1967 – 2004.*

Statistics show that the major group in the Norwegian labour market is the wage earners – the employed – while the self-employed are quite few in number. Moreover, the number of self-employed has been falling while the number of wage earners has been growing strongly since 1970. The proportions between the groups are illustrated by these figures from Statistics Norway:

Table 4. Number of wage earners and self employed, 1970, 1980, 1990, and 2000 – in 1000.[90]

Category	1970	1980	1990	2000
Wage earners	1 350,8	1 695,9	1 839,5	2 147,1
Self employed	290,6	253,0	219,3	172,4

Wage earnings and income from self-employment are the main types of labour income. In addition, there are codes in the DSF for a few other types: labour income earned from employment or from self-employment on Spitsbergen Island and income earned by diplomats serving abroad are identified by separate codes. My focus is the mainland labour market, and hence income from outside mainland Norway has to be deselected. This also excludes income from diplomats while serving abroad.

MATCHING PANEL POPULATION AND RATE OF LABOUR MARKET PARTICIPATION

Now the first real steps towards an assembled physical data matrix can be taken. Each cohort, still in separate files, must be matched with the employment history for relevant years. The general intention behind matching data sets is to transfer one or more variables from one data set, which we may call a 'donor data set', to another data set, which we may call the 'recipient data set' (see Figure 6).

Each match will increase the record length of the output data set, as new fields are added each time, e.g. when the employment history code for each year is added to the panel population, but the number of observations will be exactly the same for each cohort after matching. Figure 6 shows the principles for matching two data sets. Other aspects connected to matching data sets are discussed in the chapter 'Consistency and match results'.

The panel population still consists of the gross number of observations, and the data matrix now contains both the dependent variable, i.e. a code for *labour market participation*, and two

[90] Statistics Norway (2008): Statistical Yearbook 2008. Table 214. Oslo: Statistisk sentralbyrå.

explanatory variables, *birth cohort* and *gender*. The structure of the first version of the data matrix can be seen as the four initial fields shown in Table 37.

This data matrix is now ready for extension by two more explanatory variables: *residential characteristics* and *geographic mobility*. To achieve this, a couple of auxiliary variables must also be added, and the population has to be purified from observations without a valid geographic connection as specified below.

GEOGRAPHIC CONNECTION

The spatial aspect of this enquiry requires that the panel population must be connected to a local residential unit – a municipality – throughout the observation period. Before any of the planned geographic variables can be added, a number of municipality identifiers must therefore be connected to each observation in the panel. In this context the municipality identifier will serve as an auxiliary variable for record linkage, but this identifier must also be kept in the data matrix both as a selection criterion and for analysis purposes. The data source for municipality identifiers is the DSF income tables where municipality identifiers appear annually for all observations with labour income amounts (see the file description in Table 33).

There is not necessarily a complete series of 20 consecutive municipality identifiers for all observations in the gross panel population, as only individuals with labour income are entered into the DSF income tables each year. Observations will also be omitted from the DSF income tables due to emigration. For the 20-year observation period, a minimum requirement is therefore that each observation in the panel population must have at least two municipality identifiers, but specified to be one identifier during the first 10 years and one during the second 10 years. This requirement is needed in order to deal with the emigration issue.

For observations with lacunas in the series of municipality identifiers during the observation period, an additional and alternative source is the TRP. Registration in the TRP is not dependent on presence of labour income, and hence the coverage of the variable *municipality identifier* is generally higher in the TRP than in the DSF. Linkage with the TRP is attempted for observations with lacunas in the series of municipality identifiers. When some observations still appear without the required municipality identifiers, the reason is most likely emigration during the observation period, and these observations are deselected from the population.

Table 5. Gross and net panel population.

Gross panel population	1 372 926
Observations without valid geographic connection	141 314
Net panel population	1 231 612

The procedure just reviewed connects the identifier for residential municipality to each observation for a maximum of 20 years, but observations lacking one or a few identifiers in between are still accepted as part of the panel population.

A simple match between the gross population and the population that fulfils the geographic connection requirements creates the net population; see the technical solution in SAS data step number 13. The output from this match is a new file in which all the fields have been retained, but with a smaller population. At this stage, the panel population is 'frozen' and will not change. The number of observations that failed the geographic connection test is displayed in Table 5. When tested by year of birth, the excluded observations are quite evenly distributed, totalling about 3 per cent of each birth cohort. The distribution is also even between genders.

RESIDENTIAL CHARACTERISTICS EXPRESSED BY 'MUNICIPALITY CLASSIFICATION'?

Entering *residential characteristics* as an explanatory variable is based on the idea that regions, counties and municipalities are different both in terms of ideological and cultural characteristics, and in terms of basic qualities which in turn affect opportunities related to education, public services, communication, and eventually labour market characteristics and labour market participation.

In the context of the present study, *municipality* is an important entity and an important variable. Groups of municipalities based on specific characteristics are also meaningful in this context. The municipality is probably the best classified, the best established and a very much referenced entity in many social scientific as well as historical analyses in Norway (see e.g. the Rendalen and the Ullensaker projects referred to above).

According to the regression model above, a variable that could express different qualities attached to local labour markets must be operationalized. Desirable options would be *profession*, *employment in private or public sector,* etc. at individual level. Such variables are definitely to be found in the Population and Housing Censuses, but only every ten years, which implies broken continuity, and as already stated, not accessible for research purposes as an individual variable due to legislation.

Another possible data source is the Employer/Employee register (E/E register) which is administered by the Norwegian labour and welfare authorities. This system contains information about all employments, i.e. employees and employers, in Norway since 1978. Technically, there is one table for 'employments' and one for 'employer', and a similar pair of tables with 'employment history' and 'employer history' in this register.

In addition, the E/E register includes an auxiliary table containing codes and code explanations for *profession*. However, the variable *profession* was not entered into this system from the beginning. Actually, codes for *profession* have only been systematically entered into the E/E register since the preparations for the 2001 Population and Housing Census, initiated and actually recorded by Statistics Norway. Hence the coverage of the coded profession in the E/E register is reasonably good from this point and onwards, but still in need of improvement (see, e.g. the discussion and conclusion by Ole Villund[91]), but historical or inactive employments were never systematically coded. Besides, self-employed persons are not included in the E/E register as they are not in an employer-employee situation. A large number of observations from the panel population would thus have been excluded from being given the *profession* variable. Nevertheless, for the purpose of the experiment I transferred the code for *profession* to the panel population from the table 'Employment history' in the E/E system. However, the coverage of the variable appeared to be very low; only 60 per cent of the wage earners in the panel population had a valid value for this variable. My conclusion is that this data source unfortunately is not adequate for establishing a reliable expression for the desired *profession* variable due to low coverage.

A third alternative is to go the opposite way: to add local labour market characteristics from municipality level to individual level based on information about each individual's residential connection. The idea is to apply the municipality classification developed by Statistics Norway. Statistics Norway introduced a municipality classification based on the 1946 Population and Housing Census. Originally this was merely a distinction between urban and rural municipalities. Later, a more nuanced classification was developed. The 'Standard Classification of Municipalities' has been revised subsequent to each Housing and Population Census, but following the same main principles. Relevant issues for the present case study are the standard classifications from 1974, 1985, 1994 and 2003, based on the censuses from 1970, 1980, 1990 and 2001, respectively. Each edition of the Standard Classification of Municipalities has a basic classification expressed by a one-digit-code, as well as more sophisticated sub-classifications within each basic class.

The main municipality classification includes 7 different code values (9 codes in 1974). As an experiment, it is worthwhile to investigate if the classification codes can be applied as compensation

[91] Villund, Ole (2006): Kvalitet på yrke i registerbasert statistikk: resultater og videre utfordringer. Oslo: Statistisk sentralbyrå. Notater 2005/14

for the more desirable data sources like the census files. Obviously the municipality classification codes are only applicable in the regression model above municipality level; at municipality level this is a constant, but even if it turns out not to contribute to the regression model, this will still be a useful auxiliary variable.

According to Statistics Norway, the basic classification is based on the main criteria *industrial structure*, *degree of urbanisation* and *centrality* of each municipality. *Industrial structure* is an expression of which of the primary, secondary or tertiary industries are most fundamental to the economy of the municipality. *Industrial structure* indicates which industries are the most prominent in providing employment for the residents of the municipality. It is very noticeable that industrial structure is based on how the *residents* in each municipality are distributed by industry. Statistics Norway emphasizes that the classification code might have been different if the classification had been based on the municipality in which an individual had his or her place of work, which is not necessarily the residential municipality[92]. The principle followed by Statistics Norway, however, is also followed for the panel population as residential municipality, and not the municipality for place of work, will be the key for transferring the classification code. A central consideration for the classification is also how the working part of the population is measured (see above concerning the LSF and the census definitions, respectively).

Degree of urbanisation is measured by the relative share of the population who live in a densely populated area. Population density is expressed as the percentage of the population who lived in densely populated settlements at the time of the census[93].

Finally, *centrality* is a measure of a municipality's geographic location in relation to a centre where functions of high order are found. Central functions are mainly found in urban settlements. *Centrality* is also an expression of the possibilities the municipality population has of commuting to one or more of the urban settlements. The distance from a given municipality to an urban settlement is measured by travelling time, and the commuting distance is calculated for the fastest means of transportation.

The way the Statistics Norway classification is constructed requires considerable awareness of chronology when this variable is added to the panel population. When classification is made, the parameter status is as of the census date. To make sure that *labour market participation* for each observation may be influenced and affected by residential qualities, and not vice versa, *residential characteristics*, i.e. the municipality classification, has to be added to the panel population with a delay in relation to the point in time when the classification was made.

[92] Statistisk sentralbyrå (1994): *Standard for kommuneklassifisering 1994.* Oslo: Statistisk sentralbyrå, p. 15.
[93] Ibid., p. 12

It would have been desirable to have 100 per cent comparable classification for the whole period, but a few differences between the 1974 standard and later classifications must be dealt with. While 'agricultural municipalities' and 'fishing industry municipalities' were classified separately in 1974, they were merged into one class from 1985 onwards in the class 'primary industry municipalities'. To harmonise the classification system across these standards, I have converted 'Fishing industry municipalities' from the 1974 standard to the joint group 'Primary industry municipalities'.

There are also some other differences between the 1974 and later classifications which make a conversion of 1974 classes to the new classification desirable. The 1974 classification had a residual class called 'Other municipalities'. According to Statistics Norway's additional comments on this class, these were municipalities with low centrality[94]. My conclusion is therefore to convert 'Other municipalities' in 1974 to 'Less central service industry municipalities'. One important remark is that in none of the classification versions does the service concept distinguish between private or public service industries.

Since 1985, the notation of the basic classes has remained unchanged from one revision of the standard to the next. If structural changes take place in a municipality, the municipality itself will be moved to a different class, e.g. moved from class 5 in 1985 to class 6 in 1994. Table 6 shows the main classes from 1985 and onwards.

In any case, the consecutive series of municipality identifiers collected from the DSF income tables raises the issue of selection. As the classification is only revised every ten years and not annually, there is no point in merging each municipality identifier 20 times with the classification tables.

Table 6. Classification of municipalities – main classes.

Value	Classification
1	Primary industry municipalities
2	Mixed agriculture and manufacturing municipalities
3	Manufacturing municipalities
4	Less central, mixed service industry and manufacturing municipalities
5	Central, mixed service industry and manufacturing municipalities
6	Less central service industry municipalities
7	Central service industry municipalities

[94] Statistisk sentralbyrå (1974): *Standard for kommuneklassifisering.* Oslo: Statistisk sentralbyrå, p. 28.

For some observations the residential municipality is the same throughout the 20 years, while for other observations it will have changed once or more. The variable deduced for geographic mobility below is suitable for counting how many observations actually have the same municipality identifier for parts of, or even the whole period. In the event of a residence change during the period, one municipality identifier must be chosen to represent the residential municipality. A solution for this is to split the 20-year observation period into two halves, the first and the last 10 years, and add a classification code at 'year 10' and at 'year 20' respectively, which implies that each cohort has to be matched against two editions of the standard classification. As an example, for the 1937 cohort, residence as of 'year 10' equals residence as of 1976, which means that the classification based on the 1970 census comes prior to the first point of residential measurement. This principle is followed for each of the cohort groups (see Table 35). In the case of a missing municipality identifier as of year 10 or as of year 20, steps have to be taken to ensure that the 'counting' residential municipality is never older than the date of classification; see the chapter 'Data steps for geographic variables', Appendix 1.

The municipality structure itself is not a static quality over time. Historically, municipalities have been merged, and sometimes also split up again. The technical implication of this is that the identifier for a municipality may change over time. In the 1974 issue of the standard, there are some municipality identifiers which do not exist in later issues, and vice versa. To keep track of the merger and demerger of municipalities, I have followed the historical overview of changes in municipal and county boundaries in a report by Dag Juvkam[95].

The matching procedure is technically identical to other matches: adding one new variable – *classification code* – from the donor data sets, which in this procedure is one file based on the standard classifications for 1974, 1985, 1994 and 2003 respectively.

The municipality identifier itself is an auxiliary, and not an explanatory, variable, but it is needed for the linkage between each person's residential municipality and the classification code for that municipality. Later, this identifier will be applied to create data sets for each county and each municipality, and finally the municipality identifier will be useful to identify municipalities that may have some specific characteristics, in addition to the core quantitative analysis.

GEOGRAPHIC MOBILITY

Figures for domestic migration in Norway by 1993 show that about 4 per cent of the population migrates across municipality borders. Migration within municipalities is more frequent, but is incompletely registered and documented in official statistics. Hence nearly all official statistics and

[95] Juvkam, Dag (1999): *Historisk oversikt over endringer i kommune- og fylkesinndelingen.* Oslo: Statistisk sentralbyrå. Rapporter 99/13

analyses deal with migration between municipalities and regions. These are conclusions drawn by Halvard Skiri and Kjetil Sørlie in Sosialt utsyn 1993[96].

Based on the assumptions about geographic mobility and labour market participation outlined above, migration from one municipality to another is assumed to involve a change in employment, possibly also a change from being unemployed to being employed. This assumption is based on observed demographic and structural changes; 'the major redistribution of employment from commodity-producing industries to service industries in the period 1962 – 1991'[97], which to a large extent also involved population redistribution from rural to central areas, noticeably across municipality borders.

By comparing residential municipality year by year, it is possible to deduce migration between municipalities. A new variable, *geographic mobility*, expresses the number of countable different municipality identifiers, i.e. a move from one residential municipality to another residential municipality, and will theoretically have values ranging from 0 to a maximum of 19, i.e. one municipality change per year. However, the actual observed maximum proves to be a total of 16 municipality changes, counted for only one observation. This is later confirmed by a CPR check. The distribution of geographic mobility by gender is shown in Table 14.

As stated earlier, the consecutive series of municipality identifiers may be broken for some observations due to absence from the income tables in the DSF, one or more years in between. The logic for calculating the *geographic mobility* variable must be adapted to the fact that municipality identifiers may be missing for some observations within the series of 20 valid identifiers. Missing identifiers appear as blank spaces. If this is not accounted for, a valid municipality identifier compared with a missing identifier would count as one municipality change.

Operationalization of the variable *geographic mobility* is explained in further detail in Appendix 1 - 'Data steps for geographic variables'. The structure of the data matrix with this variable added may be found in Table 37.

FAMILY RESPONSIBILITIES

This chapter deals with how to operationalize the variable *family responsibilities* which is intended to express parental obligations – basically support or non-support of children - and in this particular context for men and women, respectively, as the observation unit is the individual and not the family.

[96] Skiri, Halvard og Kjetil Sørlie (1993): *Befolkning*. Oslo: Statistisk sentralbyrå. Sosialt utsyn 1993. Ch. 2.1.
[97] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå. Ch. 9.

One component has been essential in many studies of the basic structural changes in the post-industrial labour market in the last few decades: the general growth in employment has been explained by the increasing number of women entering the labour force, with a special focus on the increase among 'married women'. The wider implication of this would be that *marital status* might have an explanatory effect. The variable *marital status* exists in several electronic administrative registers and could easily be transferred to the panel population. However, this is not a good option for an explanatory variable in a longitudinal study, for several reasons.

The technical aspect of marital status in a longitudinal study would be difficult to operationalize, both as a diachronic and as a metric variable. In the CPR, *marital status* is expressed by nine codes and qualifies as an ordinal or categorical variable, though nine different values would perhaps make it acceptable as a metric variable. More importantly, for a lot of observations *marital status* would change during the observation period, from unmarried to married, from married to widowhood, from married to divorced and perhaps married again, and so on. In a diachronic perspective this is troublesome. Marital status at one point in time would not necessarily be representative of marital status for the entire observation period. Besides, one of the trends for the last 30 – 40 years is the fact that an increasing number of couples, parents included, live together without being formally married. Cohabitation is not reflected by the formal *marital status*. An explanatory variable would be required to express the dynamics in changing family and household status during the observation period, which can in no way be expressed by the formal *marital status*.

Surely *marital status* had a direct impact on women's labour market participation in earlier times. In the inter-war period, employers might dismiss married women as well as pregnant women. But when we enter the 1970s, marital status as such does not seem to have any effect on employment or the rate of employment for either women or men. The possible effect of being married or not on labour market participation is also considered irrelevant by Maria Stanfors in her book on the relationship between work and family in the 20th century[98].

In more recent studies on female employment in the last few decades, the concept of *married women* seems to be replaced by that of *mothers*. While formal marital status will change for a lot of people over the years, having children is an obvious diachronic and continuous obligation. Whether or not a person is supporting children (either as a mother or as a father) is the final explanatory variable to be operationalized and added to the panel population. In a longitudinal study this variable should ideally be operationalized in terms of *number of years with children,* counted from the first child's year of birth to the youngest child's 18th birthday, for instance, still as a variable linked to each observation. In accordance with the longitudinal perspective of the present enquiry, the objective is to measure the same parents over a long period of time, and in this perspective the age of the children at one point in time is not essential. As there is reason to fear that data sources for a possible variable

[98] Stanfors, Maria (2007): *Mellan arbete och familj.* Stockholm: SNS Förlag, p. 86.

number of years with children might be inaccessible, using *number of children* instead may be an acceptable replacement.

In official statistics, family responsibilities are often classified by reference groups like 'parents (single or couples) with young children', 'parents with older children' and 'parents with adult children'[99]. Such classifications are often applied in time-series analyses based on cross-sectional data, but in such cases the objective is basically to compare groups of parents over a 20-year period, and the longitudinal aspect of parenthood is absent. If measured over a period of 20 years, parents would pass through all cycles from having 'young children' to having 'adult children'.

Marital status and family responsibilities are basically Central Population Register information, but as of 2011 very little digital CPR information was in the custody of the National Archives of Norway. Besides, such data are inaccessible due to formal restrictions, which is why alternative sources must be identified. The demographic information in the DSF population table is mainly based on CPR information, but there is no field for family obligations in this register. Due to data exchange, however, CPR information can also be found in other public information systems, among them the TRP which appears to be a possible data source for the variable *number of children*.

Taxation legislation and regulations imply various taxation regulations and benefits for parents: marital status, number of children, age of children, etc., are important parameters for tax assessment. Therefore, information about family responsibilities is included in the TRP in various ways. *Number of children* below a specified age limit has been included in the TRP since the income year 1970.

A variable which is possible to obtain from this data source therefore seems to be *family responsibilities* expressed by *number of children*. In any case, this variable will be added by matching the panel population with the available volumes of the TRP. For unmarried parents, *number of children* can be transferred directly to the panel population through one single match, but this is not the case for married parents. The entity in the TRP is actually a taxpayer unit which in the case of married couples might be one observation comprising two, or even more, persons. For tax purposes, it has been sufficient to attach *number of children* to only one person in the taxpayer unit in the case of married couples, and this person is the reference person – normally the father. If spouses were jointly assessed, and only one person had labour income, they would constitute one observation in the TRP with individual variables only for the reference person. Joint assessment was common in the first years after 1967, but gradually changed and almost ceased after the tax reforms in the second half of the

[99] Cf. e.g. Statistisk sentralbyrå (1995): *Statistisk årbok 2008*. Oslo: Statistisk sentralbyrå: Classifications - in Table 64.

1980s[100]. This clearly represents an element of uncertainty for the coverage of the variable *number of children*.

This is an example of administrative practice which raises a designation problem in a context where the entity is the person. Single parents represent no problem for the panel population, but family, or couple, variables must be converted to individual variables. For married couples in a taxpayer unit, this means that *number of children* must also be transferred from the reference person to the spouse.

If spouses were jointly assessed, the second spouse – i.e. normally the female – is simply not present as an individual observation in the TRP unless she had individual income or property. For the purpose of a match, the risk is that there is no female to transfer the variable to, and the coverage of the *number of children* variable might be too low for the female part of the panel population. In cases of couples with children, the result might be a high number of married men with children, and a high number of married women without children. Considering that joint tax assessment was more common at the beginning of the observation period than towards the end, there is reason to fear that the *number of children* variable will be skewed in disfavour of the oldest female cohorts. This is a serious problem which might put the desired variable *number of children* at risk. The TRP is obviously a more inaccurate data source for this variable than the CPR. The coverage of the variable *number of children* and match results between the panel population and the TRP will be discussed in the chapter 'Historical criticism'. The technical solutions for adding this variable to the panel population can be found in Appendix 1. Table 37 shows the structure of the final data matrix with the variable *number of children* included.

FINALISING THE DATA MATRIX

To finalise the data matrix the panel population must be anonymised. The national ID number is no longer needed once gender and cohort information has been extracted from it, and the final match has been carried out. Since the data matrix in this particular case is produced and analysed inside the secured zones of the National Archives of Norway, and in any event will be deleted when the analyses are finished, this step is not strictly required. But similar data may of course be assembled and applied in other research projects by researchers who need to transport the final research data set from a secured zone and into their own 'laboratory' for analyses, in which case measures must be taken to make data anonymous. The panel population can easily be anonymised by replacing the ID number with a counter.

[100] Strand, Børge (1992): *Personlig inntekt, formue og skatt 1980 – 1989.* Oslo: Statistisk sentralbyrå. Rapporter 91/18, pp. 31.

With the panel population fixed and all variables added, it is time to summarise what kind of variables (X and Y) it was possible to operationalize:

- *Gender*
- *Birth cohort* (year of birth) – which also equals the observation period
- *Labour market participation / employment history*
- *Residential characteristics* – as of year 10 and as of year 20
- *Geographic mobility*
- *Number of children*

Variables, in social science in particular, may be more or less precise. The data matrix just created will contain both nominal, ordinal and metric variables. The dependent variable – *employment history* – will range from 0 to 20, the intervals are measurable and the distance between each value is one year. Calculating an average value would be meaningful for this variable. This is a true metric variable.

Birth cohort ranges from 1 to 22, the distance between each value is one year and this is also a metric variable. *Geographic mobility* (the number of countable changes between different residential municipalities), and *number of children* are both metric variables.

Gender is a nominal variable with only two values, and thus a dichotome variable. In regression analysis, a dichotome variable is accepted if treated as a dummy variable. A dichotome variable will express a possible difference in level between the two categories.

The *residential classification* is an ordinal type variable. The municipality classification codes range from 1 to 7. This is debatable, but it seems to be accepted that an ordinal variable with 7 or more categories may be treated as a metric variable in regression analyses[101]. Tests will later reveal if this particular variable is applicable in the regression analysis.

If imported to a data base, the anonymised data matrix would appear as in the table extract shown in Table 7 where the counter and the analysis variables are displayed. The primary key is now a counter running from observation number 00000000001 to observation number 00001231612. The table example is sorted by ascending *counter*. The column for *gende*r contains either '0' for male or '1' for female. The *birth cohort* column contains a number to represent each year of birth from 1937 to 1958.

Observations in this example range from number 59 457 to number 59 461. From this table we can e.g. deduce that observation number 59 458 is a woman who belongs to birth cohort number 9, which tells us that she was born in 1945. She has no employment history (as opposed to her

[101] Midtbø, Tor (2007): *Regresjonsanalyse for samfunnsvitere. Med eksempler i SPSS.* Oslo: Universitetsforlaget, p. 33.

neighbours in the table) and her classification code for municipality tells us that as of year 20 she lives in a 'central, mixed service industry and manufacturing municipality'. She has probably lived in this type of municipality for 20 years, as no migration between municipalities is counted for this observation. Finally, we can also read from the table that she has one dependent.

Table 7. Extract from the final data matrix with a few random observations and selected variables.

Counter	Gender	Birth cohort	Employment history	Residence classification as of year 20	Geographic mobility	Family responsibilities
00000059457	0	9	20	5	0	3
00000059458	1	9	0	5	0	1
00000059459	1	9	15	5	0	3
00000059460	0	6	20	7	0	2
00000059461	0	7	20	6	1	2

Even if all the information in the data matrix is coded and anonymous, it is quite possible to 'read' individual stories – each row tells a personal story. Alternatively, all codes could have been replaced with code explanations to make the data readable for the human eye; however, this data set is not intended to be analysed visually, but by the computer. This example also documents that although it is impossible to identify individuals in the data matrix, the entity is still the single person.

HISTORICAL CRITICISM

So far the main focus has been on the technical aspects of creating the panel population and the variables of the data matrix. Possible data sources have been identified and relevant information has been extracted from the sources to be implemented in the data matrix. With the research data set completed, it is time to return to the data sources and ask critical questions about their evidential power, quality and applicability. General criticism may be discussed at records creator and system level, but more specific criticism must be applied at data set level, file or table level, and ultimately at variable and observation level.

An electronic administrative system is constructed to serve a defined administrative purpose, and a system may run for decades after its implementation. In the course of its life cycle any system will be subject to technical and other changes, but will still understood as one system. The Central Population Register, e.g., is still regarded as one system, despite technical and administrative changes that have taken place since 1964. As explained above, a system as such is never preserved for the archives; what is preserved is information from a system in the shape of files or tables extracted from the system accompanied by necessary technical metadata. The transfer from technology-dependent to technology-independent format is in itself a possible source of error.

In general, this criticism implementation is based on common practice and principles, which may be found in various textbooks, such as '*Fortida er ikke hva den en gang var*' by Knut Kjeldstadli[102]. Kjeldstadli recommends for a start a systematic classification of the actual sources based on a set of characteristics: linguistic – non-linguistic, narrative – non-narrative, private – public, formal – informal, primary – secondary, first hand – second hand, etc.[103]

As a characteristic feature of electronic administrative registers is the extensive use of coded information, the question is whether such sources can be classified as linguistic. The examples in Figure 2 and Table 7 illustrate that this kind of source material definitely looks different from the plain text information that we normally find in paper-based sources. We must bear in mind that this appearance is typical for the type of digital archives classified as registers.

My conclusion is that electronic administrative registers are both linguistic and narrative even if their information – the story told by them – is presented in a coded form. All of them have the potential of being linguistic in the real sense of the word. A simple example of the linguistic content is the information about gender which is included in the national ID number and then deduced as a separate variable with the code values '0' and '1'. Of course this can be converted to understandable text like 'male' and 'female', instead of the digits '0' and '1'. Another example is the municipality classification, where classifying a municipality by the code '4' based on the 1985 standard is just a

[102] Kjeldstadli, Knut (1992): *Fortida er ikke hva den en gang var.* Oslo: Universitetsforlaget
[103] Ibid., pp. 161.

way of expressing that this municipality is a 'less central, mixed service industry and manufacturing municipality'. And as the comments on Table 7 show, individual stories as well as collective stories are actually told by the collection of codes. These sources do have a narrative aspect; the challenge is to utilize it.

Despite the exactness expressed by codes, there is room for interpretation of what the codes actually express. All codes are assigned on the basis of some human assessment: in general when code systems and standards are developed and, in the case of individual decisions, at the moment a piece of information is actually coded, e.g. when transferring information about profession from plain text to a code system. Underlying this decision there is an assessment. Nuances that can be expressed in plain text are lost in a code system, which allows for interpretation. Anyway, in the context of electronic registers, a given code system is always applied during the creation process, by the records creators themselves. This differs clearly from computerised archives where plain text information, which is basically unstandardized, both conceptually and in spelling, is transcribed and subsequently converted to codes for research purposes.

It seems as though a huge amount of data is available within the framework of digitally created archives, and measured by size there is, but in fact the most basic source material that was finally chosen for this particular enquiry came from a very limited number of records creators:

- The taxation authorities
- The social security and welfare authorities
- Statistics Norway

Furthermore, these few records creators share the common feature of being public agencies. The data sources applied from the two first records creators above are also subject to privacy restrictions. From Statistics Norway, however, only the municipality classification is applied. This is published material with the municipality as the entity, and accordingly is not subject to any restrictions on use.

The major part of the data for this analysis has been collected from a number of data tables in two systems, the Tax Register for Personal Taxpayers and the Central System for National Social Security. Both these systems and the archived data sets must be classified as formal: the systems were developed by governmental institutions to serve fundamental, legally-based functions in the welfare state, i.e. calculation of taxes and documentation of pension rights, respectively.

How are the systems filled with data and values? Basically there are four ways of filling a system with data:

- By manual data entry

- By optical input (e.g. OCR-scanning)
- By electronic input (e.g. selected tables or selected fields imported from another system)
- By generating values through system procedures

For the data sets applied for this case study, all methods have been used. The different methods are often attached to various stages in the life cycle of a system. At the start, the TRP is filled with demographic information transferred from the Central Population Register (via another system). Then the various amounts for income, property and deduction are added optically, or electronically by matching procedures, and are also entered manually. However, this has changed over time from mainly manual and optical input to more and more computer-based data exchange. Finally, taxes are calculated and generated by system procedures. Demographic information in the DSF also comes from the Central Population Register, while pensionable income amounts are transferred electronically from the TRP annually. Codes are generated by the system or entered manually.

For some systems there may be a 1:1 relationship between a system and a records creator, but more commonly an electronic administrative system is updated by several instances: by both local and central authorities, by private or public actors individually, and to a large extent by data exchange. Actually, we see this kind of multi-provenance in most central administrative systems. From a historical criticism point of view, this raises the question of mutual influence and dependency between data sources.

By way of introduction, the importance of data exchange between systems and between records creators was emphasized. Data exchange affects the relationship and creates dependency between different systems. Sometimes it is difficult to decide where information appears in the first instance (i.e. the values of a given variable). The starting point, and the question of what is the actual origin of the source, are determined by the direction and succession of the data flow. Answers to questions about relationships and dependency are found in flow charts and other technical metadata. A flow chart documents where a variable gets its value, or by what procedure a value is calculated, or derived, and whether a given variable is exchanged with other systems. Data flow may be one-way or two-way. In the latter case, it can be difficult to decide which source is the origin.

To a large extent, it is possible to compare key figures from the data sources applied above with public statistics, mainly from Statistics Norway. An important remark is that Statistics Norway uses both the Tax Register for Personal Taxpayers and the Central System for National Social Security as data sources for statistics. The TRP is used directly as the data source for tax and income statistics, while the DSF is the data source for various social welfare statistics. Moreover, both systems are used indirectly in several other statistics. In other words, there is dependency in many directions. It is a fact that all data sources applied by Statistics Norway are subject to extensive quality control, logical and consistency tests, etc.

A major difference between the data sources transferred to Statistics Norway and to the National Archives of Norway, respectively, is the time lag. Statistics Norway receive their volumes of data as soon as possible from the records creator; e.g. tax statistics for 2010 are produced in 2011 and onwards based on a TRP version as of autumn 2011, still for the income year 2010. As mentioned earlier, taxpayers are entitled to put forward complaints for 10 years, and in a few cases even for 13 years. Possible changes in the course of these years make the TRP a living set of data for at least 10 years. Consequently, a final version of the TRP does not exist until the time limit for complaints has expired. The TRP version for the National Archives of Norway is transferred after all possibilities for changes and updates have been closed. Thus there will actually be figures based on different versions that are compared in my case, but after all, changes are marginal if compared at macro level.

Basically all information in the TRP comes from multiple sources, but they are collected, assembled and processed in a system where the output is the annual data set. In this process new information is generated: amounts for the various types of taxes are calculated and stored as separate fields in the data set, as well as lots of other systems-generated variables. The final version of the data set is a mix of data of different origin: from administrative registers, from taxpayers, from employers, from banks, etc. Thus the TRP is a collection of both unique and redundant information[104], but the internal junction and consistency establish this system as a unique set of archives where the records creator is the taxation authorities.

Information about pensionable income – wages and salaries and income from self-employment – is extracted annually from the TRP together with the national identifier and a few other fields, and then transferred to the social security and welfare authorities to be accumulated and stored in the DSF. In this case the data flow is normally one-way, from the TRP to the DSF, and not the opposite way. Data may be corrected and changed in the DSF, but not reported back to the TRP.

The data exchange offers an opportunity for consistency and quality control. Any observation with an amount for pensionable income should be found with exactly the same amount in both the TRP and in the DSF, and this may be checked electronically. Inconsistency may appear, though, due to complaints and asynchronous updating. An example of such comparison is shown in SAS data steps 23 and 24.

To sum up, applying traditional, historical criticism to electronic registers generally poses no problem. However, some of the practical criticism has to be performed electronically, which is a major difference from paper-based sources, but also a major advantage. Furthermore, this kind of criticism obviously requires the prior existence of a physical data set.

[104] I.e., identical information existing in several other systems.

AUTHENTICITY AND RELIABILITY

When archives are transferred to an archival institution, whether they be paper-based or digital material, the reception procedure is normally the same: there is a direct contact between the records creator and the archival institution, and each acquisition is always subjected to close examination and control. Nevertheless, the production of an electronic archival package is a critical moment in the life cycle of all digital archives, as errors may be produced during the extract from a native system. For long-term preservation in an archival institution, format conversions may be needed from time to time, and this is also a critical moment – human and technical errors can occur.

The National Archives of Norway normally receives table extracts from electronic registers directly from the records creator without any intermediary. Table extracts from electronic registers for the archives are collected as packages from reliable bodies. The information in the table extracts must be authentic, in the sense that the archived data must be identical with its origin at the records creator.

In active systems, the information is continuously monitored and controlled by the records creators and potentially by everyone who is registered in the system; a quality check of individual information about taxes and pension rights is performed by possibly every taxpayer in the country.

When a data base or a register extract arrives at the National Archives of Norway, there is a procedure for quality and integrity control. Acquisitions are provided with a set of checksums by the records creator to make changes traceable: any change of bits or bytes in a data set will change the checksum. The National Archives of Norway's routines for such controls have changed a lot since the first attempts in the 1980s and up to the present. Today's routines are systematic, and theoretically and technically solidly based, all with the intention of securing the authenticity and integrity of the archives. However, the new routines are applied only to present and future acquisitions. Transfers and deposits in the past were not subjected to the same kind of quality controls, which makes historical criticism even more necessary.

The quality checks on data sets extracted for the archives are never aimed at reporting errors in order to make the records creators change or 'improve' their data. When errors are encountered, which happens all the time, the central issue is to clarify whether these can be found in the native system, or if they were caused by the extract procedures. Genuine errors must not be corrected, but remain unchanged in the archival version, like the invalid ID numbers discovered in the DSF, when their existence is confirmed as genuine by the records creator. In the case of electronic registers – administrative or research registers – some controls are more or less standardised (ID number validity check, duplicate checks etc.). Invalid ID numbers are discovered in almost all systems, normally not many, but if this is in accordance with the native system and confirmed by the records creator, this will of course be conserved in the archived data set. But the question of how to deal with invalid ID numbers in a specific research context is an issue for the given historical or scientific criticism. My

solution for handling invalid ID numbers was to deselect such observations, which was an easy choice anyway as there were few in numbers.

Due to data exchange, the same data may occur in several public systems which represents a possibility for discovering errors, inconsistency and discrepancies. Authorities continuously monitor registers by means of data exchange and 'register cleansing', i.e. matching registers in order to identify discrepancies.

Current international discussions are very much concerned with authenticity in digital archives. Technical solutions for securing authenticity are continuously under development, but this is mainly for the benefit of future acquisitions for the archives. For historic transfers, the likelihood of authenticity must be analysed by means of historical criticism. For paper-based sources there is an unbreakable connection between the information itself and the information carrier – the medium which carries the information, such as a sheet of paper, a protocol, etc. Together, the information carrier and the information itself constitute an original set of archives for paper-based sources.

On the other hand, authenticity alone is no guarantee of the reliability of the information, and does not take away the researcher's responsibility for criticism. There is no doubt about the authenticity of the church register from which the image in Figure 1 was taken. In this list of baptisms there is a child named Hans, born on 5 July 1821. Hans was a so-called illegitimate child, and as the father of this child is listed Erich Erichsen Dystvold–Eie. But even if authenticity is unquestionable, it is still an open question if Erich Erichsen really was the 'father in the flesh' in this case. Ultimately, this question is left to the judgement and the criticism of the genealogist. This example is put forward to underline the timelessness of the authenticity and reliability issue. However, digital archives add an extra dimension to this issue.

The concept 'original' is a tricky issue in a digital archives context. With digital archives the connection between information and information carrier, i.e. the storage medium, is different in nature. A discussion about 'originals' must be decided on the basis of information about provenance, logical controls, consistency and whether information can be traced back to a verifiable origin. The census data bases created by Statistics Norway since 1960 are of course accepted as 'originals' in the historical sense. Censuses are increasingly based on input from administrative registers and still appear as 'originals'. In general, new original material can be created by combining and matching register information. In fact, the data matrix for the present enquiry is also a new original; information has never been assembled in exactly this combination before. Yet this data set is to be deleted, but with the knowledge that it may be recreated in identical form at any time.

With paper-based sources, the storage medium, such as a protocol, can provide a great deal of information about authenticity. We can study paper type, ink and handwriting and make judgements about authenticity. With digital archives, we cannot turn to the storage media and deduce anything about authenticity or decide a possible original status. The connection between digital information and

storage medium is only of a transitory nature. Transfer from one storage medium to a new one is the only way of keeping digital archives alive and accessible over time. The CD-R which used to be common storage media a few years ago will be replaced by new storage media within a short time. The question of authenticity has to be considered independently of storage medium when digital archives are in question.

If we return to electronic case handling systems for a moment, the question of a 'signature' – handwritten or electronic – is crucial in the assessment of a single document's authenticity. But for administrative registers or research registers with information in structured form, such as rows, columns and often coded information, there is no place to put a signature. Considering the TRP as an example, separate fields should have been furnished with a signature for each of the individuals involved: the taxpayer, the employer, a bank representative, the taxation authorities, etc. But this is never the case, and of course is not needed for the administrative function of the register. By the way this is also a common situation for structured information in paper-based card files, questionnaires, etc.

When data flow from one system to another, there should always be some kind of consistency between the systems involved. Based on information about data exchange and data flow, obvious authenticity and reliability controls consist simply of comparing information in different sources to identify possible discrepancies. Data exchange offers an important means of evaluating sources by matching and comparison, through procedures which may be called 'computer-based source criticism'. Identical information about the same entity should be found in different systems from different records creators.

Administrative registers are subject to quality and consistency checks in many instances throughout their active service. In technical terms, reuse of information from a given data set is a duplication of data. Unlike paper-based material, however, a duplicate of electronic archives is a duplicate of bits and bytes, and does not lose readability. A more correct term for this duplication is actually a clone process; the duplicates are 100 per cent identical. The real problem connected with this process is potential human error in creating source code for the retrieval process, selection process, etc., or in my case, e.g. erroneous programming code during the construction of the data matrix. A specific researcher's handling and processing of raw data constitute an obvious risk of corrupting data, which is why consistency checks and comparison procedures throughout the research process are very important. Key figures in the final research data set and the native data should really be comparable, and any differences should be explicable. The risk of errors made by the researcher is of course equally present whether sources are paper-based or digitally created, but such errors may be investigated more efficiently when sources are digital.

CONSISTENCY AND MATCH RESULTS

One way of controlling register information is to check consistency through logical tests. As an example, a tax amount is calculated as a percentage of a given income amount. For the sake of consistency, the reversed calculation might be performed. From the tax amount it is possible to calculate what the income should be, and then any discrepancy between the calculated and the empirical amount may be identified.

Principles and methods for checking data quality and consistency between administrative registers, and between total populations and population sample data by record linkage, are easily found in literature. One example is the report 'Measuring working hours in the Norwegian Labour Force Survey' by Ole Villund [105]. This report deals with comparison and consistency control between the survey-based data for the Norwegian Labour Force Survey (LFS) (see above), and a central government register with data about employment reported by employers, the E/E system, also referred to above. Villund's report is of interest beyond the specific labour force perspective, as the method of micro-level record linkage is applicable in general, though in that particular case a main problem was the linkage key itself.

The two main data sources for my enquiry are the Tax Register for Personal Taxpayers and the Central System for National Social Security. These two systems may be compared at both macro and micro level. At macro level, some figures can be calculated and compared, while at micro level information can be compared at individual level by record linkage. If compared within the same income year, the same observation should have the same labour income amount in both systems. This kind of comparison is both important and effective as part of the historical criticism when the sources are electronic administrative registers. Discrepancies may occur, but measurements of their extent will give information about their possible influence and effect on the research result.

A logical approach to consistency control is to identify the origin of the information in these two systems. All pensionable income amounts in the DSF are transferred from another central national system: the TRP. The origin is always the tax returns from taxation authorities from which the annual pensionable income amount is transferred to the social security authorities and stored in the DSF, which accumulates pension entitlements. The data flow is always one way in this case. Hence there is a dependency between these two systems. An important difference between them, though, is that the DSF may be updated, or corrected, without a parallel update of the TRP. For corrections of the DSF older than 10 years, there would no longer be any TRP to update. An effect of this is that some discrepancies must be expected.

---

[105] Villund, Ole (2009): *Measuring working hours in the Norwegian Labour Force Survey: a pilot study of data quality using administrative registers.* Oslo: Statistisk sentralbyrå. Reports 2009/3

Both population and amounts may be compared in a match between these systems. Basically, the annual versions of the TRP comprise a larger population than the annual version of pensionable income tables in the DSF simply because many taxpayers have taxable income which is not pensionable (e.g. only capital interest) and hence such observations are not transferred to the DSF income tables. This means that in principle all individuals from the DSF income tables should have a match in the TRP, but not vice versa. A very important difference, however, is the difference in entities: the taxpayer unit in the TRP and the personal unit in the DSF.

To obtain a match between two data sets in general, one condition is that the data sets have some part of their population in common. Normally, however, each data set will have a remaining number of observations which do not match; in short, both a 'joint population' as well as 'separate populations' will exist, where the latter is the population that only exists in one of the data sets. The proportion of the joint population is one of the decisive conditions for the coverage of the variable(s), which are transferred during a match. A mismatch will cause a missing value for that specific variable. Furthermore, the coverage of the variable(s) in the donor data set is also decisive for the outcome: the match may be good, as reflected in a high rate of joint population, but the coverage of the variable in the donor data set determines how good or bad the coverage of the transferred variable in the output data set will be.

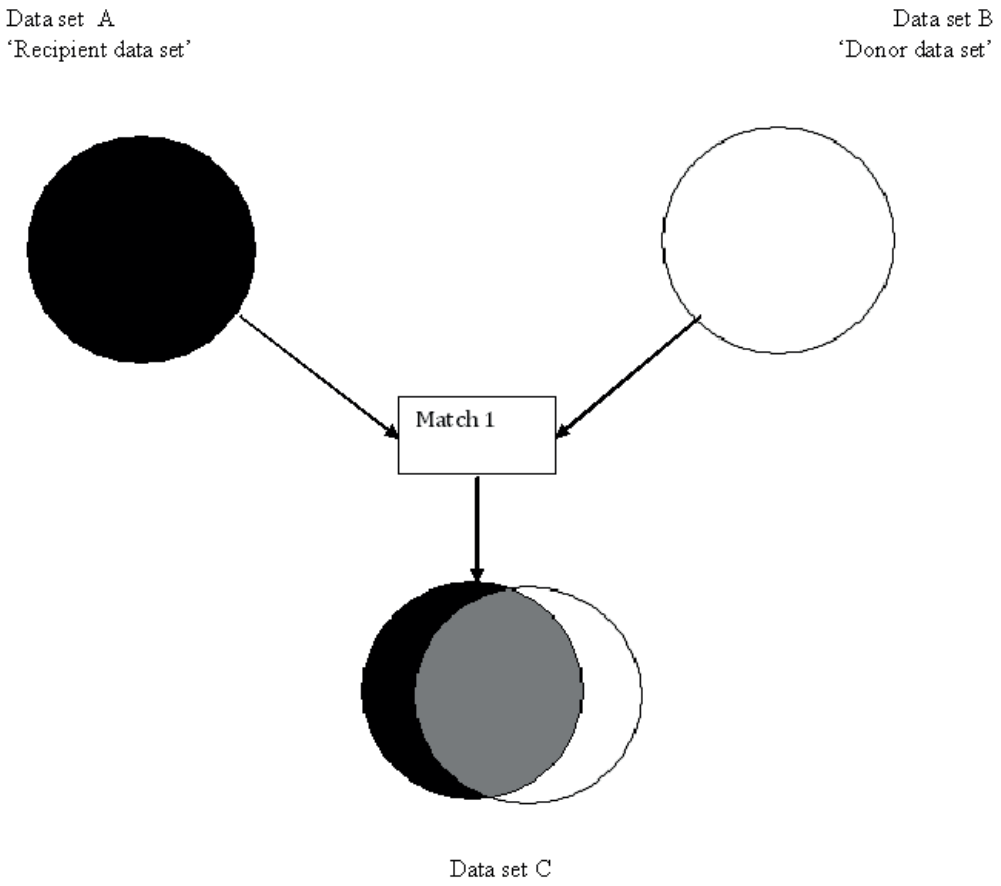
When two data sets are matched, e.g. by the national ID number, there are in principle three possible outcomes:

1. The same ID number will appear in both data sets, and there is a match. A joint population exists and variables can be transferred from one observation in the donor data set to the same observation in the recipient data set. The optimal situation is a 100 per cent match.
2. An ID number appears in only one of the data sets, and there is a mismatch. Variables cannot be transferred. The worst-case scenario is a complete mismatch – 100 per cent separate populations.
3. A mix of 1 and 2, which is the normal situation. In this case there will be missing values for given variables for parts of the population.

Figure 6 illustrates the normal situation. The grey segment in data set C represents the joint population, i.e. observations which have a match, while the black and the white represent the separate populations, i.e. observations without a match. The more the circles overlap, the better the match. Variables from the donor data set, symbolised by white, can be transferred to the joint population, symbolised by grey, but not to the remaining population, symbolised by black. This is typical for a match between the panel population in the present data matrix and the TRP.

Sometimes there may be a discussion about which population should be included in the output data set, i.e. data set C in Figure 6. The number of observations and which population is to be continued in the output data set from a match should be decided in advance, and the source code written on the basis of that decision. Both the DSF and the TRP are total populations, but from the first system a limited population was selected for this enquiry, based on year of birth as well as some other criteria. Once the population requirements are fulfilled and the panel population is finalised, this population should not be changed. With reference to data set C in Figure 6, the black and the grey population represent the panel population. This population continues to the next match, while the white never does and so on.

Figure 6. Match between two data sets with partly a joint population and partly separate populations.



A perfect match in the sense that all observations from data set A have a match in data set B is quite rare, but this might be the case where data set A is a population sample, while data set B consists of a total population, like a CPR data set.

The fact that the Tax Register for Personal Taxpayers was only preserved at five-year intervals up to and including 1990 reduces the number of possible comparisons. As an example, and as a real check, a comparison between the 1990 TRP and the 1990 DSF income table is presented in Appendix 1. The intention with this match is to check both population consistency as well as consistency of pensionable income amounts. Technical details about this match are presented in SAS data steps 19 and 20.

A general experience from matches between the panel population and the TRP is that the match is better with the more recent generations of the TRP, which is partly an effect of a more individually oriented taxation system entailing a transition from joint tax assessment to individual tax assessment and hence a transition from taxpayer units to individual units, see the information on taxation reforms from 1987 and onwards above. This is partly also a consequence of the increase in labour market participation, especially by women. Income from labour – pensionable income – is always attributed to the individual, and a higher number of people with labour income necessarily leads to a higher number of individual registrations in the TRP.

VALIDITY AND COVERAGE OF THE VARIABLES

The validity assessment concerns whether a variable really measures what it is intended to measure. This assessment is closely linked to the coverage issue and how representative the variable is. The question of coverage concerns both the panel population itself and all variables. This involves an examination of value range for each of the variables, identification of possible missing or invalid values, or lacunas in the population, etc. For some of the variables, the coverage issue has already been clarified through the initial requirements. *Gender* and *cohort* have already proved to have a 100 per cent coverage in the data matrix. The linkage key at individual level was the national ID number, which in the case of the panel population also has a 100 per cent validity and coverage. More problematic is of course the coverage of explanatory variables that were added to the population from other sources through matching. The quality of a match is essential: a match between two independent data sets can be good or bad depending on the quality of the linkage key in both data sets. In this case study all the donor data sets represent some uncertainty.

A final step in the discussion about source criticism is simply to run some tests on the data matrix, and thus evaluate the variables one by one. Since all variables are coded, the first step will be to check if all variables have a legal value range, and the possible occurrence of values outside a given

range or simply missing values where there should not be any. Gender should only appear with the values '0' or '1', the dependent variable should have values from '0' to '20', etc.

Table 8. Overview of observations and value range of variables in the data matrix.

Descriptive Statistics						
	Number of observations	Range	Minimum	Maximum	Mean	Standard Deviation
Employment history	1231612	20	0	20	14.79	6.483
Gender	1231612	1	0	1	.49	.500
Birth cohort	1231612	21	1	22	12.35	6.091
Municipality classification by year 10	1231612	6	1	7	5.06	1.938
Municipality classification by year 20	1231612	6	1	7	5.16	1.825
Geographic mobility	1231612	16	0	16	.70	1.140
Number of children	1231612	12	0	12	1.41	1.192
Valid cases	1231612					

All tests in this chapter are also tests of my own source code and processing in building the data matrix. Any suspicious results might be the result of erroneous source code rather than weaknesses in the data sources. A general check of both consistency and my own source code is that some key figures are the same all the way, e.g. number of observations, number of men and women, number in each cohort. The frequency tables are based on SPSS procedures, which implies that all present values for any variable are displayed, and not only predefined values. An overview of observations, variables and value range is presented in Table 8.

The descriptive statistics procedure in SPSS provides the number of valid cases, the mean, and the standard deviation for each variable in the analysis. The most important information to be drawn from this table is that all cases are valid, the value range of each of the variables is as expected, and the standard deviation figures are very small. The variables are examined more closely below.

THE ID NUMBER

The basic variable is actually the national identity number which serves a lot of functions. Firstly, it simply separates one individual from another in a table, and secondly, it serves as the primary key in many public information systems, data bases and registers. It is national in the sense

that it makes matching between various data bases possible, a condition for exchanging individual information nationwide across systems and across institutions. This identifier is also a condition for computer-based construction of panel data with the individual as the entity.

The ID number is constructed according to an algorithm which makes it possible to run a validity check on the identifier, as was done on the DSF population table above. The validity test showed that only 12 out of more than 6 million ID numbers failed the test. Nonetheless, these 12 observations may be examined more closely. For most of them it is obvious that the errors are found in the personal number, i.e. the last five digits of the ID number, while the date of birth is valid. Three of these observations have a year of birth within the range set for my panel population, and accordingly they should have been included in the panel population. But there is really no problem at all in omitting three observations from a total of more than 1.2 million observations. Consistency within the DSF system is actually proved by the fact that these invalid ID numbers also occur in the income tables.

My conclusion is that the ID numbers in the panel population are 100 per cent valid. Accordingly, the variables deduced from the ID number, *gender* and *birth cohort*, will be valid, provided that the source code for deduction is correct. Of course, the implication of this is also that any match with another data set will necessarily be a match between valid ID numbers. Bearing this fact in mind, the occurrence of invalid ID numbers in any of the donor data sets really does not matter: they will not match with the recipient data set, which in this case is always the panel population.

GENDER

The quality of the ID number is decisive for the quality of variables derived from it. As long as the ID number is valid, a valid gender code will always be derived, provided that the program code used to derive it is correct. This result can of course be measured. For the panel population the value of *gender* was set to '0' for male and '1' for female. As we can see from Table 9, there are no invalid or missing values for gender. Any such value would have appeared in the frequency table.

Table 9. Panel population by code for gender.

Code value	**Frequency**
0 (male)	629 947
1 (female)	601 665
Other	0
Total	1 231 612

Considering the ID number validity check above, this result is as expected and also confirms that my source code for deriving *gender* is correct.

BIRTH COHORT

Birth cohort is deduced from the ID number as demonstrated in SAS data step 4. Behind each ID number there is a physical individual, i.e. everyone should be found in the Central Population Register with name, address, and family situation, history of marital status and migration. The Central Population Register is the origin for demographic information. Demographic information based on the DSF should be comparable with official statistics which are based on the CPR, at least partly, and discrepancies should be explicable. In Table 10, the net panel population is distributed by birth cohort groups and compared with similar figures from official statistics. In this test the birth cohort number is replaced with the actual year of birth. The values in the panel population for birth cohort range from 1937 to 1958. No values outside this scope are discovered, which also confirms that my source code for deriving and calculating the year of birth is correct.

Table 10. Net panel population and population in official statistics by birth cohort groups.

Birth cohort group	Panel population based on DSF	Population in official statistics[106]	Difference
1937 -1940	164366	183673	19307
1941 - 1945	257636	280334	22698
1946 - 1950	313470	329432	15962
1951 - 1955	306596	312390	5794
1956 - 1958	189544	190219	675
Total	1231612	1296048	64436

Table 10 reveals some discrepancies which call for an explanation. Figures from the official statistics are generally a little higher than the figures from the DSF. In the column 'panel population based on DSF', the figures comprise individuals born in the years in question and still alive by the age of 50 which was one of the requirements for the panel population, while the figures in 'Population in official statistics' comprise any person born in Norway in one of these years independently of when a person died, or possibly migrated. We must assume that a percentage of each birth cohort died before

[106] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Ch. 3, Table 3.13. Oslo: Statistisk sentralbyrå.

the age of 50, and that some persons also disappeared due to emigration, which explains the lower number in the panel population for the older cohorts. The net panel population was required to have at least one valid municipality identifier during the first half and the second half, respectively, of the observation period. This is a requirement which reduces the number of observations in the panel population.

For the official statistics, data was collected from paper-based sources until 1964 and include all births for each year, i.e. all births in Norway, but births in Norway only. Persons born abroad are not included in the figures for births in the official statistics. On the other hand, it is not a condition to be born in Norway to be entered into the social security system. People born in other countries are entered into the DSF when they become members of the social security system. This explains why the difference is smaller for the younger age groups.

It is also important to note that the number of observations in the net panel population was final and static once all requirements were fulfilled. The number of observations in the panel population had to be kept unchanged during all the matching procedures. Matching against data sets with a higher number of observations has been the main rule, as the donor data sets comprise total populations without any prior selection of birth cohorts. Each match only added a new variable to the members of the panel population; it was never the intention to increase the number of observations. Anyway, the variables *gender* and *birth cohort* are reliable and valid.

RATE OF LABOUR MARKET PARTICIPATION

This variable would logically have values ranging between 0 and 20. Table 11 shows that no observations appear to have a value outside this range. There are no official statistical figures directly comparable with the figures in this table in the sense that the data sources are different. Official statistics are not based on panel populations in general, and are not specified for the same age groups as in my panel population, but for the labour force in general. However, if the dependent variable is a valid measurement, the general trends observed in the panel population should definitely coincide with the general trends depicted by other sources, as in the official statistics.

The standard deviation for the total panel population indicates that the result is significant within a scope of +/-6 observations. The median is generally higher than the mean which indicates that the variable is not normally distributed in the panel population. This might cause problems for the analysis, as the model works better when values are normally distributed, but this is nonetheless a reflection of the empirical situation. For the male population the median is actually 20 years of employment, which indicates that more than 50 per cent of the male panel population has a continuous employment history. The median and the mean clearly differ between genders, but more importantly, the differences coincide with general observations and expectations.

Table 11. Panel population and labour market participation: maximum, minimum and average. Total, men and women.

	Number of observations	*Minimum value*	*Maximum value*	*Mean*	*Median*	*Standard deviation*
Total	1231612	0	20	14.79	18.0	6.5
Female	601665	0	20	12.23	14.0	6.8
Male	629947	0	20	17.24	20.0	5.1

Though there are no identical figures to be found in official statistics, a possible reference is the major trends observed for the period. A possible consistency check is therefore to see if general labour market trends based on official statistics in this period also are reflected by the deduced variable *labour market participation*. Figure 7 and Table 12 show the same trends that have already been reported and confirmed by other research: increasing female and stable, high male employment. The major difference between men and women can also be seen in the share of the panel population with a maximum of 20 year participation in the labour market as shown in Figure 7. The value range of this variable is as expected and it appears to be consistent with the main labour market trends observed at national level.

Table 12. Panel population by number of years of labour market participation 1967 – 2007. Total, male and female.

Years of labour market participation	**Total**	**Male**	**Female**
0	73491	15322	58169
1 – 5	91176	23992	67184
6 – 10	130930	33437	97493
11 – 15	189509	57860	131649
16 – 20	746506	499336	247170
Total	1231612	629947	601665

Table 12 shows distribution by rate of labour market participation and by gender at national level for the total panel population. As expected, both tables show an evident difference between men and women. Among both men and women we find observations without any labour market participation at all, as well as with the maximum of 20 years' continuity. But the most apparent difference between men and women is found in the number of observations with the maximum 20 years of labour market participation: around 60 per cent of the male panel population is counted as

having the maximum of 20 years, while only 30 per cent of the female panel population reached this maximum. The cohort differences within the female population are clearly illustrated by Figure 7 and Figure 8.

It is appropriate to ask if this difference is due to the construction of the dependent variable. However, there is reason to believe that this difference between men and women would have been even more striking with a more rigid measurement, e.g. with the 4 G income line as discussed earlier.

If observed along the time axis, there is a clear trend towards gradually increasing labour market participation for women, as already shown in Figure 7. Turning to the other extreme, i.e. observations with no traceable employment history, the picture is basically confirmed by the opposite trend. Figure 8 shows the relative proportion of men and women, respectively, without labour market participation. For the male birth cohorts, the share without any labour market history is less than 5 per cent for each cohort. This share remains almost unchanged throughout the period. On the other hand, the proportion of women without labour market participation is high – above 20 per cent - for the oldest cohorts, but gradually falling to roughly 6 per cent.

Figure 7. Panel population and relative share of each birth cohort with maximum employment by gender. Per cent.

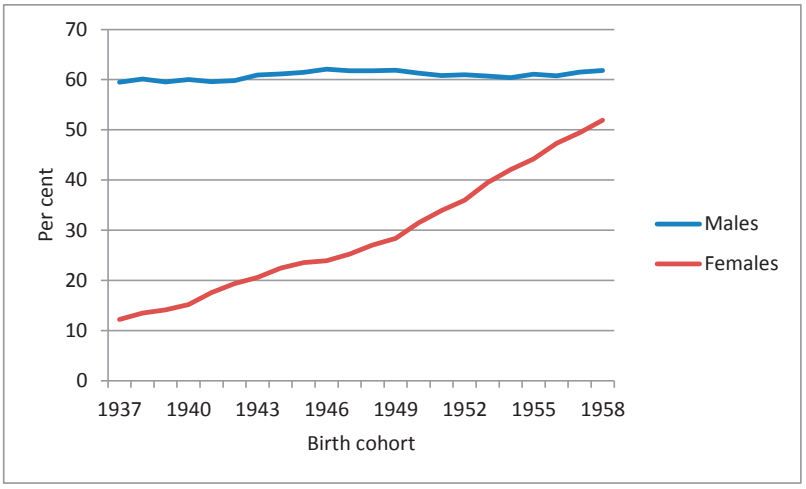
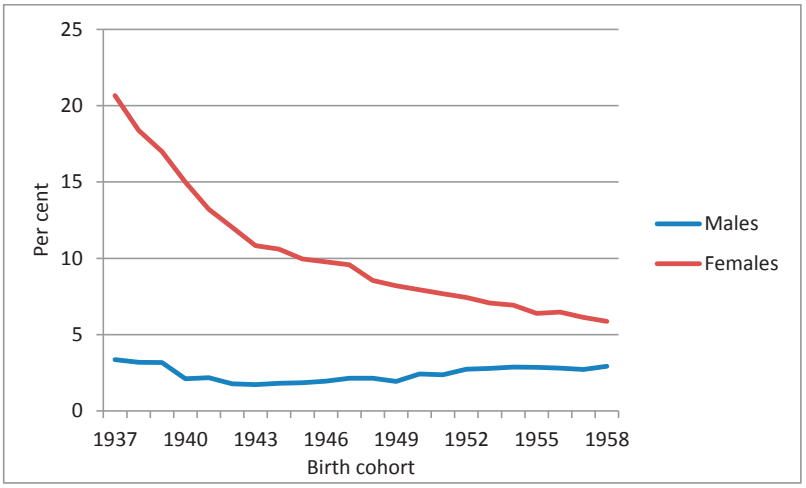


Figure 8. Panel population and relative share of each birth cohort without labour market participation. Male and female. Per cent.



The descriptive overview displays quite uniform trends and so far coincides with the existing picture of the post-industrial labour market at national level, which supports the validity of the dependent variable. My conclusion is that *labour market participation* as defined and deduced for the panel population is a valid measurement for the theoretical concept in Figure 3.

GEOGRAPHIC VARIABLES

The municipality identifier is actually an auxiliary variable, but the coverage of this variable in the panel population decides the quality of the match with the municipality classification tables, and thus also the coverage of classification codes. A high coverage of municipality identifiers in the panel population is a condition for a high coverage of the classification codes. This will in turn affect the result of the analyses. The presence of a classification code in the data matrix necessarily indicates a valid municipality identifier. No observations were allowed by the population requirements to contain only missing values during the first half or the second half of the 20-year observation period, respectively. The check shows that all observations have a municipality classification at both points of measurement, as confirmed by Table 13.

The distribution of observations is clearly uneven between the municipality classes. Account must be taken of the fact that less central municipalities have a smaller proportion of the nation's population than central and urban municipalities. The largest share of observations in class 7 refers to Oslo alone.

There is a reduction in the number of observations in primary industry municipalities from 'year 10' to 'year 20'. This is a reflection of a demographic change: a redistribution of population from rural to central areas. The actual numbers of municipalities within each class are shown in Statistics Norway's 'Standard for municipality classification' – 1974, 1985, 1994 and 2003 – respectively.

Table 13. Panel population by municipality classification.

Classification	Year 10	Year 20
Primary industry municipalities	83988	71442
Mixed agriculture and manufacturing municipalities	85757	60265
Manufacturing municipalities	125579	112847
Less central, mixed service industry and manufacturing	89990	104768
Central, mixed service industry and manufacturing municipalities	295860	341788
Less central service industry municipalities	92366	90422
Central service industry municipalities	458072	450080
Other	0	0
Total	1231612	1231612

There is no reason to question the coverage of the variable *residential characteristics* in the data matrix, but the decisive question related to utilizing the municipality classification is whether this will be a valid measurement for the desired *residential characteristic* variable. The model requirement is an explanatory variable that expresses influence on *labour market participation* by locally distinctive characteristics, and furthermore a variable with continuity and comparability over time.

Among available sources the only remaining option was the municipality classification as a possible measurement of these characteristics. By Statistics Norway's own assessment, the classification is considered useful for e.g. analysis of the settlement, urbanisation and industrial structure of municipalities[107]. A pragmatic conclusion so far is that the municipality classification is a measurement that is interesting to include in the case study as the explanatory variable *residential characteristics*. The classification meets the requirement as regards comparability over time, and there is no reason to question the assessments behind the classification as such. Nevertheless, transfer of qualities from municipality level to individual level is definitely 'noisy'. Ultimately, further analyses will reveal how appropriate this variable is, and whether it is possible to apply it in the regression analyses.

[107] Statistics Norway (1994): Standard Classification of Municipalities 1994. Oslo. Ch. 1 – 2.

According to the requirements for the variable *geographic mobility*, this measurement expresses the number of changes in residential municipality during the observation period. Theoretically this coincides with the definition of internal migration used by Statistics Norway[108].

The highest number of changes in residential municipality observed in the data set is 16, (see Table 8). This is an extreme value which occurs only once.

Migration figures for the panel population may be compared with national figures for internal migration from Statistics Norway population statistics, which counts migration between municipalities, as e.g. in Historical Statistics 1994[109]. The table in question shows average figures for annual migration between municipalities for 1967 to 1992, which covers parts of my observation period. The total migration figures for both sexes for these years range from 59.9 per 1 000 population to 38.8 per 1 000 population. For the panel population, an average number for annual migration for the whole observation period is calculated to be 18.1 per 1 000 observations in the panel. The figures seem reasonable considering that the panel population is limited to age groups 30 to 50 years old. More frequent migration is documented for the younger age groups, i.e. 30 years of age and below.

A research data set with more than 1.2 million observations may seem overwhelming. However, a main objective for this thesis is to split this data set in units for lower geographic levels and run analyses at these levels, i.e. county and municipality level. A large share of Norwegian municipalities has a very small number of residents. As of 1 January 2008, the smallest municipality was Utsira with 214 residents, while the second smallest was Modalen with 351 residents[110].

Table 14. Panel population by number of municipality changes 1967 – 2007. Total, male and female.

Number of municipality changes	Total	Male	Female
0	754019	364183	389836
1 - 5	469787	260477	209310
6 - 10	7514	5080	2434
10 and above	292	207	85
Total	1231612	629947	601665

The birth cohorts in the panel population equal the sum total of these particular birth cohorts within each municipality, but may result in a very small number of observations when the panel population is split by municipality: in fact the number of observations among the panel population is below 1 000 in 210 of the total of 435 municipalities, and below 200 in 17 municipalities. The smallest municipality by number of observations in the data matrix is Utsira with only 53 observations in the

[108] Cf. Skiri and Sørli, op. cit.

[109] Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå, Table 3.29.

[110] Statistics Norway: *Population*. StatBank.

panel population, and the second smallest is Modalen with 81 observations. For the smallest municipalities there is an obvious risk that analysis results may be insignificant, but this will eventually be revealed by the margins of error produced by the analytic software.

One final comment must be attached to the municipality issue. Some panel observations are identified by municipality identifiers not existing as of 'year 10' or 'year 20', due to the restructuring and merger of municipalities. This is why a test on possible discontinued identifiers must be taken into account before municipality-level analyses are performed. The technical solution for redirecting observations is dealt with in SAS data step number 22, where the county of Vestfold is chosen as an example because the municipality structure in this county was extensively reorganised as of 1 January 1988. Despite this precaution, there may be a few observations that remain excluded from the municipality-wise analysis due to historical municipality identifiers not possible to redistribute to new municipalities.

FAMILY RESPONSIBILITIES

So far the variables have been directly or indirectly derived or transferred from the Central System for National Social Security. The remaining variable – *number of children* – was transferred from a different system, the TRP.

The highest number of children of one parent observed in the data matrix is 12. This is confirmed as correct by the CPR. A further quality check of this variable is simply to compare it with public statistics. Table 15 shows the average number of children by selected cohorts from official statistics compared with the average number of children by the same cohorts based on the variable transferred from the TRP.

Table 15. Average number of children for selected cohorts by gender. Official statistics[111] and TRP.

Cohort	Females by TRP	Females by official statistics	Males by TRP	Males by official statistics
1940	0,7	2,4	0,9	2,3
1945	1,2	2,2	1,4	2,1
1950	1,8	2,1	1,6	2,1
1955	1,9	2,0	1,6	2,0

[111] Statistics Norway: *Births*. StatBank

Official statistics do not display figures for all birth cohorts, besides which official statistics present more continuous figures for women than for men. The most comparable figures from public statistics are found for women of the given birth cohorts at age 45, and for men of the given birth cohorts at age 50. Age 50 is actually the final status for both men and women in the panel population, but there are no official statistics with comparable figures for women at age 50.

The conclusion from this comparison seems clear, but for the experiment in question very disappointing, yet important. Discrepancies are quite high between official average figures and the average figures for number of children that was possible to transfer from the TRP. Besides, the figures from the TRP show opposite trends compared to official figures especially for women: seemingly a rising fertility rate instead of the general trend of falling fertility, which is overwhelmingly documented. In fact, this only reflects improved matching quality for the younger generations of the TRP. The lower average figures from the TRP, especially for women, reflects the problem of joint tax assessment and the taxpayer unit. A high share of the older female generations are simply not included as individual observations in the TRP; see the chapter 'Data steps for adding family obligations – number of children' in Appendix 1.

In general, these findings indicate that this variable is not reliable and should not be included in the regression analysis, and that the model has to be adjusted. The most probable outcome if this variable is included in the regression analysis would be a positive association between *number of children* and the dependent variable, which is the opposite of what is anticipated based on the official figures. The assumption might of course be wrong, but in this case a positive effect would most likely be a result of an incomplete variable. In other words, the operationalization of this variable did not result in a reliable measurement for the desired variable *family responsibilities*.

The widespread practice of joint tax assessment resulted in a clear underrepresentation of women in the oldest TRP volumes, which in turn resulted in a high mismatch between spouses when this variable was operationalized. These circumstances necessarily make it difficult to transfer variables between spouses for the oldest generations. Moreover, the birth cohort requirements excluded a match between spouses where one was born outside the birth year range 1937 - 1958.

Based on the assessments and results above, it is time to summarize the question of validity, which is obviously a more complicated issue for some of the variables than for others. Operationalization of variables as theoretically depicted in the model above is aiming at the establishment of variables as both valid and reliable expressions and measurements of what is intended by the regression model. Obviously *gender* is a valid measurement of gender, and *birth cohort* based on year of birth is equally obviously also a valid measurement. However, some of the other variables had to be operationalized by means of quite toilsome deduction – indirectly from accessible data – and the question of validity is not as obvious. Nevertheless, a final conclusion remains to be drawn after the model fit is tested. The explanatory variable *number of children* has now been proved to be unreliable. However, the exercise

of creating these variables was nevertheless useful. In general, this will be the situation when electronic administrative registers are applied in a research context, as they are designed for administrative purposes and not primarily for research.

Income and size of income collected from administrative registers are very commonly applied directly in research, but not with such a degree of interpretation and deduction as for the present purpose. The labour income concept is definitely a reflection of labour market participation. In my judgment, size of labour market income by actual NOK amount would not express the rate of labour market participation that was desirable for the theoretical model. Instead I chose to deduce an indicator based on the income amount, but in such a way that size of income would not cause any gender bias. The measurement that was finally deduced coincides very well with the trends and development of labour market participation that are seen in research based on other sources, and thus seems to be a measurement which is a valid expression of individual *labour market participation*, as well as being sustainable and comparable over years and across the gender dimension. Final conclusions about the applicability of the variables are eventually to be drawn after tests of the model fit.

TESTING THE MODEL FIT

A test of the theoretical model on real data is logically a part of the historical criticism process. The variables are operationalized on the basis of a theoretical model, but there is a risk that at least some variables do not totally correspond with the theoretical requirements, both in the technical sense and in the conceptual sense. A possible outcome of this test is that one or more variables do not fit with the theoretical model, which may necessitate a model adjustment.

The following tests are run on the total panel population which reflects the national level. A multivariate, or multiple, regression model includes several explanatory variables. An important remark at this point is that a major objective of this enquiry is to investigate geographic levels other than national. The model will be applied in analyses on data sets where the data matrix is split by counties, by birth cohorts, by municipalities and by municipality classes. The effect of the explanatory variables may very well be different, stronger or weaker or even with different direction, when these data sets are analysed separately.

The ANOVA test is a test of the statistical acceptability of the model. The regression row in Table 16 displays information about the variation in the dependent variable which is accounted for by the model, while the residual row displays information about the variation in the dependent variable that is not accounted for by the model. The residual is high and indicates either that variables not included in the model actually have a higher effect on the variation in individual *labour market participation* than variables included in the model, or this reflects inaccuracy in one or more of the

explanatory variables. A high residual is generally to be expected in social science models[112]. Nevertheless the result is statistically significant. The number of observations in the data matrix is very high, which always speaks in favour of a significant result. SPSS also reports that all requested variables are included, i.e. if any of the variables were to have missing values, appear as constants, or in any other sense not meet the statistical requirements, the variable would have been excluded from the test.

Table 16. Analysis of variance (ANOVA) test.

ANOVA[b]					
Model	Sum of Squares	df	Mean Square	F	Significance
Regression	8918382.155	5	1783676.431	51278.573	.000[a]
Residual	42840243.94	1231606	34.784		
Total	51758626.10	1231611			
a. Predictors: (Constant), migration, birth cohort, classification as of year 20, gender, classification as of year 10					
b. Dependent Variable: employment history					

The results displayed in Tables 17 – 19 are basically tests of the model fit. The results are both analyses of each of the explanatory variables' contribution to the model, as well as statistical acceptability of the results. For a start the model fit is tested by running the multivariate analysis to consider a possible redefinition and improvement of the model by excluding any explanatory variables that do not contribute to the model. The outcome of this test is ideally a revised model purified from variables that make no contribution to the model.

While the ANOVA table is a test of the model's ability to explain variation in the dependent variable, the coefficients' table expresses the strength of the relationship between the dependent variable and the explanatory variables. The constant in Table 17 is difficult to comment on, as not all variables have a zero-value (see Table 8). There are no non-significant coefficients (see the Significance column), and the t-values are well above the critical limit (+2/-2). But again, the chance of significance is supported by a large population.

The standardized Beta coefficients determine the relative importance of the variables. Compared to the unstandardized coefficients, the standardized coefficients do not change the rank very much in terms of effect. First of all, *gender* contributes most to the model, followed by *birth cohort*. *Residential characteristics* (in the table *classification by year 10* and *classification by year 20*, respectively) do not contribute much to the model, nor does *geographic mobility*. Moreover, the effect

[112] Midtbø, Tor (2007): *Regresjonsanalyse for samfunnsvitere. Med eksempler i SPSS.* Oslo: Universitetsforlaget, p. 76.

of *geographic mobility* is negative, though very weak. This is the opposite of what was anticipated. It is appropriate to ask if the effect of this variable actually is absent. But it is too soon to draw any conclusion about excluding this variable from the model. The result is significant and the t-value is acceptable.

Table 17. Coefficients. Dependent variable: Employment history.

Model	Unstandardized Coefficients		Standardized Coefficients	t-values	Significance
	B	Std. Error	Beta		
(Constant)	14.502	.020		727.069	.000
Gender	-5.048	.011	-.389	-472.984	.000
Birth cohort	.145	.001	.136	164.957	.000
Classification by year 10	.074	.005	.022	16.091	.000
Classification by year 20	.141	.005	.040	28.855	.000
Geographic mobility	-.187	.005	-.033	-39.856	.000

Table 18. Test for multicollinearity. Dependent variable: Employment history

Model	Correlations			Collinearity Statistics	
	Zero-order	Partial	Part	Tolerance	VIF
(Constant)					
Gender	-.386	-.392	-.388	.992	1.008
Birth cohort	.138	.147	.135	.992	1.008
Classification by year 10	.060	.014	.013	.354	2.822
Classification by year 20	.055	.026	.024	.356	2.812
Geographic mobility	.005	-.036	-.033	.989	1.011

Table 18 investigates if there is any problem with multicollinearity, i.e. to check if some, or much of the variation in a given variable can be explained by the other variables. The tolerance column is the percentage of the variation in a given variable that cannot be explained by the other variables. When tolerances are close to 0, there is high multicollinearity. In other words, a tolerance far from 0 is desirable. For the two residence classification variables, the tolerances are clearly closer to 0 than for the rest of the variables. This means that much of the variation in labour market participation that is explained by residential characteristics is also explained by the other variables. For

the remaining variables, tolerances of 0.9 show that only a small part of the variation in these variables can be explained by the other variables.

A Variance Inflation Factor (the VIF column in Table 18) > 2 is usually considered problematic. This is the case for the two residential characteristics variables. Otherwise all other VIF values are smaller than 2 with the smallest VIF for *gender*. Apart from the residential characteristics variables there is no problem with multicollinearity in the model.

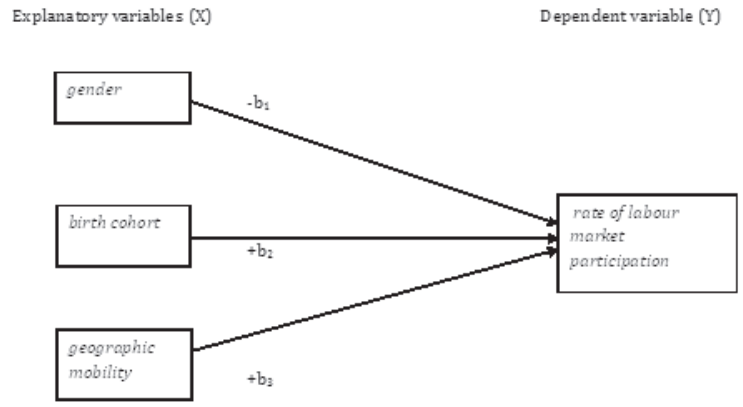
The conclusion at present is that the residential characteristics variables and *number of children* should be omitted from the model, though for different reasons. Thus the remaining explanatory variables in the adjusted model are *gender*, *generation (birth cohort)* and *geographic mobility*. The findings above necessitate adjustment of the hypothesis and the model.

Table 19. Hypothesis about labour market participation in post-industrial Norway. Adjusted.

The null hypothesis	***Alternative or research hypothesis***
Gender had no influence on the rate of labour market participation	*Gender made a difference: women and men had different rates of participation in the labour market – male accounted for a higher rate of participation than female*
Generation had no impact on the rate of participation in the labour market	*Generation made a difference: younger generations had a higher rate of participation than older generations*
Geographic mobility did not affect labour market participation	*Geographic mobility made a difference: the effect of geographic mobility on labour market participation is different from zero*

Geographic mobility appears to have a negative effect on *labour market participation*, contrary to expectation, but the main result is that the effect is very weak. Nonetheless, the result is significant and with t-values above critical level, and so far there is not sufficient reason to exclude this variable from the model. The direction of the effect is uncertain, however, as the results from the analyses by municipality class revealed, see below. Thus the effect cannot be specified as either positive or negative in the adjusted model.

Figure 9. Adjusted model for explaining variation in labour market participation.



THE THEORETICAL MODEL APPLIED ON PHYSICAL DATA

CORRELATION ANALYSIS

For a start, a descriptive overview is useful for reference and consistency purposes. Some general figures have already been presented: the net panel population distributed by birth cohort is displayed in Table 10. Table 11 showed the minimum, maximum and average rate of labour market participation for men and women, respectively, and Figures 7 and 8 confirmed some general development trends: a gradual increase in labour market participation by the female birth cohorts 1937 - 1958 for the years 1967 – 2007 is obvious. While 79 per cent of the 1937 female cohort has at least one year of labour market participation during their living years from age 30 to age 50, this is the case for 94 per cent of the 1958 cohort. For the intervening cohorts, there is a steeper rise in labour market participation for the oldest generations; from the 1937 cohort to the 1943 cohort the share of observations with employment participation increased from 79 per cent to 89 per cent.

Table 20. Multivariate correlation. Nation.

Correlations					
		Employment history	Gender	Birth cohort	Geographic mobility
Employment history	Pearson Correlation	1	-.386**	.138**	.005**
	Sig. (2-tailed)		.000	.000	.000
	Number	1231612	1231612	1231612	1231612
Gender	Pearson Correlation	-.386**	1	.000	-.087**
	Sig. (2-tailed)	.000		.933	.000
	Number	1231612	1231612	1231612	1231612
Birth cohort	Pearson Correlation	.138**	.000	1	.008**
	Sig. (2-tailed)	.000	.933		.000
	Number	1231612	1231612	1231612	1231612
Geographic mobility	Pearson Correlation	.005**	-.087**	.008**	1
	Sig. (2-tailed)	.000	.000	.000	
	Number	1231612	1231612	1231612	1231612
**. Correlation is significant at the 0.01 level (2-tailed).					

A correlation analysis will display association between variables, the strength of the association and whether the association is positive or negative.

The results of a multivariate correlation analysis, with all remaining variables from the model included, are shown in Table 20. The coefficients are expressed by Pearson's Correlation coefficients which range from -1 to +1. The closer the value of the coefficient is to 1, the stronger the correlation, while a weak or absent correlation will have a coefficient close to 0. The direction of any correlation is expressed by the sign as positive or negative.

The correlation table is based on the total net panel population, which represents the national level[113]. The strongest correlation is found between *labour market participation* (i.e. *employment history* in the table) and *gender*. As expected, this correlation is negative. For a dichotome variable, like *gender*, the coefficients will express a difference in the level of labour market participation between the two categories, i.e. the level of *labour market participation* is lower for women than for men.

Furthermore, there is a positive correlation between *birth cohort* and *labour market participation*. A positive correlation was also anticipated – younger cohorts were anticipated to have higher labour market participation than older cohorts. The correlation is weak, though, but the result coincides with the pattern shown in Figures 7 and 8.

As already indicated, the correlation between *labour market participation* and *geographic mobility* is very weak, but in this analysis the effect is positive.

Important in a multivariate correlation analysis is to reveal a possible relationship between variables that are intended to be included as explanatory in the regression analysis; see also the multi-collinearity test. Any such relationship would indicate that there might be an underlying factor which affects both explanatory variables, and that the effect on the dependent variable is spurious. However there is no problem with any such correlations between the explanatory variables. There is a negative correlation between *gender* and *geographic mobility*, but very weak.

The correlation analysis also reports that all correlations are significant at 0.01 level in a two-tailed test[114]. The significance test in this case is two-tailed because the results above indicated that the direction of the effect might be different than anticipated, at least for one of the variables. Anyway there is a 99 per cent probability that the associations found in the panel population are not a result of chance.

When the same multivariate correlation analysis is run separately for municipalities grouped by municipality class, the results are basically the same. The effect of *geographic mobility* is weak in all groups of municipalities, from 0.026 in 'primary industries' municipalities, to – 0.02 in 'manufacturing industries' municipalities. The interesting result is that the effect is positive or

[113] It makes no difference if the coefficients are read along the horizontal or the vertical axis in Table 20. The diagonal shows the variables correlated with themselves and is always = 1.
[114] See the double asterisks in Table 20.

negative depending on the industrial structure in the municipalities, but in any case the effect is very weak.

A preliminary conclusion is that this analysis supports the following hypotheses about variation in *labour market participation,* at national level:

- *Gender* explains – i.e. there is a difference in the level of labour market participation between genders, and the effect of *gender* is negative - as anticipated.
- *Generation* explains - i.e. the direction of the association is positive as expected - the younger the cohort, the higher the rate of labour market participation.

On the other hand, the correlation analysis can hardly be said to support the third hypothesis: *geographic mobility* seems to have no effect on *labour market participation* - this variable does not explain. A few more results are needed to finally conclude whether to omit this variable.

REGRESSION ANALYSIS

To begin with, the regression analyses are run on the total panel population. A multivariate, or multiple, regression model includes two or more explanatory variables. A multiple model increases precision and credibility, and makes it easier to distinguish more important explanations from less important explanations. Initially though, the remaining explanatory variables are analysed separately, in a bivariate analysis. The following bivariate regression analysis intends to measure in what direction and with what strength the X variables - one by one - affect the size of the Y variable:

- *Labour market participation (Y)* and
  - *gender (X)*
  - *cohort (X)*
  - *geographic mobility (X)*

The regression coefficient quantifies the influence of *gender*, *birth cohort* and *geographic mobility* on the rate of *labour market participation*. When the results in Tables 21 a – 21 c are ranked, the *gender* model has the best explanatory effect with a standardized Beta coefficient like -.386, i.e. close to 39 per cent of the variation in *labour market participation* is explained by *gender*, while *birth cohort* explains about 14 per cent of the variation. But an effect of *geographic mobility* is almost absent.

Table 21 a. Coefficients. Labour market participation and bivariate effect of gender.

Coefficients[a]					
Model	Unstandardized Coefficients		Standardized Coefficients	t-values	Significance
	B	Std. Error	Beta		
(Constant)	17.237	.008		2287.405	.000
Gender	-5.003	.011	-.386	-464.069	.000
a. Dependent Variable: Employment history					

Table 21 b. Coefficients. Labour market participation and bivariate effect of birth cohort.

Coefficients[a]					
Model	Unstandardized Coefficients		Standardized Coefficients	t-values	Significance
	B	Std. Error	Beta		
(Constant)	12.972	.013		991.614	.000
Birth cohort	.147	.001	.138	155.178	.000
a. Dependent Variable: Employment history					

Table 21 c. Coefficients. Labour market participation and bivariate effect of geographic mobility.

Coefficients[a]					
Model	Unstandardized Coefficients		Standardized Coefficients	t-values	Significance
	B	Std. Error	Beta		
(Constant)	14.771	.007		2158.677	.000
Geographic mobility	.031	.005	.005	6.036	.000
a. Dependent Variable: Employment history					

Turning to the model summary tables, the rank between the models is unchanged if compared with the coefficients tables. Table 22 displays selected figures from the bivariate analysis of the remaining explanatory variables. The R square in Table 22 is Pearson's R square. This will vary between 0 and 1. The 'Standard error of the estimate' in this table is a measurement of the scattering around the regression line; the higher the standard error, the less variation is explained by the model. The standard error of the estimate does not show much difference, but there is consistency in the rank between the models: the *gender* model has both the highest R square and the lowest standard error,

while the *geographic mobility* model has both the lowest R square and the highest standard error. For the *geographic mobility* model the R square is 0, which means that *geographic mobility* simply does not explain any of the variation in labour market participation, and this null hypothesis cannot be rejected.

Table 22. Model summary - extract. Labour market participation and bivariate effect of gender, birth cohort and geographic mobility, respectively.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
Gender	.386[a]	.149	.149	5.981
Birth cohort	.138[a]	.019	.019	6.420
Geographic mobility	.005[a]	.000	.000	6.483

A final test of change statistics with *gender* and *birth cohort* included in the model, and *geographic mobility* excluded and included, respectively, shows a minimal R square change. It is no longer possible to insist on including *geographic mobility* in further analyses. The remaining option is therefore actually a trivariate analysis – still at national level with the variables *gender* and *birth cohort.* The result is displayed in Table 23.

Table 23. Labour market participation and trivariate effect of gender and birth cohort.

Coefficients[a]					
Model	Unstandardized Coefficients		Standardized Coefficients	t-values	Significance
	B	Std. Error	Beta		
(Constant)	15.416	.013		1174.515	.000
Gender	-5.003	.011	-.386	-469.372	.000
Birth cohort	.147	.001	.138	168.451	.000
a. Dependent Variable: Employment history					

The unstandardized coefficient (B) expresses the estimated effect of each of the explanatory variables on variation in *labour market participation*. Based on the standardized coefficients, *gender* contributes most to the model; *labour market participation* is clearly negatively affected by *gender*. Furthermore, the dependent variable is positively affected by *birth cohort*: when year of birth increases, *labour market participation* also increases.

Table 24. Model summary. Trivariate regression. Nation.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
	.410[a]	.168	.168	5.913
Predictors: (Constant), Birth cohort, Gender Dependent Variable: Employment history				

The significance column is an expression of the probability of getting high values on the explanatory variables by chance. In this case, the significance test is two-tailed at 1 per cent significance level. The coefficient is 0.000, i.e. there is less than 1 per cent probability that the effects - by *gender* and by *birth cohort* - are results of random factors. In other words, the results observed for the panel population indicate that the effect different from 0 is not a result of chance, and the null hypotheses should be rejected.

The model summary table reports the strength of the relationship between the remaining explanatory variables and the dependent variable. When both *gender* and *birth cohort* are included in the model, the explanatory effect is increased from about 15 per cent for the bivariate *gender* model to about 17 per cent for the trivariate model. Still, this is not a high explanatory effect; most of the variation is actually explained by other variables which ideally should have been included in the model, or rather which should have been satisfactorily operationalized.

Since *gender* now appears to be the explanatory variable that contributes most to the existing model, the main analyses will be run as bivariate with *gender* as the explanatory variable. The *gender* model is also expected to be the most interesting model to explain variation in *labour market participation* in local labour markets. A major point of interest is now to measure possible differences in the *gender* effect between populations at lower geographic level – county level for a start, and eventually at municipality level. Besides, a bivariate analysis is a technical tool to come to terms with output from 435 municipalities in as compact a form as possible.

But, to begin with, the bivariate effect by *gender* is analysed cohort by cohort. The results are collected in Table 25. First of all, these figures confirm that for all cohorts *labour market participation* is negatively affected by *gender,* and secondly that this effect is decreasing cohort by cohort.

The *gender* effect on *labour market participation* is quite strong for the 1937 cohort; for the 1958 cohort the effect is clearly weaker, but still noticeable. The cohort column is sorted by ascending year of birth, but the ranking of cohorts would actually not have changed if the Beta coefficients had been the sorting criteria instead. Indirectly this also indicates the linear effect of *generation*.

Table 25. Bivariate effect of gender by birth cohort. Nation.

Cohort	Standardized Coefficients Beta	t-values	Significance	Number of observations
1937	-.596	-146.060	.000	38806
1938	-.578	-142.411	.000	40497
1939	-.558	-137.671	.000	42012
1940	-.541	-133.434	.000	43051
1941	-.516	-122.747	.000	41604
1942	-.492	-122.874	.000	47273
1943	-.477	-122.862	.000	51270
1944	-.454	-122.309	.000	57632
1945	-.445	-121.601	.000	59857
1946	-.441	-126.884	.000	66567
1947	-.417	-115.900	.000	63855
1948	-.396	-107545	.000	62140
1949	-.378	-100.625	.000	60628
1950	-.347	-90.796	.000	60280
1951	-.325	-83.385	.000	58847
1952	-.305	-79.295	.000	61111
1953	-.286	-74.307	.000	61840
1954	-.267	-68.986	.000	61912
1955	-.255	-66.088	.000	62886
1956	-.244	-63.440	.000	63669
1957	-.235	-60.640	.000	62714
1958	-.220	-56.636	.000	63161

The two columns 't-values' and 'Significance' express the reliability of the results, while the final column shows the number of observations in each cohort. The result of this analysis is significant at less than 1 per cent level for all cohorts. The t - values are also far from being critically low.

BIVARIATE EFFECT OF GENDER AT COUNTY LEVEL: EVIDENT REGIONAL DIFFERENCES

For the following analysis, the data matrix has been split into a number of data sets – one for each county, one for each of the different municipality classification codes, one for each birth cohort and eventually one for each municipality. A map of Norway with county names is included in Appendix 4.

Table 26 sums up the results of a bivariate regression analysis at county level. Separation of the data matrix by county is based on municipality status in year 20 in this table. A similar analysis

based on separation by municipality status in year 10 basically shows the same results, which is why these results are not commented on any further.

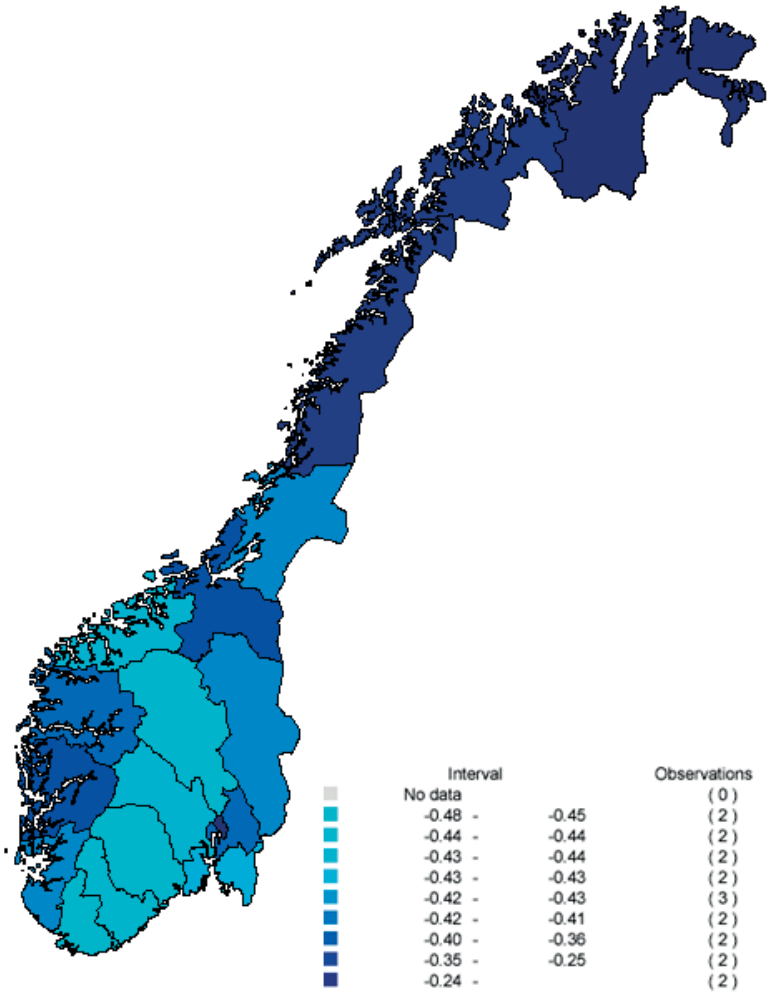
The results display evident regional differences: *labour market participation* is negatively affected by *gender* in all 19 counties, but the strength of the effect differs clearly between counties. The strongest *gender* effect is found in Vest-Agder with a coefficient of –0.478, while the weakest is found in Oslo with –0.194. By the way Oslo is the only regional unit which is both a county and a municipality which is why Oslo appears both in the county and in the municipality tables. Also in Finnmark, as well as in Troms, the two northernmost counties, the effect of *gender* appears to be weak.

Table 26. Bivariate effect of gender by county. Sorted by Standardized Beta coefficients.

County	Standardized Coefficients Beta	t-values	Sig.	Number of observations
Oslo	-.194	-72.174	.000	132630
Finnmark	-.239	-35.273	.000	20486
Troms	-.302	-65.099	.000	42265
Nordland	-.350	-96.083	.000	66284
Sør-Trøndelag	-.378	-109.285	.000	71814
Hordaland	-.396	-145.940	.000	114538
Sogn og Fjordane	-.417	-76.759	.000	28056
Akershus	-.418	-169.679	.000	136165
Rogaland	-.422	-144.982	.000	96877
Nord-Trøndelag	-.422	-87.261	.000	35168
Hedmark	-.422	-108.243	.000	54162
Vestfold	-.426	-115.535	.000	60365
Østfold	-.428	-126.889	.000	71658
Buskerud	-.432	-124.596	.000	67694
Møre og Romsdal	-.433	-123.034	.000	65587
Aust-Agder	-.435	-81.549	.000	28502
Oppland	-.437	-111.021	.000	52235
Telemark	-.456	-110.638	.000	46664
Vest-Agder	-.478	-110.467	.000	40462

The county-wise differences are also depicted in Figure 10 where the darkest blue colours represent counties with weak *gender* effects and the light blue colours represent counties with stronger *gender* effects.

Figure 10. Labour market participation and bivariate effect of gender. Standardized Beta coefficients. County.



Municipality-wise analyses generally show weak *gender* effects in the biggest cities – especially Oslo, Bergen and Trondheim - which in turn affect the figures for their respective counties: Oslo which is both a county and a municipality, Bergen affects the result for Hordaland and Trondheim for Sør-Trøndelag.

Table 27. Bivariate effect of gender by municipality class. Classification as of year 20.

Classification	Standardized Coefficients Beta	t-values	Sig-nificance	Number of observations	Number of munici-palities
Manufacturing municipalities	-.477	-182.163	.000	112847	66
Mixed agriculture and manufacturing municipalities	-.469	-130.239	.000	60265	60
Primary industry municipalities	-.427	-126.346	.000	71442	93
Central, mixed service industry and manufacturing municipalities	-.428	-277.115	.000	341788	78
Less central, mixed service industry and manufacturing municipalities	-.400	-141.216	.000	104768	74
Less central service industry municipalities	-.334	-106.663	.000	90422	30
Central service industry municipalities	-.320	-226.365	.000	450080	34

County-wise differences are products of how each county is composed by attributes of each county in terms of urbanisation, population density, industrial structure, etc. of its municipalities. Somehow residence matters as regression coefficients clearly differ between counties, and between municipalities as shown below. But the *residential characteristics* chosen for the individual observations proved to be unsuitable for the regression analyses. Though the classification codes seem unsuitable for a regression analysis, they may very well serve as a selection variable to split up the panel population and run separate analyses for each of the municipality classes.

The results shown in Table 27 are based on separation of the data matrix by municipality class as of year 20. A similar analysis run on a data set broken down by municipality class as of year 10 showed very small differences compared to the result above, which is why the result based only on the status as of year 20 is presented.

The negative effect of *gender* on *labour market participation* is strongest in 'manufacturing municipalities' and weakest in 'central service industry' municipalities. This might be interpreted as a confirmation of the close interaction between the growth of service industries and increased female

employment, but it is actually not possible to go behind the figures. We can only say that these figures show that women have entered the labour market to a larger extent in 'central service industry' municipalities than in other municipalities, but we cannot measure to what extent women entered into the service industries or into other industries. But the differences also indicate that the growth of service industries has not been evenly distributed around the country.

In all municipalities, there are basic service industries – a local administration, education institutions and health care institutions – whether the industrial structure in a municipality is classified as primary industries, manufacturing industries, service industries etc. Considering the gender-segregated labour market, there must probably be additional service industry components, e.g. governmental institutions for education or health or both, in a local labour market to even out the level of labour market participation between genders and thus weaken the *gender* effect. The extent of additional service industries differs regionally and locally. It must also be assumed that if an additional service industry component is public rather than private, this will equalize the level of labour market participation between genders even more. The presence, or the absence, of one type of service industry or one type of service institution might be very decisive for the *gender* effect locally. With reference to Håland and Daugstad[115], for instance, the presence of a governmental health service institution or an educational institution in addition to the basic local institutions in a given local labour market would account for higher female labour market participation, and thus a weaker *gender* effect. On the other hand, with a possible closure of governmental institutions of this kind in the future, there would be an obvious risk of a reversed *gender* effect.

Obviously, an essential confounding variable in this picture is *education*. This will be further considered when municipality results are ready.

MUNICIPALITIES AND BIVARIATE EFFECT OF GENDER: CONSIDERABLE LOCAL DIFFERENCES

All municipalities are now subject to identical bivariate analysis, but some technical decisions must be made in advance: Due to the restructuring of municipalities over years, a decision must be taken according to which municipality structure and status by which to run the analyses and present the results. A major municipality restructure took place during the 1960s, but changes have also occurred from time to time since then. Measures must be taken to find a municipality status to represent the whole period from 1967 to 2007. By 1960 the number of municipalities was 732, by 1970 the number was 451 and by 2005 the number is 433[116]. My choice is the municipality status by

[115] Håland, Inger og Gunnlaug Daugstad (2003): *Den kjønnsdelte arbeidsmarknaden.* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 6/2003

[116] Statistics Norway (2010): *Statistical Yearbook 2010*. Oslo: Statistisk sentralbyrå. 2010. Table 2.

1994 with 435 municipalities[117]. In general there was not much change of the municipality structure between 1974 and 2003, apart from the restructuring in Vestfold county by January 1rst 1988.

Data sets for analyses on municipality level have been prepared by selection on the municipality identifier. Referring to the two municipality identifiers in the data matrix (cf. table 37), the selection is based on municipality identifier by 'year 20' in the observation period. For the cases of expired municipality identifiers additional selection criteria has been included as displayed in SAS data step 18. The intention with this data step is to redirect and move observations from expired municipalities to municipalities existing by 1994. A few observations were not possible to redirect and hence these were not included in the analyses, which means that the total number of observations municipality wise differs slightly from the total number in the panel population, but this is only a marginal difference.

Finally a decision about which coefficient that should represent the municipality wise analyses is needed. The regression tools in SPSS produce several coefficients and other expressions of the analyses results. It is obvious that a simplified measurement has to be applied for the comparison of 435 local labour markets. The final decision is to apply the Standardized Beta coefficients, presumably this is more comparable than the unstandardized Beta coefficients.

The significance results for the municipality analysis (cf. the Significance column in the coefficients tables above) are not displayed, simply because all results have proved to be significant on less than 1 per cent level in a two - tailed significance test for all municipalities, with a very small number of exceptions: A few municipalities, mainly with extremely small numbers of residents and consequently an even smaller number of observations in the panel population failed the significance test, and besides the t-values were critically low. These municipalities are Kautokeino and Karasjok in Finnmark county, and Kvitsøy and Utsira in Rogaland county. As there was reason to fear – the result from Utsira e.g., with only 53 observations in the panel population is insignificant.

The overview of the results from the remaining 431 municipalities display that the standardized Beta-coefficients range from -.154 in both Vardø and Tana municipalities in Finnmark, to -.620 in Iveland municipality in Aust-Agder county.

All municipality results - coefficients, t-values etc.- are imported into a data base table. This table is then sorted by the standardized Beta-coefficients. Thus a country wide rank of the municipalities is possible, but only the 'high ten' and the 'low ten' are displayed and commented in the tables below. The municipalities in these tables are identified by municipality identifier and by name. However, all municipality results are displayed graphically in Figure 11.

---

[117] Statistics Norway (1994): *Standard Classification of Municipalities 1994.* Oslo: Statistisk sentralbyrå. P. 27.

Labour market participation is negatively affected by *gender* in all municipalities, but the major point is that the strength of the effect varies clearly between municipalities. In local labour markets where the *gender* effect is very weak there is not much difference in the level of *labour market participation* between men and women.

Table 28 displays the ten municipalities with the strongest *gender* effects. Most of these municipalities are located in the south-west of Norway: County identifiers '08' to '11' – i.e. Aust-Agder, Vest-Agder and Rogaland. The regression coefficient is closer to 1 than to 0 for these municipalities, which expresses that female labour market participation in these municipalities is low throughout the total observation period from 1967 to 2007. The number of observations in all these municipalities is quite small, but the result is significant in all of them.

Table 28. Ten municipalities with the strongest gender effects.

Municipality	Standardized Coefficients Beta	t-values	Significance	Number of observations
0935 Iveland	-.620	-13.726	.000	303
0121 Rømskog	-.593	-9.643	.000	173
1112 Lund	-.590	-20.388	.000	781
1545 Midsund	-.581	-15.991	.000	505
0811 Siljan	-.581	-18.759	.000	693
1145 Bokn	-.580	-9.736	.000	857
1111 Sokndal	-.571	-20.362	.000	981
1027 Audnedal	-.570	-13.529	.000	383
0817 Drangedal	-.568	-23.528	.000	1164
1029 Lindesnes	-.563	-23.363	.000	1179

Iveland municipality in Aust-Agder points out with the highest Beta coefficient, though decimals really do not differ much between these municipalities. Iveland was classified as 'agricultural' by 1974, as a 'mixed agriculture and manufacturing' municipality by 1985 and re-classified as a 'manufacturing' municipality by 1994 and later. In general a reclassification may be due to municipality merging, which in turn may affect the industrial structure etc. in the new unit. This is not the case for Iveland, however. The additional classification of Iveland is that the production industries employ more people than the service industries[118].

[118] Statistisk sentralbyrå (1994): *Standard for kommuneklassifisering 1994.* Oslo: Statistisk sentralbyrå. P. 20 and p 30.

Mainly the municipalities in table 28 are classified as either 'primary industry', 'mixed agriculture and manufacturing' or 'manufacturing' municipalities. None of them are classified as 'central, service industry' municipalities. Among these ten municipalities Siljan probably represents the most post-industrial labour market if considered by the official classification: Siljan is classified as a 'central, mixed service industry and manufacturing' municipality where service industries employ more people than the production industries. Midsund is a 'primary industries' municipality throughout the whole period – in 1974 specified as a 'fishing industries' municipality.

Both Audnedal and Bokn are classified as 'mixed agriculture and manufacturing' municipalities - specified as occupying more people in production industries than in service industries. An additional classification for Bokn is that by 1994 there are more employees in agriculture and forestry than in manufacturing industries.

Sokndal, Drangedal, Rømskog and Lindesnes are all 'manufacturing industry' municipalities where production industries employ more people than the service industries. Lund is also a 'manufacturing industry' municipality, with additional classification as 'single-industry municipalities' mainly manufacture of wood products.

Apart from Siljan, a feature common to these municipalities is that production industries employ more people than service industries. In this sense, none of these local labour markets are actually post-industrial. In Siljan, *labour market participation* should perhaps have been expected to be more even between men and women, especially considering the widespread commuting to the neighbouring city Skien which has a high rate of service industries.

The other end of the scale where *gender* effects are generally weak is mainly represented by municipalities in the northernmost county, Finnmark (municipalities starting with the county identifier '20' in the table). Apart from Oslo where the *gender* effect is weak, as expected, the result shows that seven of the ten municipalities in Table 29 are located in Finnmark county, which belongs to the 'action zone' referred to earlier. The labour market in the action zone is characterised as both gender and educationally segregated. Men are predominantly employed in primary industries, construction and transportation, while women are predominantly employed in the service industries, in education and health care in particular[119].

Statistical figures for the action zone show that in general a higher share of persons employed within this zone work in state or municipal service industries. The male employment rate in Finnmark is lower than the national average, while female employment is higher than male employment in seven Finnmark municipalities.[120]

[119] Stortingsmelding nr. 8 (2003-2004): *Rikt mangfold i nord. Om tiltakssonen i Finnmark og Nord-Troms.* Kommunal- og Regionaldepartementet. Ch. 3.3.2.
[120] Op. cit. ch. 3.3.1.

A glance at the classification for the municipalities appearing in Table 28 show a structural difference from the ten municipalities in Table 29. There is a predominant service industry, or in other words a post–industrial component, in all municipalities in Table 29 with the exception of Lærdal, which is classified as a 'primary industry' municipality. Apart from Oslo which is a 'central service industry' municipality, centrality is low in all municipalities in Table 29, which indicates that centrality seems to be less important than the service component for the strength of the *gender* effect.

Table 29. Ten municipalities with the weakest gender effects.

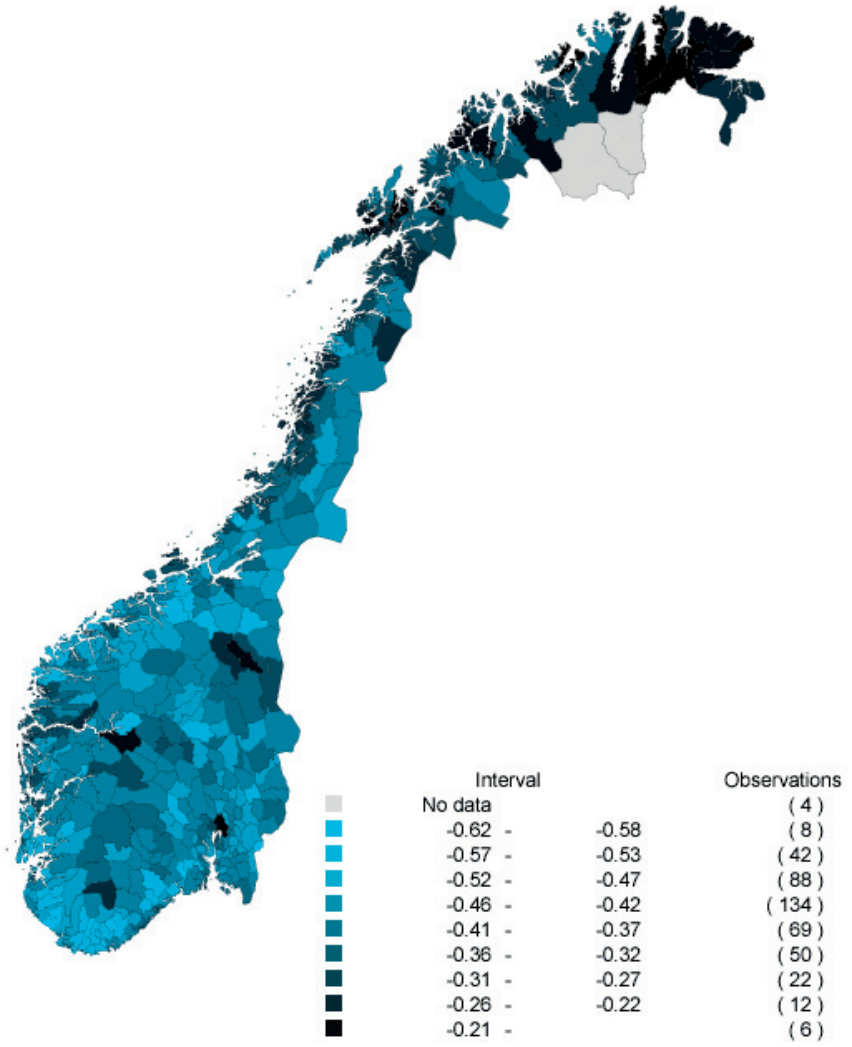
Municipality	Standardized Beta Coefficients	t - values	Significance	Number of observations
2002 Vardø	-.154	-4.341	.000	783
2025 Tana	-.154	-4.685	.000	910
2004 Hammerfest	-.182	-9.617	.000	2709
1911 Kvæfjord	-.191	-6.109	.000	985
0301 Oslo	-.194	-72.174	.000	132630
2022 Lebesby	-.200	-4.387	.000	464
2027 Nesseby	-.218	-3.839	.000	296
1422 Lærdal	-.222	-5.562	.000	600
2020 Porsanger	-.227	-8.187	.000	1233
2028 Båtsfjord	-.230	-6.078	.000	665

An interesting feature of Table 29 is the appearance of Lærdal municipality in Sogn og Fjordane county. In the 1970 municipality classification Lærdal was given a residual code – 'other municipalities' – but with the additional information that service industries had a higher share of employed than manufacturing industries. Since 1985 Lærdal has been classified as a 'primary industry' municipality. Counted by number of residents and by number of observations in the panel population, Lærdal is a small municipality with a total of 2199 residents in 1994, while the panel population in Lærdal counts 312 males and 288 females. Considering the 'primary industry' classification and the generally strong *gender* effect for this municipality class, a stronger *gender* effect should also have been expected for Lærdal. An additional public service institution in the form of a county hospital was established in Lærdal in 1970[121] which may explain the weak *gender* effect. Based on the general picture of the gender-segregated labour market, this might be an obvious

[121] Aschehoug og Gyldendal (1982): *Store norske leksikon.* Oslo: Kunnskapsforlaget

association, but this could only be further confirmed by individual *profession* and *employer* information from the E/E system or the census data bases.

Figure 11. Labour market participation and bivariate effect of gender. Standardized Beta coefficients. Municipality.

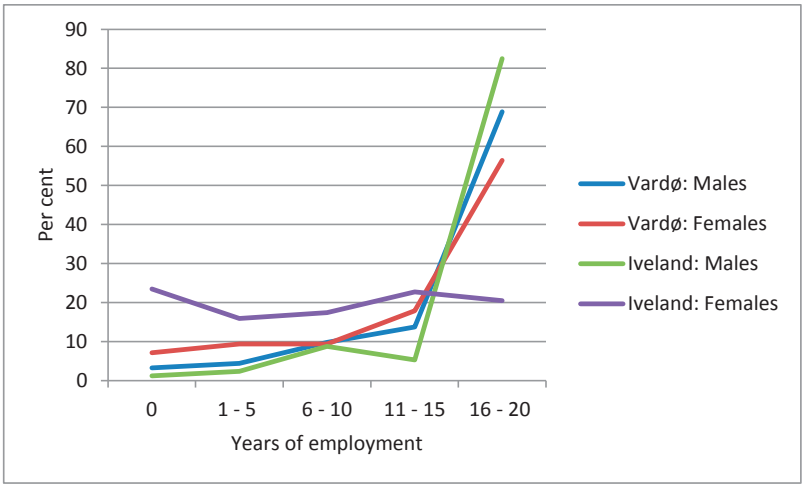


In Figure 11, the darkest blue polygon in the southernmost part of Norway is Bygland municipality in Aust-Agder county with a coefficient like -.277 which is quite exceptional for the region. In the various classifications this is a 'primary industry' municipality, but the service industries

employ more people than the production industries. The service industries in this municipality are mainly public, in heath care and education.

The four municipalities selected by Ragnhild Steen Jensen were Elverum, Nord-Odal, Årdal and Rana. Among these four municipalities, Rana and Årdal represent the most single-industry, manufacturing-related labour markets. The regression coefficients for these municipalities vary between -.407 and -.533, which means that the *gender* effect in general is quite strong in all of them. The strongest effect is found in Nord-Odal and in Årdal. The weakest effect is measured in Elverum which is classified as a 'central, mixed service industry and manufacturing municipality'. This municipality is also a location for additional public service industries in the form of governmental education and health institutions.

Figure 12. Vardø and Iveland. Males and females by number of years of employment. Per cent.



Municipalities with insignificant results appear as grey polygons in Figure 11. Two of these were the Finnmark municipalities Kautokeino and Karasjok which both cover large areas in km2 and appear as the largest grey polygons. Anyway, the main trends and differences between north and south are clearly depicted in the map. In the southern part of Norway, the municipalities Lærdal, Oslo and also Tolga (with a Beta coefficient of -.236 which is just outside the range of Table 29) stand out in the map as the darkest blue polygons. Tolga municipality in Hedmark county was merged with Os municipality in 1966, but separated again in 1976, and has been a separate unit since. Tolga is a 'primary industry' municipality in all classifications since 1974. But Tolga is also a municipality with extensive commuting. Regional public service institutions are located in the neighbouring municipalities Tynset and Røros, both of which are within commuting distance.

A weak *gender* effect expresses that there is not much difference in the level of labour *market participation* between men and women, and vice versa, a strong effect expresses differences in the level of labour market participation for men and women. If the effect is negative, the level is lower for women than for men. If labour market participation was higher for women than for men, the effect would have been positive, but this is not observed in any of the present analyses.

In Figure 12 the weak gender effect in Vardø is contrasted to the strong gender effect in Iveland. This picture confirms that in Vardø the gender effect reflects that the employment rate is equally high for men and women, while in Iveland the gender effect reflects a low employment rate for women and a high employment rate for men. The curve for females in Iveland is almost flat and in fact shows that the share of the female panel population with no *labour market participation* is slightly higher than the share with 16 – 20 years of employment history.

A desirable explanatory variable was *education*, and of special interest, the distribution of higher education. To some extent, the lack of individual education information can be compensated for by figures from official statistics, but only to some extent, because levels of measurement are different. Aggregated education numbers do not reflect the individual, gender or generation specific distribution in combination with labour market participation.

Figure 13 shows the county-wise distribution of higher education in 1970. Oslo and Akershus are the two counties with the highest share of the population with higher education. In general, the education level is higher among the population in the Agder counties than in Finnmark.

Figure 13 also confirms that the education level is higher among the female population in Aust-Agder and Vest-Agder than in Finnmark as of 1970. In 2000 the county-wise distribution of education is basically the same. As stated above, education is not available as a variable at individual level for this particular case study. Although the aggregated education figures do not necessarily reflect the individual association between education and labour market participation, the general picture makes it difficult to explain the difference in gender effects between Finnmark and Vest-Agder in particular, by education – level and distribution.

Figure 13. Persons 16 years and older with tertiary education by county. Men and women. Per cent. 1970. [122]

THE ASSOCIATION BETWEEN POST-INDUSTRIALISM AND FEMALE LABOUR MARKET PARTICIPATION

The question to be answered by the regression analysis was: what explains variation in labour market participation in post-industrial Norway – at national, regional and local level?

The theoretical model in Figure 3 had to be adjusted. At any rate, the adjusted model in Figure 9 does have explanatory power. The preliminary answers from the theoretical model to the question above are confirmed by the regression analyses on the empirical data. The analyses have confirmed systematic associations between *individual labour market participation* and *gender* and *generation*. The results have been proved to be significant with very few exceptions.

The conclusion is that the null hypothesis should be rejected: *gender* and *generation* had an effect on *labour market participation*, and this effect was observed at national, county and municipality level. The direction of the effect of each of the two explanatory variables also coincided

[122] Statistics Norway (1975): *Statistical yearbook 1975. Table 434.* Oslo: Statistisk sentralbyrå.

with the theoretical expectations in the model: *labour market participation* is negatively affected by *gender* and positively affected by *birth cohort*. The effect of *gender* was negative at absolutely all geographic levels, but the strength of the effect differed clearly between local labour markets. The *gender* effect was very weak in some local labour markets, and quite strong in others. *Generation* also clearly affected variation in *labour market participation*: the effect was positive, but clearly weaker than the *gender* effect.

The adjusted model included the explanatory variables *gender*, *birth cohort* and *geographic mobility*. There was apparently no effect of *geographic mobility* on *labour market participation*, and this null hypothesis could not be rejected. Altogether - testing the hypotheses on empirical data supports the preliminary answers about *gender* and *generation* effects from the adjusted theoretical model.

Table 24 summed up the explanatory effect of the trivariate model with *gender* and *birth cohort*. The R square (0.168) from the trivariate model was higher than the R square for each of the bivariate models in Table 21. The analyses also revealed that the residuals were high, which means that much of the variation in labour market participation is explained by other variables, e.g. *education*, *profession*, *employment in private or public sector*, in addition to explanatory variables that were intentionally included in the model, but appeared to be insufficiently operationalized. The test of the model fit revealed that the two *regional characteristics* variables were closely associated with each other, and at the same time did not contribute to the model, which is why these variables were omitted from the model.

However, the effect of *gender* was clearly stronger than the effect of any of the other explanatory variables included in the model, which makes a bivariate analysis defensible for local labour market analyses. As shown in Table 21 a, *gender* explained about 39 per cent of the variation in *labour market participation*, but for some local societies *gender* alone explained more than half of the variation.

At national and county level the number of observations was very high, but for a few municipalities the number of observations was critically small. These few municipalities failed the significance test, and for these municipalities the t-values were also below the critical line. Apart from this, the results at national level, county level, municipality class level, generation level and municipality level were all within acceptable significance and tolerance limits. The significance tests for all municipality analyses showed a monotonous result of 0.000. The effects of *gender* and *generation* on labour market participation that were measured are not random.

With reference to Ragnhild Steen Jensen and the question of whether the national labour market is just a summarised version of a number of local labour markets, and particularly as regards 'women's transition into paid work', the obvious answer is that the national picture is far from being a

large-scale mirror of a number of local labour markets. The speed and degree of women's transition into paid work has been very different locally.

The map in Figure 11 also expresses the axis of time from 1967 to 2007. The status in 2007 is that in some local labour markets the level of *labour market participation* is almost equal between men and women, while other local labour markets are characterised by distinctly lower *labour market participation* for women than for men**.** The pattern from Table 27 fits with the pattern from Table 28, and indicates that the labour market in e.g. 'manufacturing industry' and 'primary industry' municipalities enhanced male employment rather than female employment.

There is seemingly an interaction between post-industrialism and a high rate of female labour market participation, but a closer examination reveals that there are also exceptions when characteristics of local labour markets are considered by the municipality classification. Municipalities within the classes 4 – 7 (see Table 6) all have an important service industry component, and thus constitute more post-industrial labour markets than the other municipalities, and hence the *gender* effect should generally be expected to be weaker than in municipalities classified as 'primary industry' or 'manufacturing industry' municipalities. However, among both the most post-industrial and the least post-industrial local labour markets there are examples of local labour markets with a strong *gender* effect as well as with a very weak *gender* effect. A strong *gender* effect among 'central service industry municipalities' is e.g. found in Fet municipality in Akershus county with a coefficient like -.502. On the other hand, a weak *gender* effect in the 'manufacturing municipalities' group is found in Fedje municipality in Hordaland county with a coefficient of -.278. At the start of the 21st century some local labour markets may still be as manufacturing-oriented as they were 40 or 50 years ago, yet high female labour market participation can be found in such labour markets as well, although these are exceptions from the general trends.

In general, post-industrialism has been regarded as a condition for women's entrance into the labour market, expounded almost to the extent of a social law: women's new position in the labour market is intimately connected with the transition from an industrial to a post-industrial labour market, and this is a general feature of all modern, industrialised countries as this subject is e.g. summarised by e.g. Edgeir Benum[123]. But the examples above show that high female labour market participation may very well appear in labour markets which are not characterised as post-industrial and the 'law' that post-industrialism is a condition for high female labour market participation must be nuanced.

Local and regional differences are obviously not explained entirely by more or less technical measurements, e.g. classifications, indexes or other quantifiable sizes. Non-measurable conditions certainly do play a part, perhaps even a more important part than factors that can be measured,

[123] Benum, Edgeir (1998): *Hverdagsrevolusjonen, fritiden og kulturen. Aschehougs Norgeshistorie.* Vol. 12. Oslo: Aschehoug & Co, p. 170.

quantified or analysed quantitatively. But the measurements are still useful instruments for identifying general structural differences as well as for singling out the most different units.

Some local differences and trends depicted in Figure 11 must be considered to be a result of industrial and regional political measures. On the other hand, a few findings do not correspond with expectations and anticipations. Further explanations should also be traced among non-material and non-quantifiable forces, such as local attitudes, traditions and ideological factors which can hardly be counted or quantified.

As mentioned above, this enquiry has suffered from variables that were not possible to operationalize, not available or not accessible as individual variables, e.g. *education*. Since 1999 Statistics Norway has made a Gender Equality Index (GEI) for Norwegian municipalities. *Education* – both level of education and distribution of education - is accounted for in the GEI prepared by Statistics Norway since 1999. This index was revised in 2009 and according to the Statistics Norway definitions:

> (…) the indicators are kindergarten coverage for children aged 1-5 years, percentage of female municipal council members, number of women per 100 men aged 20-39 years, percentage of women/men aged 16 and above with higher education, percentage of women/men aged 20-66 in the labour force, and average income for women/men[124].

Two of the components in the GEI are related to, but not quite identical with the dependent variable *labour market participation*: 'Being in the labour force' and 'average income'. As of 2010 the equality status based on the index is summarised by Statistics Norway as follows:

> The municipalities with most equality are mainly found in the central and inland area of Eastern Norway. Many municipalities with the most equality are also found in the municipalities of Sogn og Fjordane, Troms and Finnmark. All of the six largest towns and cities - Oslo, Kristiansand, Stavanger, Bergen, Trondheim and Tromsø – are also among the municipalities with most equality[125].

The present enquiry definitely also has a gender equality aspect: increasing female labour market participation and hence a transition into paid work must be seen as one of many components of the general women's liberation process in the last decades of the 20th century. In many ways, the results of this enquiry coincide with findings based on Statistics Norway's more recent gender equality

[124] Statistics Norway (2010): *Statistics by subject – Gender Equality.* Oslo: Statistics Norway at www.ssb.no at www.ssb.no

[125] Hirsch, Agnes Aaby (2010): *Lokal likestilling – målt på en ny måte.* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 1/2010, p. 5.

index. High gender equality is found both in Finnmark and Troms, where the *gender* effect on *labour market participation* was generally weak, while at the other end of the scale Statistics Norway analyses also show low gender equality for the southern Norway counties Aust-Agder and Vest-Agder. This is the GEI status as of 2010, but based on the panel population results, the regional and local differences in terms of female labour market participation are definitely observed from the late 1960s.

Statistics Norway has ranked municipalities in groups in the 2010 GEI according to given values of the index[126] from 'Most equal municipalities' to 'Municipalities with least gender equality'. Nine of the ten municipalities from Table 28 – with strong *gender* effects and hence low gender equality - are found among the municipalities with lowest values on the gender equality index, i.e. with least gender equality. The only exception is Lund municipality in Rogaland county. From Table 29 Vardø, Oslo and Lærdal are found among the municipalities with the highest values in the 2010 GEI.

The GEI values observed for the Agder municipalities are discussed by Dag Ellingsen in the article '*Sørlandet fortsatt på etterskudd'* [127] from 2010. This article deals with the fact that in the municipalities in the Agder counties, *('Sørlandet'*), gender equality has been low for all of the ten years in which the GEI has been produced. This picture did not change when the GEI was revised. One of Ellingsen's observations for this region is that women have a low rate of labour market participation, and his explanations are basically that this is due to widespread conservative attitudes towards gender equality and traditional family values.

The panel population is composed of the birth cohorts 1937 – 1958. But, as stated earlier, these cohorts were not the only actors in the labour market during the period 1967 – 2007. The *gender* effect found in the panel population should also be found for the birth cohorts older than 1937 and younger than 1958, with an expected stronger *gender* effect for the oldest cohorts and a weaker effect for the younger birth cohorts.

RESEARCH FINDINGS IN OTHER COMPARATIVE STUDIES

In Norway the ambition of high female employment is an unquestioned political objective. With reference to the Norwegian political objectives about increased female labour market participation, and the 2000 Lisbon Strategy (above), an obvious question by now is then – to what

[126] Statistics Norway (2010): *Regional differences in gender equality*. Oslo: Statistics Norway at www.ssb.no

[127] Ellingsen, Dag (2010): *Sørlandet fortsatt på etterskudd.* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 1/2010, p. 15.

extent are these objectives reached? Measured at national level in Norway as of 2008 the objective seems to be quite close, e.g. the *gender* effect is apparent, but not very strong as seen from Table 21, but broken down on municipality level, the *gender* effect is quite strong in a high number of municipalities, which means there is still a long way to go. The local perspective has demonstrated that some local communities need more focus than others, in terms of political measures for equalization.

The final model with *gender* and *generation* as the remaining X-variables proved to have explanatory power. Due to the auxiliary, geographic variables the model could be applied on county level, municipality class level and on municipality level. Thus the picture of participation in the labour market appears as much more versatile and nuanced than the national picture can give – regional and local labour markets with gender based differences clearly exist.

Nationally and internationally several studies argue that within this subject of research there are special fields which still remain to be examined more closely: in particular comparative studies across regions within nations and across nations. A comparative enquiry about regional female labour force participation in Austria is conducted by Martin Falk and Thomas Leoni[128]. Their study is based on data sources from 2001 – 2002 and investigates and compares the differences between 121 political districts in Austria, and a major finding is that there is a high degree of variation between the regions when degree of female labour force participation is measured. They find that regional determinants (of which some are comparable to the components in the municipality classification by Statistics Norway – e.g. population density) do affect female labour force participation. Though the residential qualities as operationalized by municipality classification in the data matrix for the case study, could not be included in the regression analyses, there are regional and local characteristics with an impact on labour market participation in Norway as well, as seen from Table 27 and commented in the previous chapter.

Comparable with the present case study is also the enquiry by Thomas P. Boje[129] presented in 2007 (based on data sources for 2000 – 2001). In this study household strategies for combining paid work, unpaid work and care in four European countries[130] are analyzed, and Boje concludes that there is a clear gender effect in all four countries in this aspect. Furthermore he finds that in the Scandinavian countries mothers have higher employment rates than in all other European countries. Besides – in these countries women maintain their employment also when having small children[131]. The present case study, however, have identified local communities, e.g. in the Agder counties, where

[128] Falk, Martin and Thomas Leoni (Undated): 'Regional female labour force participation: empirical evidence for Austria'. *Vienna: Austrian Institute of Economic Research WIFO*.
[129] Boje, Thomas P. (2007): 'Welfare and work. The gendered organisation of work and care in different European countries'. *United Kingdom: European Review, Vol. 15, No. 3, p. 373-395*.
[130] Denmark, the Netherlands, Sweden and the United Kingdom.
[131] Ibid. p. 379

women have a low rate of labour market participation, and supported by the qualitative studies by Magnussen et al.[132] , show that having small children is an incentive to leave or reduce employment (see below, and cf. also Olivier Thévenon og Vlasblom et. al. above).

Another study to be referenced is an enquiry on determinants of regional female labour market participation in the Netherlands by An Lieu and Inge Noback[133]. This is a study of specific female labour market participation on municipality level including 278 of totally 496 municipalities in the country, based on data for 2002. The authors find that there are significant regional differences in female labour market participation as of 2002, ranging from 34 per cent to 70 per cent. An incentive for their enquiry is that they claim there is no 'national' labour market[134] , which is why municipality data must be analyzed and compared. This study carries several similarities with the present case study in terms of method, analysis and type of data, however the Dutch study is cross-sectional by 2002 and specifically focused on female labour market participation. Their findings support the basic idea that the national labour market is an abstraction, concealing various submarkets with individual characteristics, in accordance with findings from the case study.

As Vlasblom et al.[135] points out in their European based enquiry - there are changes in behaviour that have driven the increase in participation rates for females over the last decades. But also in EU-countries there is a long way to go before the objective from the 2000 Lisbon Strategy is reached, as comparative studies have revealed significant difference between countries. Olivier Thévenon reveals disparities between European countries from north to south based on data from 1992 and 2005. The Scandinavian countries (Denmark, Sweden and Norway) have the highest employment rates for women at both points of measurement. The lowest rates are found in eastern and southern European countries. However, he also finds that the growth from 1992 to 2005 is high in some Mediterranean countries, and Spain in particular[136].

The municipality wise analyses in the case study uncovered that the *gender* effect was clearly stronger in some municipalities than in others, i.e. women have a clearly lower level of employment than men where the effect is strong. In particular there is a concentration of municipalities with strong gender effects in the southernmost parts of Norway. This feature is commonly explained by reference to local culture, tradition and perception[137]. The data sources for the present enquiry have limitations in the sense that they do not reveal the underlying perceptions, attitudes and possibly local culture - the qualitative aspect of the findings - that might explain the local differences. However, attitudes and

---

[132] Op. cit.

[133] Liu, An and Inge Noback (2010): Determinants of regional female labour market participation in the Netherlands. The Netherlands: University of Groningen.

[134] Ibid. p. 642.

[135] Op. cit., abstract.

[136] Thévenon, Olivier, op. cit., p.238

[137] E.g. Ellingsen, Dag (2010): 'Sørlandet fortsatt på etterskudd'. *Oslo: Statistisk sentralbyrå. Samfunnsspeilet 1/2010, p. 15.*

perceptions in the Agder counties are closely investigated by Magnussen et. al[138] who presented a report in 2005 with special attention to mothers' labour market participation in these counties. Their report focuses on women's labour market participation, as this issue is considered particularly relevant for gender equality[139]. Their investigation is a qualitative, interview-based enquiry. Women in this enquiry are limited to mothers[140], as the impact of children are supposed to be the most influential single factor on labour market participation for women. One of the characteristics of the Agder counties is that they differ from the nation as a whole in terms of religious beliefs. The report also points out that these counties have a higher rate of disability pension than the national average.

In terms of motherhood and role expectations[141], Magnussen et. al. confirms that perceptions, attitudes and culture locally affect behaviour related to participating in the labour market. A common perception among the interviewees is that it is seen as desirable by both men and women that mothers stay home with the children. Among the interviewees is also a widespread acceptance of lower economic standard of living – if that is the price to pay for staying home with the children, in other words a one-income household seems to be widely accepted. It is actually regarded as selfish to fulfil one's own ambitions through a professional career. To be a 'workaholic' and a good mother at the same time is considered impossible, but on the other hand it seems to be quite possible to be a good father and a 'workaholic' at the same time[142] . Nevertheless – a major point is that the mothers claim that this is intentionally and in accordance with their own choice and preferences.

The questions to be answered by the regression analyses as stated page 45, are summarized above in the chapter 'The association between post-industrialism and female labour market participation'. The *gender* model proved to be the model with the highest relative importance. The effect of *gender* was negative whatever the level of analyses – national, county wise or municipality wise. But the association between *generation* and *labour market participation* is also clear, though weaker than the *gender* effect. The linear effect of *generation* can be seen indirectly from Table 25: the younger the generation – the higher the labour market participation. The *generation* effect must be explained by women's behaviour more than men's behaviour, as men's labour market participation were quite stable during the entire observation period.

Residential qualities matter for female labour market participation. The *gender* effect differs between the various municipality classes as shown in Table 27. A reasonable interpretation is that centrality and in particular the presence of service industries are determinants for the rate of female labour market participation. Apparently also political measures have an effect on female labour market

[138] Op. cit.
[139] Ibid. p. 207.
[140] Ibid. p. 19.
[141] Ibid. pp. 147
[142] Ibid. p. 163

participation, cf. the results and findings from the 'action zone' above, but immaterial values do play a part, and even counteract political measures, and more so in some local societies than in others.

A common feature for the international enquiries of the post-industrial labour market cited above, is that they have a short historical horizon, i.e. from the 1990s and onwards. To a large extent this is due to the digital data sources that are available. Several EU-oriented enquiries about labour force participation are based on data from The European Union Labour Force Survey (EU LFS). This data collection started in 1983 for the member states at that time. New member countries are represented only from their membership entry.

The Gender Equality Index (GEI) by Statistics Norway was introduced in 1999 as a municipality wise index, and the findings based on the GEI coincide to a high degree with the findings from my enquiry, cf. above. But the GEI is cross-sectional and the time series give a short historical overview. A major point with the present enquiry is that the regional and local differences that can be observed in recent studies, have existed for many decades. The case study has also demonstrated that data sources actually exist for conducting research along a historical axis starting from the late 1960s, and including population born during the interwar period. The subgroup of digital archives represented by electronic registers is source material about total populations – about 'all' of us and allows historians to select which individuals or which groups that should be the subject for research. In Norway and also the rest of the Scandinavian countries this type of data sources have a long history compared to the international situation. This is a situation that requires attention from several academic disciplines, above all for historians, cf. the following chapter.

DIGITAL ARCHIVES IN HISTORICAL RESEARCH – RESULTS AND EXPERIENCES

The experimental case study now just closed mainly served the purpose of illustrating and demonstrating the usability of digitally created archives, namely the informational value of electronic administrative registers in a historical research context, and the technical and legal conditions that must be coped with when the sources appear as raw data, which is the required format for long-term storage in archival institutions. This enquiry has demonstrated some weaknesses and limitations, but also strengths and possibilities offered by this kind of source material. The study will probably also have shown that the threshold for using such material from 'raw data' format is quite high, but nevertheless quite possible.

The fact that archived electronic register information must be used from 'raw data' format is also recognized by national authorities in 'Digitaliseringsmeldingen'[143], a white paper in which a strategy for the preservation and distribution of the digital cultural heritage is presented. A major implication of this is that utilization of digital archives of the register category from raw data format is not dependent on a future technical solution with a more user-friendly interface; in other words, we do not have to wait for such a solution to be developed.

An increasing number of electronic administrative registers are about to become very interesting for historical research, at least for contemporary history, as they gradually reflect a longer axis of time. In general, these systems store information dating not only back to their commencement 40 - 50 years ago, but also information about entities in the systems which may be very much older: the oldest person entered into the CPR was born in 1855 as mentioned above. As a consequence, there is no difficulty in establishing research populations based on registers of people born in the early 1900s or late 1800s, for that matter. Establishing a nationwide panel population for the birth cohorts 1937 – 1957 was no problem.

Each of the subcategories of digitally created archives outlined in the chapter 'Definitions and concepts' requires different solutions for access, recreation and application in a research context. Hopefully, this enquiry has contributed to clarification, awareness and consciousness of digital archives and their different subcategories, and helped to prevent misconceptions and confusion with computerised archives.

---

[143] St. melding nr. 24 (2008 - 2009): *Nasjonal strategi for digital bevaring og formidling av kulturarv.* Oslo: Kultur- og kirkedepartementet. Ch. 4.9.

Numerous assessments and decisions had to be made during the construction of the research data set. Operationalization of several variables required compromises and a need for alternative sources - many more than expected when this project started. The sources allowed some variables to be successfully operationalized while others were not.

In the present case study, the standardised national identifiers played a very important part. Without the national ID number, matching at individual level would have been impossible, and the variables *gender* and *birth cohort* could not have been derived. Selecting the panel population presented no technical challenges and was definitely successful, as were the demographic variables *gender* and *year of birth*. These variables could easily be operationalized and they passed consistency, coverage and validity tests. This kind of basic register information will probably be very much in demand in future research.

Some variables were laborious to operationalize, especially the dependent variable, but in the end this turned out to be a valid and representative variable. This conclusion is strengthened by the results of the analysis which coincided closely with comparable official statistics. Other variables were unsuccessful and turned out as not being either valid or reliable. A metric variable for *residential qualities* such as that intended for the regression model was impossible to obtain. But the fact that a variable was not successfully operationalized is not just a result, it is an important result. All in all, the operationalization of all the variables offered useful experience.

One of the difficulties in building the research data set was to combine variables which by nature are longitudinal with variables which by nature lack longitudinal and diachronic qualities. Such experiences will be common also in the future, and compromises must be made, a situation which by the way also applies to paper-based archives. In general, however, electronic administrative registers allow panel design, and this is basically no more resource-intensive than cross-sectional and time-series design.

Another problem was the question of the different entities that can be found in the various administrative registers: person – family – and especially the taxpayer unit. *Number of children* in the formal family constellation in the Tax Register for Personal Taxpayers is denoted only on one part of the taxpayer unit. This required a tedious effort to convert it to an individual variable, which in the end still turned out to be unreliable. Application of personal information from electronic administrative registers requires strict awareness of the entity which is defined for administrative purposes for a given system. At the same time, this is an illustration of a general challenge posed by secondary use of information constructed for administrative purposes.

A further comparison between the computerised data exemplified by the 'Ullensaker project' and the present case study comprises some similarities, but also several differences. This is basically social history, the entity is the individual person, and the analysis methods are quantitative and electronically performed, but the road from primary data to the final data set for the analysis is very

different. Digital archives, categorised as electronic administrative registers, are already coded and, most importantly, they contain national identifiers for basic entities, which makes computer-based linkage at individual level possible. While the identification and linkage process was recognised as very time-consuming and expensive by the Ullensaker project, this part of a research process represents no problem when electronic registers are the source material.

When dealing with digital archives, the number of observations to handle does not really matter – it makes no difference whether the population involves a few hundred or millions of observations.

The data matrix itself appeared to be flexible for various types of analyses, no missing values were reported by the analytic software, and e.g. 'all requested variables entered' were reported for absolutely all subdivisions of the data matrix.

The main instrument for analysing the data was all the way linear regression. Whether this was the best approach may be debatable. Other solutions, e.g. descriptive statistics, might have produced results that were just as good or even better, but analyses at low regional level including all regional units require a compact and comparable presentation of results which is obtainable by the regression coefficients. An obvious advantage in applying regression analysis is that the data set to be analysed can be defined quite accurately very early in the research process. The regression model specifies clear conditions and requirements for the research population and the variables to be operationalized. This is also very helpful in the heuristic process and gives a direction for what kind of source material to look for.

Data at aggregated level may be found in abundance e.g. in official statistics, but in general there is a problem with comparability over years in official statistics. Even statistics based on the population and housing censuses struggle with comparable measurements and definitions over longer time spans, e.g. 40 – 50 years or more. Sample surveys cannot provide figures at low regional level, and almost never at municipality level, unless one or a few municipalities are particularly selected for research. In any event, the results in terms of regional and local differences based on the panel population coincided very well with the more recent Gender Equality Index developed by Statistics Norway, which is obviously based on total population data.

Through this case study the migration strategy is actually demonstrated to be functional. According to this strategy, information from an active, electronic system must be extracted for the archives and stored in technology-independent format. The extracted data set must always be accompanied by technical metadata. During the heuristic stage the technical metadata must be studied carefully in order to identify sources in terms of relevant data sets, tables and fields. Though metadata information is readable to the human eye, basic computer skills are needed to understand the information. Together, the data tables and the technical metadata constitute a package of information, which is enough for future retrieval and use.

The main data sources for this enquiry (the TRP and the DSF) were created on a mainframe platform from the 1960s with the native character set EBCDIC. Archival extracts from these systems have been transferred to the National Archives of Norway as archival packages containing both data tables and technical metadata since the mid-1980s. From each of these packages it has been possible to identify relevant variables, and retrieve, reuse and process data on a technological platform quite different from the native environment. Information stored in technology-independent format, i.e. raw data, is not a user-friendly format, but is nevertheless an accessible format, as the examples have demonstrated. The data matrix could be established without regenerating any of the native data bases. Once the data matrix was ready, it could be analysed by user-friendly procedures in SPSS which was my software choice.

Personal information is strictly protected by legislation. The National Archives of Norway started appraisal and actually received table exports from some main systems by state administration in the 1980s. All these table exports came from central systems with history from the 1960s and onwards, and practically all table exports contain personal information. The release date for free distribution of personal information is normally 60 years, but in some cases as much as 100 years, e.g. for information collected according to the Statistical Act[144]. Thus the 1960 Population and Housing Census is open for free distribution from 2060. Of course protection of privacy applies to any kind of archives. Neither paper-based nor digital archives with personal information can be distributed freely before they reach the release date prescribed by law, but since all digital archives are young, they are excluded from free distribution for many years to come.

Though access to digital archives is restricted, there are exceptions: for research purposes access may be granted based on specified exceptions in the Norwegian Public Administration Act. An important consideration in this context is that researchers who have been granted access to confidential material are themselves legally bound to confidentiality[145]. Researchers may never publish information in such a way that it may be traced back to a specific individual. All these regulations of course apply to my own access to the relevant archives. This particular case study is above all an example of utilisation of digital archives for research purposes within the legal framework for access for research purposes as defined in the Public Administration Act. To meet these requirements, the ID number was replaced with a neutral counter when it was no longer needed, and eventually the data matrix was deleted to finalise and close this project.

Administrative registers often contain information which earlier only existed in paper-based forms, card files etc. in large numbers and large volumes. In paper-based form, such archives could hardly be used at all due to their large volume, which is why they often were not preserved, or only

[144] Statistikkloven - LOV-1989-06-16-54, § 2-7.
[145] Forvaltningsloven - LOV-1967-02-10, § 13e.

preserved as samples. But in electronic version such information may very well be utilized, which in fact opens the door for an extended range of research subjects.

Although public electronic registers and data bases were constructed to serve administrative purposes, there has always been awareness of their potential for secondary use, e.g. for statistics and for research purposes (see the information about the construction of population and housing censuses above), and hopefully to a growing degree for historical research purposes. Obvious subjects of research would be e.g. demographic studies, economic studies and, in general, regional and local studies. Electronic registers offer vast possibilities for tracing individuals over decades in panel studies, they are suitable for cross-sectional studies and for extracting sample populations. But any use of digital archives belonging to the register category for future research has to be handled from raw data format, as this case study has demonstrated. It is not realistic to assume that more user-friendly and ready-made versions will be prepared in the foreseeable future. Thus the method developed for this enquiry can be transferred to other historical as well as other research projects based on electronic registers.

Methods for the long-term preservation of digitally created archives are heavily debated and elaborated in contemporary archival theory. As mentioned earlier, a key concept for the operationalization of methods for transferring archives from the records creators to the archival institutions, the long-term storage itself and, finally, distribution is 'package'. Archival packages appear with different names and different contents depending on which of these stages they are related to. A Submission Information Package (SIP) is a package produced by a records creator for transfer to an archival institution. Inside the secured repository of an archival institution, the SIP is received and when authenticity has been verified, each package is stored as an Archival Information Package (AIP). Information from an AIP must be made accessible in terms of a Dissemination Information Package (DIP). The data matrix assembled for the present enquiry is an example of a DIP. This particular DIP contains information from several different AIPs: each of the annual volumes of the TRP is a separate AIP, as well as each annual income table from the DSF. Thus the data for this particular DIP has been collected from a large number of AIPs. The final DIP is anonymous, but the entity is still the individual person.

Normally, a DIP must be exported from the archival institution to e.g. a researcher for utilisation and analysis, a process that parallels the methods applied by Statistics Norway, for instance. In order not to violate privacy restrictions, any DIP with the single person as the entity must be made anonymous inside the secured repository of an archival institution. As a consequence, any such DIP cannot be matched to any other individual data. The application of a given DIP in a different research context is thus very limited. The basic rule is therefore that any new research context requires a new, 'tailor-made' DIP from the archival institution, but according to specifications and requirements given by the researcher.

The major strength of electronic administrative registers in research is the numerous ways of combining data from different systems and from different records creators at individual level. It is not possible to predict which combinations of data that will be demanded. It is debatable whether a user-friendly solution for access is possible without losing flexibility – the flexibility to combine data from different records creators and different data sets – lengthwise and crosswise as attempted through this case study. Any solutions to meet the future demand for research data sets must always recognize that the construction and contents of any DIP is basically unique and unpredictable.

An article by Siw Ellen Jacobsen published in the periodical *Bladet Forskning* in 2010 deals with the paradox that though Norwegian administrative and research registers '(...) are the best in the world' [146], these registers are only to a small extent actually applied in social science research. This is partly due to privacy restrictions, but also to the problem that there is no 'recipe' for how to access them for research within applicable legal frameworks. Hopefully, this thesis points out solutions for a wider use of electronic registers in social science and historical research in the near future.

ELECTRONIC REGISTERS, ADMINISTRATIVE STRUCTURES AND EMPIRICAL FINDINGS

Electronic administrative registers emerged in Norwegian state administration from the 1960s as outlined in the chapter 'Definitions and concepts'. Administrative registers and systems have been developed ever since, in increasing numbers and in a context of changing conditions in terms of economy, politics and technological complexity. Though many of the administrative registers are established to support long-existing functions as stated in the chapter 'Digital archives versus computerised archives', it is necessary to consider that the establishing, contents and functions of electronic administrative systems are influenced by political decisions, legislation, economy, technology and systems developers and suppliers. In the early days the state institutions were developers themselves, but gradually professional systems developers have entered the market with possibilities to influence design and functionality of a given system, and thus possibly the picture of the past reflected by a given system. Nevertheless political decisions and legislation play a fundamental part in establishing, design and contents of administrative registers. Some electronic, administrative registers are results of debatable and controversial decisions, but registration in administrative registers are broadly accepted by the public. Register coordination and record linkage as tools to reveal fraud – tax fraud, social security abuse etc. is continuously carried out without much public debate.

The Data Protection Authority *'shall facilitate protection of individuals from violation of their right to privacy through processing of their personal data'*[147]. This institution monitors the handling

[146] Jacobsen, Siw Ellen (2010): *Samfunnsvitere må bruke tall.* Bladet Forskning 2/2010, pp. 16.
[147] Quotation from the web pages of the institution.

of personal information according to the Personal Data Register Act[148] , e.g. type of data to be entered and stored in a system, and distribution and secondary use - in particular linkage of personal information. Anonymising data or deletion of systems with sensitive, personal information when demanded by the Data Protection Authority has sometimes conflicted with the National Archives' policies in preservation of personal data and primarily preservation in authentic shape, see also above about authenticity. The conflicting interests are dealt with in the Archives Act[149], but so far never tested to the limits.

Statistics Norway and the censuses have played an important part in the development of basic data registers. As mentioned the 1960 Population and Housing census was the basic input for the CPR, the 1970 census for a research register on the highest completed education of the Norwegian population, the 1980 census formed the basis for the Cadastre, and the 2001 census introduced the household number through the expanded numeric address code (cf. Table 34). This census also introduced codes for 'profession' in the E/E-register on a permanent basis from 2001.

Technologically a strict framework for data entry is required, and the user interface 'dictates' data entry to a higher degree than what is the case in a manual system - see also the review about data entry and how systems are filled with data in the chapter 'Historical criticism'. A system will e.g. deny the current user to overrule conditions for data entry – an attempt to enter an invalid personal ID-number e.g. would normally be stopped by validity checks in the software. Individually characterized data entry would be quite possible in a manual system.

Systems are also influenced by standards – national as well as international. Standardization is generally desirable in a perspective of data exchange and efficiency. Standards are not static and will be developed over time, causing discrepancies between old and new versions, which may cause problems for comparison in the historical perspective, cf. e.g. the changes in standard for municipality classification in the chapter 'Residential characteristics expressed by municipality classification?'

The latest version of the NOARK-standard developed by the National Archives of Norway, is a standard for recordkeeping. This standard regulates recordkeeping and makes preservation more automated, i.e. the entire life cycle of the archives is influenced by this standard, in other words - influenced by the National Archives – the archival institution. A debatable issue is whether this intervention by the archival institution is too strong. There is a difference between systems developed according to the records creators own terms – only, and systems developed according to terms by an archival institution which poses influence throughout the entire life cycle of the archives, and thus leaving less 'free space' for influence by the records creator. The question is whether the intervention by the archival institution affects the future digital archives and thus the picture of the past too

[148] Personopplysningsloven – Lov om behandling av personopplysninger. LOV-2000-04-14-31
[149] Arkivloven – Lov om arkiv. LOV-1992-12-04-126, § 9.

strongly. Anyway the NOARK- standard is basically connected to systems in main group 1A above, and possible case handling functions in future subject-specific systems. The contents of subject-specific systems applied in the case study are not influenced by this standard. The principles for appraisal internationally, and thus nationally, seem to give more and more precedence to the evidential value of the archives (see the chapter 'Informational value versus evidential value'). Taken to the limit this might result in preservation only of systems with basically evidential value, while systems with low evidential value, but high informational value will not be preserved (like subject-specific systems), alternatively that information from subject-specific systems is preserved in formats which make computer based analyses impossible. A consequence would be reduced possibilities for research based on the informational value of the archives.

Administrative needs and legislation are continuously changing and can make comparability over time difficult. The layout of the annual TRP files e.g. are affected by annual as well as periodical changes in tax legislation. When the surtax on high income was introduced from 1988, this required new fields to be added in the TRP.

The political and technological assumptions for any electronic administrative register, also puts forward conditions for the future use of the registers as historical sources, like all other administrative sources by the way. Paper based parish registers, tax rolls, crew rolls, probate records etc. were all created for primary functions in an administrative context. A general condition for electronic, administrative registers is that information is collected for administrative purposes, and not for giving answers to questions that researchers will ask in aftermath. Any appliance in historical research of such records represents a secondary use of information collected for administrative purposes. Then there is always a question about how a given administrative source fits with the intention in a specific research context as discussed in the chapters 'Representation of the concepts from the theoretical model', 'Historical criticism' and 'Digital archives in historical research – results and experiences'.

Administrative registers are collections of various factual and formal information about a population. They reflect a formal status, which is not necessarily the empirical status, cf. the discussion about marital status and cohabitation. The thousands of persons who stay illegally in the country are not captured by the registers, etc. Furthermore – the administrative registers do not reflect personal attitudes, perceptions, beliefs and the like. These are examples of questions that would be able to ask in an interview based enquiry. In the present case study this is particularly a shortage for the findings of the Agder counties where the differences between male and female labour market participation are among the highest in the country, as discussed above.

Registers are not neutral or flawless, there will always be errors in data sources that are never discovered and corrected, but on the other hand the Archival Act actually insists on preserving the errors as long as they are genuine and authentic, cf. the presence of invalid ID-numbers in the DSF

population table discussed above. Thus interpretation of the data sources will always include uncertainty. These are all issues that require careful attention and treatment during source criticism. Dependency between data sources due to data exchange and reuse of data involves the risk that errors are repeated from one data source to the next, cf. the chapter 'Historical criticism', where this issue is closer discussed. It is also noticeable that the purpose for applying electronic data sources differs between e.g. quantitative analyses on the one hand and single person queries e.g. for documentation of individual rights on the other hand. For the latter category possible erroneous information in single fields may have serious consequences, but one single error among thousands of observations hardly affects the errors of margin in quantitative analysis.

All in all the circumstances under which electronic registers are established and conserved will affect the appliance in research and accordingly research results. Nevertheless – there is much demand for these data sources for research. Compared to use of questionnaires the registers have obvious advantages, but also disadvantages. With a few exceptions (e.g. censuses) questionnaires can only comprise sample populations and hence there will be selection problems, statistical uncertainty and error margins etc. which is avoided by total populations as found in electronic registers. Design and contents of questions for a questionnaire, as well as interpreting the answers will involve critical judgements. Besides there is always a risk of missing answers from the respondents due to attrition and apostasy. Generally a database will have a better 'memory' than most people in terms of historical factual information. Very few people will e.g. remember their exact income amounts 10, 15 or 20 years back in time, or the registration number of their previous cars etc. Such information can easily be retrieved from databases as long as they exist.

Preservation of archives, whether paper based or digital, is an important part of administrative structures which form the basis for any research that requires use of historical source material. Records, information and documentation from public administration – the transactions, decisions, case handling etc. are eventually the responsibility for archival institutions. The question about preserving digitally created information emerged from the early 1980s when the National Archives of Norway realised that such archives should be considered the same way as traditional, paper based archives. Systematic assessments of digital archives in state administration have been carried out for decades, and resulted in decisions about preservation or not by the National Archivist of Norway. Such decisions are continuous, this is a selection process which ultimately decides what kind of archives that will be accessible for any research. The decisions are solidly based on archival theory, legislation and regulations, but nevertheless – these are decisions that affect what kind of source material which is actually preserved, and basically what kind of picture future researchers will get about the past. Together with practice in making records accessible or not for research, these are frameworks that will affect research results, cf. also the discussion about the NOARK-standard above.

The capacity of handling incoming deposits and transfers is certainly also a limitation, for the National Archives of Norway as well as for archival institutions in general. Information from the oldest subject-specific systems, and in particular sub group 2A above, were prioritised for preservation by the National Archives, but still – as of 2014 – the discrepancy between what is decided to be preserved and what is actually transferred or deposited from various records creators, is only increasing. This is basically the situation for all archival institutions.

The ideal situation for the present case study would have been unlimited access to a complete CPR extract, to the census databases 1960 – 1990, an unbroken chain of the TRP extracts for the years 1967 – 2008 to mention the most desirable data sources. But this is unfortunately not the case for reasons outlined above. Sources must be found among what is actually transferred to the National Archives of Norway, and to what can legally be accessed.

In the present case study the results are affected by several limitations and administrative conditions. One example is the different entities from 'person' in the DSF to 'taxpayer' in the TRP. This difference possibly gave a poorer quality to the variable *family responsibilities* in terms of *number of children*. One of the most probable effects on results of the analysis is that *number of children* would have been a variable with much better coverage if the CPR and the census files had been available. However, this does not necessarily mean that the effect of the variable would have been significant, or that the null hypothesis should have been rejected, but the variable would most likely have been included in the analysis. It is probable that also the variable *geographic mobility* would have been possible to deduct from the migration history tables in a complete CPR extract. Still it is uncertain if the coverage would have been much different from the same variable deducted from the DSF. The time frame for the observation period appeared to be limiting for possible data sources. A later start for the observation period would have opened up for more data sources, but the cost would have been a shorter observation period.

The present discourse about digital archives seemingly emphasises as more and more important also to preserve queries from the administrative use of a given system, in order to repeat the same queries after transfer to an archival institution. Obviously this might play a part in the question about authenticity, but it is important to have in mind that future research will not repeat the administrative queries, but on the contrary ask new questions, apply the systems in new combinations and new constellations different from the administrative use. Thus there is a timeless link between use of digital sources and the numerous combinations of various historical sources in previous and present historical research, as done by several historians e.g. in the 'Ullensaker project', and in the study 'Marriage decisions in a peasant society' referenced above.

APPENDIX 1: SAS DATA STEPS

Logically, the data steps documented below should have been presented in between the various sections in the chapters 'Building the data matrix' and 'Historical criticism', respectively, but to enhance general readability these data steps have been collected and displayed in one chapter.

For the preparation of the data matrix, the software choice has been SAS. Inclusion of SAS code snippets in this thesis is required for several reasons: this is necessary documentation of how raw data sources are handled, how variables are operationalized and altogether leaves an open window to the method and how the sources are handled. The data matrix for this particular research project had to be deleted when analyses were finished, which makes documentation by source code, table description etc. even more necessary. The data sources are not open to access for everyone, and the most accurate way to document the selection of the research population and operationalization of the variables is by means of the source code. In many ways this is the most important chapter in this book.

On the other hand, this is not a tutorial in SAS programming. Many of the code examples might have been written in a more compact and advanced form, e.g. data step number 5 below could alternatively have been written with a loop construction, but perhaps that would have been less informative for the common reader.

The SAS software package offers both programming resources as well as analytic resources. Basically, programming resources and analytic resources are distinguished as DATA STEPS and PROCEDURE STEPS, respectively. In technical terms data steps are recognized by the key word DATA, while procedure steps are recognized by the key word PROC. Only in terms of PROC steps do we talk about standard procedures, e.g. PROC SORT, PROC CORR, PROC REG etc. SAS DATA steps, on the other hand, are never standard, but source code written individually for a given task. In general, DATA steps are required to handle raw data as preparation for a research data set (from an AIP to a DIP) which is later to be analysed electronically. In this particular enquiry, a large amount of preparation had to be done before standard procedures could be applied, a situation which will be common for archived raw data extracts from electronic registers. This chapter deals with methods and techniques for establishing a DIP based on one or more AIPs stored in the National Archives of Norway's secured repository. With reference to the secured zones mentioned above, all basic DATA steps must be executed behind the fences in the secured zones, while PROC steps belong to the individual researcher's 'laboratory'.

Though the DATA steps are written as separate modules, several data steps may be appended in the logical order and submitted as one job. During the preparation of the data matrix several SAS-data sets and text files were produced. All temporary files were deleted when inspection and checks verified the result and these files were no longer needed. This was a useful way to monitor the process,

and to avoid starting from the very beginning if something 'derailed' somewhere during the process. Temporary files enable a restart from where a possible mistake happened.

Only DATA steps are presented below. To create the 'tailor-made' data set – the DIP - for this enquiry, a total of more than 900 DATA steps were required. Among these I have selected the most crucial DATA steps to serve as documentation for the creation of the data matrix, and to provide ideas for future use of electronic registers.

DATA STEPS FOR ESTABLISHING THE PANEL POPULATION

The structure and contents of the demographic table from the Central System for National Social Security (DSF) are shown in Table 30. With reference to the choice of variables from this table, I wrote the code shown in SAS data step 1. The input file is named *person.dat* in the SAS code.

Table 30. Description DSF – demographic table. [150]

Start position	Field length	Data type	Field name
1	11	Character	National identity number – primary key (pk)
12	26	Character	Name (last name – first name)
38	1	Character	Code for type of pension. Only for current pensions – otherwise '0'
39	8	Numeric	Possible date of death. YYYYMMDD Actual date or '0'

In this data step two separate files are created: *data.population* and *data.d_number*. The log for this data step shows that roughly 6.5 million observations were read from the infile *person.dat*. This data step also includes a deselection of D-numbers. The data set *data.population* has about 5.8 million observations, and the data set *data.d_number* has slightly more than six hundred thousand observations. In the study 'The D-number population', the D-number population numbered about four hundred thousand observations as of 1995[151]. The status of the DSF population table in this case study

[150] All table descriptions and file names by my translation.
[151] Strand, Børge (1996): *D-nummerpopulasjonen.* Oslo: Statistisk sentralbyrå, Notater 96/39, p. 6.

is as of December 2007. The data set with the D-number population is created solely for checking and monitoring purposes, but is then deleted.

SAS data step 1. First selection of population.

```
LIBNAME data 'c:/data/';
DATA data.population
       data.d_number;
  INFILE 'D:\ARKIV1\person.dat' LRECL=46;
  INPUT
            @1 pk       $char11.
            @39 yod     $char4.
            ;
IF substr(pk,1,1)<4 THEN OUTPUT data.population;
      ELSE OUTPUT data.d_number;
RUN;
```

The input data set (INFILE) *person.dat* is read directly from the storage media – the CD-R in this case – with the drive letter D:\. The catalogue structure and data set notation on the storage media follow standards and regulations developed by the National Archivist of Norway[152]. The *.dat* extension is in accordance with the standards.

SAS data step 2. Validity check of national ID number.

```
DATA data.invalid
       data.valid;
SET data.population;
%INCLUDE 'C:\My SAS Files\9.2\SAS_MACRO\pkcheck.sas';
 %pkcheck(pk,return);
 IF return = '1' THEN output data.valid;
      else output data.invalid;
RUN;
```

[152] Forskrift om utfyllende tekniske og arkivfaglige bestemmelser om behandling av offentlige arkiver. FOR 1999-12-01 nr 1566: VIII. Bestemmelser om elektronisk arkivmateriale som avleveres eller overføres som depositum til Arkivverket

After deselecting the D-numbers, the remaining population must be squeezed through a few more cleansing procedures that will reduce the population: a validity check of the ID number (the primary key - pk) as well as a duplicate check.

Data step 2 contains a macro call for the validity check. (The source code for the validity check is protected by copyrights and may not be published). This step returns one file with invalid ID numbers, and one file with valid ID numbers: *data.valid.* Next, the file with valid ID numbers has to be checked for duplicates. The duplicate check requires that the file is sorted on the primary key. After this procedure, the data set *data.dup_free* is further processed, while the file *data.duplicate* is deleted after visual and logical inspection.

SAS data step 3. Duplicate check.

```
DATA data.dup_free
      data.duplicate;
      SET data.valid;
      BY pk;
      IF FIRST.pk AND LAST.pk THEN OUTPUT data.dup_free;
      ELSE OUTPUT data.duplicate;
      ;
RUN;
```

Below is an example of a (fictitious) national ID number with the main components separated to illustrate the structure:

Table 31. Components of the national ID number.

Date of birth	Personal number	
	Individual number	Control digits
170860	744	19

The variable *gender* has to be deduced from the ninth digit in the national ID number and added to the research data set as a separate field. Moreover, a four-digit year of birth is desirable for the population selection.

The date of birth in the ID number does not display the century in which a person is born as the format is DDMMYY. To separate by year of birth, a four-digit year of birth with century included is desirable. Strictly speaking, the selection of birth cohorts might have been based on two digits as there is little risk of confusion with a year of birth outside the 1900s, but theoretically the DSF

population table could contain persons born in the years 1855 - 1858. As a precaution the four-digit year of birth is preferable.

The four-digit year of birth must be calculated as a combination of individual number and the actual year within the century. The following set of rules is applied for persons born between:

- 1854-1899: individual number in the series 500 – 749
- 1900-1999: individual number in the series 000 – 499
- 1940-1999: also individual number in the series 900 - 999
- 2000-2039: individual number in the series 500 - 999.

These rules are set out in section 2-2 of the National Population Register Regulations relating to national ID number[153]. According to these rules, the example ID number above – *17086074419* – would have been issued to a woman (digit number nine is an even number) who was born in 1860 (individual number 744). To derive the century digits, it is simply enough to 'wrap' the SAS code around the set of rules from the regulations as in data step 4.

SAS data step 4. Derive gender and calculate four-digit year of birth.

```
DATA _NULL_;
SET data.dup_free;
LENGTH yob $4;
FILE 'C:\DATA\century.txt' lrecl= 20;
if substr(pk,9,1) in('1' '3' '5' '7' '9') then gender = '0';
 else if substr(pk,9,1) in('0' '2' '4' '6' '8') then gender = '1';
if '5' <= substr(pk,7,1) <= '9' and substr(pk,5,2) < '40' then
SUBSTR(yob,1,2) = '20';
else if '0' <= substr(pk,7,1) <= '4'
       then SUBSTR(yob,1,2) = '19';
else if  substr(pk,7,1) = '9' and substr(pk,5,2) > '39' then
SUBSTR(yob,1,2) = '19';
else if '5' <= substr(pk,7,1) <= '7' and substr(pk,5,2) > '54' then
SUBSTR(yob,1,2) = '18';
SUBSTR (yob,3,2) = SUBSTR(pk,5,2);
```

[153] Forskrift om folkeregistrering. FOR-2007-11-09-1268, § 2-2.

```
PUT
            @1 pk             $char11.
            @12 yod                 4.
            @16 gender              1.
            @17 yob                 4.
            ;
RUN;
```

In this code snippet the new field *gender* is added to the temporary text file *century.txt* with the value = '0' for male and '1' for female. The field *year of death (yod)* is kept for later use, when the criterion 'alive at 50 years' will be applied.

The text file *century.txt* is produced in data step number 4. The structure of this text file may be read directly from the code above (the PUT section) with the fields *primary key (pk), year of death (yod), gender* and *year of birth (yob)*. The file *century.txt* is the basis for selecting the gross panel population and splitting by cohort. When *century.txt* is sorted by year of birth, the lowest year of birth appears to be 1860 and the highest is 2007.

SAS data step 5. Selecting population cohorts.

```
DATA
data.cohort1937   data.cohort1938   data.cohort1939 (...) data.cohort1958;
INFILE 'C:\DATA\century.txt' lrecl= 20;
INPUT       @1 pk       $char11.
            @12 yod           4.
            @16 gender        1.
            @17 yob           4.
            ;
DO;
IF    yob = 1937 and (yod = 0 or yod > 1987) THEN OUTPUT data.cohort1937;
ELSE IF yob = 1938 and (yod = 0 or yod > 1988) THEN OUTPUT data.cohort1938;
ELSE IF yob = 1939 and (yod = 0 or yod > 1989) THEN OUTPUT data.cohort1939;
(...)
ELSE IF yob = 1958 and (yod = 0 or yod > 2008) THEN OUTPUT data.cohort1958;
END;
RUN;
```

The gender procedure is submitted before selecting the birth cohorts, only for technical reasons: each birth cohort has to be treated as a separate file for several data steps to come. Eventually these files will be appended to form one new file when the dependent variable is finally established.

Keeping the cohorts separate for the next data steps is a technical solution as each cohort is going to be matched against different generations of income tables to add the labour market participation history. For instance, the 1937 cohort has to be matched with DSF income years 1967 – 1986, while the 1958 cohort will be matched with DSF income years 1988 – 2007. This principle of successive matching is applied in general, in order to attach variables to each individual within the 20 years between 30 and 50 years of age for each cohort.

From the file *century.txt* the gross panel population is selected, on the basis of the following requirements: selection by year of birth to meet the requirement 'year of birth 1937 – 1958' and by year of death to meet the requirement 'alive at 50 years age'. This requirement is tested by the values in the field *year of death (yod)*.

Each cohort is selected on the basis of the same logic, but with ascending year of birth and with ascending (possible) year of death. The value of the field *year of death* is either '0' for living persons, otherwise the value of the field is the actual year of death in the format YYYY. This data step writes each cohort to a separate file for technical purposes. Later these cohort tables will be appended into one file. The code in SAS data step 5 was applied for the cohort selection.

Data step number 5 produces one file for each birth cohort. The field *year of death (yod)* is now omitted from the cohort tables as this criterion is no longer needed. The temporary file *century.txt* is also deleted. The variable *birth cohort* is derived from the year of birth by following this principle: Year of birth = 1937 is converted to birth cohort 1, year of birth = 1938 converted to birth cohort 2, etc. Thus the value range for this variable will be 1 – 22 which should make it applicable as a metric variable in the regression analysis. Each of the cohort tables has the following structure:

Table 32. Structure of each cohort table.

Start position	Field length	Data type	Field name
1	11	Character	pk
12	1	Numeric	Gender: 0 = male, 1 = female
13	2	Numeric	Birth cohort

DATA STEPS FOR THE DEPENDENT VARIABLE

The dependent variable – *labour market participation* – is deduced from the labour income amount in the DSF income tables. The structure and contents of the income tables from the DSF are shown in Table 33. There is one income table for each year since 1967. The municipality identifier from this table will be used later on when the geographic variables are added.

Table 33. Description DSF – income tables.

Start position	Field length	Data type	Field name
1	11	Character	National ID number (pk)
12	4	Numeric	Income year
16	1	Character	Code for type of income B = wage earnings C = self employment in other industries D = self employment in primary industries X = Correction amount (negative)
17	4	Numeric	Municipality identifier
21	9	Numeric	Income amount in NOK

In cases where a wage earner has several employers in the course of a year, the amounts are already summed up in the DSF. But when a person has different income types (codes are 'B', 'C' or 'D', respectively), he or she will occur more than once in the income tables, with one observation for each income type. Possibly three different income amounts have to be summed up to obtain one amount for each observation.

By selecting the income types one by one, and then running a match between the data sets, there will be only one observation for each person per income year, and the amounts can be summed up. In this way a 1:1 relationship is established between the population data set and each income data set.

The following data steps document how the dependent variable is operationalized, starting by selecting one file for each of the possible income types: *wages*, *income from self-employment in primary industries* and *income from self-employment in other industries*. This data step produces one temporary file for each income type for each income year.

SAS data step 6. Select income by type and by income year.

```
LIBNAME data 'c:/data/';
DATA data.wage67;
  INFILE 'c:\data\income67.dat' LRECL=46;* Read DSF income file 1967
  INPUT
            @1    pk                $char11.
            @16   code              $char1.
            @21   wage67            9.
            ;
if code = 'B' then output data.wage67;
RUN;
```

SAS data step number 6 has to be repeated once for each income year, creating data sets for wage earners, and for each of the two types of self-employment. These files are going to be merged in the next data steps. The code for income type is used only for this selection. The distinction between self-employment in primary industries and in other industries has not been present from the beginning. A change of definitions took place in 1982.

SAS data step 7. Match labour income by type and sum up income fields.

```
LIBNAME data 'c:/data/';
DATA data.match1_67;
MERGE data.wage67  (IN=inn1 keep  = pk wage67)
       data.selfemp67  (IN=inn2 keep  = pk selfemp67);
BY pk;
IF inn1 or inn2 THEN OUTPUT data.match1_67;
run;
data data.match2_67;
MERGE data.match1_67  (IN=inn1 keep  = pk wage67 selfemp67)
       data.primary67   (IN=inn2 keep  = pk primary67);
BY pk;
IF inn1 or inn2 THEN OUTPUT data.match2_67;
run;
```

```
data _NULL_;
SET data.match2_67;
FILE 'C:\data\income67.txt' LRECL=21;
income67 = SUM(wage67,selfemp67,primary67);
PUT
            @1    pk                $char11.
            @12   income67          10.
            ;
RUN;
```

After a split by income type, the files are ready to be matched in order to sum up three possible income types for each observation. Before any match all data sets must be sorted on the linkage key – the national ID number in this case. Technically it is a golden rule never to match more than two data sets at one time. The SAS key word for match is MERGE. To make sure that all observations are written to the text file, there has to be an IF – OR statement in the match.

For each income year a text file is produced where the field *incomeYY* (*income67* in the example) is the total labour income – in the example for the income year 1967. The amount in this field will be tested according to the income line settled for each income year (2 G) in the following data step. In advance the text file *income67.txt* is imported to a SAS file by a separate data step.

SAS data step 8. Test income amount and replace with code for employment.

```
LIBNAME data 'c:/data/';
DATA _null_;

SET data.income67;
FILE 'c:\data\coded67.dat' LRECL=13;
IF income67 > 10800 THEN coded67 = 1;
 ELSE coded67 = 0;
  PUT
            @1    pk                $char11.
            @12   coded67           2.
            ;
RUN;
```

In data step number 8, exemplified by the income year 1967, the actual income amount is replaced with the value ‘0’ or ‘1’ according to the given income line. This data step is repeated for each of the income years, but with an income line adjusted for each year according to the limits

established by 2 G as shown in Appendix 3. The output from this procedure is a text file for each income year 1967 – 2007, which later has to be linked to the panel population.

In the course of the following data steps the panel population is matched with the dependent variable. The principles are exemplified by the SAS code from data step 9. This data step creates the data set *data.match67*, which is the 1937 cohort with one new field added – *coded67* – which is the code for labour market participation in 1967. This principle is followed in the next step, and so on:

SAS data step 9. Matching the 1937 cohort with employment code for 1967.

```
DATA data.match67;
MERGE data.cohort1937 (IN=inn1 keep  = pk gender cohort )
       data.coded67  (IN=inn2 keep  = pk coded67);
BY pk;
IF inn1 THEN OUTPUT data.match67;
RUN;
```

The number of observations in the output data set *data.match67* is always kept constant. The if-test 'IF inn1 THEN OUTPUT ' determines that only observations from the data set *inn1* (the 1937 birth cohort in the example) are written to the output data set. The output data set *data.match67* combines data from both files, the donor data set and the recipient data set, which in SAS data step 10 is matched with the following generation of employment history (1968 in this instance).

SAS data step 10. Matching the 1937 cohort with employment code 1968.

```
DATA data.match68;
OPTIONS MISSING = 0;
MERGE data.match67 (IN=inn1 keep  = pk gender cohort coded67 )
       data.coded68  (IN=inn2 keep  = pk coded68);

BY pk;
IF inn1 THEN OUTPUT data.match68;
run;
```

For each match a new file is created with one more field added. This sequence is repeated until all 20 years of labour market history have been added. After the final matching step for the 1937 cohort, where the *data.match86* is created, the series of employment history codes are summed up and the output is written to a temporary text file - *coded37.txt* – (see data step number 11).

SAS data step 11. Summing up coded fields for employment history.

```
data _NULL_;
SET data.match86;

FILE 'c:\data\coded37.txt' LRECL = 16;

two_G =     sum(coded67, coded68, coded69, coded70, coded71, coded72,
            coded73, coded74, coded75, coded76, coded77,
            coded78, coded79, coded80, coded81, coded82, coded83, coded84,
            coded85, coded86);
PUT
            @1    pk          $char11.
            @12   gender      1.
            @13   cohort      2.
            @15   two_G       2.
            ;
RUN;
```

The last field *two_G* is simply the sum total of the values of each *codedYY* field where the values are either '0' or '1'. The value will range from 0 to 20 in the sum field *two_G*. In other words, this is the dependent variable and how it is constructed, in this case with the 1937 birth cohort as an example.

This procedure is repeated for each birth cohort, creating one text file for each: *coded38.txt*, *coded39.txt* etc. The structure is identical for all these files. The source code must be adapted according to ascending cohorts, and with ascending years of employment history. No selection according to birth cohorts is needed from the income tables. All income tables will have a higher number of observations than the cohort tables, but this is no problem for the matching procedures later on, as the number of observations in the panel is constant; see the discussion of matching data sets above.

After the final match, a new version of the data matrix can be constructed by appending the birth cohort files. All temporary files from previous data steps can be deleted. The dependent variable is now added to the panel population. The structure and contents of the first version of the data matrix can be read from Table 37, as the first four fields.

DATA STEPS FOR GEOGRAPHIC VARIABLES

Next, the identifier for residential municipality, and temporarily also residential county, must be added to each observation as a new field. This will serve several purposes: one purpose is to identify and deselect observations without a valid geographic connection, another purpose is to add the variable *municipality classification* to the panel population, and a third purpose is to calculate the variable *geographic mobility*. The county identifier and the municipality identifier are components in the numeric address code, see table 34.

The municipality identifier has to be transferred to each observation from the DSF income tables according to the same principles as applied in SAS data step 9 and 10 when the *labour market participation* code was added to the population. Again, the panel population has to be split by cohorts and matched with the municipality history tables. During these data steps the extracted municipality history tables are treated as the donor data sets, while the cohort data sets are the recipient ones. Otherwise this process is quite similar to the procedure used when the dependent variable was added to the data matrix.

Table 34. Structure of numeric address.

Position	Length	Data type	Field name
1	2	Character	County identifier
1	4	Character	Municipality identifier
5	5	Character	Street number
10	4	Character	House number in the street
14	4	Character	Code for entrance.
18	3	Character	Possible number for a section
21	1	Character	Code for type of floor
22	2	Character	Floor number
24	2	Character	Dwelling number within the floor

The municipality identifier is basically a character type variable which cannot be summed up or subjected to mathematical operators, but for the purpose of selecting municipality identifiers within the legal range (0101 – 2030), a temporary numeric variable is introduced. Technically this is done by defining the county identifier - the two initial digits of the municipality identifier - as numeric and applying the operator < which requires a numeric variable. Only counties with values < 21 are then selected from each of the DSF income tables. It is sufficient to perform this test at county level: a county identifier above 20 necessarily reflects a municipality identifier outside the valid range.

SAS data step 12. Read municipality identifier and select valid identifiers.

```
LIBNAME data 'c:/data/';
DATA data.muni67;
  INFILE 'c:\data\income67.dat' LRECL=46; *Read DSF income table;
  INPUT
            @1    pk              $char11.
            @17   county          2.
            @17   muni67          $char4.
            ;
if county < 21 then output data.muni67;
RUN;
```

With only valid municipality identifiers included, observations with missing identifiers for either the first 10 years or the last 10 years of the observation period now have to be identified. With the county identifier still defined as a numeric field, the series of county identifiers can be added up to obtain a temporary variable *sum*. Observations with *sum* = 0 within the 10- year period in question do not occur with a valid geographic connection. Hence these observations are excluded from the panel population, while all remaining observations have at least one valid county identifier and thus represent the net panel population. It is not considered necessary to display the SAS code for this test.

The remaining panel population now has at least one valid municipality identifier during the first half and the second half of the observation period, respectively. The municipality identifiers as character fields cannot be summed up like the income codes, but will appear temporarily as a series of 20 consecutive identifiers – one for each income year. From this series two identifiers must be selected. The municipality identifier selected will be number 10 and number 20 within the sequence of 20 identifiers for most of the observations. But for some observations residential municipality for the first half of the observation period may actually be as of year 8 or 9, or as of year 11 or 12, and residential municipality for the second half of the observation period may be represented by residence year number 19 or 18. This is a technical solution to make sure that the prevailing residential municipality for any observations is younger than the classification it is supposed to be an effect of. Otherwise a possible influence between residential characteristics and labour market participation might be reversed. There is also a risk that the prevailing municipality as of year 10 or as of year 20 may not be the most representative place of residence for all observations. It is not considered necessary to display source code for these selections.

Duplicates occur in the DSF income tables due to different types of labour income as outlined above. By law it is only possible to have one residential municipality for each year. In the case of duplicates, each of the duplicates will have the same municipality identifier within the same year. It

does not really matter which of the duplicates is selected for the final population. Running a duplicate check after the final match is sufficient. The duplicate check is basically the same as in SAS data step 3, and also splits the municipality history files in three: one file with observations without duplicates, one file with the first duplicate and one file with duplicates number two and above. The last file is deleted after inspection, while the first and second files are appended. Thus the number of observations in each cohort is identical to the number of observations in *data.matrix_1* above; in other words, a 1:1 relationship is established between the recipient and the donor files.

SAS data step 13. Match between gross and net population.

```
LIBNAME data 'c:/data/';
PROC SORT DATA=data.matrix_1; BY pk; run;
PROC SORT DATA=data.validresidence; BY pk; run;
DATA data.matrix_2;
MERGE data.matrix_1          (IN=inn1 keep  = pk gender cohort employ )
      data.validresidence    (IN=inn2 keep  = pk);
            BY pk;
            IF inn2 THEN OUTPUT data.matrix_2;
RUN;
```

A new version of the data matrix is produced in this step, but the only difference between version one and version two is the number of observations.

The municipality classification codes must now be added to the data matrix with the municipality identifier as the linkage key. Before matching, the population has to be split by groups of cohorts to be matched with different versions of the standards for municipality classification. The cohorts are split according to the principles from SAS data step 5, but this time in groups of birth cohorts. Finally, classification codes are matched with the cohort groups according to the principles shown in Table 35. This solution makes it possible to treat the classification codes as one variable for all cohorts – for each of the two periods, even if the periods are different for different cohorts.

Table 35. Cohort groups and edition of municipality classification for match.

Cohort groups	Municipality identifier by year 10	Municipality identifier by year 20
1937 – 1943	Matched with classification code 1974	Matched with classification code 1985
1944 – 1950	Matched with classification code 1985	Matched with classification code 1994
1951 – 1958	Matched with classification code 1994	Matched with classification code 2003

The precautions taken to avoid missing municipality identifiers as of 'year 10' and as of 'year 20' above will also avoid coincidence between point of measurement and point of classification. Again, this procedure is applied twice for the cohort groups displayed in Table 35 in order to add the classification code for the two 10-year periods respectively. The cohort groups are then appended, and finally the two new fields are added to the new version of the data matrix, which is now identical to the final version in Table 37, apart from the final field *number of children*.

Geographic mobility was defined as migration between municipalities. The variable *geographic mobility* was created by comparing municipality identifiers in pairs resulting in the code value '0' when identifiers were identical, or the code value '1' in all other cases, for a series of 20 years for each observation. The consecutive series of codes were then summed up in one field and added to the next version of the data matrix. This version of the data matrix now contains five new fields: r*esidence1* and r*esidence2* which equals the identifier for residential municipality as of 'year 10' and as of 'year 20' in the observation period, respectively. The two fields - *classification1* and *classification2* - contain the municipality classification code as of 'year 10' and as of 'year 20' and expresses *regional characteristics*. Finally, the *geographic mobility* field is the sum total of all possible municipality changes in the period.

DATA STEPS FOR ADDING FAMILY OBLIGATIONS – NUMBER OF CHILDREN

The only data source available for this variable appeared to be the Tax Register for Personal Taxpayers. The oldest generations of this system which have actually been transferred to the National Archives of Norway are stored in their native format, which means EBCDIC character set[154], and mainly compressed numeric fields. It was also common at the time to compile several record types, or tables, with different layouts in one file, with a code for each record type for identification. This format and structure were accepted by the National Archivist of Norway for long-term storage at the

[154] The EBCDIC character set was accepted for long-term storage by the National Archivist of Norway in the 1980s and 1990s.

time of acquisition. A study of the TRP file description will show in which record type a requested variable exists, and in this instance the relevant variables are found in record type '2' which is the main table. The example below shows SAS code written to import data to SAS format from the 1980 TRP.

The interesting field from the TRP is *number of children below 17 years of age* (the field *child17* in the code example). The given age limit for children may sometimes change according to legislation: in 1975 and 1980, the age limit was under 17 years, in 1985 under 16 years, and in 1990 under 19 years of age.

Technically, some auxiliary variables are needed for the procedure of allocating children to individual observations: *Family number* (*famno* in the example) and *personal code* (*pcode* in the example); see also the sub-chapter 'National identifiers and coded information'.

SAS data step 14. Reading 1980 generation of TRP.

```
LIBNAME data 'C:/data/';
DATA data.trp80;
/*Appending data files - this volume is split in 10 files due to its size*/
FILENAME INN                ('C:\data\FILE1_1980.dat'
                            'C:\data\FILE2_1980.DAT'
                            'C:\data\FILE3_1980.DAT'
                            'C:\data\FILE4_1980.DAT'
                            'C:\data\FILE5_1980.DAT'
                            'C:\data\FILE6_1980.DAT'
                            'C:\data\FILE7_1980.DAT'
                            'C:\data\FILE8_1980.DAT'
                            'C:\data\FILE9_1980.DAT'
                            'C:\data\FILE10_1980.DAT');
INFILE INN LRECL=500 RECFM=F ;
INPUT
      @1    rectype          $EBCDIC1.
      @6    pk               S370FPD6.
      @97   famno            S370FPD6.
      @103  pcode            $EBCDIC1.
      @229  child17          S370FZDL2.
      ;
OPTIONS MISSING = 0;
IF rectype = '2' THEN output data.trp80;
RUN;
```

The personal code reflects whether a person is a reference person, a spouse or a dependent (code values are '1', '2' and '3', respectively) in a nuclear family. For single parents there is no family member with a 'spouse' personal code. Unmarried adults have their own ID number as their family number, and the same personal code as a reference person. The code values are always based on the formal marital status originating from the Central Population Register, but in the TRP marital status is always as of 1 November of the year prior to the income year in question.

It is a complex exercise to operationalize *number of children* as a variable connected to the individual observation. The auxiliary variables *personal code* and *family number* will be utilized in order to copy *number of children* for married couples from only one parent to a variable for the spouse as well in each case. The INFILE in data step number 15 is a restructured version of the SAS file *data.trp80,* created in SAS data step 14.

As a first step, the reference persons and the spouses must be disjoined into separate data sets, by selection on personal code. Observations with a personal code that indicates they are dependents are deselected in this data step; due to their young age (18 or under) they also fall outside the age range of the panel population.

SAS data step 15. Data step for splitting by personal code.

```
DATA        data.pcode1
            data.pcode2
            data.pcode3;
INFILE 'C:\DATA\trp80.txt' lrecl= 27;
INPUT
      @1    pk                $CHAR11.
      @12   famno             $CHAR11.
      @23   pcode             $CHAR1.
      @24   child17           3.
      ;
if pcode = '1' then output data.pcode1;
      else if pcode = '2' then output data.pcode2;
      else output data.pcode3;

RUN;
```

The next step (data step number 16) matches couples by the family identifier (*famno* from data step 15) to make a temporary spouse data set with the husband's variables and his spouse's variables in the same row, with one half of the row for each. The *number of children* – for the couple, but attributed to the father – is then copied as a new field on the spouse part of the record, before the couple is split into individuals again. The RENAME command in the code is needed to avoid

confusion of variable names as all variables will now appear in the same record; the first half of each record represents the variables for the reference person, and the second half for the spouse. Table 36 shows the description for the temporary spouse table.

Table 36. Description for spouse table assembled from TRP.

Field number	Start position	Field length	Data type	Field name
1	1	11	Character	pk
2	12	11	Character	Family identifier
3	23	1	Character	Personal code = 1
4	24	2	Numeric	Number of children
5	26	11	Character	pk2
6	37	11	Character	Family identifier
7	48	1	Character	Personal code = 2
8	49	2	Numeric	Number of children from other relationship
9	51	2	Numeric	Number of children = field number 4 + field number 8

The *family number* (*famno*) is the linkage key for this match: for each married observation the *family number* is the ID number of the reference person and thus links the couple. This would also be the case for all observations with dependents' personal code, which is why they are deselected from this data set. For unmarried observations with a personal code as a 'reference person', the family number equals the primary key for that observation. For observations with personal code as a 'reference person', only field number 1 (*pk*) and field number 4 – *child17* - are needed in the output. For observations with a 'spouse' personal code, only field number 5 and field number 9 will be transferred.

To make this picture even more complex, in an existing marriage both parents may have the custody of children from earlier relationships. In such cases the possible child, or children, are attributed separately to the biological parent, either the mother or the father. That is why, where the mother is concerned, this field must be summed up with the field transferred from the reference person, but not vice versa. Where the father is concerned, it is not possible to distinguish between children from a present marriage and those from a possible earlier relationship. The *gender* field only serves auxiliary and control purposes.

Finally, *number of children* is now a field for both married parents, and observations and fields can be reorganized to fit into the individual pattern again. These observations are now appended to one file where one person constitutes one observation for each generation of the TRP, together with a file with single parents and their number of children. For each of the TRP generations available there is now a table extract with the two fields *primary key* and *number of children*.

SAS data step 16. Creating spouse data set from TRP.

```
DATA  data.spouses
      data.unmarried;
MERGE data.pcode1 (IN = inn1 keep  = pk famno pcode child17 gender)
      data.pcode2 (IN = inn2 keep  = pk famno pcode child17 gender
                                        RENAME = ( pk = pk2
                                              pcode  = pcode2
                                              child17 = child172
                                              gender = gender2));
BY famno;
IF inn2 THEN OUTPUT data.spouses;
ELSE OUTPUT data.unmarried;
RUN;
```

Now the variable *number of children* is added to the panel population through a final match. In order to compensate for a possible low coverage of this variable especially for the older birth cohorts, the panel population was eventually matched with all available generations of the TRP. The solution presented in data step 16 actually created separate fields with *number of children* for each TRP generation. Among these, the field with the highest number of children was finally selected. Anyway this issue was dealt with in detail in the sub-chapter 'Validity and coverage of the variables', and the final conclusion was to exclude this variable from the model as the operationalization was not successful.

After this long detour to compensate for data sources not existing, not presently available or not accessible for privacy reasons, a reminder about the intention of this troublesome exercise might be in place: the aim was to equip the panel population with the explanatory variable *number of children.* This is also the final variable to be added, and the data matrix is now complete both in terms of population, and in terms of variables.

DATA STEPS FOR FINALISING THE DATA MATRIX

So far the national ID number has served as the primary key. This is no longer needed once all information has been deduced from the ID number and all matches are completed. The data matrix may now be anonymised by replacing the ID number with a neutral counter. The following data step creates the final, anonymous version of the data matrix.

SAS data step 17. Create anonymous data set.

```
DATA _NULL_;
set data.matrix_4;
FILE 'C:\DATA\matrix_anonymous.dat' LRECL=31;
PUT
            @1    counter          Z11. /*the 'z' adds leading 0*/
            @12   gender           1.
            @13   yob              2.
            @15   employ           2.
            @17   muni10           $CHAR4.
            @21   muni20           $CHAR4.
            @25   class1           1.
            @26   class2           1.
            @27   migrate          2.
            @29   children         3.
      ;
counter + 1;
RUN;
```

Until now the data matrix has been sorted by *year of birth*, but as a further means of making data anonymous, it should be sorted on a field different from *year of birth* to randomise the counter. My choice is to sort by *municipality classification - year 10 (class1* in the data step*)*. The counter is then issued independently of *year of birth*. For technical purposes (i.e. to avoid moving positions), the variable length of 11 digits for the counter is kept. The first digits in this field will appear as leading zeros. Table 37 shows the structure and layout of the final data matrix with the number of children added and the national identifier replaced by a counter.

For the municipality-wise analyses, the data matrix has to be split into separate files by municipality identifier. Due to the restructuring of municipalities during the observation period, there is a risk that some observations might be registered as residents of expired or discontinued municipalities. Data step number 18 is an example of redirecting observations from discontinued municipalities to municipalities existing as of 1994. The output data sets from this data step are identified by the name of the municipality.

Table 37. Description – final data matrix.

Start position	Field length	Data type	Field name
1	11	Character	Counter
12	1	Numeric	Gender, 0 = male, 1 = female
13	2	Numeric	Birth cohort. 1 = 1937, 2 = 1938 etc.
15	2	Numeric	Employment history. Value range from 0 to 20.
17	4	Character	Code for residential municipality by year 10 *('residence 1')*. Value range from '0101' to '2030'
21	4	Character	Code for residential municipality by year 20 *('residence 2')*. Value range from '0101' to '2030'
25	1	Numeric	Municipality classification code by year 10 *('classification 1')*. Value range from '1' to '7'
26	1	Numeric	Municipality classification code by year 20*('classification 2')*. Value range from '1' to '7'
27	2	Numeric	Geographic mobility. Actual value range from 0 to 16.
29	3	Numeric	Number of children. Actual value range from 0 to12.

SAS data step 18. Redirect population from expired municipalities.

```
LIBNAME data 'c:/data/';
data data.horten data.holmestrand data.tonsberg data.sandefjord data.larvik
      data.svelvik data.sande data.hof data.vaale data.ramnes data.andebu
      data.stokke data.notteroy data.tjome data.lardal;
set data.matrix_anonymous;
if muni20 = '0701' or muni20 = '0703' or muni20 = '0717' then output
data.horten;
if muni20 = '0702' then output data.holmestrand;
if muni20 = '0704' or muni20 = '0705' or muni20 = '0721' then output
data.tonsberg;
if muni20 = '0706' then output data.sandefjord;
if muni20 = '0708' or muni20 = '0707' or muni20 = '0709' or muni20 = '0725'
or muni20 = '0726' or muni20 = '0727' then output data.larvik;
if muni20 = '0711' then output data.svelvik;
if muni20 = '0713' then output data.sande;
if muni20 = '0714' then output data.hof;
if muni20 = '0716' then output data.vaale;
if muni20 = '0718' then output data.ramnes;
if muni20 = '0719' then output data.andebu;
if muni20 = '0720' then output data.stokke;
if muni20 = '0722' then output data.notteroy;
if muni20 = '0723' then output data.tjome;
if muni20 = '0728' then output data.lardal;
RUN;
```

EXAMPLE DATA STEPS FOR THE HISTORICAL CRITICISM PROCESS

In the sub-chapter 'Authenticity and reliability', the concept 'computer-based source criticism' was introduced. Technically these controls must be run before the data matrix is anonymised. In the example below, the 1990 versions of the TRP and the DSF have been selected. Before the match between the DSF and the TRP, both data sets were subjected to a duplicate check and a validity check of ID numbers, and D-numbers were deselected from both data sets. Moreover, only the relevant birth cohorts are included in these matches. The first match is a population consistency test, while the second match is an amount consistency test.

SAS data step 19. Match between DSF income table 1990 and TRP 1990 with SAS log.

```
LIBNAME data 'c:/data/';
PROC SORT data=data.dsf_90; BY pk; run;
PROC SORT data=data.trp_90; BY pk; run;
DATA data.match
      data.only_dsf
      data.only_trp;
MERGE data.dsf_90   (IN=inn1 keep  = pk dsfinc)
       data.trp_90    (IN=inn2 keep  = pk trpinc);
BY pk;
IF inn1 and inn2 THEN OUTPUT data.match;
IF inn1 THEN OUTPUT data.only_dsf;
IF inn2 THEN OUTPUT data.only_trp;
OPTIONS MISSING = 0;
RUN;

SAS LOG:
NOTE: There were 1 119 949 observations read from the data set DATA.DSF_90.
NOTE: There were 1 115 635 observations read from the data set DATA.TRP_90.
NOTE: The data set DATA.MATCH has 1 112 275 observations and 3 variables.
NOTE: The data set DATA.ONLY_DSF has 1 119 949 observations and 3 variables.
NOTE: The data set DATA.ONLY_TRP has 1 115 635 observations and 3 variables.
```

From the DSF based file with a total of 1 119 949 observations there were only 7 674 mismatching observations; in other words, there was a 99.3 per cent match with the TRP. From the TRP based file with a total of 1 115 635 observations there were only 3 360 mismatching observations, or a 99.7 per cent match with the DSF. The obvious conclusion is that this is a very high degree of population consistency. The few mismatching observations may be explained partly by different units (taxpayer versus person), and partly by lack of synchronisation between the archival extracts of the data sets. In any event, this is definitely a very good match, and hence there is very high population consistency between the two systems.

Among the matching observations in the data set *data.match*, it is interesting to identify any discrepancies in pensionable income amounts. Ideally, the amounts should be identical for each matching observation in both data sources. The following test is based on the file *data.match* from data step 19 with pensionable income from the TRP and from the DSF respectively, in separate fields: *dsfinc* and *trpinc*.

SAS data step 20. Calculate difference between pensionable income in DSF and in TRP.

```
LIBNAME data 'c:/data/';
data data.difference
      data.nodifference;
SET data.match;
FILE 'c:\data\difference.dat' LRECL = 20;
diff = (trpinc - dsfinc);
PUT
      @1  pk      $char11.
      @12 diff    9.
      ;
IF diff NE 0 THEN OUTPUT data.difference;
 ELSE OUTPUT data.nodifference;
run;
```

```
NOTE: There were 1 112 275 observations read from the data set DATA.match.
NOTE: The data set DATA.DIFFERENCE has 6505 observations and 4 variables.
NOTE: The data set DATA.NODIFFERENCE has 1105770 observations and 4 variables.
```

The log from data step 20 displays that 6 505 observations have different amounts. A closer inspection of these differences clearly reveals the tendency that, if different, the amount is generally higher in the TRP than in the DSF. For 95 per cent of the 6 505 observations the highest amount comes from the TRP. Mainly the differences are very small, but for 10 observations the difference is over NOK 1 million with the highest amount in the TRP. Such discrepancies can generally be explained by erroneous data entry or data transfer, which have been corrected in the DSF, but not in the TRP.

In this example, the 1990 generation of the two data sets were matched, but during the process of finalising the data matrix, several generations of the TRP and the DSF were matched. A lot of internal matches between the panel population with origin from DSF population table and the annual DSF income tables were also required. Only observations with a labour income amount or a required code are included in the annual DSF income tables. As a result, a number of mismatches between the DSF population table and the DSF income tables have to be expected, but that does not represent any problem in the context of constructing the dependent variable. As there was a general increase in labour market participation over the years, and hence an increasing number of observations in the DSF income tables, a gradually better match had to be expected. In fact, a better match for the earlier generations would have been suspect. The results from all matches between the DSF demographic table and the income tables for 1967 to 2007 confirm this picture: the percentage of matches ranges

from 58 in the first year to 99 in the final years. The same trends could be observed in matches between the DSF and the TRP, due to a reduction of joint tax assessment for married couples.

APPENDIX 2: EXTRACT OF TECHNICAL METADATA FOR THE TAX REGISTER FOR PERSONAL TAXPAYERS

Below is an extract of technical metadata for the main table of the Tax Register for Personal Taxpayers for the income year 1995. The metadata extract is presented according to the standardised format required by the ADDML standard developed by the National Archivist of Norway. This table contains a total of 188 fields. The field 'Production_date' is the initial field in the table extract in Figure 2.

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE addml SYSTEM 'addml.dtd'>
<addml version='7.3'>
 <reference>
  <archives ar_id='Tax register for personal taxpayers 1995'>
  </archives>
  <system sy_id='Main_table'>
   <sy_name>Maintable_ordinary</sy_name>
   <startdate>1995</startdate>
   <enddate>2005</enddate>
  </system>
 </reference>
 <structure>
  <dataset ds_id='TRP'>
   <ds_descr>TRP_for_personal_taxpayers – main table</ds_descr>
   <charset>ISO-8859-1</charset>
   <format>FIXED</format>
   <recsep>CRLF</recsep>
   <nu_files>1</nu_files>
   <file name='TRP1995_main_table' path='F:\ARKIV1A\INNENBYS.dat'>
    <reclength>1035</reclength>
    <nu_rectypes>1</nu_rectypes>
    <rectype name='TRP_1995'>
     <primkey>MunicipalityID+NationalID</primkey>
          ( … )
     <fieldtype name='Production_date'>
      <ft_descr>Date for last change of correction - format YYMMDD</ft_descr>
      <startpos>29</startpos>
      <endpos>37</endpos>
      <ft_fixlength>9</ft_fixlength>
      <datatype>STRING</datatype>
     </fieldtype>
          ( … )
```

Appendix 3:'Basic Amount' (G) - and average pensionable income 1967 – 2007 [155]

Year	Basic amount			Average pensionable income[156]		
	G[157]	G*2	G*4	Males and females	Males	Females
2007	65505	131010	262020	304200	358 400	243 300
2006	62161	124322	248644	285200	335 800	228 700
2005	60059	120118	240236	273300	323 100	218 000
2004	58 139	116 278	232 556	261600	308 400	209 600
2003	55 964	111 928	223 856	253500	299 700	202 200
2002	53 233	106 466	212 932	245200	291 800	193 100
2001	50 603	101 206	202 412	232700	278 700	181 000
2000	48 377	96 754	193 508	220300	263 900	171 100
1999	46 423	92 846	185 692	216200	259 700	166 900
1998	44 413	88 826	177 652	203800	245 500	156 300
1997	42 000	84 000	168 000	190800	229 800	146 300
1996	40 410	80 820	161 640	180400	216 500	139 100
1995	38 847	77 694	155 388	170200	204 700	130 600
1994	37 820	75 640	151 280	160100	192 600	122 700
1993	37 033	74 066	148 132	155500	187 400	118 800
1992	36 167	72 334	144 668	149600	180 300	114 100
1991	35 033	70 066	140 132	143462	173 000	109 000
1990	33 575	67 150	134 300	137253	166 600	102 700
1989	32 275	64 550	129 100	129007	157 300	95 400
1988	30 850	61 700	123 400	125470	154 500	90 800
1987	29 267	58 534	117 068	118434	146 600	84 500
1986	27 433	54 866	109 732	109060	135 600	76 800
1985	25 333	50 666	101 332	98686	123 000	68 500
1984	23 667	47 334	94 668	90722	113 200	62 400
1983	22 333	44 666	89 332	87004	108 300	59 500
1982	20 667	41 334	82 668	80758	100 900	54 300
1981	18 658	37 316	74 632	72897	91 400	48 300
1980	16 633	33 266	66 532	65005	81 700	42 300
1979	15 200	30 400	60 800	58736	73 900	37 700
1978	14 550	29 100	58 200	56879	72 000	35 900
1977	13 383	26 766	53 532	51915	65 700	32 200
1976	12 000	24 000	48 000	46836	58 700	29 100
1975	10 800	21 600	43 200	40868	51 200	24 900
1974	9 533	19 066	38 132	35338	44 200	21 200
1973	8 500	17 000	34 000	32558	39 100	20 600

[155] All amounts in NOK – current value.

[156] Statistics Norway: Tax Statistics 1967 – 2007

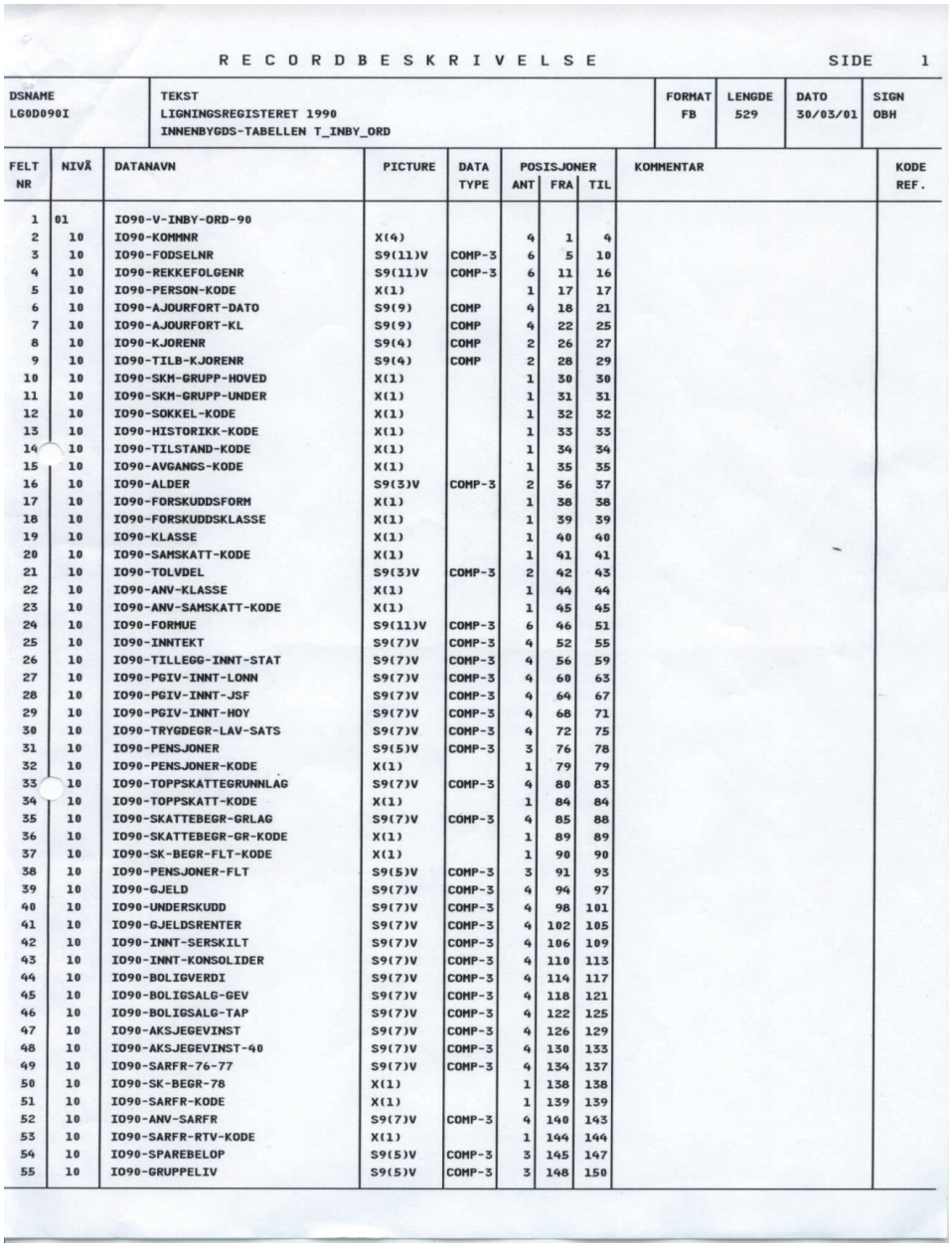
[157] www.nav.no

1974	9 533	19 066	38 132	35338	44 200	21 200
1973	8 500	17 000	34 000	32558	39 100	20 600
1972	7 900	15 800	31 600	29969	35 900	18 900
1971	7 400	14 800	29 600	28126	33 500	17 600
1970	6 800	13 600	27 200	24790	29 100	15 900
1969	6 400	12 800	25 600	22927	26 600	15 000
1968	5 900	11 800	23 600	21344	24 700	14 000
1967	5 400	10 800	21 600	21315	24 000	14 700

APPENDIX 4: NORWAY – COUNTIES AND MUNICIPALITIES[158].

[158] Norwegian Mapping Authority/Statens Kartverk 712064-1662 by Nordeca AS

APPENDIX 5: TECHNICAL METADATA FOR 1990 TAX REGISTER FOR PERSONAL TAXPAYERS, PAGE 1.



R E C O R D B E S K R I V E L S E SIDE 1

DSNAME	TEKST	FORMAT	LENGDE	DATO	SIGN
LG0D090I	LIGNINGSREGISTERET 1990 INNENBYGDS-TABELLEN T_INBY_ORD	FB	529	30/03/01	OBH

FELT NR	NIVÅ	DATANAVN	PICTURE	DATA TYPE	POSISJONER ANT	FRA	TIL	KOMMENTAR	KODE REF.
1	01	IO90-V-INBY-ORD-90							
2	10	IO90-KOMMNR	X(4)		4	1	4		
3	10	IO90-FODSELNR	S9(11)V	COMP-3	6	5	10		
4	10	IO90-REKKEFOLGENR	S9(11)V	COMP-3	6	11	16		
5	10	IO90-PERSON-KODE	X(1)		1	17	17		
6	10	IO90-AJOURFORT-DATO	S9(9)	COMP	4	18	21		
7	10	IO90-AJOURFORT-KL	S9(9)	COMP	4	22	25		
8	10	IO90-KJORENR	S9(4)	COMP	2	26	27		
9	10	IO90-TILB-KJORENR	S9(4)	COMP	2	28	29		
10	10	IO90-SKM-GRUPP-HOVED	X(1)		1	30	30		
11	10	IO90-SKM-GRUPP-UNDER	X(1)		1	31	31		
12	10	IO90-SOKKEL-KODE	X(1)		1	32	32		
13	10	IO90-HISTORIKK-KODE	X(1)		1	33	33		
14	10	IO90-TILSTAND-KODE	X(1)		1	34	34		
15	10	IO90-AVGANGS-KODE	X(1)		1	35	35		
16	10	IO90-ALDER	S9(3)V	COMP-3	2	36	37		
17	10	IO90-FORSKUDDSFORM	X(1)		1	38	38		
18	10	IO90-FORSKUDDSKLASSE	X(1)		1	39	39		
19	10	IO90-KLASSE	X(1)		1	40	40		
20	10	IO90-SAMSKATT-KODE	X(1)		1	41	41		
21	10	IO90-TOLVDEL	S9(3)V	COMP-3	2	42	43		
22	10	IO90-ANV-KLASSE	X(1)		1	44	44		
23	10	IO90-ANV-SAMSKATT-KODE	X(1)		1	45	45		
24	10	IO90-FORMUE	S9(11)V	COMP-3	6	46	51		
25	10	IO90-INNTEKT	S9(7)V	COMP-3	4	52	55		
26	10	IO90-TILLEGG-INNT-STAT	S9(7)V	COMP-3	4	56	59		
27	10	IO90-PGIV-INNT-LONN	S9(7)V	COMP-3	4	60	63		
28	10	IO90-PGIV-INNT-JSF	S9(7)V	COMP-3	4	64	67		
29	10	IO90-PGIV-INNT-HOY	S9(7)V	COMP-3	4	68	71		
30	10	IO90-TRYGDEGR-LAV-SATS	S9(7)V	COMP-3	4	72	75		
31	10	IO90-PENSJONER	S9(5)V	COMP-3	3	76	78		
32	10	IO90-PENSJONER-KODE	X(1)		1	79	79		
33	10	IO90-TOPPSKATTEGRUNNLAG	S9(7)V	COMP-3	4	80	83		
34	10	IO90-TOPPSKATT-KODE	X(1)		1	84	84		
35	10	IO90-SKATTEBEGR-GRLAG	S9(7)V	COMP-3	4	85	88		
36	10	IO90-SKATTEBEGR-GR-KODE	X(1)		1	89	89		
37	10	IO90-SK-BEGR-FLT-KODE	X(1)		1	90	90		
38	10	IO90-PENSJONER-FLT	S9(5)V	COMP-3	3	91	93		
39	10	IO90-GJELD	S9(7)V	COMP-3	4	94	97		
40	10	IO90-UNDERSKUDD	S9(7)V	COMP-3	4	98	101		
41	10	IO90-GJELDSRENTER	S9(7)V	COMP-3	4	102	105		
42	10	IO90-INNT-SERSKILT	S9(7)V	COMP-3	4	106	109		
43	10	IO90-INNT-KONSOLIDER	S9(7)V	COMP-3	4	110	113		
44	10	IO90-BOLIGVERDI	S9(7)V	COMP-3	4	114	117		
45	10	IO90-BOLIGSALG-GEV	S9(7)V	COMP-3	4	118	121		
46	10	IO90-BOLIGSALG-TAP	S9(7)V	COMP-3	4	122	125		
47	10	IO90-AKSJEGEVINST	S9(7)V	COMP-3	4	126	129		
48	10	IO90-AKSJEGEVINST-40	S9(7)V	COMP-3	4	130	133		
49	10	IO90-SARFR-76-77	S9(7)V	COMP-3	4	134	137		
50	10	IO90-SK-BEGR-78	X(1)		1	138	138		
51	10	IO90-SARFR-KODE	X(1)		1	139	139		
52	10	IO90-ANV-SARFR	S9(7)V	COMP-3	4	140	143		
53	10	IO90-SARFR-RTV-KODE	X(1)		1	144	144		
54	10	IO90-SPAREBELOP	S9(5)V	COMP-3	3	145	147		
55	10	IO90-GRUPPELIV	S9(5)V	COMP-3	3	148	150		

LIST OF ACRONYMS

ADDML	*Archival Data Description Markup Language*
AIP	*Archival Information Package*
AKU	Arbeids Kraft Undersøkelse – *Labour Force Sample Survey*
4GL	*Fourth-generation programming language*
ASCII	*American Standard Code for Information Interchange*
CD-R	*Compact Disk – Recordable*
CPR	*Central Population Register*
CPU	*Central Processing Unit*
DBMS	*Data Base Management System*
DIP	*Dissemination Information Package*
DSF	Det Sentrale Folketrygdsystem – *Central System for National Social Security*
EBCDIC	*Extended Binary-Coded Decimal Interchange Code*
G	Grunnbeløp - *Basic amount*
GB	*Giga Byte*
GEI	*Gender Equality Index*
ILO	*International Labour Organisation*
ISO	*International Standard Organisation*
LFS	*Labour Force Sample Survey, (see AKU*)
Lrecl	*Logical record length*
MB	*Mega Byte*
MoReq	*Model Requirements – EU standard*
NAN	*National Archives of Norway*
NOARK	Norsk Arkivstandard
NOK	*Norwegian Kroner*
NSD	*The Norwegian Social Science Data Services*
OAIS	*Reference Model for an Open Archival Information System*
OCR	*Optical Character Recognition*
OECD	*Organisation for Economic Co-operation and Development*
PK	*Primary Key*
SIP	*Submission Information Package*
SQL	*Structured Query Language*
XML	*Extended Markup Language*

SOURCES

DIGITAL SOURCES

Records creator	System origin	Period
National Insurance Administration (RTV)	Central System for National Social Security	1967 – 2007 Population table
	Employer/Employee Register	1967 - 2007 Income tables
		1978 - 2005
Norwegian Directorate of Taxes (SKD)	Tax Register for Personal Taxpayers	1970
		1975
		1980
		1985
		1990
		1995
Statistics Norway	Standard Classification of Municipalities[159]	1985
		1994
		2003
Statistics Norway	PxMap2[160]	2006
Statistics Norway	StatBank Norway at ww.ssb.no: Statistics on - – Population – Labour market – Income – Education – Statistical Yearbook – Labour Force Survey/AKU – Tax statistics	

[159] Table parts only.

[160] Source: Norwegian Mapping Authority/Statens Kartverk.

PAPER-BASED SOURCES

Records creator	Records/system origin	Period
Statistics Norway	Classification of the municipalities of Norway	1974
Statistics Norway	Tax statistics	1967 – 2007
Statistics Norway	Historical Statistics	Chapter 3, 11,
Statistics Norway	Statistical Yearbook	1967 - 2009

LEGISLATION AND REGULATIONS

Arkivloven - Lov om arkiv. LOV-1992-12-04-126 *(Archives Act)*

Enhetsregisterloven – Lov om Enhetsregisteret. LOV-1994-06-03-15 *(Act relating to the Central Coordinating Register for Legal Entities)*

Folketrygdloven - Lov om folketrygd. LOV-1997-02-28-19 *(National Insurance Act)*

Folkeregisterloven - Lov om folkeregistrering. LOV-1970-01-16-1 *(National Population Register Act – in Norwegian only)*

Forvaltningsloven - Lov om behandlingsmåten i forvaltningssaker. LOV-1967-02-10 *(Public Administration Act – in Norwegian only)*

Offentleglova - Lov om rett til innsyn i dokument i offentleg verksemd. LOV-2006-05-19-16 *(Freedom of Information Act)*

Personopplysningsloven - Lov om behandling av personopplysninger. LOV-2000-04-14-31 *(Personal Data Register Act)*

Statistikkloven - Lov om offisiell statistikk og Statistisk Sentralbyrå. LOV-1989-06-16-54 *(Act relating to Official Statistics and Statistics Norway)*

Skatteloven - Lov om skatt av formue og inntekt. LOV-1999-03-26-14 *(Norwegian Taxation Act – in Norwegian only)*

Forskrift om offentlege arkiv. FOR-1998-12-11-1193.

Forskrift om folkeregistrering. FOR-2007-11-09-1268 *(Regulations relating to the National Population Register – in Norwegian only)*

Forskrift om utfyllende tekniske og arkivfaglige bestemmelser om behandling av offentlige arkiver. FOR-1999-12-01-1566 *(Regulations concerning further technical and archival provisions regarding the handling of public records – only Ch. VIII in English)*

BIBLIOGRAPHY

Acker, Joan, (1992): *The Future of Women and Work: Ending the Twentieth Century.* Pacific Sociological Association

Allison, Paul D. (1995): *Survival Analysis Using the SAS System. A Practical Guide.* SAS Institute Inc., NC, USA.

Amblie, Svein (2001): *Elektronisk dokumentbehandling.* Oslo: Kommuneforlaget AS

Aschehoug (1997): *Store norske leksikon.* Oslo: Kunnskapsforlaget

Aschehoug og Gyldendal (1982): *Store norske leksikon.* Oslo: Kunnskapsforlaget

Balleer, Almut, Ramón Gómez-Salvador and Jarkko Turunen (2009): *Labour Force Participation in the Euro Area. A Cohort Based Analysis.* Germany: European Central Bank - Working Paper Series, No 1049/May 2009

Benum, Edgeir (1998): *Hverdagsrevolusjonen, fritiden og kulturen. Aschehougs Norgeshistorie, bd. 12.* Oslo: Aschehoug & Co

Benjamin, Daniel K. and Levis A. Kochin: Searching for an Explanation for Unemployment in Interwar Britain. *Journal of Political Economy*, 87, 1979, pp. 441 – 478

Boje, Thomas P. (2007): Welfare and work. The gendered organisation of work and care in different European countries. *United Kongdom: European Review, Vol. 15, No. 3, p. 373-395.*

Bull, Hans Henrik (2006): *Marriage decisions in a peasant society.* Oslo: Faculty of Humanities, University of Oslo Unipub

Norwegian Mapping Authority (2010): *Matrikkel/Adresse/Bolignummer*. Virtual document at www.statkart.no.

Clausen, Hans Peter (1968): *Hva er historie?* Oslo: Gyldendal Norsk Forlag

Coll, Line M og Claude A. Lenth (2000): *Personopplysningsloven - en håndbok.* Oslo: Kommuneforlaget

Dahl, Ottar (1970): *Norsk historieforskning i det 19. og 20. århundre.* Oslo: Universitetsforlaget

Dahl, Ottar (2002): *Grunntrekk i historieforskningens metodelære.* Oslo: Universitetsforlaget

Danielsen, Rolf (1995) *Norway: A History from the Vikings to Our Own Times, Part IV*. Oslo: Universitetsforlaget

Danielsen, Rolf (1991): *Grunntrekk i norsk historie fra vikingtid til våre dager.* Oslo: Universitetsforlaget

Danmarks Statistik (1982): *Personstatistik på registergrunnlag.* København: Danmarks Statistik

Davies, Stephen (2003): *Empiricism and History.* Basingstoke: Palgrave

Dyrvik, Ståle (1983): *Historisk demografi: ei innføring i metodane.* Bergen: Universitetsforlaget

Dørum, Anne-Mette (1999): *Prosjekt utviklings- og handlingsplan for arkivtjenesten i trygdeetaten.* Oslo: Forvaltningsinfo AS

Daasvatn, Liv (1996): *Håndbok i SAS.* Oslo: Statistisk sentralbyrå. Interne dokumenter 96/16

Daasvatn, Liv og Kristian Lønø (1995): *SAS som tabellverktøy.* Oslo: Statistisk sentralbyrå. Interne dokumenter 95/1

Eikemo, Terje Andreas og Tommy Høyvarde Clausen (2007): *Kvantitativ analyse med SPSS: en praktisk innføring i kvantitative analyseteknikker.* Trondheim: Tapir akademisk forlag

Ellingsen, Dag (2010): Sørlandet fortsatt på etterskudd. Oslo: Statistisk sentralbyrå. *Samfunnsspeilet* 1/2010

Ellingsen, Dag og Ulla-Britt Lilleaas (2014): *'Noen vil ha det slik'. Tradisjonelle kjønnsroller og svake levekår på Sørlandet.* Kristiansand: Portal forlag, p 57.

Engen, Ingvar (2002): *Sletting eller bevaring av personsensitiv arkivinformasjon. Utviklingstrekk de siste 20 år. I Med Clio til Kringsjå.* Oslo: Novus forlag

Falk, Martin and Thomas Leoni (Undated): Regional female labour force participation: empirical evidence for Austria. *Vienna: Austrian Institute of Economic Research WIFO*.

Feinstein, Charles and Mark Thomas (2002): *Making History Count. A Primer in Quantitative Methods for Historians.* Cambridge: Cambridge University Press

Fonnes, Ivar (2000): *Arkivhåndboken for offentlig forvaltning.* Oslo: Kommuneforlaget

Frees, Edward W. (2004): *Longitudinal and Panel Data. Analysis and Application in the Social Sciences.* Cambridge: Cambridge University Press

Fredriksen, Dennis Finn (1993): *Strukturelle endringer på arbeidsmarkedet 1950 – 1990*. Oslo: Statistisk sentralbyrå. Sosialt utsyn 1993, kap. 5.3.

Furre, Berge (1992): *Norsk historie 1905 – 1990.* Oslo: Det Norske Samlaget

Furre, Berge (2000): *Norsk historie 1914 – 2000.* Oslo: Det Norske Samlaget

Furuvold, Per (2001): *Systemområde Infotrygd.* Oslo: Rikstrygdeverket

Gjelseth, Martha Cecilie (2000): *Relasjonsdatabaser som verktøy i en historisk-demografisk studie.* Oslo

Godfrey, Donald G. (ed.) (2006): *Methods of Historical Analysis in Electronic Media.* Mahwah, N.J.: Lawrence Erlbaum Associates

Grønlie, Tore (2004): *Grunntrekk i norsk historie: tiden etter 1945.* Oslo: Pensumtjeneste

Gullikstad, Berit (2002): *Kvinnelig livsoppgave - mannlig lønnsarbeid? Kjønn og arbeid under velferdsstatens oppbygging ca. 1945 – 1970.* Trondheim: Historisk institutt, Senter for kvinne- og kjønnsforskning, NTNU

Halvorsen, Bjørn (1992): *KIRUT, om databasen - presentasjon av data for 1989 – 90.* Oslo: Rikstrygdeverket

Haskins, Loren and Jeffrey Kirk (1990): *Understanding Quantitative History.* Cambridge, Mass.: MIT Press

Hirsch, Agnes Aaby (2010): *Lokal likestilling - målt på en ny måte.* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 1/2010

Hudson, Pat (2000): *History by Numbers: An Introduction to Quantitative Techniques.* London: Arnold

Hufton, Olwen (1997): Women, gender and the Fin de siècle. *In: Michael Bentley (ed.) Companion to Historiography. London and New York*

Hunnes, Arngrim, Jarle Møen og Kjell G. Salvanes (2008): Wage Structure and Labor Mobility in Norway, 1980 – 2007. Bergen: Samfunns- og næringslivsforskning. Working Paper No 19/08

Håland, Inger og Gunnlaug Daugstad (2003): *Den kjønnsdelte arbeidsmarknaden.* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 6/2003

Jacobsen, Siw Ellen (2010): *Samfunnsvitere må bruke tall*. Bladet Forskning 2/2010

Jarausch, Konrad H. and Kenneth A. Hardy (1991): *Quantitative Methods for Historians: A Guide to Research, Data and Statistics.* Chapel Hill, N.C: University of North Carolina

Jaumotte, Florence (2004): *Labour Force Participation of Women: Empirical Evidence on The Role of Policy and Other Determinants in OECD Countries.* OECD Economic Studies, Vol. 2003/2

Jensen, Ragnhild Steen (2004): *Sted, kjønn og politikk: Kvinners vei inn i lønnsarbeid.* Oslo: Unipax Institutt for samfunnsforskning

Jule, Allyson(2014): Gender Theory. *Trinity Western University, Canada*

Juvkam, Dag (1999): *Historisk oversikt over endringer i kommune- og fylkesinndelingen.* Oslo: Statistisk sentralbyrå. Rapporter 99/13

Kaldal, Ingar (2002): *Frå sosialhistorie til nyare kulturhistorie*. Oslo: Samlaget

Kiberg, Dag og Birger Skilbrei (1998): *Trygdedata. Håndbok for brukere av informasjon om trygdesystemet.* Bergen: Norsk Samfunnsvitenskapelig Datatjeneste

Kjeldstadli, Knut (1992): *Fortida er ikke hva den en gang var.* Oslo: Universitetsforlaget

Kjelland, Arnfinn (2009): *Norsk lokalhistorie og 'nyare' mikrohistorie.* Heimen, 46/2009

Kjelstad, Randi (1998): *Kvinner og menn 1998: Hvor likestilte er vi?* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 5/1998

Kvitastein, Olav A. og Kjell G. Salvanes (2003): *Effekten av arbeidsmarknadstiltak på regional arbeidsløyse. Bruk av norske registerdata for perioden 1993 – 2000.* Bergen: Samfunns- og næringslivsforskning. SNF-rapport nr. 20/2003

Lange, Vilhelm, Dag Mangset og Øyvind Ødegaard (2001): *Privatarkiver.* Oslo: Kommuneforlaget

Langholm, Sivert (1974): *Historie på individnivå.* Historisk Tidsskrift, 3/1974

Lappegård, Trude og Turid Noack (2009): *Familie og jobb i ulike kvinnegenerasjoner.* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 1/2009

Liu, An and Inge Noback (2010): Determinants of regional female labour market participation in the Netherlands. A spatial structural equation modelling approach. *Springerlink.com. University of Groningen. The Netherlands*

Lind, Gunner (1994): *Data Structures for Computer Prosopography.* Odense: Odense University Press and the authors

Lønø, Kristian (2000): *Håndbok i SAS. Del 2: Oppslag.* Oslo: Statistisk sentralbyrå. Håndbøker

Magnussen, May-Linda, Trond Stalsberg Mydland og Gro Kvåle(2005): *Arbeid ute og hjemme: Sørlandske mødres valg og vurderinger. Rapport fra prosjektet Likestilling og arbeidsliv på Agder.* Kristiansand: FoU-rapport nr. 5/2005.

Mathisen, Helge (1998): *Systemer og systemsammenheng.* Oslo: Rikstrygdeverket. Prosjekt nr. 2014. År 2000-prosjektet.

Midtbø, Tor (2007): *Regresjonsanalyse for samfunnsvitere. Med eksempler i SPSS.* Oslo: Universitetsforlaget

MoReq2 (2008): *Modular Requirements for the Management of Electronic Records.* MoReq Specifications. Luxembourg: Office for Official Publications of the European Communities

Mykland, Liv og Kjell-Olav Masdalen (1987): *Administrasjonshistorie og arkivkunnskap.* Oslo: Universitetsforlaget

Ona, Lars Erik Hjorthaug (2005): *Deltid og vekst: deltidsarbeid blant kvinner 1945 – 60.* Oslo: L.E.H. Ona

Osmundsen, Terje (1991): *Ny tid: Norge - industrinasjonen som forsvant?* Oslo: Cappelen

Pampel, Fred (2011): Cohort Changes in the Socio-demographic Determinants of Gender Egalitarianism. *Social Forces 89(3) 961–982, March 2011*

Pettersen, Silje Vatne og Randi Kjelstad (2008): *Er det plass til mødre i det nye arbeidslivet?* Oslo: Statistisk sentralbyrå. Samfunnsspeilet 3/2008

Pryser, Tore (1974): *Thranittene i Ullensaker: en sosialhistorisk analyse.* Oslo

Riksarkivaren (1992): *Håndbok for Riksarkivet*. Oslo: Ad Notam Gyldendal

Riksarkivaren (1999): *Noark-4. Norsk arkivsystem.* Oslo: Kommuneforlaget

Riksarkivaren (2002): *Rapport fra Bevaringsutvalget.* Oslo: Riksarkivaren. Rapporter og retningslinjer nr. 10

Riksarkivaren (2002): *Bevarings- og kassasjonsbestemmelser for skatteetatens elektroniske arkivmateriale. 10.09.2002.* Oslo. Riksarkivaren

Riksarkivaren (2009): *Noark 5, versjon 2.0.* www.arkivverket.no

Rikstrygdeverket (2001): *Infotrygd - Registreringsinstruks. Versjon 2.0.* Oslo: Rikstrygdeverket

Sande, Per (1978): *Småbrukerne i Ullensaker 1835-1865: en sosialhistorisk analyse.* Oslo.

Schellenberg, Theodore (1956): *Modern Archives: Principles and Techniques.* Chicago: The University of Chicago Press

Scott, Joan Wallach (1991): 'Women's History'. *In: Peter Burke (ed.) New perspectives on historical writing. Polity Press*

Scott, Joan Wallach (1986): 'Gender: A Useful Category of Historical Analysis'. *American Historical Review 91, No. 5, pp. 1053–75*

Sejersted, Francis (2002): *Er det mulig å styre utviklingen? Teknologi og samfunn.* Oslo: Pax Forlag A/S

Sejersted, Francis (2005): *Sosialdemokratiets tidsalder.* Oslo: Pax Forlag A/S

Sirevåg, Trond (2002): *De elektroniske arkivene - Hva har vært Arkivverkets strategi? Resultater og utfordringer fremover.* Oslo: Riksarkivarens skriftserie 13

Sirevåg, Trond (2005): *Multiproveniens i arkivsystemer og andre utviklingstrekk - Arkivteoretiske og praktiske implikasjoner.* Oslo: Riksarkivarens Rapporter og retningslinjer nr. 18

Sjørbotten, John Erik (1996): *Programeksempler i SAS.* Oslo: Statistisk sentralbyrå. Interne dokumenter 96/13

Sjørbotten, John Erik (2003): *Programeksempler i SAS utgave 2.* Oslo: Statistisk sentralbyrå. Interne dokumenter 2003/12

Skatteetaten (2010): *Folkeregistrering.* http://www.skatteetaten.no/

Skilbrei, Birger (1992): *Oversikt over dataregistre i Tygdeetaten.* Oslo: Rikstrygdeverket

Skiri, Halvard og Kjetil Sørlie (1993): *Befolkning*. Oslo: Statistisk sentralbyrå. Sosialt utsyn 1993, kap. 2.1.

Skog, Ole-Jørgen (2004): *Å forklare sosiale fenomener: en regresjonsbasert tilnærming.* Oslo: Gyldendal akademisk

Skrede, Kari (2006): *Hovedtrekk ved inntektsutviklingen for kvinner og menn i perioden 1982 – 2003.* Oslo: Statistisk sentralbyrå. Økonomiske analyser 2/2006

Skrede, Kari og Kristin Tornes (red.) (1983): *Studier i kvinners livsløp.* Oslo: Universitetsforlaget

Slagsvold, Britt og Svein Olav Daatland (red.) (2006): *Eldre år, lokale variasjoner. Resultater fra den norske studien av livsløp, aldring og generasjon (NorLAG) - runde 1.* Oslo: Norsk institutt for forskning om oppvekst, velferd og aldring

Stanfors, Maria (2007): *Mellan arbete och familj. Ett dilemma för kvinnor i 1900-talets Sverige.* Stockholm: SNS Förlag

St. melding nr. 24 (2008 - 2009) (2009): *Nasjonal strategi for digital bevaring og formidling av kulturarv.* Oslo: Kultur- og kirkedepartementet

Statistics Norway (2009): *Activity Plan for 2010. Initiatives and Priorities.* Oslo: Statistisk sentralbyrå. Planer og meldinger 2010/04

Statistics Norway (2003): *Norwegian Standard Classification of Education - C751.* Oslo: Statistisk sentralbyrå.

Statistics Norway (2010): *Statistics by subject – Gender Equality.* Oslo: Statistics Norway at www.ssb.no.
Statistics Norway (2010): *Regional Differences in Gender Equality.* Oslo: Statistics Norway at www.ssb.no.
Statistisk sentralbyrå (1974): *Standard for kommuneklassifisering.* Oslo: Statistisk sentralbyrå
Statistisk sentralbyrå (1985): *Standard for kommuneklassifisering.* Oslo: Statistisk sentralbyrå
Statistisk sentralbyrå (1991): *Registerbasert inntekts-, formues- og skattestatistikk for personer.* Notater 91/21. Oslo: Statistisk sentralbyrå
Statistisk sentralbyrå (1994): *Standard for kommuneklassifisering 1994.* Oslo: Statistisk sentralbyrå
Statistisk sentralbyrå (1995): *Historisk statistikk 1994.* Oslo: Statistisk sentralbyrå
Statskonsult (2002): *IKT i det offentlige 2002.* Oslo: Notat 2002:4
Statskonsult (2002): *Offentlige registre - grunnlaget for den elektroniske forvaltningen.* Oslo: Statskonsult. Rapport 2002:02
Stortingsmelding nr. 8 (2003-2004): *Rikt mangfold i nord. Om tiltakssonen i Finnmark og Nord-Troms.* Kommunal- og Regionaldepartementet.
St. melding nr. 24 (2008 -2009): *Nasjonal strategi for digital bevaring og formidling av kulturarv.* Oslo: Kultur- og kirkedepartementet
Strand, Børge (1992): *Personlig inntekt, formue og skatt 1980 – 1989.* Oslo: Statistisk sentralbyrå. Rapporter 91/18
Strand, Børge (1996): *D-nummerpopulasjonen.* Oslo: Statistisk sentralbyrå, Notater 96/39
Strand, Børge (1996): *Kobling av adresse-registrene i DSF og GAB. Dokumentasjon og resultater.* Oslo: Statistisk sentralbyrå, Notater 96/7
Strand, Børge (2003): *Noen erfaringer fra et digitalt ordningsarbeid.* Oslo: Riksarkivaren. Arkivmagasinet 1/2003
Sunde, Elisabeth (1999): *Fagavdelingenes innsyns- og styringsmuligheter ved utvikling av saksbehandlingssystemer. - En studie av systemutvikling i Rikstrygdeverket og Skattedirektoratet.* Oslo: Universitetet i Oslo. Avdeling for forvaltningsinformatikk.
Sønneland, Helge M. (1987): *Samtidens arkiver - fremtidens kildegrunnlag.* Oslo: Universitetsforlaget
Sørlie, Kjetil (2000): *Rekruttering av kvinner til kyst- og bygde-Norge. Sammenhengen mellom innflytting, jobbtilknytning og familieforøkelse.* Oslo: Norsk institutt for by- og regionforskning. Notat 2002:114.
Teigen, Mari (2006): *Det kjønnsdelte arbeidslivet: en kunnskapsoversikt.* Oslo: Institutt for samfunnsforskning
TemaNord (1996): *To Preserve and Provide Access to Electronic Records.* Copenhagen: TemaNord 1996:549

Thévenon, Olivier (2009): 'Increased Women's Labour Force Participation in Europe: Progress in the Work-Life Balance or Polarization of Behaviours?' *Institut national d'études démographiques (INED), Paris.*

Thorvaldsen, Gunnar (1999): *Databehandling for historikere.* Oslo: Tano Aschehoug

Utne, Harald og Erik Vassnes (1995): *Kopling av A/A- og LTO-register: dokumentasjon av kvalitet og konsistens i begrep.* Oslo: Statistisk sentralbyrå. Notater 95/2

Vaage, Odd Frank (2005): *Tid til arbeid: arbeidstid blant ulike grupper og i ulike tidsperioder, belyst gjennom tidsbruksundersøkelsene 1971–2000.* Oslo: Statistisk sentralbyrå. Rapporter 2005/15

Villund, Ole (2006): *Kvalitet på yrke i registerbasert statistikk: resultater og videre utfordringer*. Oslo: Statistisk sentralbyrå. Notater 2005/14

Villund, Ole (2009): *Measuring Working Hours in the Norwegian Labour Force Survey: A pilot Study of Data Quality Using Administrative Registers.* Oslo: Statistisk sentralbyrå. Reports 2009/3

Vlasblom, Jan Dirk and Joop J. Schippers (2004): 'Increases in Female Labour Force Participation in Europe: Similarities and Differences'. *European Journal of Population (2004) 20, pp 375–392*

Yerkes, Mara (2006): *Diversity in Work: the heterogeneity of women's labour market participation patterns 2006-44*. Amsterdam: University of Amsterdam

Warren, Hannah (2007): 'Using Gender-Analysis Frameworks: Theoretical and Practical Reflections. Gender and Development'. GB.Taylor & Francis Vol. 15, No. 2, pp. 187-198.

INDEX

action zone;42
ADDML-standard;184
administrative register;22
archival format;20
Archival Information Package;152
Archival packages;152
authenticity;98; 99; 100; 152
auxiliary variable;82
Basic Amount;78
Bevaringsutvalget;22
binary field;38
bivariate;37
case handling systems;21
Central Population Register;27
Central System for National Social Security;65
Central Taxation Authorities;14
centrality;85
checksums;98
Civil Registration;27
Cohabitation;89
computerised archives;15
confounding variables;48
Continental Shelf;30
control digits;27
coverage;50
cross sectional design;36
Data Base Management System;71
data exchange;26
data matrix;51
data model;67
data set;13
degree of urbanisation;85
dependent variable;37
deposit;20
diachronic;51
dichotome variable;38
digital archives;15
Digitalarkivet;26
Directorate for Seamen;28
Dissemination Information Package;152; 190
D-number;26; 28
donor data set;81
dummy variable;38
dwelling unit;29
EBCDIC;73; 190
electronic archives;15
Electronic record keeping systems;21
emigration;82; 108
employment history;45
evidential value;24
explanatory variable;37
family obligations;45
family responsibilities;88
field;14
file;13
Gender Equality Index;142
geographic mobility;88
heuristic;11; 33; 35; 150
household identifier;28; 29
income line;77; 78
independent variable;37
individual number;27
individual rate of labour market participation;45

industrial structure;85
information carrier;99
information system;13
informational value;24
integrity;20; 98
International Labour Organisation;39
Joint assessment;90
joint population;102
joint tax assessment;91
labour force;39
Labour Force Sample Surveys;75
labour income;76
labour market;39
linear regression analyses;33
local labour market characteristics;84
logical tests;101
logistic regression;39
long term storage;13
longitudinal;10
mainframe;21
marital status;89
matching;14
merging;14
metric variable;38
migration;87
migration strategy;10; 71
MoReq;21; 190
multicollinearity;118
multiple;37
multi-provenance;96
multi-variate correlation;122
municipality classification;84; 92; 173
municipality restructure;131
national identifiers;26
national identity number for persons;26
National Library of Norway;24
National Social Insurance System;41
NAV;65
NOARK;21
nominal variable;38
Norwegian Archives Act;24
Norwegian Labour and Welfare Service;65
Norwegian Social Science Data Services;23
nuclear family;28
numeric address code;29
optical input;96
ordinal variable;38
organisation number;30
organisational number for legal entities;26
Other systems;22
panel design;36
panel population;69
pension rights earnings;66
pensionable income;66; 76
personal code;28
Personal number;27
post-industrial;39
primary key;51
production format;20
profession;84
prospective longitudinal design;37
provenance;25
published material;24
raw data;10; 18; 39; 71
recipient data set;81
record linkage;14
records creator;10
redundant information;97
reference person;28; 90; 91; 175; 176
register cleansing;99

research registers;21; 22
residual;116
retrospective longitudinal design;37
Rikstrygdeverket;65
sample surveys;22
SAS;72
SAS notation;14
self - employed;80
separate populations;102; 103
significance;117
Skattedirektoratet;61
SPSS;72
Statistics Norway;22
Subject-specific systems;21
Submission Information Package;152; 190
system;13
table;13
Tax Register for Personal Taxpayers;65
taxpayer unit;90; 91; 102; 115
Technical metadata;66
technology dependent storage;13
technology independent format;10
time series design;37
total populations;22
transfer;20; 81
trivariate;125
t-values;117
Ullensaker project;31
validity;28; 104; 149
validity check;28
variable;14
wage earner;80