UiT The Arctic University of Norway

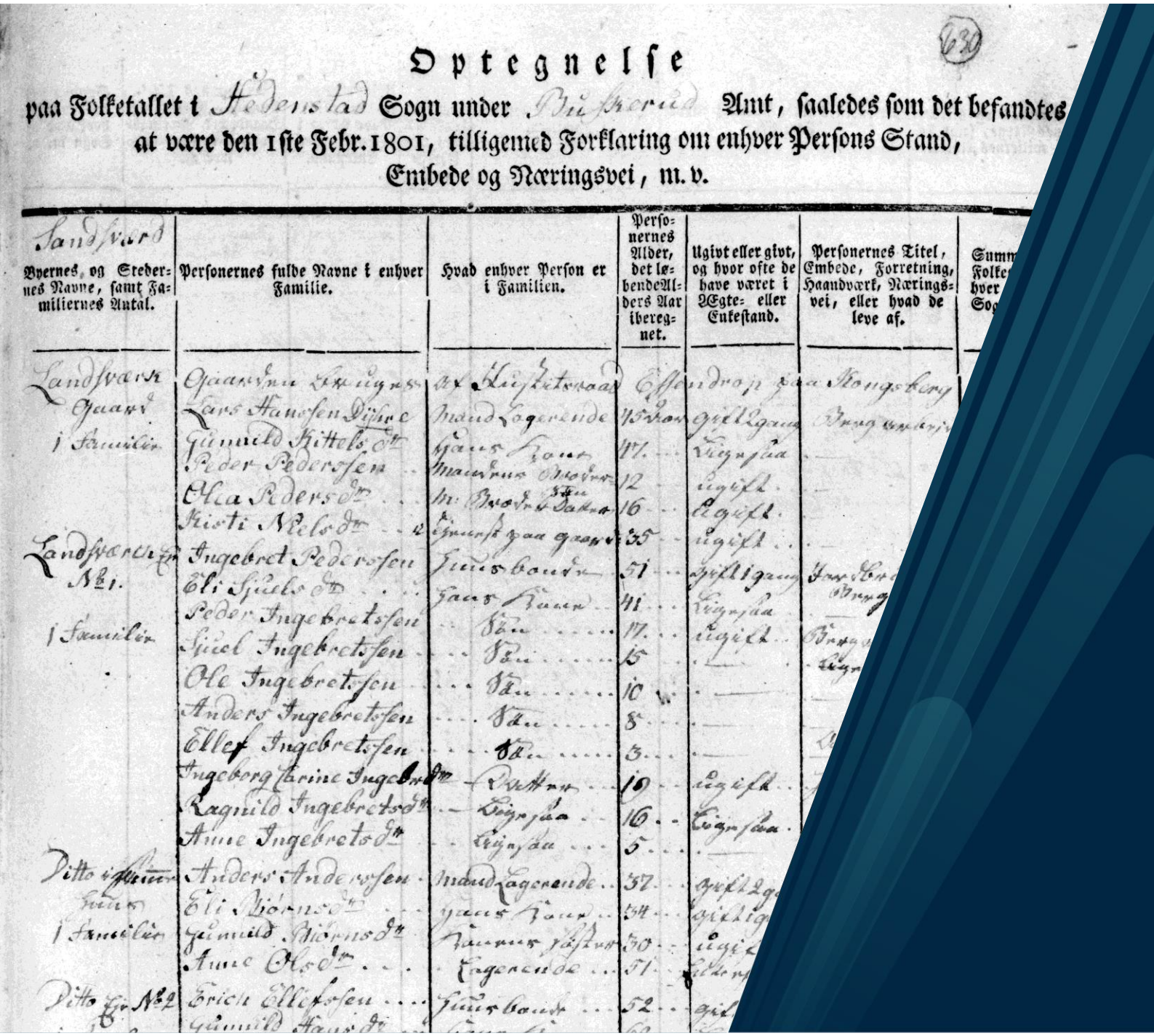
Faculty of Science and Technology

Department of Computer Science

Record linkage of Norwegian historical census data using machine learning

Narae Park

INF-3990 Master's thesis in Computer Science, August 2022



Cover image: List from the census in Hedenstad parish 1801

(Author: Unknown, Institution: The state archive in Kongsberg, Sandsvær bygdebokkomité, License: CC BY-SA, from National Archives Digital Photo Archive(https://www.digitalarkivet.no/))

Abstract

The Historical Population Register (HPR) is a project to build the longitudinal life history of individuals by integrating the historical records of the people in Norway since the 19th century. This study attempted to improve the linking rate between the 1875-1900 censuses in HPR, which is currently low, using machine learning approaches. To this end, I developed a machine learning model for linking that is suitable for the Norwegian census and tested various algorithms, feature sets, and match selection options. I compared the results in terms of performance and match size, and also examined their representativeness to the entire population. The study results showed that the linking rate of HPR can be significantly improved by machine learning approaches while maintaining high accuracy. In addition, this study presented a reference for future use by demonstrating how the performance varies depending on the feature set and match selection. On the other hand, this study also revealed that linked data generally do not represent the population of the census, and the characteristics and degree of bias vary depending on the linking algorithm, suggesting that caution is needed when using linked data for research.

Keywords: historical record linkage, Norwegian census, Historical Population Register (HPR), machine learning

Acknowledgements

This study is indebted to the help of many people to whom I am grateful.

First of all, I would like to express my sincere gratitude to my supervisor, Lars Ailo Bongo, for his guidance, insightful advice, and inspiring discussions. His enthusiasm often lifted me up, and his unspoken encouragement kept me on the move. I would also like to thank my co-advisor, Hilde Leikny Sommerseth, for her warm welcome, expert insights, valuable advice, and useful references she has provided me. Bjørn-Richard Pedersen deserves my deep gratitude. I appreciate all the advice and comments he gave as an advisor, the discussions we had as colleagues, and the encouragement and support he gave as a friend. Thanks for proofreading the thesis, too. I would also like to express my gratitude to Trygve Andersen for all the materials he gave me and for always answering my many questions, along with my respect for the work he has done so far. Without them, it would have been impossible for me as a foreigner to do research on the historical records of Norway.

I would like to thank Gunnar Thorvaldsen for giving me useful advice and comments with the knowledge and wisdom from his expertise at the end of the study. My thanks also go to the HDL Lab members who gave useful feedback at the seminars, including the A208 members who used to spend time working and resting together. I thank Jan Fuglesteg and Kai-Even Nilssen, the department staff who helped me a lot from the beginning to the end of the master's program.

I am grateful to my friends here in Norway and in Korea for their kindness and encouragement which gave me the courage to continue this journey. I would also like to express my gratitude and love to my family who always support and believe in me.

Finally, I would like to thank the numerous anonymous and non-anonymous people who share their knowledge with the public. Without them, my study would not have been completed. I hope that my study can also be a small stepping stone to someone who will come next.

Table of Contents

List of Tables

List of Figures

Abbreviations and Acronyms

HPR Historical Population Register [Historisk befolkningsregister (HBR)]

NHDC Norwegian Historical Data Center [Registreringssentral for historiske data (RHD)]

Chapter 1

Introduction

1.1 Motivation

The life history of individuals is both personal and social because they reflect the times and society in which the individuals lived. The late 19th and early 20th centuries are the beginning of modern life that continues to this day, when industrialization and modernization brought fundamental changes to people's lives. The life histories of this period show the impact of historical and social changes on people's lives, and are therefore still an important resource of data in a variety of fields including history, sociology, economics, and public health.

However, it is not easy to construct the life history of individuals during this period because the existing microdata sources of the population are not integrated. The traditional event-based parish population registers such as church books were maintained, and the census, a nationwide systematic and simultaneous population record, was also introduced during this period. However, it is challenging to identify individuals across these sources because there are no reliable identifiers for individuals such as today's social security numbers. Solving this problem, that is, identifying individuals in multiple historical records and linking the records to individuals is called historical record linkage. Historical record linkage studies have been done in various fields over the past 80 years[1]. In the early days of these studies, manual linking using some samples from the local area was the main method,

[1] This topic, the problem of identifying the same object across multiple sources, has been studied under various terms in different fields:
record/data linkage, record/data/entity/object/data/fuzzy/citation/reference/computer matching, data/information integration, deduplication, duplicate detection, entity/co-reference resolution, entity/reference/data reconciliation, entity extraction, file linking, object identification, re-identification, object consolidation, merge/purge problem, list washing, etc [1][2].

but with the digitization of historical resources and the development of computer technologies, the scope of the linking expanded and automated methods gradually became common. This has led to great strides being made in historical record linkage studies, but there are still issues waiting to be resolved.

In this study, I will deal with the problem of linking historical censuses in Norway from the late 19th century to the early 20th century. Norway is one of the countries that started researching historical record linkage early. The parish registers (church books) dating back to the 17th century have been well preserved, and a nominal census was first introduced in 1801. Social security numbers that can identify individuals were introduced in 1964, and since then, the population records have been managed as the Central Population Register [3]. At the same time, scanning and transcribing historical population records began in the 1980s, followed by a project to build the Historical Population Register (HPR)[2] by linking these records from 1800 to 1964, which is still ongoing. The HPR will be integrated with the Central Population Register in the future, completing a unified Norwegian population register since the 19th century [4].

The currently available HPR[3] is linked based on the concept of family reconstitution[4] using church book records for the areas where these have been transcribed, and otherwise is linked by finding the same families between the censuses for the areas where church book records have not been transcribed yet [6][4][7].[5] It is highly accurate because it is based on explicit family records (marriage, baptism), with ambiguity removed using family information, and by using an approach that minimizes false links. However, since it is based on family matching when linking between the censuses, the link rate is reduced during periods with a large time span when families change, such as the 1875-1900 period shown in Figure 1.1. In addition, since the current HPR was built on family matching, it may have introduced biases such as people who live with the same family members for a long time are more likely to be linked.

[2] 'Historisk befolkningsregister (HBR)' in Norwegian

[3] It is available to the public on the web: https://histreg.no/, https://rhd.uit.no/

[4] Family reconstitution is a demographic technique that reconstructs the life histories of individuals and families by linking the records of demographic events of the parish, such as baptisms, burials, and marriages, with names of persons. It was initially developed by Swedish demographers and mainly advanced by French researchers including Henry and his colleagues in the 1950s [5]. The current HPR is linked with some modifications to the original family reconstitution method.

[5] See chapter 2.3.2 for details

Therefore, in this study, I explore a way to improve the 1875-1900 census link rate, which is currently the bottleneck in establishing the longitudinal population set in HPR. Since the current HPR contains high-accuracy links, a machine learning approach using them is an attempt with potential. At the same time, since there is a concern about selection bias in HPR links, the bias in the results linked by the machine learning model trained on them is also investigated. Through this, this study seeks to contribute to constructing a reliable and useful longitudinal population set in Norway.

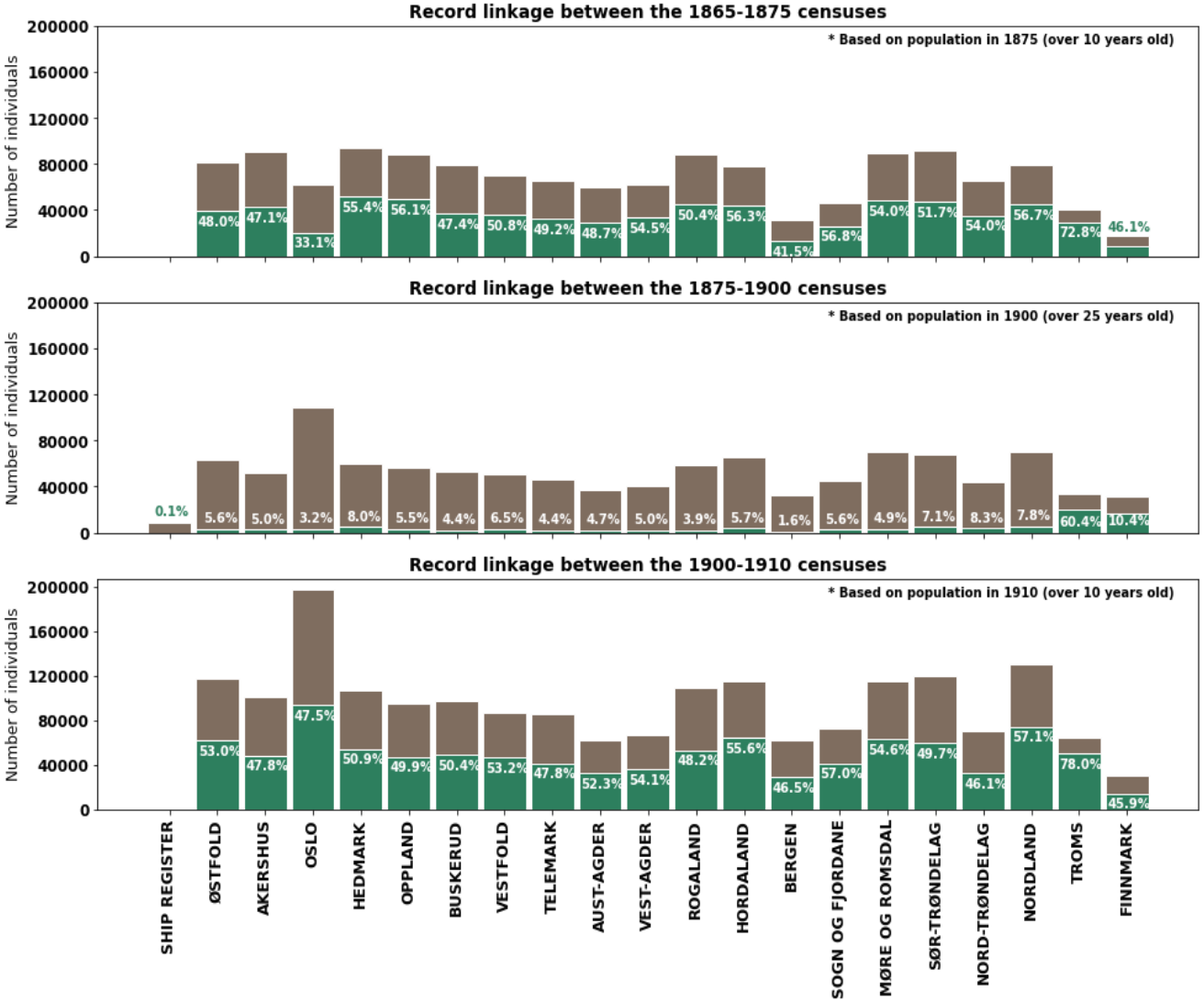


Figure.1.1. Link rates between historical censuses in the current HPR, by county.[6] The green area in each bar represents the proportion of people whose records are linked in that county.

[6] It includes the two largest cities of the time, Oslo (its name at the time was Kristiania) and Bergen.

1.2. Challenges

Linking the Norwegian censuses in the late 19th and early 20th centuries has both the general challenges of census linking and the unique challenges of the Norwegian censuses during this time.

If someone tries to find the same person in two censuses, he or she will intuitively try to search for one person whose characteristics such as name, gender, age (or year of birth), place of birth, residence, etc. are the same in both censuses. However, several challenges may arise from the fundamental reason that there is no solid basis for identifying the same person.

- The same person with slightly different characteristics: It is quite common for the two records to be the same person even though they do not have exactly the same characteristics due to misreporting, misspelling, enumeration errors (e.g., duplication, omission), transcription errors, or changes in characteristics over time (e.g., change of the last name due to marriage, change of residence due to moving), etc.

- Multiple people with the same characteristics: Conversely, there are cases where one of the match candidates cannot be identified because there are multiple people with the same characteristics. This may often occur when the person lived in a large city with a large population or had a common name. For example, there were ten "Ole Olsen"s born in 1854 and lived in Oslo in 1875.

- The trade-off between matching rate and precision: The above issues create a trade-off in determining the linking method. Allowing some flexibility in filtering match candidates can increase the number of matches by including the correct one in the candidates even with slightly different characteristics, but it can also increase false matches because more people who are not correct are included in the candidates (increasing type 1 errors[7] or false positives). On the other hand, if the criteria for filtering match candidates are strictly set, the possibility of false matches is reduced, but even the correct one is excluded from the candidates, thereby reducing the overall match size (increasing type 2 errors or false negatives).

- Limited 'Ground Truth': Because it is usually not possible to ascertain the same person in the historical records, the ground truth is unknown or difficult to get in historical

[7] Type 1 error means that records that should not be linked are linked, that is, an incorrect match (false positive), and type 2 error means that records that should be linked are not linked, that is, missing a match (false negative).

record linking. This causes difficulties in developing linking methods or evaluating the performance of linking methods based on it. Usually, the links manually reviewed by humans are regarded as a kind of ground truth, but manual linking by humans is expensive and time-consuming. Besides, since it is also only a method of linking[8], it may differ from the real answer. If the manual linking is biased, there may be problems in developing or evaluating linking methods based on it.

- Computational complexity: In order to match the same person in two datasets, each record in one dataset has to be compared with all the records in the other dataset. Therefore, as the size of the dataset increases, the computational complexity increases quadratically. This challenge is usually handled by indexing (blocking) to eliminate comparison pairs that are unlikely to match, but indexing has the limitation of filtering out some true matches as well. In addition, when complex comparisons are performed on comparison candidates, since the comparison algorithm is usually expensive, computational complexity still remains a potential challenge as the dataset size increases.

- Lack of representativeness: Since the method of identifying the same person is based on matching the characteristics of the census, people who maintain the same characteristics for a long time can be linked more easily. (e.g., men - who do not change names due to marriage; children or married adults - who continue to live with the same family members; people who live in the same place for a long time; people who have the same job for a long time, etc.) As a result, these groups may be more included in the linked population, resulting in differences from the original population distribution. That is, there may be a selection bias that does not represent the original population as it is.

- The trade-off between matching rate and representativeness: The issue of representativeness in the linked population creates a trade-off between matching rate and representativeness. If only characteristics that do not change over time and have few missing values in the census are used for linking, the problem of representativeness in which people with certain characteristics are more easily linked can be reduced, but the total number of linking is also reduced because information useful for linking cannot be used. On the other hand, if characteristics such as family members, residence, and occupation are also used for linking, it helps to find the correct one among multiple

[8] Abramitzky compared the performance of several linking algorithms including hand linking, and showed that hand linking is also on the frontier of a trade-off between accuracy and efficiency [8].

match candidates, but representativeness issues increase because people with changes in those characteristics are less likely to be linked.

In addition to these general challenges, there are challenges unique[9] to Norwegian censuses in the late 19th and early 20th centuries.

- Changes of last names: In Norway, several naming practices for surnames have been used, such as a patronymic based on the father's first name (e.g., **Johan** Olsen's son would be Hans **Johan**sen, and his daughter would be Anna **Johans**datter.), an inherited last name from the father (e.g., Johan **Olsen**'s son would be Hans **Olsen**, his daughter would be Anna **Olsen** and his wife would be Marie **Olsen**.), and a last name based on the place of residence or farm name (e.g., if Johan **Olsen** moves to the farm **Berg**, his name could become Johan **Berg**.) (IPUMS, 2010)[9]. As permanent and inherited last names gradually became more common, the inheritance of fixed last names became mandatory by law in 1923 [9]. During this transition (1865-1920), some people's last names were changed due to changes in naming practices, which causes difficulties in identifying the same person based on name matching.

- Uses of middle names: Norwegian names often include middle names. Although middle names are not unique to Norwegian names, they are often stated in different ways in the census such as the first name with the middle name (e.g., Johan Hans Olsen), the first name with the initial of middle name (e.g., Johan H. Olsen), only the first name (e.g., Johan Olsen), or even only the middle name (e.g., Hans Olsen). This also causes difficulties in identifying the same person based on the same name in the census.

- Changes of birthplaces/municipalities: The boundaries of municipalities, which are the basis for a place of birth and residence in the Norwegian census, changed often between the mid-19th and early 20th centuries. For this reason, the place of birth of a person, which should not be changed, may change between censuses, and the municipality of residence may change even though they have not actually moved. This is one cause of confusion in identifying the same person in Norwegian censuses during this period.

- Long intervals between some censuses: The Norwegian census was generally conducted every ten years, but there are rather long intervals between several censuses (1800-1865: 65 years, 1875-1900: 25 years) due to problems such as numerical surveys,

[9] Some of these are not necessarily specific to the Norwegian census, but are challenges that need to be dealt with in order to address the linking of Norwegian censuses at this period.

non-nationwide coverage, and incomplete transcription. This long interval of more than one generation creates difficulties in matching the same person between censuses.

Many of these challenges have to do with trade-offs, so it is hard to find a simple answer. Thus, ascertaining the real impact of these trade-offs on outcomes can be a starting point for handling the challenges.

1.3. Research questions

In Norway, building the Norwegian Historical Population Register (HPR), microdata on the Norwegian population since the 19th century, has been ongoing since scanning and transcribing historical population microdata began in the 1980s. HPR is a longitudinal population register of the entire country that spans over 200 years and will be useful to researchers in a wide range of fields. This study aims to explore a census-based record linking method that can help build the HPR for the period prior to 1964, when unique identifiers were introduced. To achieve that goal of the study, I will address the following research questions.

1. **Could the linking rate between the 1875 and 1900 censuses in HPR be improved by using machine learning models trained on the HPR links?**

The current HPR contains links manually reviewed in previous studies and links created by using family event records (baptism, marriage) in church books, both of which have high accuracy. If machine learning models are trained using these high-accuracy links, it may be possible to link the entire censuses with high accuracy by these models.

To verify this, I add the steps of training and applying machine learning models to the normal record linking process. After preprocessing the Norwegian census data to handle the challenges of the Norwegian census linking, filtering potential match candidates through indexing, and comparing these candidate pairs, comparison vectors of each candidate pair are generated. Using the HPR links as training data for building machine learning models, an optimal machine learning model is obtained through testing several machine learning algorithms. By classifying comparison vectors into match or non-match with the obtained machine learning model, linking results are gained. To evaluate the impact of feature sets used for linking, I train a time-invariant feature-based model and an extended feature based model. And to evaluate the impact on how the final matches are selected, I get the result of

taking only unique matches and the result of also taking the best matches in multiple matches. These linked results are evaluated in several respects, such as link rates, performance (precision and recall), and differences in results.

2. **Are the populations linked by the machine learning models from question 1 representative of the entire population of the census?**

The high-accuracy links in the current HPR are based on the concept of family reconstitution. Thus, there may be an issue of selection bias in which the linked population does not sufficiently resemble the original population and over-/under-represents people with certain characteristics. Since machine learning models learn a pattern for classification from training data, linking results by machine learning models trained on the HPR links may have the same selection bias problem. Because sample representativeness can be an important issue depending on the topic of the study, it is necessary to examine the bias in the results linked by machine learning models.

To explore this, I statistically compare characteristics of the population of the census and populations linked by the machine learning models. And to check whether the results are due to the machine learning approach affected by the training data, I also compare the population linked by the traditional rule-based method. Also, I evaluate the impact of representativeness in the linked population by comparing how changes in some characteristics differ over time in populations linked in different ways.

1.4. Contributions

This study has the following contributions.

- It describes the unique characteristics of Norwegian census linking and applies them to develop a machine learning model suitable for Norwegian census linking.
- It links the 1875-1900 Censuses with a low link rate nationwide with a machine learning model using existing HPR links without the effort of building new training data. As a result of linking, the linking rate has been improved by up to 35% without a significant decrease in quality. Although the HPR links used as training data were created by supporting additional sources such as parish records, this study shows the benefits of the

machine learning approach by obtaining performance close to that using additional sources even with census-based linking by a machine learning model trained on them.

- It demonstrates the difference between the results linked by machine learning models according to the feature set and match selection in terms of match size and performance. Based on these results, methods and parameters for census linking can be selected and adjusted in the future by the needs of researchers or the evaluations of experts.
- It examines the representativeness of linked populations by comparing the characteristics of the populations linked in different ways to the original population. As a result of comparing, characteristics that are easy to link generally tended to be overrepresented in the linked populations and there were differences in characteristics between linked populations according to the linking algorithm. These results indicate that when using the linked population for research, a close review or additional processing of data is required depending on the purpose of the study.

1.5. Structure of the thesis

This thesis consists of a total of six chapters.

Chapter 1 introduces the motivation, challenges, research questions, and contributions of this study.

Chapter 2 provides theoretical and technical background for historical record linkage studies. This serves as a basis for the design of this study.

Chapter 3 describes the data sources, methods, and processes used to address the research questions of this study.

Chapter 4 presents and evaluates the results of the study obtained as an answer to the research questions of this study.

Chapter 5 discusses the findings and insights gained through the study.

Chapter 6 summarizes the conclusion of the study and suggests future work.

Chapter 2

Background

This chapter provides theoretical and technical backgrounds on historical record linkage. I first introduce the general process of historical record linking, and then review the linking approaches used in previous historical record linkage studies in terms of algorithms and variable sets. I also describe the history of historical record linkage studies in Norway and the Norwegian Population Register as a result of them. These backgrounds based on literature review serve as the basis for the study design in Chapter 3.

2.1. General record linking process

The general process of matching records from two datasets[10] is shown in Figure 2.1 [2][11]. After preprocessing the original datasets, candidate pairs that are likely to be matched are filtered out through indexing. Then, the similarity of each candidate pair is calculated, and each pair is classified into a match or non-match according to the degree of similarity. I will illustrate the work performed at each step in more detail in the following sections.

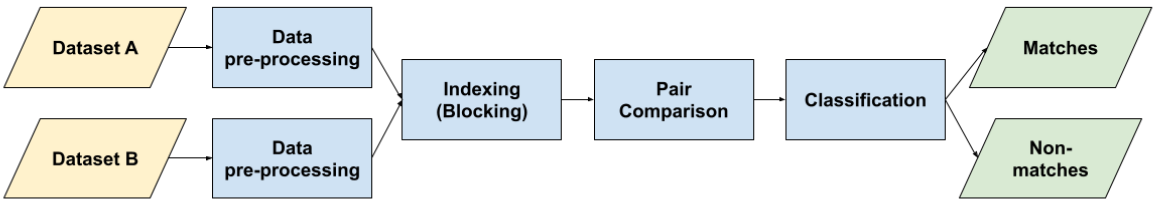


Figure 2.1. The general process of matching records from two datasets

[10] For linking multiple historical datasets longitudinally, an approach can be taken that continues linking the two datasets. Linking multiple datasets simultaneously requires further research [10].

2.1.1. Data preprocessing

In the data preprocessing, the record formats are prepared to be the same so that the records of the two datasets can be compared. Accordingly, data inspection, standardization, and creation of derived variables are generally performed at this stage.

Historical datasets are often constructed by transcribing original (handwritten) historical documents. In order to minimize the loss of information in this process, the raw values in various formats of the original document are often registered as they are. As a result of this, errors in the dataset can stem from both the enumerator who took the census and the transcriber who digitized the values. Therefore, first, the data is checked for errors, and if possible, the errors are corrected. Missing values are also processed as needed, such as removing or filling in estimated values. Through this, it is also checked whether the attribute values of the data are valid. Next, the data is converted into a standardized format. If necessary, a standardized format can be created for the purpose. Standardizing data representation facilitates comparison and analysis of data. For example, a person's date of birth can be expressed in various ways. Standardizing this to, for instance, a four-digit representation of the birth year allows it to be used in analysis as a numerical variable. Deriving the variables necessary for linking from raw data is also a useful approach for effective linking. It is possible to create derived variables by extracting, combining, or inferring from raw data. For example, each person's family members can be extracted from the household list of the census and used as a new variable.

Data preprocessing is a preparatory step for subsequent steps, but it is also a step that takes a lot of time and effort in practice. The preprocessing might need to be repeated several times after new issues have been discovered in subsequent steps, requiring modification to the process.

2.1.2. Indexing

In the indexing process, potential matching candidate pairs are filtered out from the records of the two datasets. In order to fully compare the records of the two datasets, as many record pairs as the Cartesian product of the size of the two datasets ($M \times N$) must be compared. As the data set size increases, the computational complexity increases quadratically. Moreover, this operation is quite inefficient, since most of the record pairs being compared have little chance of a match. For example, one record A from the 1875 census would theoretically match at most one of all records from the 1900 census. The majority of the 1900 census

population do not share the same attributes as A. Accordingly, an indexing is introduced in which the majority of comparison pairs with no matching potential are filtered and candidate pairs with matching potential are extracted.

For indexing, a method of extracting records with the same attributes by using a block key consisting of the attribute values of the records is mainly used.[11] However, the true match may be omitted from the candidate block due to errors in the data or changes in attribute values over time. Although obtaining candidates using multiple block keys is one way to compensate for this, in practice, some true matches are inevitably omitted due to indexing, which may introduce bias, so caution is required [12]. In order to minimize the omission of true matches by indexing, attributes that have few errors or missing values, are evenly distributed across the population, and do not change between the two datasets are generally used as block keys.

2.1.3. Comparison

Next, the two records are compared for each match candidate pair obtained from indexing. By comparing how similar the attribute values of two records are, it is possible to estimate whether two records refer to the same person. Attributes used for comparison are mainly the attributes expected to be consistent for the same person, such as name, year of birth, and place of birth. However, in order to increase the accuracy of matching, attributes that can change over time such as family members, occupation, and residence are sometimes additionally used.

Ideally, the attribute values of the two records of the same person should be the same (for immutable attributes), but since various errors may be involved in the data, the similarity between the attribute values of the two records is generally measured as a numerical value indicating the degree. The more attributes of two records are similar to each other, the more likely they are to refer to the same person.

Various methods are used to calculate the similarity between attribute values. Usually, numerical differences are used for numerical attributes such as birth year, and several types of string similarity algorithms are used for string attributes such as names. Using these methods, the similarities between attributes to be compared are calculated for each matching candidate pair. As a result, it is possible to obtain a comparison vector composed of similarity values.

[11] For this reason, indexing is often referred to as blocking.

2.1.4. Classification

Finally, each comparison vector of the candidate pairs obtained through the comparison process is classified into match or non-match. If necessary, it can be classified into three classes: match, non-match, and cases requiring manual review.

One of the naive methods to classify comparison vectors is to calculate the sum of comparison vectors (or the sum weighted to specific attributes) and classify it into match or non-match based on a threshold.

In machine learning methods, these comparison vectors are used to train a classifier that classifies matches or non-matches by supervised or unsupervised learning. This trained classifier decides matches or non-matches in an automated way for new matching candidate pairs.

Since this classification is made for each match candidate pair, multiple matches can be obtained for one record as a result[12]. That is, for one record in one dataset, there are two or more records having similar attribute values in another dataset. When linking censuses, an individual in a census must match up to one person from another census (assuming there are no duplicate records). Therefore, for multiple match results, additional adjustments are needed, such as discarding all the multiple matches or selecting only the first match having a sufficiently large value compared to the second match.

[12] In order to overcome this problem, studies on collective matching were also attempted to determine matching at the global level by using linking information of other records [2]. However, it was not covered in this study because of its high computational complexity that makes it difficult to apply to large-scale datasets.

2.2. Approaches for record linking

While following the general record linking process, a variety of approaches have been attempted so far regarding which sources and variables to use for linking, and what algorithms to use to classify matches and non-matches. Each of them has its own strengths and weaknesses, so there have been several discussions, and they can be selected according to the researcher's needs and interests. Since this study will also examine the influence of linking variables and linking algorithms, this section introduces the discussions related to linking algorithms and linking variables from a historical point of view.

2.2.1. Linking algorithms

In the early days of linking historical records, historians manually searched population records to trace each individual or household. Record linking studies at the time were conducted at the local level, as they relied on the manual linking of historians, which was time-consuming and expensive, and national microdata was not yet ready for use. As computers have developed and historical records have been transcribed, the amount of available data has increased and the scope of linking has expanded nationwide. The IPUMS (Integrated Public Use Microdata Series) project, which publicly release census microdata including the US historical census, led by the Minnesota Population Center at the University of Minnesota in the early 1990s, has contributed significantly to the development of automated linking studies [13].

Ferrie [14] laid the foundation for an automated linking algorithm by introducing a rule-based linking process based on phonetic coding of names, age, state of birth, and family members using IPUMS data. The basis of **a rule-based method** is to utilize the knowledge and methods that historians and experts use when manually linking records, into a set of rules and processes that can be followed when conducting automated record linking For example, Ferrie classified a candidate with the same phonetic prefix of the name, the same state of birth, living with the same family members, and having a birth year difference of less than 5 years as a match, and if the match results for the same record were more than two, chose the ones with the smaller age difference [14]. The rule-based method has the advantage of being intuitively understandable[13] and easy to apply, and the knowledge of domain experts can be reflected in the linking [15]. Since the 2000s, as computational power has increased and national micro-datasets have been released, the number of studies on record linkage have

[13] However, as the rules become more complex, it can become difficult to understand the consequences.

greatly increased. During this period, the use of automated linking algorithms became common, and researchers generally adopted rule-based automation methods that adjusted and extended Ferrie's method. As one of the rule-based automation algorithms widely used today, Abramitzky's ABE-JW [16][17] basically links individuals based on the similarity of birth year, place of birth, and names.

On the other hand, record linkage studies by **probabilistic approaches** rather than deterministic approaches have also been developed. The Fellegi-Sunter's model [18], which is based on the probability that two records represent the same individual and the probability that they do not, was used as a main probabilistic record linkage model until the 2000s. Inheriting the Fellegi-Sunter model, Winkler [19] discussed ways to improve record linkage with a probabilistic methodology. Mill [44] proposed an unsupervised automated linking algorithm that does not require training data using the Expectation–maximization (EM) process, and based on this, Abramitzky [20] proposed an automated probabilistic linking method using the EM algorithm. The probabilistic method has the advantage that it is well supported theoretically based on statistics, is not affected by human bias and does not depend on training data, but also has the disadvantage of relying on modeling assumptions that are difficult to verify [21][22].

With the development of machine learning technology since the 2000s, various **machine learning-based linking algorithms** have been introduced. The machine learning approach used for linking is usually supervised learning, where a linking model is trained using training data in which matches and non-matches are pre-classified. The model learns how to classify matches and non-matches from training data given with correct answer labels, and uses the learned method to classify data given without answers. Being able to learn patterns from training data is both an advantage and a disadvantage of the machine learning approach, so good performance can be expected when high-quality training data is given. Goeken[23] built IPUMS-LRS (IPUMS Linked Representative Samples) using Support Vector Machines (SVM) algorithm, and Antonie [38] also used SVM algorithm to link Canadian censuses in the late 19th century. Feigenbaum [21] applied several machine learning algorithms such as Logit, OLS, SVM, and Random Forest (RF) to link censuses, compared their performance and draw implications. Record linkage studies using machine learning approaches continues to be conducted in recent years[24–26]. Since machine learning algorithms greatly depend on the quality of the training data, Price [26] introduced methods of acquiring high-quality links through Wikipedia-style genealogy platforms and using them as training data for machine learning.

2.2.2. Linking variables and data sources

Variables and additional sources used for linking records are also one of the factors affecting the linking result. When finding the same person in the records of two censuses, using many attributes makes it easier to find the right person among multiple candidates, which results in improved accuracy and match size. However, this also entails concerns that some groups that are more advantaged (or disadvantaged) for maintaining certain attributes may be over-represented (or under-represented) in the linked results, leading to distortion of past population estimates. For this reason, there have been several discussions and approaches to date regarding variables and sources used for linking .

In a person's census record, attributes such as name, birth year, birthplace, race, are permanent attributes of that person. Thus, they do not change over time, helping to identify the same person at different points in time. And the use of these immutable attributes to identify people is relatively free from concerns about the bias of the linked population. As an early rule-based algorithm, Ferrie [14] used not only **time-invariant variables** such as name, year of birth, and state of birth, but also variables that could be helpful in linking censuses such as family members, in the linking process. However, since it was argued that the use of family members, place of residence, and occupation information for linking could introduce selection bias and distort estimates [27], most record linking algorithms have only used time-invariant characteristics such as name, sex, birth year, and birth place.

However, since the time-invariant attributes recorded in the census are limited, often many people share the values in common. This results in multiple match candidates in identifying the same person. In this case, examining additional attributes such as family members, place of residence, and occupation and using them to identify the same person is very helpful in finding the right match. Several studies used these **extended variables** including mutable attributes for linking to improve match accuracy and match size. Fu [11] showed that integrating individual and household linkage helps to improve accuracy when linking population data in Lancashire, UK. Antonie [28] also showed that when family member information is used for linking the entire Canadian census, the overall linking rate can be greatly increased due to disambiguation without significant additional bias. Recently, in the linking of the entire US Census, Helgertz [24] proposed a method to obtain a high linking rate and accuracy while hardly compromising representativeness by using the link between households.

It is also possible to use **additional historical data sources** other than the census for linking censuses. A historical source at the microdata level, widely used in European historical studies, is the parish register. This is a book with detailed records of people's religious ceremonies, such as baptism, marriage, and burial, administered by the parish church. Family reconstitution methods, which had a significant impact on European historical demographic studies including the construction of HPR in Norway, are based on these parish registers. Scandinavian countries generally have well-preserved parish registers, so studies using them in the linking process or verification of linking have often been conducted [29][30][31][7]. In addition to parish records, genealogy platforms where descendants can directly participate and verify links are also data sources that can be used for census linking. As mentioned in Section 2.2.1, Price [26] used high-accuracy links from a genealogy platform to train a machine learning linking model. These genealogy platforms where high-quality links can be expanded by people's participation have room for future use in historical record linkage studies.

2.3. Historical record linkage studies in Norway

This study explores ways to improve the links in the Norwegian Historical Population Register (HPR), which integrates the historical population records of Norway since the beginning of the 19th century. As a background to understand it, this section describes the history of historical record linking studies in Norway and one of its main achievements, the HPR.

2.3.1. History

Norway is one of the countries where historical record linkage research has been developed early on because of the well-preserved parish registers (church books), population censuses, and other historical records at the microdata level. Early record linkage studies in Norway were attempted at the local community level and were influenced by family reconstitution. Family reconstruction is the technique of linking records developed by Louis Henry in the 1950s to obtain demographic statistics based on event records in church registers. This method links data from different sources with individuals and families as linking points [6]. Norway's abundant church book records provided a good environment for using family reconstitution for historical population studies.

In the 1970s and 1980s, individual and family reconstitution studies on Kristiania and Ullensaker regions were conducted by researchers at the University of Oslo, and a demographic analysis and migration study using population reconstitution for Rendalen in eastern Norway were done by Sogner (1979). Also, individual and family reconstitution studies on regions in western Norway, microdata transcription and semi-automatic record linking studies were carried out by researchers at the University of Bergen. Studies on linking of historical records from Rendalen parish continued until the 2000s, extending the individual and family reconstitution database to the period 1733-1900 [32]. Fure (2000) built a population database for Asker and Bærum in eastern Norway using DemoLink, a semi-automated record linkage program that constructs individual life courses by linking historical sources. Sunde (2001) studied the people who migrated from the fjords of western Norway to the prairies of the midwestern United States by manual linking using family reconstitution. As for northern Norway, Thorvaldsen (1995) studied the migration in Troms county in the late 19th century through linking of censuses using a historical record linkage software. [6][4][3]

The establishment of the Norwegian Historical Data Center (NHDC)[14] in 1981 can be said to be a milestone in historical population records linkage in Norway. The NHDC, led by the University of Tromsø (UiT), began to transcribe historical population records such as national censuses, parish registers, and other microdata sources and build an integrated database of Norwegian population records. The NHDC has established an integrated National Historical Population Register (HPR), covering the entire population of Norway since 1800 by linking transcribed historical records longitudinally on an individual basis, and this is now available to the public on the web [4][7]. Building the HPR is an ongoing project, and even today untranscribed census and parish registers are transcribed and continuously updated in its database. On the basis of these Norwegian historical microdata sources, demographic studies on various topics such as migration, mortality, nuptiality, fertility, and family history have been actively conducted [3].

2.3.2. Historical Population Register (HPR)

The Norwegian Historical Population Register (HPR) is the fruition of a long-term effort to build a database that integrates Norway's historical population records vertically and horizontally. It is a comprehensive database of Norwegian historical population records since 1800, which can be used for research in various fields of study. The development, main principles and processes of the HPR are detailed in [33] and [7]. Here, I introduce the historical record linking method of the HPR, which is the focus of this study.

The basic approach when building the HPR is to utilize all available sources, but only adopt reliable links. The HPR was built on the basis of family reconstitution or similar family-based linking using Norway's plentiful church book records.

The way individuals' records are linked in the HPR is as follows. A unique ID is assigned to each individual record from all archived sources, and the IDs of records found to belong to the same person through the linking process are merged into one. Therefore, by indexing a person's ID, various historical records linked to it (decennial censuses, church book records such as birth, marriage, burial, etc.) can be identified, and an individual's life course can be reconstructed from these. As the main sources available for individual linking, historical censuses are mostly available in transcribed form, and church book records are partially available because they are currently being transcribed or being post-processed.

[14] Registreringssentral for historiske data (RHD)

For regions where church book records are available, the NHDC has developed an automatic linking program based on a family reconstitution method using church records. This program identifies the bride and the groom as a married couple using marriage records from church books, and identifies the bride, the groom, and their parents using the information of the bride's parents and the groom's parents in the church book. And then, it also identifies the couple and their children using the baptism records from church books. In this way, a person and his/her parents are linked by assigning a father's ID and a mother's ID to each individual record of the archived population sources. By tracking these parent-child links, both family composition and intergenerational connections can be established. The records of church books from Troms were first transcribed and used for program development, and the results of linking are now included in the HPR. This is why Troms county has a high link rate of 60-80%[15] in linking historical censuses. (see Figure 1.1)

For regions other than Troms, where church book records are not available, the HPR currently uses census data to find and link the same family on a family-by-family basis. This is done by finding and linking families that have the most family members in common, based on family member information from the census[16]. Since it is rare for any two families to have the same family member information, this is a useful individual identification method during periods when family members do not change significantly. The 1865–1875 and 1900–1910 census linking, using this method, show linking rates of 40-60%. However, when the time interval between censuses is large, the family members usually change, so this method may have some limitations, as shown by the low linkage rate in the 1875 and 1900 censuses (see Figure 1.1).

The HPR also includes in its database the results of previous record linkage studies, such as the individual and family reconstitution database of Rendalen, and the population linked results of the NAPP (North Atlantic Population Project)[17], which is a project to harmonize and distribute census data of the North Atlantic countries in a machine-readable database format[18] [4].

---

[15] Linkage rate based on population aged over 25 in 1900

[16] For example, create a string with 3 letters of first name, 3 letters of last name, sex, year of birth, and place of birth of each family member, and use it as an identification key for each family member. When comparing two families from two censuses, the number of the same family member key is counted as the number of common family members.

[17] Not all linked results of the NAPP are included, but only the results verified by the NHDC are included in HPR, so it does not completely match the links of the NAPP.

[18] https://www.nappdata.org/napp/intro.shtml,
https://www.nappdata.org/napp/progress_report_oct08_may09.shtml

Chapter 3

Methods

This chapter describes the methods and processes carried out to address the research questions of this study. I begin with a description of the Norwegian census as data sources used in the study, and specify the process of linking census records nationwide by a machine learning approach and the method of examining the representativeness of the linked data.

This study puts a lot of weight on the development of census linking using machine learning suitable for the Norwegian census. Therefore, a detailed description of the process is included in this chapter. Since the record linking process consists of several steps, and the output of each step will be described together as an input to the next step, a large portion of this chapter is devoted to the process of linking the Norwegian census data.

3.1. Data sources

This study uses the Norwegian national population census as historical records for linking individuals. I obtained Norwegian population census datasets from the Norwegian Historical Data Center (NHDC)[19].

3.1.1. Norwegian population census

A census is the complete enumeration of the population in a country or region at a particular time.[20] It is a snapshot containing a wealth of information about the population and households in an area at a point in time. Norway's first public census was taken numerically[21]

[19] Norwegian population census datasets can be obtained by requesting it from the Norwegian Historical Data Center (NHDC) or from the Integrated Public Use Microdata Series (IPUMS) platform: https://rhd.uit.no/, https://international.ipums.org/international/

[20]https://www.unfpa.org/census, https://www.unfpa.org/census

[21] It indicates the aggregated number of inhabitants sorted by district, age, profession, etc. (https://www.arkivverket.no/en/find-your-ancestors/public-censuses)

in 1769, and the first national nominal[22] census took place in 1801. Norwegian nominal census data now fully transcribed and available are for the years 1801, 1865, 1875, 1900 and 1910.[23] Table 3.1 shows the size and regional distribution of the currently available census data, at the county level (including two largest cities, Oslo and Bergen). This study focuses on the linking between the 1875 and 1900 censuses, so they are used as the main sources.

Table 3.1. The size and regional distribution of Norwegian population censuses currently available

	1801		1865		1875		1900		1910	
	population	%	population	%	population	%	population	%	population	%
ØSTFOLD	50,141	5.7	98,846	5.9	107,183	6.0	139,629	6.1	155,723	6.3
AKERSHUS	56,968	6.5	105,959	6.3	118,597	6.6	113,957	5.0	132,033	5.3
OSLO	9,212	1.1	53,651	3.2	77,676	4.3	231,895	10.1	247,209	10.0
HEDMARK	61,064	7.0	120,161	7.1	123,567	6.9	130,292	5.7	140,671	5.7
OPPLAND	66,454	7.6	124,974	7.4	115,080	6.4	119,371	5.2	122,943	5.0
BUSKERUD	64,680	7.4	95,051	5.6	104,419	5.8	115,299	5.0	128,306	5.2
VESTFOLD	39,099	4.5	85,341	5.1	91,454	5.1	107,631	4.7	111,996	4.5
TELEMARK	47,503	5.4	81,738	4.9	85,482	4.7	98,178	4.3	111,287	4.5
AUST-AGDER	28,875	3.3	66,356	3.9	77,918	4.3	78,292	3.4	79,223	3.2
VEST-AGDER	39,758	4.5	75,454	4.5	79,869	4.4	82,625	3.6	84,867	3.4
ROGALAND	44,398	5.1	104,846	6.2	116,173	6.5	125,587	5.5	144,675	5.9
HORDALAND	60,448	6.9	105,356	6.3	102,446	5.7	138,514	6.0	150,859	6.1
BERGEN	18,125	2.1	31,797	1.9	39,694	2.2	68,482	3.0	78,015	3.2
SOGN OG FJORDANE	52,601	6.0	86,759	5.2	60,257	3.3	91,143	4.0	92,806	3.8
MØRE OG ROMSDAL	57,326	6.5	104,356	6.2	117,916	6.5	142,567	6.2	149,539	6.1
SØR-TRØNDELAG	60,503	6.9	105,521	6.3	117,203	6.5	136,451	6.0	154,707	6.3
NORD-TRØNDELAG	42,691	4.9	82,693	4.9	83,266	4.6	86,210	3.8	89,239	3.6
NORDLAND	52,207	5.9	89,620	5.3	104,072	5.8	157,919	6.9	172,786	7.0
TROMS	19,218	2.2	45,331	2.7	54,868	3.1	77,092	3.4	85,349	3.5
FINNMARK	7,707	0.9	20,334	1.2	24,566	1.4	35,469	1.6	41,059	1.7
SHIP REGISTERS	-	-	-	-	-	-	17,294	0.8	-	-
TOTAL	878,978	100.0	1,684,144	100.0	1,801,706	100.0	2,293,897	100.0	2,473,292	100.0

[22] It provides detailed information, such as first and last name of each individual. (ibid.)

[23] The census has been conducted approximately every ten years since 1801, but the 1815, 1825, 1835, 1845, and 1855 censuses were numerical, and the 1885 census was only carried out in towns. The 1891 census has recently been transcribed but has not yet been published, and the 1920 census is not yet ready to be made available nationwide. From the 1930 census onwards, it will be released successively, in accordance with the Norwegian Statistics Act (1989), which restricts its use for 100 years.
(ibid., https://www.arkivverket.no/slektsgranskning/folketellinger, NHDC)

In census data obtained from the NHDC, there are both original variables from the transcribed census and some variables constructed by the NHDC. The variables (often referred to as attributes or fields) registered in the census have some differences depending on the census year[24], but the variables that the censuses have in common are not very different from those in other countries. They include names (first name, last name), sex, year of birth, place of birth, family relationship, marital status, occupation, residence, which are typically used for linking records. Table 3.2 shows the common variables in the census data. Among these, I used the variables that have few missing values and are standardized or easily standardizable for linking, which are shown in bold in Table 3.2.

Table 3.2. The common variables in the census data and the variables used in this study (in bold). The original Norwegian field names are shown in square parentheses.

- **Municipality[25] number[26] [Kommunenummer]**
- **Enumeration district number [Kretsnummer]**
- **Residence list number [Bostedsnummer]**
- **Apartment list number (Towns only) [Leilighetsnummer]**
- **Person number[13] [Personnummer]**
- Household number[27] [Markering av hushold]
- **First name [Fornavn]**
- **First name (standardized)[28] [Fornavns]**
- **Last name [Etternavn]**
- **Last name (standardized) [Etternavns]**
- **Sex [Kjønn]**
- **Year of birth (Date of birth) [Fødeselsår (Fødelsdag)]**

[24] For a list of fields for each of currently available censuses, see: https://rhd.uit.no/folketellinger/rubskj_b.html, https://rhd.uit.no/koding/kodetdata_e.html

[25] In Norwegian census datasets, the place of birth or residence is coded on a municipal basis. Norway is divided into counties [fylker], subdivided into municipalities [kommuner]. The number of municipalities in Norway in the past censuses is 433 in 1865, 495 in 1875, 595 in 1900, and 659 in 1910. (For detailed information about each name and code, see: https://rhd.uit.no/koding/fs_koder.html) A municipality is a relatively small geographic unit, so having a person's birth place at the municipality level greatly helps to identify the same person. Vick and Huynh [34] showed that Norway's geographic unit (municipality) defining the block of candidates is much smaller than that of the United States (state), which helps to create non duplicated links.

[26] Municipality number, Enumeration district number, Residence number, Apartment number, and Person number are organized hierarchically in the datasets. They are indicated in order of 'municipality>enumeration district>residence>apartment>person', for example, '0432>001>0001>00>001', '0432>001>001>00>002'. They are designed and assigned to each census record by the NHDC.

[27] Household number is not provided in the datasets obtained from the NHDC. So I derived the variable from the census data myself. See Section 3.2.1

[28] The original names registered in the census were standardized by the NHDC and assigned to each record. See Section 3.2.1 for related details.

- **Place of birth (coded)**[29] **[Fødested]**
- **Family relationship (coded)**[30] **[Familiestatus]**
- **Marital status [Sivilstatus]**
- Occupation [Yrke][31]
- **Residence name [Bostednavn]**
- Religion[32] [Religion]
- Infirmity[33] [Sykdom]
- Ethnicity [Etnisitet]

[29] In the dataset, an individual's place of birth and community of residence are coded into 4-digit numbers. It was developed by the NHDC and assigned to each individual record. For the table of municipalities corresponding to codes, see: https://rhd.uit.no/koding/fs_koder.html

[30] In the dataset, family relation (Relation to household head) is also coded into a 4-digit number. It was assigned to each record coded by the NHDC corresponding to the RELATE variable of NAPP/IPUMS. (https://international.ipums.org/international-action/variables/RELATE#codes_section)

[31] Occupation is one of the variables with high potential utilization in linking or linked results, but the values are not standardized in the dataset and the missing value rate of the 1875 census is high at 42.44%, so I excluded it from use.

[32] Religion and ethnicity fields were rarely registered.

[33] Infirmity is not provided in the datasets obtained from the NHDC.

3.2. Linking censuses using machine learning

For linking the 1875 and 1900 censuses, I followed the general record linking process introduced in Section 2.1, and added some steps to train and use machine learning models. An overview of the overall process is shown in Figure 3.1. Each step of the process is presented in detail in each section below.

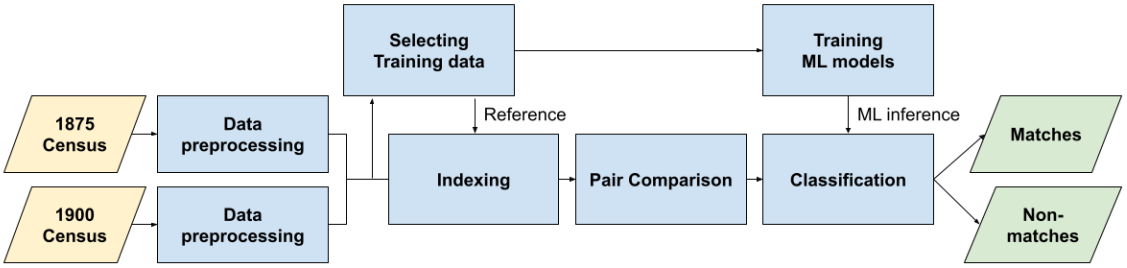


Figure 3.1. The overview of the linking process in this study

3.2.1. Data preprocessing

The attribute values of the census dataset are mixed with the raw values of the original census transcribed and already standardized values. The raw values of the census dataset are transcribed from scanned census pages, so they are expressed in various forms. Therefore, it is necessary to standardize these forms before starting the linking process so that they can be compared with each other during the linking process.

The ***Sex*** field is usually expected to have only two values - man [*mann*] and woman[*kvinne*], but actual values in the census vary due to the registrant, transcribers, or transcription notations[34,35]. Therefore, it is necessary to convert them into the standardized form 'm' and

[34] The meanings of some transcription notations are as follows.
- '!!': the information is missing or the information in the source is obviously wrong.
- '*': the information was not given to the original source, but was added by the registrant based on surrounding information.
- '%%': the information in the source is struck out. Marked before and after the struck through text.
- '@': a separator between two reading options that cannot be determined.
- '??': the original information is uncertain or difficult to read.

https://rhd.uit.no/histform/felles.html, https://www.familysearch.org/en/wiki/Norway_Census

[35] As an example, the values of the 'sex[kjonn]' field of the 1875 census are as follows.

K	m	K	!!	m*(space)*	m*	M	M!!	K!!
k*	k*(space)*	m %k%	%m%	m!!	K %M%	M %K%	K %m%	s
n	ug	mug	?	h	e	k!!	M!! %K%	

‘k’, by referring to the raw values and transcription signs. During the process, invalid values are treated as missing values. As a result, the number of missing values slightly increases after preprocessing.

The ***Birth year*** field also has raw values in a variety of forms in the census, so they should be converted into a standardized 4-digit year of birth. Invalid values are treated as missing values.

The ***Marital status*** field should also be standardized with five fixed abbreviations 'ug' (unmarried), 'g' (married), 'e' (widow / widower), 's' (separated) or 'f' (divorced) according to the census guidelines. Invalid values are set as missing values.

As for ***First name*** and ***Last name***, there can be many variations on one name[36]. The same person's name may be registered with small variations in spelling due to misreporting or misspelling, making it difficult to identify the same person. For this reason, several methods for standardizing names and comparing the similarity of name strings have been used in previous studies. Phonetic encoding[37] that converts a string to its pronunciation [14][17][26], or a standard name dictionary that defines standard names for variant names [23][34] are such methods. Norwegian researchers including Gulbrand Alhaug and Bente Ramsvik of Tromsø University created a standard name dictionary for all unique first and last name strings of the 1865, 1875 and 1900 Norwegian census[38][34][4]. Current Norwegian census datasets contain names standardized using this dictionary along with the original names.

Standardizing names corrects for minor spelling variations, but there is a trade-off in that names that are merely similar to another can be erroneously standardized, losing the chance to identify different people. Regarding the effect of name standardization, Vick and Huynh [34] showed that standardizing first names can prevent false links from being included in the matched links, and also increase the overall linkage rate in Norway. Gjelseth [32] also mentioned the possibility that standardizing names would be useful in Rendalen's

[36] As an example, some of the original names standardized as the name 'Ingeborg' in the 1875 census are as follows.

Ingeborg	Ingebor	Engeborg	Ingborg	Engebor	Ingbor	Ingebaar
Ingebørg	Ingeburg	Ingbør	Ingebaarg	Ingbaar	Engebaar	Ingebord
Engbor	Engborg	Ingebrg	Ingeborr	Ingebog!!	Ingaborg	Engebord
Yngeborg	Engebørg	Jngeborg	Igeborg	Jengebor	Ingebør	Inbebor

[37] Phonecitc encoding algorithms include Soundex, Phonex, Phonix, NYSIIS, ONCA, Double Metaphone, Fuzzy Soundex, etc.[2]

[38] This work was also published as a dictionary of Norwegian first names [35].

population linking study. Also, when I did a preliminary test of linking using original names and linking using standardized names, the linkage rate using standardized names was higher. In this study, I used both original and standardized names for linking.

In addition, for the names, illogical values are kept. Cases marked as illogical (marked with transcription notation '!!') include, for example, a case where a man's last name should be 'Johannesen (Johannes's son)' according to the patronymic convention, but registered as 'Johannesdatter!! (Johannes's daughter)'. This has partially meaningful information, so I thought that it would be more useful to utilize partial information rather than treating it as a missing value.

Additionally, extra spaces included in the name strings were removed for accurate string comparisons later.

Table 3.3 shows the before and after standardizing some variables through data cleaning.

Table 3.3. The number of unique values and missing values by variable before and after data cleaning.

Field		1875 census		1900 census	
		Before (%[39])	After (%)	Before (%)	After (%)
Sex	unique[40]	26	3	11	3
	missing[41]	172 (0.01)	202 (0.01)	13 (0.00)	39 (0.00)
Birth year	unique	4,726	115	23,422	106
	missing	2,361 (0.13)	4,055 (0.23)	3,294 (0.14)	5,212 (0.23)
Marital status[42]	unique	433	6	268	6
	missing	419,625 (23.29)	422,158 (23.43)	21,843 (0.95)	22,698 (0.99)
First name (original)	unique	155,953	155,765	175,330	175,329
	missing	3,125 (0.17)	3,125 (0.17)	10,125 (0.44)	10,125 (0.44)
First name (standardized)	unique	108,108	108,084	131,156	131,156
	missing	11,282 (0.63)	11,282 (0.63)	10,307 (0.45)	10,307 (0.45)
Last name (original)	unique	109,108	109,081	104,199	104,199
	missing	3,478 (0.19)	3,478 (0.19)	4,576 (0.20)	4,576 (0.20)
Last name (standardized)	unique	73,698	73,689	88,967	88,967
	missing	25,757 (.143)	25,757 (1.43)	4,943 (0.22)	4,943 (0.22)

[39] Percentages of the number of missing value records to the total number of population in the census (the 1875 census: 1,801,706, the 1900 census: 2,293,897).

[40] The missing value is also counted as one unique value.

[41] The number of missing values here means the number of records whose values are unknown in each attribute.

[42] The marital status variable has a fairly high percentage of missing values in 1875, so I did not use it directly as a linking variable, but indirectly used it in creating the household ID as a derived variable. (See ch.3.2.1)

Next, as a step in data preprocessing, I created derived variables that can be helpful for linking from the original census.

Adjusted birth place (coded): Birth place is coded as a 4-digit number based on the municipality where an individual was born, however, some of these municipalities were divided into smaller municipalities, merged with each other, or changed borders between censuses[43]. Therefore, it is necessary to take this into account when comparing the census records. For example, Rendal (Rendalen today), a municipality in southeastern Norway, was divided into Ytre Rendal (code: 0432) and Øvre Rendal (code: 0433) in 1880. As a result, birth places of residents born in Øvre Rendal are classified as 0432 in the 1875 census, but 0433 in the 1900 census[44]. This can affect identifying them across censuses because it is generally expected that a person's birthplace will not change. To compensate for this, for the new municipalities that were divided during 1865-1910, I assigned a unified municipality code as an adjusted birth place code[45].

Adjusted municipality (coded): Likewise, for residents of municipalities whose borders have been changed, their municipality has been changed even if they have not moved. So, I also added a unified municipality code before division, as an adjusted municipality code.

Household ID: I created unique household IDs by concatenating residence IDs and the household head IDs to extract family member information. First, I derived unique residence IDs by concatenating census year + district + residence + apartment number. However, since multiple households can reside in one residence, each household must be further divided. So,

---

[43] For historical changes in municipality and county divisions, see :
https://www.ssb.no/a/histstat/rapp/rapp_199913.pdf (Historisk oversikt over endringer i kommune- og fylkesinndelingen [Historical overview of changes in the municipality and county division])

[44] Not all of them are. For example, there are 0 residents of Øvre Rendal (code: 0433) in 1875 census (since it is before the division), but 44 people who were born in Øvre Rendal (code: 0433) in 1875 census already exist. I guess it depends on when the reference point for allocating the birth place code is. Note that there may be some errors due to this.

[45] I did not include cases where only a part of one municipality was incorporated into another due to changes in the borders of municipalities. (For example, the borders of Glemmen (code: 0132) were changed in 1867, so 2013 people, a part of Glemmen population, were incorporated into Fredrikstad (code: 0103).) This is because the changed population cannot be distinguished in the census, and I thought that in this case, the information lost by merging municipalities was greater than the gains.
One more thing to note is that although the population for this case is generally small, the number of Aker's (code: 0218) population incorporated into Oslo's (code :0301) population in 1878 was 18,970, which is quite large. There may be some errors in the values of the birth place or residence place with respect to these people.

I derived the household head IDs using the family relationship and marital status. The household head ID can be used directly as the household ID, but since there are missing values in the family relationship field, I assumed that when combined with the residence id, more accurate household information can be obtained.[46,47]

Household members: Assuming that individuals with the same household ID are one household, I created a household member field that stores information of individuals with the same household ID. Household members information consists of a concatenation of the first 3 letters of the first name + birth year + sex + adjusted birth place code[48]. For the birth year, small variations may be introduced in the census, so a ±1 year range is allowed. This variable is used to calculate the number of the common household members when comparing potential candidate pairs later.

Self-information of first and last name: Common names are more difficult to link than rare names [23][24][26]. In information theory, the self-information of an event is inversely proportional to the probability of its occurrence[49]. Thus, considering that the commonality of names would affect the linking, I counted the number of people with the same first name and last name in the census, respectively. Then, I calculated the probability of this number with respect to the total population, and added its negative logarithm as a derived variable indicating the self-information of the name.

Self-information of the birth place: Likewise, people from highly populated municipalities are more difficult to link than people from sparsely populated municipalities [23]. Because it means more people have the same birth place. Considering that the commonality of birth place also would affect the linking, I counted the number of people sharing the same birth place in the census, calculated the probability of the number with respect to the total

---

[46] Since the household ID was created by combining the residence ID and the household head ID, there may be a limitation that if a household lives in multiple residences, they are considered as separate households.

[47] The number of households and the average household size in the 1875 and 1900 census according to household IDs are as follows.
Number of households in the 1875 census: 326,289
Average household size (number of household members) in the 1875 census: 5.52
Number of households in the 1900 census: 459,752
Average household size (number of household members) in the 1900 census: 4.99
(Note that group accommodations such as hospitals and prisons are also included in households.)

[48] This refers to the method used for the current HPR linking.(See chapter 2.3.2)

[49] In information theory, the self-information (information content, surprisal, or Shannon information) is expressed as follows. I(x)= - logP(x) (https://en.wikipedia.org/wiki/Information_content)

population, and added its negative logarithm as a derived variable indicating the self-information of the birth place.

First name with middle name removed, and ***middle name***: Norwegian names often have middle names. The percentage of names registered in the census containing middle names is shown in Table 3.4. For people with middle names, there are cases where both the first name and the middle name were registered in the census, where only the first name was registered, and in rare cases only the middle name was registered. For example, a person whose name is registered as 'Ole Hans' in one census may be registered as 'Ole', 'Hans' or 'Ole Hans' in another census. Considering this, I extracted the first name without the middle name, and the middle name from the name registered as separate variables to use for indexing or matching later.

Table 3.4. The number and percentage of population with middle name in Norwegian censuses

Census	Population	Population with middle name[50] (%[51])
1865	1,684,144	323,441 (19.21)
1875	1,801,706	423,321 (23.50)
1900	2,293,897	440,211 (19.19)
1910	2,473,292	524,034 (21.19)

Initials of first, middle, and last name: Since there were cases in which names were written only as initials in the census, I extracted the initials of the names and added them each as variables. In addition, even if there are slight variations in names, the initials of names are usually maintained, so they can be used for matching and comparison.

The variables used in this study after the data preprocessing are shown in Table 3.5.

Table 3.5. Variables in the census used in this study after data preprocessing

Original variables	Derived variables
• First name (original)	• Adjusted birth place (coded)
• First name (standardized)	• Adjusted municipality (coded)
• Last name (original)	• Household ID
• Last name (standardized)	• Household members
• Sex	• Self-information of first name
• Family Relationship (coded)	• Self-information of last name

[50] Including cases with only initials for the middle name.
[51] Percentage of the total population

• Marital status • Birth place (coded) • Birth year • Residence • Municipality (coded)	• Self-information of birth place • First name with middle name removed • Middle name • First name initial • Middle name initial • Last name initial

3.2.2. Selecting municipalities for training data

For training and evaluation of machine learning models for linking, it is necessary to have training data. Since the machine learning model learns to classify matches (two records belong to the same person) or non-matches (two records do not belong to the same person) from the training data, it is crucial to build quality training data. The existing HPR contains high-quality links that are the fruition of previous record linkage studies, that is, links with high accuracy because parish registers were additionally used or manually reviewed for linking, so I chose to use these as a golden standard training dataset. The high-quality data constructed in this way can also be used as a reference to determine indexing conditions later.

First, the population records of Rendalen municipality in south-eastern Norway, from 1733 to 1828, were linked by the demography project of Sogner [36], and further up to 1900 by Gjelseth and other researchers in her project [32]. These links have been manually reviewed through the family reconstitution method using parish registers, so they are highly accurate. The results are now integrated into the HPR. I used data from this municipality in the training data.

Next, the population records of **Troms county** are largely linked by the NHDC at the University of Tromsø (UiT). These links were created by an automated method based on family reconstitution using transcribed parish registers, and some were manually reviewed [33][7]. These links, like Rendalen, are also highly accurate because they were created using explicit family records (baptism and wedding) in parish registers.[52]

However, there are differences in the extent to which census data are linked to parish registers among the municipalities of Troms. Figure 3.2 shows the percentage of people who are linked to their baptism records in Troms County at the 1900 census, and Figure 3.3 shows the percentage of married or widowed people who are linked to their marriage records in

[52] See Section 2.3.2 for how the links were created.

Troms County at the 1900 census. It can be inferred that the linked data of municipalities with a high percentage of people linked to their records in the parish register would be more accurate, as supported by records from the additional source. In order to include only those municipalities with higher accuracy in the training data, I selected only the ones where the percentage of people linked to their baptismal records exceeded the county average of 70% and those where the percentage of people linked to their marriage records exceeded the county average of 83%.

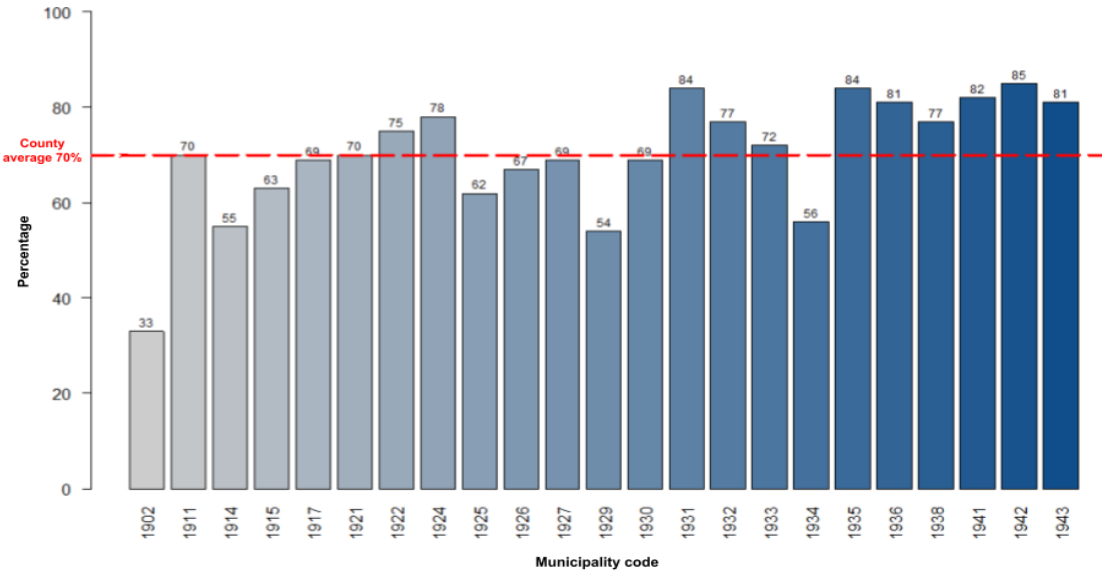


Figure 3.2. Percentage of people linked to their baptismal records in Troms at the 1900 census[53]

[53] Source: Trygve Andersen, Senior engineer at the NHDC at UiT, and the principal leader of the development of automatic record linkage algorithms.

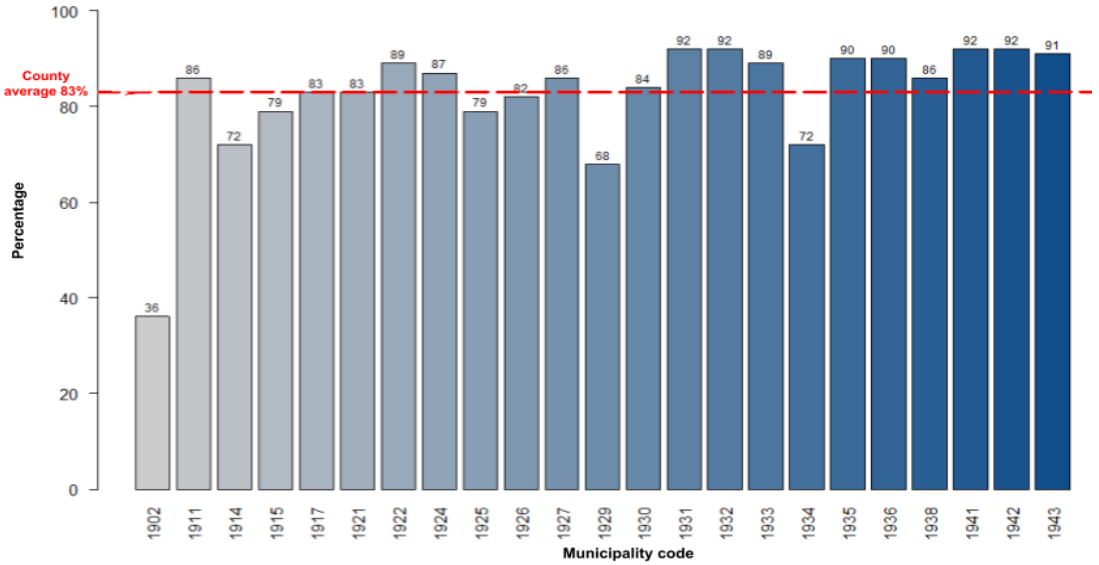


Figure 3.3. Percentage of people linked to their marriage records in Troms at the 1900 census[54]

Table 3.6 shows the nine municipalities as golden standards built from the above.

Table 3.6. Municipalities selected from existing HPR to build the training data

Municipality (code)	Linked/Total inhabitants in 1875 (%)	Linked[55]/Total inhabitants (aged 20 and over) in 1900 (%)
Rendalen (0432)	1462/3530 (41.4)	1443/2129 (67.8)
Ytre Rendal (0432)	1462/3530 (41.4)	656/959 (68.4)
Øvre Rendal (0433)	÷	787/1170 (67.3)
Bardu (1922)	592/1122 (52.8)	603/849 (71.0)
Målselv (1924)	1536/3044 (50.5)	1314/1940 (67.7)
Lenvik (1931)	2273/4490 (50.6)	2213/3249 (68.1)
Hillesøy (1930)	÷	514/873 (58.9)
Lenvik (1931)	2273/4490 (50.6)	1699/2376 (71.5)
Balsfjord (1933)	1861/3850 (48.3)	1737/2516 (69.0)
Malangen (1932)	÷	630/854 (73.8)
Balsfjord (1933)	1861/3850 (48.3)	1107/1662 (66.6)
Karlsøy (1936)	1063/2175 (48.9)	1054/1624 (64.9)
Helgøy (1935)	÷	438/659 (66.5)
Karlsøy (1936)	1063/2175 (48.9)	616/965 (63.8)
Lyngen (1938)	2327/4672 (49.8)	2241/3493 (64.2)

[54] Ibid.

[55] The number of linked individuals here is among the total population in 1900, not among the population over 20 years old in 1900.

Skjervøy (1941)	1646/3211 (51.3)	1540/2227 (69.2)
Skjervøy (1941)	1646/3211 (51.3)	1016/1454 (69.9)
Nordreisa (1942)	÷	524/773 (67.8)
Kvænangen(1943)	744/1673 (44.5)	696/1054 (66.0)

Looking at the top 10 municipalities with high linking rates in HPR as shown in Table 3.7, all 9 municipalities selected as training data are in the top 10. This suggests that the data of these municipalities with a high linkage rate with their parish records are more accurate, which leads to higher linkage rates in HPR. And this shows that the data of these municipalities are convincing as training data.

Table 3.7. Top 10 Municipalities for linking rates between 1875-1900 censuses in HPR. The municipalities of training data are shown in bold.

Rank	Municipality	Linking rate in 1875	Municipality	Linking rate in 1900
1	**Bardu (1922)**	52.8%[56]	**Bardu (1922)**	71.0%[57]
2	**Skjervøy (1941)**	51.3%	**Skjervøy (1941)**	69.2%
3	**Lenvik (1931)**	50.6%	**Balsfjord (1933)**	69.0%
4	**Målselv (1924)**	50.5%	**Lenvik (1931)**	68.1%
5	**Lyngen (1938)**	49.8%	**Rendalen (0432)**	67.8%
6	**Karlsøy (1936)**	48.9%	**Målselv (1924)**	67.7%
7	**Balsfjord (1933)**	48.3%	**Kvænangen (1943)**	66.0%
8	**Kvænangen (1943)**	44.8%	**Karlsøy (1936)**	64.9%
9	Kvæfjord (1911)	41.5%	**Lyngen (1938)**	64.2%
10	**Rendalen (0432)**	41.4%	Kvæfjord (1911)	58.8%

3.2.3. Indexing (Blocking)

To find the same person in the 1875 census and the 1900 census, it is necessary to compare the records of the two datasets. However, comparing national censuses by Cartesian product is computationally expensive and inefficient, so I used indexing to filter out potential match candidate pairs[58]. In determining the indexing conditions, I referred to the attribute distribution of matches in the training data.

[56] Percentage of total inhabitants in 1875

[57] Percentage of inhabitants aged 20 and over in 1875

[58] Two records found in the censuses that have a chance of being the same individual

In linking censuses, typical attributes for indexing are sex, year of birth, place of birth, and name[8][20][21][23][24][26]. This is because these attributes have few missing values and are time-invariant, making them useful criteria for indexing candidates with the same values over time. To determine indexing conditions for this study, I looked into the distribution of municipalities selected as training data in Section 3.2.2 with respect to general indexing attributes, as shown in Table 3.8.

Table 3.8. Match distribution of municipalities in training data, by attributes. To check if there are regional differences, Rendalen (southeastern Norway) and the municipalities of Troms (northern Norway) are separated.

	Rendalen (%)	**Troms (8 municipalities) (%)**
Population (1875): total	3,530	24,237
Population (1900): over the age of 20[59] /total	2,128 / 3,974	16,952 / 32,894
Matches between 1875-1900 Census	1,443[60] (67.81[61])	11,398 (67.24)
Matches with		
Different sexes	2 (0.14[62])	1 (0.01)
Different initials in first names	27 (1.87)	327 (2.87)
Different initials in last names	493 (34.16)	1,508 (13.23)
Different initials in first or middle names	11 (0.76)	156 (1.37)
Different initials in last or last-middle names	466 (32.29)	1,429 (12.54)
Different birthplace code (adjusted)	68 (4.71)	613 (5.38)
Different birthplace code (original)	742 (51.42)	1,575 (13.82)
Different birthplace code (adjusted, first 3 digits)	50 (3.47)	395 (3.47)
Different birthplace code (original, first 3 digits)	50 (3.47)	397 (3.48)
A 0-year difference in birth years	1,133 (78.52 | 78.52[63])	5,826 (51.11 | 51.11)
A 1-year difference in birth years	177 (12.27 | 90.78)	3,139 (27.54 | 78.65)
A 2-year difference in birth years	55 (3.81 | 94.60)	953 (8.36 | 87.02)
A 3-year difference in birth years	21 (1.45 | 96.05)	441 (3.87 | 90.88)
A 4-year difference in birth years	14 (0.97 | 97.02)	251 (2.22 | 93.09)
A 5-year difference in birth years	4 (0.28 | 97.30)	180 (1.58 | 94.67)

[59] The interval between 1875 and 1900. The possibility of error in the 5-year range in records was considered.
[60] Number of matches based on the municipality of residence in 1900. Including duplicate matches.
[61] The percentage of the population over the age of 20 in 1900.
[62] Percentage of total matches
[63] Cumulative percentage

Matches are generally consistent in ***sex***, so I decided to use the attribute sex as an indexing key.

Matches with different ***initials of the first name*** are about 2-3%. Among them, if the cases where ***the middle name's initial and the first name's initial*** are the same (e.g. 'Ole Hans' and 'Hans') are removed, the differences will be about 1%. Although the improvement was not very big, I think that it is a condition to prevent the loss of true matches, so I decided to include these cases. That is, the match between the first name's initials, and the match between the first name's initial and the middle name's initial are used as indexing keys.

Matches with different ***initials of the last name*** are found in over 30% in Rendalen and more than 10% in Troms. The main reason could be that the naming convention gradually changed during the period 1875-1900 from patronymic to inheriting the father's last name[64] [9] (IPUMS, 2010). For this reason, cases where a person's last name changes, which is generally not expected to, are observed during this period. Therefore, if the last name's initial is included as an indexing key in linking Norwegian censuses for this period, many true matches will be lost. So, I excluded the last name's initial from indexing keys, even though the similarity of last names is a typical indexing key.

For the ***birthplace code***, if the original 4-digit code from the census is used as an indexing key, the number of true matches lost due to changes in the borders of municipalities can be large. This is shown by more than 50% of true matches with different municipal codes in Rendalen, which was divided into two municipalities in 1880. Therefore, I decided to use the adjusted birth place code as the birthplace code, and to use the first 3 digits of the code as an indexing key, taking into account errors in the data.

For the ***birth year***, 91-96% of matches fall within the 3-year difference between the two censuses, and 95-97% fall within the 5-year difference between the two censuses. I decided to use the 5-year band of the birth year as an indexing key[65].

In summary, the conditions used for indexing are shown in Table 3.9. Although indexing is a very useful method to solve the computational complexity problem in large datasets, it is almost impossible for indexed potential comparison candidates to include all true matches,

[64] The practice of inheriting permanent surnames was put into law in 1923 [9].

[65] However, if the birth year gap as an indexing condition increases, cases named after siblings who died early may be included as match candidates. For these cases, more accurate estimates can be obtained if they are reviewed together with the funeral records.

so it can inevitably lose some true matches. In order to minimize this loss, I tried to use as wide indexing as possible.

Table 3.9. Conditions for indexing used in the study

• Same sex
• Same first 3 digits in adjusted birth place code
• Same first name initials, or same middle and first name initials
• Birth year difference within 5 years

When the above indexing conditions are applied, the total number of matches in the training data is 13,504 and the number of matches included in the candidate set after indexing is 12,246 (90.68%), which means about 10% of true matches are lost. This is generally similar to the results of previous studies[66] [21][24][26], but it is necessary to think about how to deal with them in the future.

Table 3.10 summarizes the application of indexing to linking the 1875-1900 censuses.

Table 3.10. Indexing applied to linking the 1875-1900 censuses

• Population at the 1875 Census	1,801,706
• Population at the 1900 Census (over aged 20)	1,279,349
• Total pairs checked for indexing	2,305,010,769,394 (=1,801,706 * 1,279,349)
• Potential match candidate pairs obtained through indexing	356,709,541
• Reduction rate of match candidate pairs by indexing	1/6450 (0.000155) (=356,709,541/2,305,010,769,394)
• Average of match candidate pairs per capita in 1875	198 (356,709,541/1,801,706)
• Average of match candidate pairs per capita in 1900 (over aged 20)	279 (356,709,541/1,279,349)

[66] The blocking strategy performance of other studies compared in [26] are 0.627 [14], 0.625[37], and 0.842 [20]. In [24] and [21, the blocking performance applied to the actual case is 0.900 and 0.909, respectively.

3.2.4. Pair Comparison

For each pair of potential match candidates obtained through indexing, it's possible to create a comparison vector by comparing the attributes of each pair. In order to get comparison vectors, I designed the features that would be used to create them. These features are used to measure the similarity between the two census records, and the number of these features becomes the number of dimensions of the comparison vector. There are two main approaches to selecting variables to use for linking: the approach that takes only time-invariant variables to minimize bias [20][23][21][8], and the approach that takes additional variables such as family members and residence to increase accuracy and linking rate [24][26][28]. In this study, I used both approaches to evaluate the differences.

First, from the variables corresponding to the persistent attributes of a person and the variables derived therefrom, I constructed **a set of time-invariant features** as shown in Table 3.11. As a person's permanent attributes I took *birth year, birth place*, and *names*.[67]

Since ***birth year*** is a numerical variable, the difference between the birth year values of two records can be used as a feature that can measure the similarity between two records.

Although ***birth place*** code is a nominal variable, the geographic or administrative distance between municipalities is associated with numerical differences in the municipal codes. For example, the first two digits of the four-digit municipal code indicate the county to which the municipality belongs, and municipalities subdivided from one are usually assigned adjacent numeric codes. Therefore, I considered that the difference in the numerical values of the birth place code was also a feature that could indirectly measure the similarity between the birth places in the two records. Both the numerical difference between the original birthplace codes and the numerical difference between the adjusted birth place codes according to the change in borders were included as features.
In addition, I thought that the commonality of the birth place, which indicates the amount of information of the birth places (how common the birthplaces of two records are), could

[67] Sex is also a permanent attribute of a person, but since the sex has already been considered in indexing and only the records with the same sex has been selected as comparison candidates, it will not be included in the comparison vector feature set.

also affect the linking. So, I included as a feature the sum of the self-information of birth places from the two records.[68]

Regarding ***names***, first, the difference between name strings can be measured using JW (Jaro-Winkler) distance[69], which is one of the methods of measuring the similarity between strings. I took JW distance between original first name strings, JW distance between standardized first name strings, JW distance between original last name strings, and JW distance between standardized last name strings as features for the comparison vector.
Also, since initial matching can be an important indicator when identifying the same person even though there are some differences in the string of names, I added as features whether the initials of the first name match each other, the initials of the last name match each other, and whether the initials of the first name and the middle name match each other.
In addition, as with the place of birth, I included the commonality of names, which indicates the amount of information of the first/last name (how common the first/last names of two records are) as features .

Next, I designed **a set of extended features** by deriving additional features from variables that are not permanent attributes of a person, but that can help identify the same person. As non-persistent variables that can help find the right candidate among multiple match candidates, I used *residence, family member information, marital status* and *family relations*.

For ***residence***, I added the JW distance between residence strings as a feature to measure the similarity between the residences in the two records.
I also used the difference in municipal codes to which the records belong as a measure of similarity between residences. As with the birth place codes, I included both the numerical difference between the original municipal codes and the numerical difference between the adjusted municipal codes as features.

Regarding ***family member information***, I counted the number of family members common in both records as the similarity of family members and added it as a feature.

As for ***marital status***, I added as a feature whether the marital statuses of the two census records are the same.

---

[68] -log(P(x)*P(y))=-log(P(x))-log(P(y)) (See footnote 24.)
This is not a feature for measuring the similarity between two records of a comparison pair, but rather a feature indicating a characteristic of the comparison pair itself. This kind of feature can also be used for machine learning models to learn match or non-match pattern classification.
[69] I used Jaro-Winkler distance as (1 - Jaro-Winker similarity).

Also, it is quite plausible that marital status changes over a period of 25 years, but it is not possible to change from married, widow(er), separated or divorced to not married. So, I added a check for such an illogical change as a feature.

As for ***family relations***, I also added as a feature whether the family relations in the two census records are the same.
It is plausible that the relationship with the head of the household would change over a span of 25 years, but any case where the head of the household becomes a child would be very rare. So, I added whether there is such an illogical change as a feature.

As described above, I built an extended feature set as shown in Table 3.11 by adding some features to the time-invariant feature set.

I created comparison vectors in which the values of the features in Tables 3.11 were calculated for each pair of comparison candidates extracted from indexing. In practice, since the time-invariant feature set is a part of the extended feature set, time-invariant feature vectors can be constructed by extracting only the time-invariant features from the extended feature vectors.

Table 3.11. Feature sets for the comparison vector

Time-invariant feature set	Extended feature set
• Birth year difference • Birth place code (original) difference • Birth place code (adjusted) difference • First name (original) JW distance • First name (standardized) JW distance • Last name (original) JW distance • Last name (standardized) JW distance • Whether first name initials match • Whether last name initials match • Whether middle name initial and first name initial match • Whether first name initial and middle name initial match • Commonality of first name • Commonality of last name • Commonality of birth place code (adjusted)	• Birth year difference • Birth place code (original) difference • Birth place code (adjusted) difference • First name (original) JW distance • First name (standardized) JW distance • Last name (original) JW distance • Last name (standardized) JW distance • Whether first name initials match • Whether last name initials match • Whether middle name initial and first name initial match • Whether first name initial and middle name initial match • Commonality of first name • Commonality of last name • Commonality of birth place code (adjusted) • Residence *(address)* JW distance • Municipality code (original) difference • Municipality code (adjusted) difference

	• Number of family members in common • Whether marital statuses match • Whether family relations match • Whether there is an illogical change in marital status • Whether there is an illogical change in family relations

3.2.5. Training machine learning models

Using the comparison vectors generated in Section 3.2.4 and the municipalities as the golden standards selected in Section 3.2.2, I constructed a dataset for training machine learning models. I obtained the match candidates of these nine municipalities through indexing and assigned match and non-match labels to each candidate pair using the HPR links as the source. Table 3.12 shows the size and class distribution of the constructed training dataset.

Table 3.12. Size and class distribution of the training dataset

Municipality	Candidate pairs	HPR links	Matched class	Unmatched class
Rendalen (0432)	377,964	1,462	0.36%	99.64%
Bardu (1922)	139,534	592	0.41%	99.59%
Målselv (1924)	292,047	1,536	0.48%	99.52%
Lenvik (1931)	848,982	2,273	0.24%	99.76%
Balsfjord (1933)	756,621	1,861	0.22%	99.78%
Karlsøy (1936)	398,596	1,063	0.24%	99.76%
Lyngen (1938)	1,048,072	2,327	0.20%	99.80%
Skjervøy (1941)	207,848	1,646	0.71%	99.29%
Kvænangen (1943)	97,191	744	0.69%	99.31%
Total	4,166,855	13,504	0.29%	99.71%

Table 3.12 shows that the class distribution is very imbalanced (0.997:0.003) because the number of true matches, i.e., the number of people identified as the same person across the two datasets, is very small compared to the number of candidate pairs. This is a general characteristic of the training set for historical record linking and needs to be considered when training linking models.

The number of data in the non-matched class is so large and, accordingly, the number of non-matched predictions is also very large. Therefore, *accuracy* ($\frac{TP+TN}{TP+TN+FP+FN}$)[70], which represents the ratio of the number of correct predictions to the total number of predictions, is not a good measure of the performance of the model. Instead, I used *precision*($\frac{TP}{TP+FP}$)[71], *recall*($\frac{TN}{TP+FN}$)[72], and their harmonic mean, *f1-score*, as metrics to measure the model's performance, focused on matched class prediction.

I split the entire training dataset into a training set and a test set in a ratio of 9:1 to evaluate the performance of the model later. Using the split training set, I trained models using a time-invariant feature set and an extended feature set, respectively. As the machine learning algorithms, I used Logistic Regression, Support Vector Machines, Random Forest, and XGBoost[73], and evaluated their performance. These algorithms are widely used in machine learning, and have also been used in previous record linkage studies that introduced machine learning approaches [23][21][38][28][24][25][26]. I performed 5-folds cross-validation[74] on the training set to obtain precision, recall, and f1 score, and selected the algorithm with the highest F1-score.

3.2.6. Classification

The model trained in Section 3.2.5 calculates the probability of being classified as a match for the comparison vector of each candidate pair. To obtain the final matches from the model's prediction, two hyper parameters should be determined.

First, I needed to determine the threshold of the probability to classify a match and a non-match. The overall match size varies depending on how the threshold is set. Lowering the threshold increases the number of correct matches, but also increases the likelihood of incorrect matches being included. On the contrary, increasing the threshold reduces the number of correct matches, but also reduces the likelihood of incorrect matches being

[70] Proportion of correct predictions in all predictions
[71] Proportion of correct predictions as positive in all predictions as positive
[72] Proportion of true correct predictions as positive in all positives
[73] I tried Bayesian hyperparameter optimization on some models, but I got no performance improvement. This may be due to the limitation of the test scope, so further examination is required in the future.
[74] The training set and test set are split 9:1.

included. To examine the effect of the threshold[75], I looked into the difference in match results by changing the threshold from 0.1 to 0.9 for the training set.

Next, since the prediction of the model is made for each pair of candidates, multiple candidates can be classified as matches for one record. Therefore, the overall match size also depends on how these multiple matches are handled. I can either select the best candidate with the highest probability, select the candidate with the highest probability only if the gap between the highest and second highest probabilities exceeds a certain threshold, i.e., only if the gap between the best and second candidates is large enough, or discard all duplicate matches. To examine the effect of the match selection option, I looked into the difference in match results by changing the ratio of the highest probability to the second highest probability as the condition for taking the candidate with the highest probability from 1 to 2, including the case of discarding all duplicate matches, that is taking only unique matches.

3.2.7. Two-way check

I performed the process from indexing 3.2.3 to classification 3.2.6 in both directions for the 1900 census based on the 1875 census, and for the 1875 census based on the 1900 census. That is, based on each person in the 1875 census, I tried to find a match among the population aged 20 and over in the 1900 census, and in the same way, based on each person aged 20 and over in the 1900 census, I tried to find a match among the population in the 1875 census. And I decided to keep only those matches that were common in the results made in both directions.

[75] The default threshold of the prediction probability of the XGBoost model is 0.5. That is, if the prediction probability calculated by the model exceeds 0.5, it is classified as a match, and if it is less than 0.5, it is classified as a non-match.

3.3. Evaluating the performance of the linking models

Since the results of linking the census cannot be verified with the correct answers, I evaluated the performance of the linking models with test sets to check the reliability of the results. As part of the evaluation of machine learning models, I additionally implemented a linking model with a rule-based method for comparison. I used these linking models to classify the test set split from the training data and the test set provided by the NHDC into match or non-match and evaluate their performance.

3.3.1. Implementation of a rule based model for comparison

To compare with the linking by the machine learning model, I implemented a linking model by the rule-based method. As rules to be used for matching, I used the JWdistance of the first name/last name, the similarity of birthplace, and the similarity of birth year, referring to the widely used ABE-JW method [20]. Since the results depend on the settings of these parameters, I tested them with multiple variations on the training set. For multiple matches, all were discarded since there was no basis for choosing a more definite match. From the test results, I implemented the model with the parameters shown in Table 3.14 that showed the best performance.

Table 3.13. Matching rules used for the rule-based linking

Matching rules	JW distance of standardized last names in two censuses	≤0.15
	JW distance of standardized first names in two censuses	≤0.15
	Difference between adjusted birth places in two censuses	0
	Difference between birth years in two censuses	≤1

3.3.2. Testing with the test set split from training data

I evaluated the performance of the models using the test set split in a ratio of 9:1 from the training data constructed in Section 3.2.5. I classified match candidates of the test set with machine learning models with different feature sets (time-invariant feature set and extended feature set) and match selections (taking unique matches and taking both unique matches and best matches for multiple matches), and the rule-based model implemented in Section 3.3.1. From the results, precision, recall, and F1 scores were calculated and compared. Added to this, I looked into the relationship between the linking results by different linking models using a Venn diagram.

3.3.3. Testing with the test set provided by the Norwegian Historical Data Center (NHDC)

The NHDC provided a test set of highly certain matches to test the performance of the trained ML models. It was generated based on the household matching method, covering the whole country. The total number of records after removing the records included in the training data in the provided dataset is 62,439. In the total dataset, the distribution of records is biased in the northern region, so I created two datasets to check the regional impact: a full test set(n=62,439) with all records and a sub test set(n=10,000) in which 500 records were randomly selected from each of 20 counties in Norway. Since the provided test set consisted of only 'matched' pairs, I created match candidate pair sets containing the records of matched pairs, for each of the full match set and the sub match set. Using the machine learning models I trained and the rule based model, I classified the candidates of the test sets and evaluated the results.

3.4. Analyzing the representativeness of linked populations

Linked data, i.e., the longitudinal life history of individuals, can be used for demographic estimation in a variety of studies. To check if there are any considerations for using linked populations for research, I attempted to examine if linked populations are sufficiently representative of the entire population. I also briefly looked into how some attributes of people actually changed between 1875 and 1900 in linked populations.

3.4.1. Comparing characteristics of linked populations

In order to examine whether the linked results are sufficiently representative of the original population, I compared the distribution of characteristics in the population of the original census and the populations linked by different methods. To evaluate whether the difference from the original population is statistically significant, I performed single-sample t-tests for continuous variables and goodness-of-fit tests for categorical variables.

3.4.2. Comparing changes in characteristics over time in linked populations

Since the individual records of the 1875 and 1900 censuses are longitudinally connected by linking, it is possible to explore how people's mutable attributes have changed over time. I examined and compared how the place of residence (municipality), marital status, and family relationship changed in linked populations by different linking methods.

Chapter 4

Results

This chapter presents the results obtained as answers to the research questions of this study. First, in order to check the reliability of the linking results, I present the performance evaluation of the models and show the results of linking the 1875-1900 censuses with these models. And then I demonstrate the representativeness of these linked data to the census population.

4.1. Performance evaluation of the models

4.1.1. Performance according to algorithms, feature sets and match selections

For a machine learning model suitable for census linking, I tested several widely used machine learning algorithms. Table 4.1 and Figure 4.1 show the performance of the machine learning models trained according to the above.

Table 4.1. Performance measurement by algorithm

Algorithms	Features	precision		recall		f1-score	
		mean	std	mean	std	mean	std
Logistic regression	time-invariant	0.74	0.01	0.47	0.01	0.58	0.01
	extended	0.81	0.01	0.57	0.01	0.67	0.01
SVM	time-invariant	0.78	0.00	0.49	0.01	0.60	0.01
	extended	0.86	0.01	0.60	0.02	0.71	0.01
Random forest	time-invariant	0.74	0.01	0.54	0.01	0.62	0.01
	extended	0.86	0.01	0.64	0.01	0.74	0.01
XGBoost	time-invariant	0.76	0.01	0.58	0.00	0.66	0.00
	extended	0.84	0.01	0.70	0.01	0.76	0.01

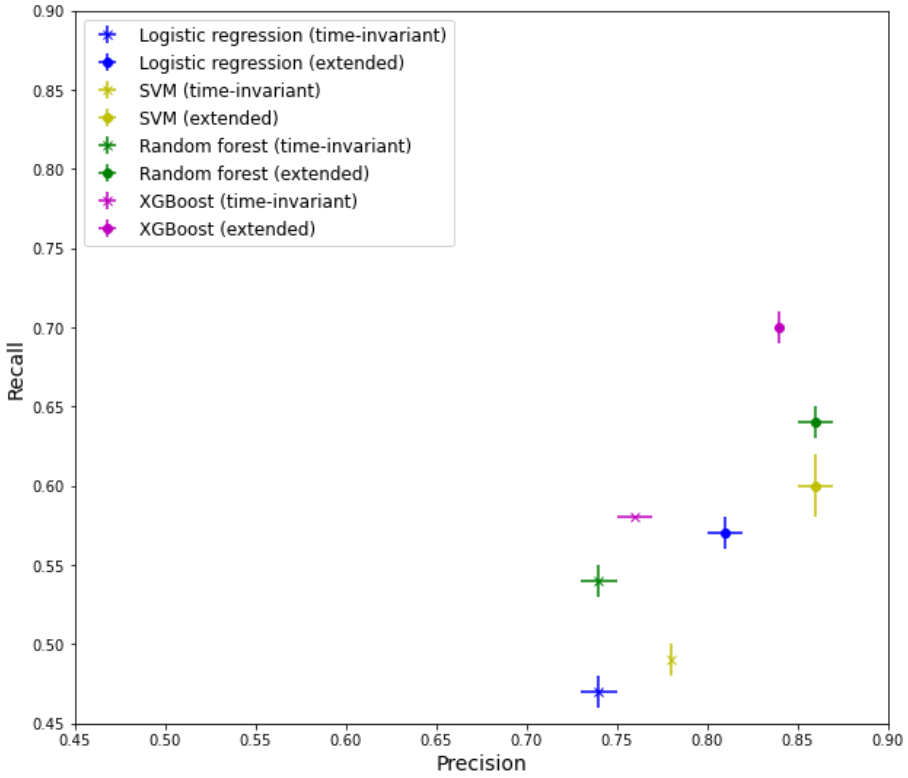


Figure 4.1. Performance measurement by algorithm. Visualization of Table 4.1. Each point and its x-axis and y-axis lines represent the mean and standard deviation in 5-folds cross-validation.

It can be seen that the performance of the models trained based on the extended feature set is better than the models trained based on the time-invariant feature set. Also, among the algorithms, the model using the XGBoost showed the best performance.

Additionally, I looked into the features that affected the prediction of the model. Figure 4.2 and Figure 4.3 show the importance of features in the trained XGBoost model. They demonstrate that differences in birthplaces and differences in last names' initials played important roles in the time-invariant feature set, and that the number of family members in common, differences in last names' initials, and differences in birthplaces played important roles in the extended feature set.

Since I placed no restrictions on last names when indexing comparison candidates, the last name factor appears to play a major role in model training. This also suggests that, although some of the population changed their last names, many still kept similar or identical last names. In addition, the birthplace factor in the Norwegian Census, which is divided into

relatively small geographic regions, also appears to play a significant role in model training. Overall, the features with high feature importance are the number of common family members, birthplace difference, gender similarity, name similarity, birth year difference, which are mainly considered features when identifying the same person in other record linkage studies.

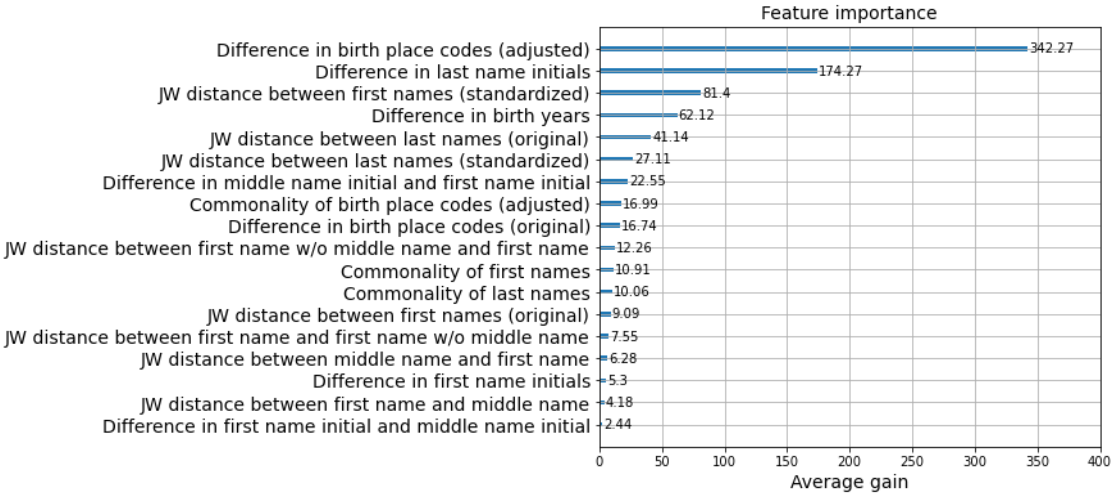


Figure 4.2. Feature importance[76] in XGBoost model trained on the dataset with time-invariant features

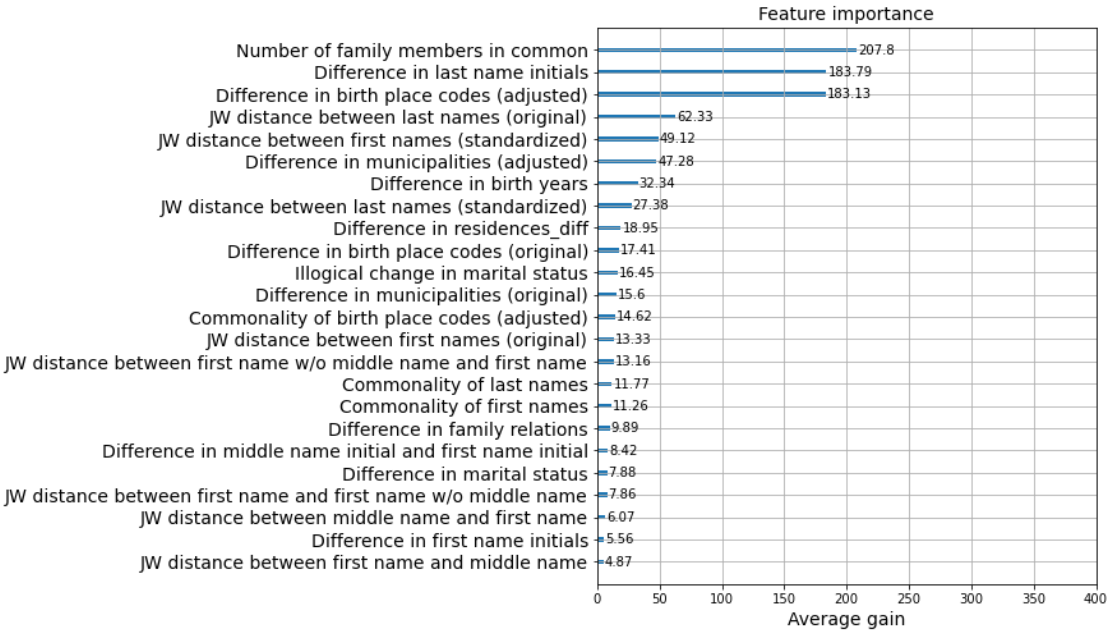


Figure 4.3. Feature importance in XGBoost model trained on the dataset with extended features

[76] The score was calculated as the average gain across all splits the feature is used in. https://xgboost.readthedocs.io/en/stable/python/python_api.html

I tested the performance on the training set[77] by changing the feature sets (time-invariant and extended feature sets) and match selection options (absolute and relative cutoffs[78] of the predicted probability) using the linking model applied with the XGBoost algorithm, which showed the highest performance. The results are shown in Tables 4.2 and 4.3.

Table 4.2. Model performance according to match selection parameters for the training set (Time-invariant feature based model). The highest value across the test and the highest value in unique matches are shown in color.

Absolute cutoff		**Relative cutoff**											
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	Unique
0.1	Precision	0.72	0.73	0.74	0.74	0.74	0.75	0.75	0.75	0.75	0.75	0.76	0.77
	Recall	0.79	0.76	0.75	0.74	0.73	0.73	0.72	0.71	0.70	0.70	0.69	0.51
	F1-score	**0.75**	**0.75**	**0.74**	**0.74**	**0.74**	**0.74**	**0.73**	**0.73**	**0.73**	**0.73**	**0.72**	**0.61**
0.2	Precision	0.77	0.78	0.79	0.79	0.79	0.79	0.80	0.80	0.80	0.80	0.80	0.81
	Recall	0.74	0.71	0.71	0.70	0.69	0.68	0.67	0.67	0.66	0.65	0.65	0.58
	F1-score	**0.76**	**0.75**	**0.74**	**0.74**	**0.74**	**0.73**	**0.73**	**0.73**	**0.72**	**0.72**	**0.72**	**0.67**
0.3	Precision	0.80	0.81	0.82	0.82	0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.83
	Recall	0.70	0.67	0.66	0.66	0.65	0.64	0.63	0.63	0.62	0.62	0.61	0.59
	F1-score	**0.74**	**0.74**	**0.73**	**0.73**	**0.72**	**0.72**	**0.72**	**0.71**	**0.71**	**0.71**	**0.71**	**0.69**
0.4	Precision	0.82	0.84	0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
	Recall	0.65	0.62	0.61	0.61	0.60	0.59	0.59	0.58	0.58	0.58	0.58	0.57
	F1-score	**0.72**	**0.71**	**0.71**	**0.71**	**0.70**	**0.70**	**0.70**	**0.69**	**0.69**	**0.69**	**0.69**	**0.68**
0.5	Precision	0.85	0.86	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
	Recall	0.58	0.56	0.55	0.55	0.54	0.54	0.54	0.54	0.53	0.53	0.53	0.53
	F1-score	**0.69**	**0.68**	**0.68**	**0.67**	**0.67**	**0.67**	**0.66**	**0.66**	**0.66**	**0.66**	**0.66**	**0.66**
0.6	Precision	0.87	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
	Recall	0.51	0.49	0.49	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
	F1-score	**0.64**	**0.63**	**0.63**	**0.62**	**0.62**	**0.62**	**0.62**	**0.62**	**0.62**	**0.62**	**0.62**	**0.62**
0.7	Precision	0.90	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	Recall	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
	F1-score	**0.56**	**0.55**	**0.55**	**0.55**	**0.55**	**0.55**	**0.55**	**0.55**	**0.55**	**0.55**	**0.55**	**0.55**
0.8	Precision	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	Recall	0.30	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
	F1-score	**0.45**	**0.44**	**0.44**	**0.44**	**0.44**	**0.44**	**0.44**	**0.44**	**0.44**	**0.44**	**0.44**	**0.44**
0.9	Precision	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	Recall	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
	F1-score	**0.23**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**

[77] I used the training set here with the intention of fitting the model. Testing on the test sets to evaluate the model performance is described in Sections 4.1.2 and 4.1.3

[78] Absolute cutoff means the threshold value that a match candidate's predicted probability must exceed to be classified as a match. Relative cutoff means the threshold of the ratio for a candidate with the highest probability to be classified as a match, if the ratio of the highest probability to the second highest probability exceeds this value, in the case of multiple matches.

Table 4.3. Model performance according to match selection parameters for the training set (Extended based model). The highest value across the test and the highest value in unique matches are shown in color.

Absolute cutoff		Relative cutoff											
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	Unique
0.1	Precision	0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.84	0.84	0.84
	Recall	0.86	0.85	0.84	0.83	0.83	0.82	0.81	0.81	0.80	0.80	0.79	0.59
	F1-score	**0.84**	**0.84**	**0.83**	**0.83**	**0.83**	**0.83**	**0.82**	**0.82**	**0.82**	**0.82**	**0.81**	**0.69**
0.2	Precision	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.88	0.87
	Recall	0.83	0.82	0.81	0.80	0.79	0.79	0.78	0.78	0.77	0.76	0.76	0.67
	F1-score	**0.85**	**0.84**	**0.84**	**0.83**	**0.83**	**0.83**	**0.82**	**0.82**	**0.82**	**0.82**	**0.81**	**0.76**
0.3	Precision	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
	Recall	0.80	0.79	0.78	0.77	0.77	0.76	0.75	0.75	0.74	0.74	0.73	0.69
	F1-score	**0.84**	**0.84**	**0.83**	**0.83**	**0.83**	**0.82**	**0.82**	**0.82**	**0.82**	**0.81**	**0.81**	**0.78**
0.4	Precision	0.91	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
	Recall	0.77	0.75	0.74	0.74	0.73	0.73	0.72	0.72	0.71	0.71	0.70	0.69
	F1-score	**0.83**	**0.83**	**0.82**	**0.82**	**0.81**	**0.81**	**0.81**	**0.80**	**0.80**	**0.80**	**0.80**	**0.79**
0.5	Precision	0.93	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	Recall	0.73	0.72	0.71	0.70	0.70	0.69	0.69	0.68	0.68	0.68	0.68	0.68
	F1-score	**0.82**	**0.81**	**0.81**	**0.80**	**0.80**	**0.80**	**0.79**	**0.79**	**0.79**	**0.79**	**0.78**	**0.78**
0.6	Precision	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	Recall	0.68	0.67	0.67	0.66	0.66	0.65	0.65	0.65	0.65	0.65	0.65	0.65
	F1-score	**0.79**	**0.79**	**0.78**	**0.78**	**0.78**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**
0.7	Precision	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
	Recall	**0.63**	**0.62**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**
	F1-score	**0.76**	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**
0.8	Precision	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	Recall	0.56	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
	F1-score	**0.71**	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**
0.9	Precision	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Recall	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44
	F1-score	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**

Across the board, it can be seen that raising the absolute cutoff improves precision, and lowering the relative cutoff improves recall. Taking only unique matches improves precision, and taking both unique and best matches (relative cutoff: 1) improves recall. Since the decrease in recall is greater than the increase in precision when the relative cutoff increases, the F1-score, which is the harmonic mean of precision and recall, is higher when both unique and best matches are taken. When it comes to the feature set used for training, it can be seen that the model using the extended feature set has higher performance than the model using the time invariant feature set.

The performance on the relative cutoff is highest when the value is 1, that is, taking the candidate with the highest probability regardless of the difference from the candidate with the second highest probability. The performance on absolute cutoff is shown in Figure 4.4.

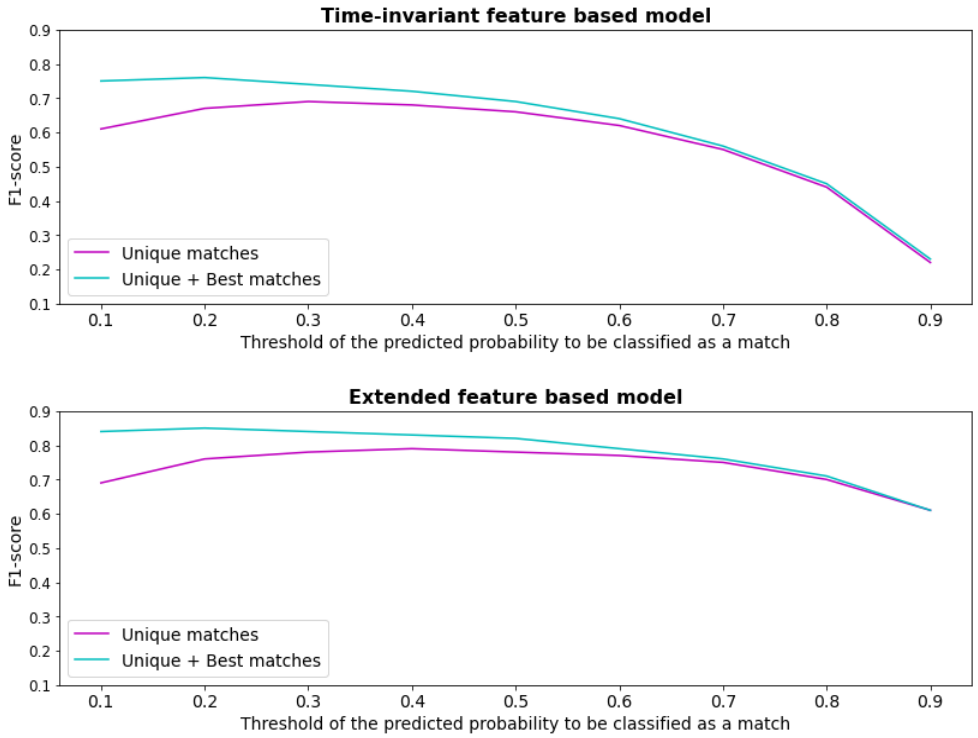


Figure 4.4. Performance of the models with the predicted probability threshold variation

Overall, it can be seen that the performance of the extended feature based model is more stable with respect to the change in the threshold of the prediction probability. Since the extended feature based model uses more information, there is less ambiguity and fewer multiple matches. Thus it appears to be relatively less sensitive to absolute cutoff changes that cause multiple matches. Also, in case of taking both unique and best matches, the performance is highest when the predicted probability threshold is 0.2, whereas in the case of taking only unique matches, the performance is the highest when the predicted probability threshold is 0.3-0.4. It can be guessed that when taking only unique matches, a slightly higher probability threshold is preferred to reduce multiple matches.

4.1.2. Performance on the test set from training data

I tested different linking models on the test set that was split from the training data . Table 4.4 shows the performance of the results.

Table 4.4. Performance of different linking models on the test set split from the training data

			Precision	Recall	F1-score
Machine Learning[79]	Time-invariant	Unique	0.67	0.67	0.67
		Unique+Best	0.60	0.75	0.67
	Extended	Unique	0.82	0.72	0.77
		Unique+Best	0.72	0.83	0.77
Rule-based	ABE-JW		0.56	0.58	0.57

The results show that the performance of the machine learning models is higher than that of the rule-based model, and the performance of the extended feature based model is higher than that of the time-invariant feature based model. As for match selection, taking only unique matches has high precision, and taking both unique and best matches has high recall. There is a trade-off between precision and recall, so F1-score is the same.

Figure 4.5 shows the relationship between the results linked by different models as a Venn diagram.

[79] Match selecting hyperparameters (absolute cutoff α and relative cutoff beta β) were selected according to the model fitting results in Section 4.1.1.
Time-invariant feature based model, with unique matches: α: 0.3, β:-
Time-invariant feature based model, with unique and best matches: α: 0.2, β:1
Extended feature based model, with unique matches: α: 0.4, β:-
Extended feature based model, with unique and best matches :α: 0.2, β:1

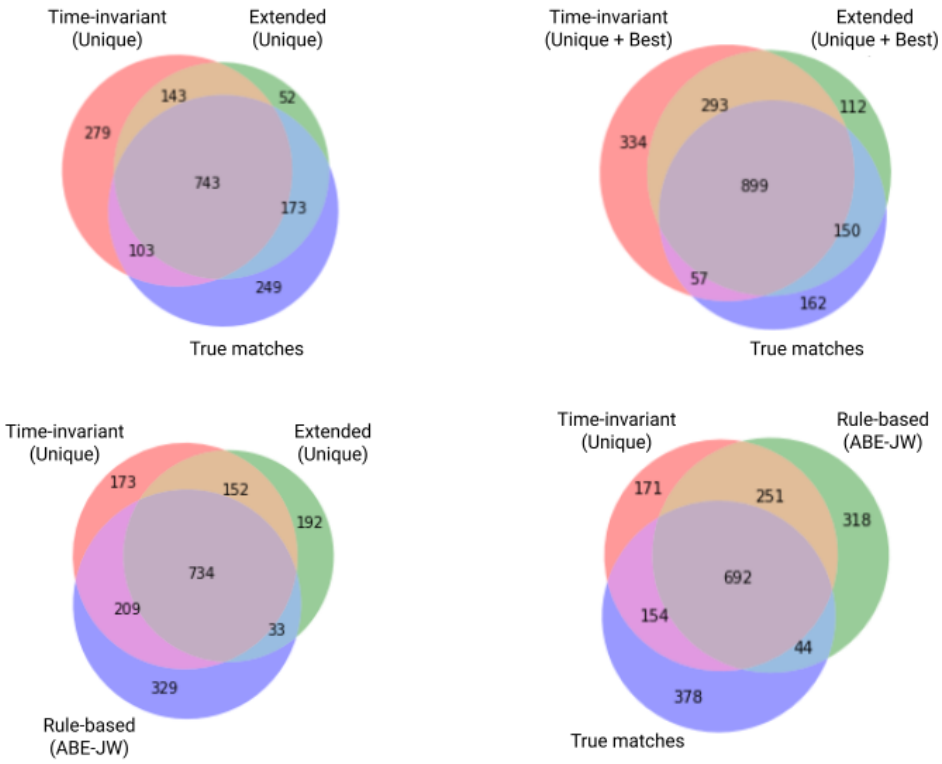


Figure 4.5. Relationship between the results linked by different models and true matches (for the test set split from training data)

The linked results of the time-invariant feature based model and the extended feature based model are somewhat different, but the area where the extended feature based model overlaps with true matches is larger. Regarding the effect of match selection, the area shared by the time-invariant feature based model and the extended feature based model is larger when taking both unique and best matches. When comparing the results of the three models, there are also some differences between the three areas. Since the features used in the rule-based model are time-invariant, the rule-based model shares a larger area with the time-invariant feature based model. The machine learning model closest to the rules of the rule-based model is a model based on time-invariant features, with only unique matches. Comparing these two results, the time-invariant feature based model has a larger overlap with true matches.

The performance according to match selection is slightly different from that of the training set, so I checked the effect of match selection parameters (absolute cutoff and relative cutoff) on the performance of models for the test set as shown in Tables 4.5. and 4.6.

Table 4.5. Model performance according to match selection parameters for the test set (Time-invariant feature based model). The highest value across the test and the highest value in unique matches are shown in color.

Absolute cutoff		Relative cutoff											
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	Unique
0.1	F1 score	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.62
0.2		0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.66
0.3		**0.68**	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	**0.67**
0.4		0.67	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
0.5		0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.64	0.64	0.64	0.64	0.64
0.6		0.62	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
0.7		0.56	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
0.8		0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44
0.9		0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23

Table 4.6. Model performance according to match selection parameters for the test set (Extended based model). The highest value across the test and the highest value in unique matches are shown in color.

Absolute cutoff		Relative cutoff											
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	Unique
0.1	F1 score	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71
0.2		0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.76	0.76
0.3		**0.78**	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	**0.77**
0.4		**0.78**	0.78	0.78	0.78	0.78	0.78	0.78	0.77	0.77	0.77	0.77	**0.77**
0.5		0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.76	0.76	0.76	0.76	0.76
0.6		0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
0.7		0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
0.8		0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
0.9		0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59

The results show that there is little difference according to the change in the relative cutoff. This is because the size of the test set is relatively small (candidate pairs = 416,686, true match pairs = 1,268), so there are not many multiple matches to deal with relative cutoff. The absolute cutoff to obtain the best performance has also been increased to 0.3-0.4.

4.1.3. Performance on the test set provided by the NHDC

With the test set provided by the NHDC for the final testing of the linking models, I evaluated the performance of different models with the same settings as in Section 4.1.2. The results for the provided full test set and the sub test set randomly selected by 500 records for each county are shown in Tables 4.7 and 4.8.

Table 4.7. Performance of different linking models on the test set provided by the NHDC (full-testset)

			Precision	Recall	F1-score
Machine Learning[80]	Time-invariant	Unique	0.98	0.63	0.77
		Unique+Best	0.94	0.74	0.82
	Extended	Unique	0.99	0.76	0.86
		Unique+Best	0.99	0.90	0.94
Rule-based	ABE-JW		0.97	0.59	0.73

Table 4.8. Performance of different linking models on the test set provided by the NHDC (sub-testset)

			Precision	Recall	F1-score
Machine Learning	Time-invariant	Unique	0.97	0.65	0.78
		Unique+Best	0.94	0.75	0.83
	Extended	Unique	0.99	0.76	0.86
		Unique+Best	0.98	0.90	0.94
Rule-based	ABE-JW		0.97	0.61	0.75

The above results show that the models have higher performance on the test set provided by the NHDC than the test set from the training data. This may be partly because the true matches provided by the NHDC were relatively easy to be classified as matches because they were highly certain based on the family matching method.

[80] Match selecting hyperparameters (absolute cutoff α and relative cutoff β) were selected according to the model fitting results in Section 4.1.1.
Time-invariant feature based model, with unique matches: α: 0.3, β:-
Time-invariant feature based model, with unique and best matches: α: 0.2, β:1
Extended feature based model, with unique matches: α: 0.4, β:-
Extended feature based model, with unique and best matches :α: 0.2, β:1

As with the test set of training data, the performance of machine learning models is higher than that of the rule-based model, but the difference is reduced. The performance of the extended feature based models is still higher than that of the time-invariant feature based models. In particular, the extended feature based model with unique and best matches significantly improves both precision and recall over 0.90. This is because matches in the test set were verified based on the family matching method, so it can be inferred that the extended feature based model using family member information has some advantages.

On the other hand, unlike the test in Section 4.1.2, the performance difference between models according to match selection is large. The performance of the model with unique and best matches is higher than the model with only unique matches. The tradeoff between precision and recall is still there, but the change in recall is larger than the change in precision.

There is little difference in performance between the full test set and the sub test set. The effect of regional bias in the records does not seem to be significant.

The results of Sections 4.1.1 and 4.1.2 show that the performance of the model is dependent on the test set. But nevertheless, the relative differences in the performance of each model remain mostly consistent. This reveals that the trained models in the study show a certain level of consistent performance over different test sets.

The relationship between the linking results of different models is shown in Figure 4.6.

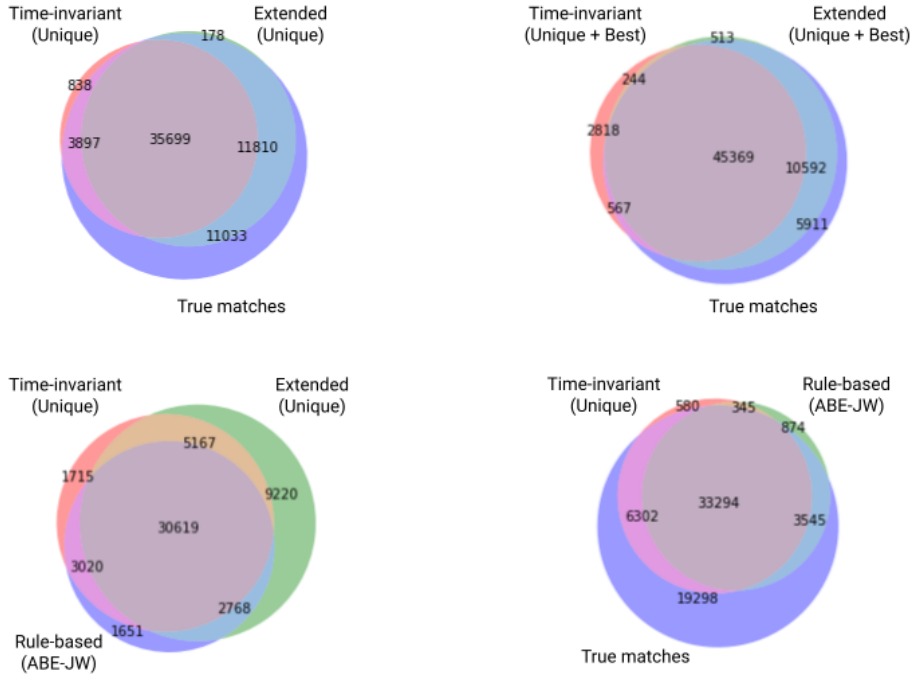


Figure 4.6. Relationship between the results linked by different models and true matches (for the test set[81] provided by the NHDC)

As can be expected from the improved performance of the models, the overlapping area is much larger than the test result of Section 4.1.2. This means that the predictions of each model are more likely to agree with each other. Another point to note is that the number of predictions by the extended feature based model has increased, and the increased predictions generally correspond to true matches. It can be inferred that there are matches in the test set that fit the classification algorithm of the extended feature based model.

Additionally, I examined the performance of the model according to the change of match selection parameters as shown in Tables 4.9-4.12.

[81] It is the results for the full test set. The results for the sub-test set are almost similar to this one, so I do not include them here.

Table 4.9. Model performance according to match selection parameters for the full test set provided by the NHDC (Time-invariant feature based model). The highest value across the test and the highest value in unique matches are shown in color.

Absolute cutoff		**Relative cutoff**											
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	Unique
0.1	F1 score	**0.84**	0.84	0.84	0.84	0.83	0.83	0.83	0.83	0.83	0.82	0.82	0.74
0.2		0.82	0.82	0.82	0.82	0.81	0.81	0.81	0.81	0.81	0.80	0.80	**0.77**
0.3		0.80	0.80	0.80	0.79	0.79	0.79	0.79	0.79	0.78	0.78	0.78	**0.77**
0.4		0.78	0.77	0.77	0.77	0.76	0.76	0.76	0.76	0.76	0.75	0.75	0.75
0.5		0.74	0.73	0.73	0.73	0.73	0.72	0.72	0.72	0.72	0.72	0.72	0.72
0.6		0.69	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
0.7		0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
0.8		0.49	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
0.9		0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25

Table 4.10. Model performance according to match selection parameters for the full test set provided by the NHDC (Extended based model). The highest value across the test and the highest value in unique matches are shown in color.

Absolute cutoff		**Relative cutoff**											
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	Unique
0.1	F1 score	**0.94**	0.93	0.93	0.92	0.92	0.91	0.91	0.90	0.90	0.90	0.89	0.76
0.2		**0.94**	0.93	0.93	0.92	0.91	0.91	0.91	0.90	0.90	0.89	0.89	0.82
0.3		**0.94**	0.93	0.92	0.92	0.91	0.91	0.90	0.90	0.89	0.89	0.89	0.84
0.4		0.93	0.93	0.92	0.91	0.91	0.90	0.90	0.89	0.89	0.88	0.88	0.86
0.5		0.93	0.92	0.92	0.91	0.90	0.90	0.89	0.89	0.88	0.88	0.88	0.88
0.6		0.93	0.92	0.91	0.90	0.90	0.89	0.89	0.88	0.88	0.88	0.88	0.88
0.7		0.92	0.91	0.90	0.90	0.89	0.89	0.89	0.89	0.89	0.89	0.89	**0.89**
0.8		0.91	0.90	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	**0.89**
0.9		0.89	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88

Table 4.11. Model performance according to match selection parameters for the sub test set provided by the NHDC (Time-invariant feature based model). The highest value across the test and the highest value in unique matches are shown in color.

Absolute cutoff		Relative cutoff											
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	Unique
0.1	F1 score	**0.85**	0.85	0.84	0.84	0.84	0.84	0.84	0.83	0.83	0.83	0.83	0.76
0.2		0.83	0.83	0.83	0.82	0.82	0.82	0.82	0.82	0.81	0.81	0.81	**0.78**
0.3		0.81	0.81	0.80	0.80	0.80	0.80	0.80	0.79	0.79	0.79	0.79	**0.78**
0.4		0.78	0.78	0.78	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.76	0.76
0.5		0.75	0.74	0.74	0.74	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.73
0.6		0.70	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69
0.7		0.63	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62
0.8		0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
0.9		0.28	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27

Table 4.12. Model performance according to match selection parameters for the sub test set provided by the NHDC (Extended based model). The highest value across the test and the highest value in unique matches are shown in color.

Absolute cutoff		Relative cutoff											
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	Unique
0.1	F1 score	**0.94**	0.93	0.93	0.92	0.92	0.91	0.91	0.90	0.90	0.89	0.89	0.76
0.2		**0.94**	0.93	0.92	0.92	0.91	0.91	0.90	0.90	0.89	0.89	0.89	0.82
0.3		0.93	0.93	0.92	0.92	0.91	0.91	0.90	0.90	0.89	0.89	0.88	0.84
0.4		0.93	0.92	0.92	0.91	0.91	0.90	0.90	0.89	0.89	0.88	0.88	0.86
0.5		0.93	0.92	0.91	0.91	0.90	0.90	0.89	0.89	0.88	0.88	0.88	0.88
0.6		0.92	0.92	0.91	0.90	0.90	0.89	0.89	0.89	0.89	0.89	0.89	**0.89**
0.7		0.92	0.91	0.90	0.90	0.89	0.89	0.89	0.89	0.89	0.89	0.89	**0.89**
0.8		0.91	0.90	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	**0.89**
0.9		0.89	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88

Since the size of the test sets is larger (candidate pairs in the full test set = 11,376,146, true match pairs in the full test set = 62,439, candidate pairs in the sub test set = 1,918,356, true match pairs in the sub test set = 10,000) than the previous one, the number of multiple matches has increased, resulting in a performance difference according to the relative cutoff. Since multiple matches are reduced when the absolute cutoff is high, there is little difference according to the relative cutoff.

Lowering the absolute cutoff increases recall when both unique and best matches are taken at the expense of precision. In a large dataset, there are likely to be many multiple matches, so

the increase in recall is large when the absolute cutoff is lowered. This results in higher performance at lower absolute cutoffs. For this reason, it can be seen that the absolute cutoff to obtain the best performance has been also reduced to 0.1-0.2 when both unique and best matches are taken.

On the other hand, when only unique matches are taken, the absolute cutoff for obtaining the best performance for the time-invariant feature based model is 0.2-0.3, and for the extended feature based model it is 0.7-0.8. This seems to be because, when extended features are used for classification for this test set, there are not many multiple matches and there are many candidates with high prediction probability, so precision can be improved in increasing the absolute cutoff.

4.2. Results of linking the 1875-1900 censuses

The main goal of this study is to increase the linking rate between the 1875-1900 censuses in HPR. Table 4.13 and Figure 4.7 shows the results of linking the 1875 and 1900 census using the linking models.

Table 4.13. Linking rates for the 1875 and 1900 censuses linked by different models.

	Machine Learning[82]					
	Time-invariant		Extended		Rule-based	HPR
	Unique	Unique+Best	Unique	Unique+Best	(ABE-JW)	
SHIP REGISTER	7.2%[83]	17.3%	0.3%	1.1%	9.0%	0.1%
ØSTFOLD	18.9%	34.0%	23.6%	36.2%	20.7%	4.7%
AKERSHUS	19.1%	36.1%	22.6%	37.8%	21.5%	4.3%
OSLO	15.2%	26.5%	7.5%	18.1%	16.6%	2.7%
HEDMARK	20.7%	40.0%	28.7%	51.3%	21.7%	7.0%
OPPLAND	20.0%	38.9%	24.4%	47.8%	20.8%	4.8%
BUSKERUD	19.8%	36.4%	23.9%	43.2%	21.1%	3.8%
VESTFOLD	22.8%	37.9%	27.1%	39.8%	23.3%	5.6%
TELEMARK	21.9%	39.4%	28.3%	43.1%	23.5%	3.8%
AUST-AGDER	24.2%	39.8%	30.7%	43.5%	24.8%	4.0%
VEST-AGDER	27.9%	44.7%	33.4%	49.2%	29.3%	4.4%
ROGALAND	22.3%	37.2%	28.8%	43.5%	23.0%	3.3%
HORDALAND	21.1%	37.2%	28.7%	46.3%	22.7%	4.9%
BERGEN	15.4%	27.4%	11.1%	25.3%	17.9%	1.4%
SOGN OG FJORDANE	17.3%	29.8%	25.5%	38.0%	17.3%	4.9%
MØRE OG ROMSDAL	23.1%	38.6%	30.3%	47.4%	22.8%	4.3%
SØR-TRØNDELAG	19.6%	33.7%	23.5%	38.8%	19.1%	6.1%
NORD-TRØNDELAG	26.7%	43.5%	34.3%	49.8%	25.6%	7.2%
NORDLAND	27.8%	44.5%	33.5%	46.8%	28.4%	6.6%
TROMS	33.6%	51.1%	37.1%	52.6%	30.5%	52.2%
FINNMARK	19.9%	35.1%	21.8%	33.7%	19.3%	9.1%
TOTAL	21.2%	36.7%	25.0%	40.2%	22.0%	6.2%

[82] Match selecting hyperparameters (absolute cutoff α and relative cutoff β) were selected based on the performance (F1-score) on the entire training data.
Time-invariant feature based model, with unique matches: α: 0.3, β:-
Time-invariant feature based model, with unique and best matches: α: 0.1, β:1
Extended feature based model, with unique matches: α: 0.4, β:-
Extended feature based model, with unique and best matches :α: 0.1, β:1

[83] The proportion of linked population to linkable population (aged over 20) in 1900

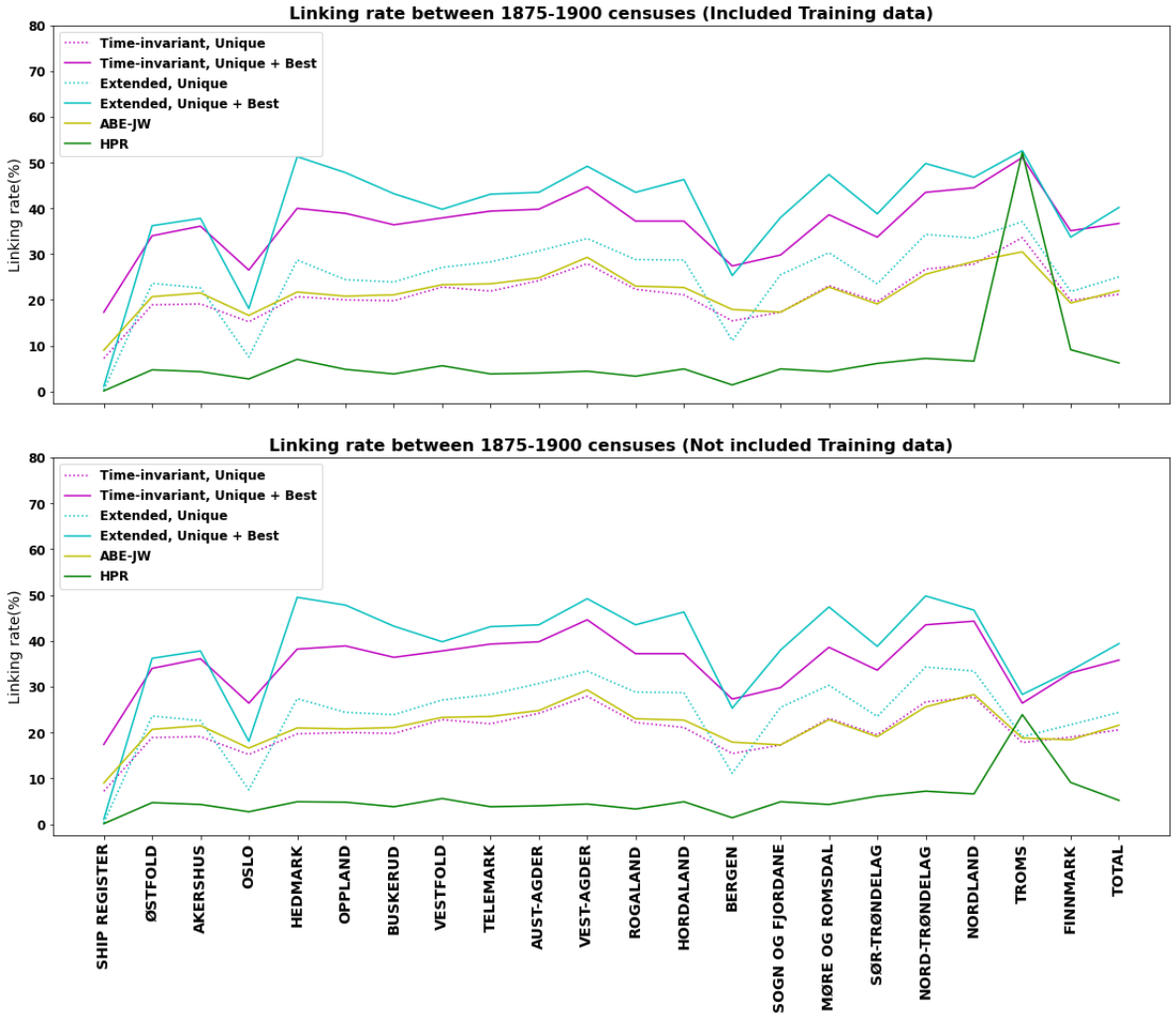


Figure 4.7. Linking rates for the 1875 and 1900 censuses linked by different models. Above is the linking rate for nationwide census data including training data to check the overall linking rate, and below is the linking rate for census data excluding training data to check the performance of the model.

The above results show that the nationwide census linking rate can be significantly improved compared to the current HPR by using machine learning models.

In the extended feature-based model, when only unique matches are taken, it is improved by about 19%, and when both unique matches and best matches are taken, it is improved by about 34%. In the time-invariant feature-based model, it is improved by about 15% when only the unique matches are taken, and by about 31% when both the unique matches and

the best matches are taken. When the rule-based model is used, it is also improved by about 15%, which is a slightly higher linking rate than when taking only unique matches, based on time-invariant features. However, as shown earlier, in terms of performance (precision and recall), the performance of the machine learning model is higher than that of the rule-based model.

When it comes to regional differences, in Oslo and Bergen, the extended feature based model shows a lower linking rate than the time-invariant feature based model. This seems to be because there is a large migrant population in these areas[84], making it relatively difficult to maintain the number of common family members or residential features used in the extended feature-based model. I was concerned that regional bias in the training data would cause regional bias in the results, but it doesn't seem to be noticeable.

The relationship between the linking results by different models is shown in Figure 4.8.

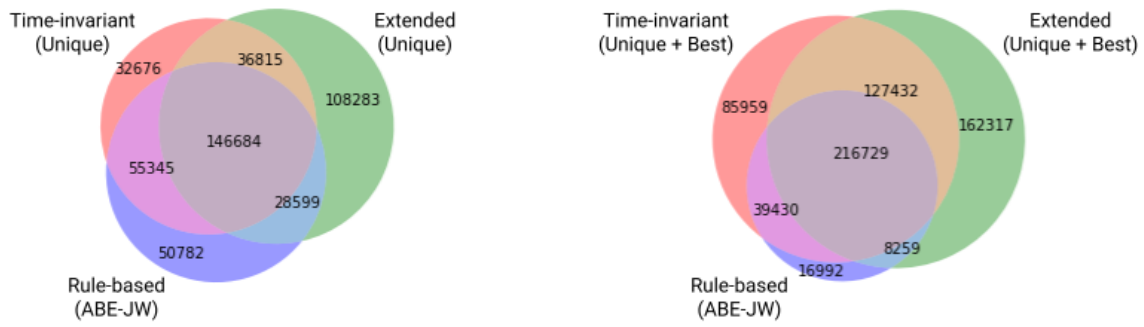


Figure 4.8. Relationship between the results linked by different models (for the 1875-1900 censuses)

It can be seen that the difference between each group is larger than in the results for the NHDC-provided test set in Section 4.1.3. It is a little closer to the results for the test set from training data in Section 4.1.2. It can be speculated that there is some difference between the candidate pairs in the NHDC-provided test set and the candidate pairs in the census, and the

[84] A large number of Akershus' population were incorporated or migrated to Oslo between 1875-1900 (footnote 45, https://www.ssb.no/a/histstat/rapp/rapp_199913.pdf), and Oslo's population increased rapidly between 1875-1900 (Table 3.1). Bergen was merged with the rural area of Bergen[*Bergen landdistrikt*] in 1877, which may also have influenced the linking of Bergen (https://www.ssb.no/a/histstat/rapp/rapp_199913.pdf)

NHDC-provided test set consists of candidate pairs that are easier to classify (more likely to agree with each other) than the actual census.

The proportion of common areas according to match selection is larger when both unique and best matches are taken, as in Sections 4.1.2 and 4.1.3. Taking only the population shared by the linking of the time-invariant feature based model and the linking of the extended feature based model as the final result is one possible linking strategy, which is more conservative considering accuracy.

In cases where the linking results are different, which result is more accurate? From an empirical examination of some cases, the linked results by the extended feature based model were more accurate, and Tables 4.4,4.7 and 4.8 also prove this.

4.3. Representativeness of linked populations

4.3.1. Characteristics of linked populations

In order to investigate whether the linked results are sufficiently representative of the original population, I examined the characteristics of the populations linked by different machine learning models and a rule-based method as shown in Table 4.14.

Table 4.14. Characteristics in the population of the census and in populations linked by different linking methods

			Machine Learning models				Rule-based
		1900 census (25+ years old)	Time-invariant		Extended		ABE-JW
			Unique	Unique+Best	Unique	Unique+Best	
Number of population[85]		1,077,990	271,520	469,550	320,381	514,737	281,410
Age	Mean (std)	46.6 (15.7)	48.3 (16.3)	47.6 (16.1)	51.4 (16.6)	49.1 (16.4)	48.4 (16.3)
	25-45	51.0%	45.9%	47.6%	37.2%	43.1%	45.5%
	45-60	26.5%	27.1%	26.9%	28.2%	27.3%	27.2%
	60-	22.5%	26.7%	24.9%	34.1%	28.8%	26.8%
Family Size	Mean (std)	6.0 (4.8)	5.8 (4.2)	5.9 (4.3)	5.6 (3.2)	5.7 (3.4)	5.8 (4.3)
	1	2.7%	2.5%	2.5%	2.4%	2.3%	2.6%
	2-4	35.7%	36.3%	35.9%	37.6%	36.6%	36.8%
	5-9	50.9%	51.8%	51.9%	51.3%	52.1%	51.2%
	10-	10.7%	9.4%	9.7%	8.7%	9.0%	9.4%
Sex	Female	52.3%	47.1%	49.3%	48.0%	49.7%	45.3%
	Male	47.7%	52.9%	50.7%	52.0%	50.3%	54.7%
Marital Status	Unmarried	22.7%	22.9%	23.0%	20.7%	22.3%	23.1%
	Married	64.6%	64.3%	64.4%	64.6%	64.2%	64.0%
	Widow(er)	12.0%	12.3%	12.0%	14.2%	13.0%	12.3%
	Separated	0.1%	0.1%	0.1%	0.0%	0.1%	0.1%
	Divorced	0.1%	0.1%	0.1%	0.0%	0.1%	0.1%
	Unknown	0.6%	0.5%	0.5%	0.4%	0.4%	0.5%
Family Relation	Householder	34.1%	40.1%	37.8%	39.8%	37.7%	41.3%
	Spouse	35.2%	30.6%	32.4%	32.2%	32.9%	29.1%
	Child	7.9%	10.3%	9.9%	11.4%	12.5%	10.2%
	Boarder	16.0%	13.4%	14.1%	11.2%	11.7%	13.8%

[85] The linked populations here are for the whole country, including data from the municipalities used in the ground truth set.

Place of Birth	Eastern[86]	41.5%	40.1%	42.2%	38.0%	41.0%	41.4%
	Southern	7.5%	9.1%	8.5%	9.1%	8.2%	9.0%
	Western	25.6%	24.5%	24.3%	26.2%	25.9%	24.6%
	Mid	10.4%	11.0%	10.6%	11.3%	11.0%	10.3%
	Northern	10.3%	14.4%	13.1%	13.9%	12.3%	13.5%
Municipality	Eastern	45.5%	41.4%	43.6%	38.9%	42.2%	42.8%
	Southern	7.2%	8.7%	8.2%	9.1%	8.2%	8.7%
	Western	25.1%	23.9%	23.7%	26.1%	25.8%	24.0%
	Mid	10.3%	10.6%	10.3%	11.2%	10.8%	9.9%
	Northern	11.1%	14.9%	13.7%	14.7%	12.9%	14.1%
Living in	City[87]	27.6%	22.4%	22.1%	14.4%	16.8%	23.2%
	Rural area	72.4%	77.6%	77.9%	85.6%	83.2%	76.8%
Living in place of birth		55.2%	64.4%	63.4%	72.4%	74.8%	64.1%

* All p-values of single-sample t-tests (continuous variables) and goodness-of-fit tests (categorical variables) for distributions of characteristics: <0.01

Across the board, it is difficult to say that linked populations are statistically representative of the population of the 1900 census. Linked populations are older, and have a higher proportion of males, people living in rural areas, and people living in their birth place than the population of the census. In relation to the household head, the ratio of children is higher and the ratio of boarders is lower than the census. The proportion of people whose place of birth or residence is northern Norway is also higher than that of the census. This difference in representativeness between the census and linked populations has already been mentioned in [22][24][28]. It can be easily inferred that characteristics that have a higher proportion in the linked population (older people, men, rural residents, residents in their birth place, children of household heads) are characteristics that are easier to link.

Regarding the high proportion of people born or residing in northern Norway, there may have been a regional bias due to the training data being mainly from the northern regions. However, similar results are obtained with the rule based linking, so there may be certain

[86] The 20 counties are grouped according to the following commonly accepted Norwegian regional divisions.
Eastern Norway: ØSTFOLD, AKERSHUS, OSLO, HEDMARK, OPPLAND, BUSKERUD, VESTFOLD, TELEMARK
Southern Norway: AUST-AGDER, VEST-AGDER
Western Norway: ROGALAND, HORDALAND, BERGEN, SOGN OG FJORDANE, MØRE OG ROMSDAL
Mid Norway: SØR-TRØNDELAG, NORD-TRØNDELAG
Northern Norway: NORDLAND, TROMS, FINNMARK
(https://en.wikipedia.org/wiki/Regions_of_Norway)

[87] Cities are assigned 0 as the third digit in the 4-digit municipality code.
(https://www.ssb.no/a/histstat/rapp/rapp_199913.pdf)

characteristics unique to northern Norway. One possible assumption is that the records of names, birth years, and places of birth used as key variables for linking may be more accurate in northern Norway. The census was filled in by each household's head in urban areas and by census managers in rural areas[88], which is likely to have been more consistent when filled in by the same person. As shown in Table 4.15, northern Norway has a lower proportion of urban areas than other regions, so the records may have been more consistent than other regions.

Table 4.15. Percentage of cities by region

	1900 census	Time-invariant		Extended		ABE-JW
		Unique	Unique+Best	Unique	Unique+Best	
Eastern	36.5%	31.9%	30.2%	20.5%	22.5%	32.4%
Southern	22.0%	19.8%	19.0%	15.7%	16.3%	20.2%
Western	24.5%	19.2%	19.5%	12.5%	15.5%	20.0%
Mid	19.6%	17.4%	16.5%	9.7%	12.1%	17.8%
Northern	8.4%	5.9%	6.0%	4.5%	4.8%	5.9%

On the other hand, it can be seen that there are differences in characteristics depending on the linking method even between the linked populations.

Populations linked by the rule-based method have a higher proportion of males than other linked populations. This also leads to a high proportion of household heads in the rule-based linked population. This is probably because the similarity of names, which is one of the matching rules of the rule-based method, is better maintained in males.

The population linked by the extended feature based model has a lower proportion of the unmarried population and a higher proportion of the widowed population than other linked populations. In relation to the household head, the proportion of children is higher and the proportion of boarders is lower than in other linked populations. In addition, the proportion of people living in the place of birth, and rural areas, is higher than that of other linked populations. This may be because the extended feature based model uses the information of family members and place of residence for linking, and these features are better maintained for people with families and those who live long in the same area and move less.

[88] https://rhd.uit.no/census/instrukser.html

4.3.2. Changes in characteristics over time in linked populations

How have the mutable attributes of people changed over time in longitudinally linked populations? I examined whether there is a difference in the pattern of change in attributes over time in the populations linked by different methods.

Changes in counties of residence between 1875-1900 in linked populations are shown in Tables 4.16, 4.17 and 4.18.

Table 4.16. Percentage of inhabitants who migrated from each county[89] in 1875 to each county in 1900 in the population linked by the time-invariant feature based model (unique + best matches). Each row sums to 100%: i.e. assuming the total population of each county in 1875 is 100%, it shows the percentage of those who moved to each county in 1900. Since most people live in the same county, to make people's movements more recognizable, I displayed the moving rate as a color density of 10% maximum.

1900	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
1875																					
01	0.5	78.7	3.8	12	0.5	0.3	0.9	1.8	0.4	0.3	0.2	0.1	0.1	0.1	0	0.1	0.2	0	0.1	0.1	0
02	0.2	3.5	62	26.8	1.5	1	1.9	1.3	0.4	0.2	0.2	0.1	0.1	0.1	0.1	0	0.1	0.1	0.1	0.1	0
03	0.4	3.2	8.7	74.5	1.8	1.3	2.5	2.3	0.7	0.6	0.7	0.4	0.2	0.7	0.1	0.3	0.8	0.2	0.3	0.1	0.1
04	0	0.7	3.4	7	83.9	2.2	0.5	0.4	0.1	0.1	0.1	0.1	0.1	0.1	0	0.1	0.5	0.2	0.2	0.3	0.1
05	0	0.3	1.4	3.2	1.9	90	1.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.3	0.3	0.2	0.4	0.1	0
06	0.3	0.9	2.1	8.8	0.4	1	81.5	3	0.8	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0
07	1.9	1.4	1.2	9.2	0.2	0.2	3.1	79	1.8	0.4	0.4	0.2	0.1	0.3	0.1	0.1	0.2	0.1	0.1	0	0
08	0.7	0.6	0.6	3.7	0.2	0.2	0.9	1.8	88.8	1.4	0.4	0.2	0.1	0.1	0.1	0.1	0.1	0	0.1	0	0
09	2	0.7	0.4	4.1	0.1	0.1	0.3	0.7	1.8	83.8	4.8	0.3	0.1	0.3	0	0.1	0.1	0	0.1	0	0.1
10	0.8	0.4	0.3	3	0.1	0.1	0.2	0.4	0.5	3	89	1.5	0.1	0.2	0.1	0.1	0.1	0	0.1	0	0
11	1	0.2	0.2	1.4	0.1	0.1	0.2	0.2	0.2	0.4	0.9	92.2	1.2	0.8	0.1	0.2	0.1	0	0.1	0.1	0.1
12	0.5	0.1	0.1	0.4	0	0.1	0.1	0.1	0.1	0.2	0.2	1.6	90.2	4.9	0.6	0.3	0.1	0.1	0.4	0.1	0
13	1.5	0.3	0.6	4.6	0.1	0.1	0.2	0.5	0.3	0.3	0.3	1.3	8.1	74.8	3.2	1.3	0.8	0.2	0.9	0.2	0.1
14	0.1	0.1	0.2	0.6	0.1	0.3	0.3	0.2	0.1	0.1	0.1	0.3	3	6.8	86.4	0.3	0.2	0.1	0.6	0.1	0
15	0.2	0.1	0.1	1.3	0.1	0.3	0.1	0.1	0	0.1	0.1	0.2	0.3	0.8	0.6	92	2.2	0.3	0.8	0.2	0.1
16	0.2	0.2	0.3	2.3	0.6	0.4	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.4	0.1	1.4	88.5	2.6	1.5	0.4	0.3
17	0.1	0.1	0.2	1	0.1	0.2	0.1	0.1	0.1	0	0	0.1	0.1	0.1	0	0.3	5.5	89.3	2	0.3	0.2
18	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.1	0	0.1	0.2	0.3	0.5	0.1	0.6	0.9	0.8	93.3	1.2	0.5
19	0.1	0.1	0.1	0.9	0.2	0.2	0	0.1	0.1	0	0.1	0.1	0.1	0.4	0.1	0.3	0.9	0.2	3.4	89.6	2.9
20	0.3	0.2	0.2	1.8	0.2	0.2	0.1	0.3	0.1	0.1	0	0.2	0.2	0.4	0.1	0.6	1.3	0.3	2.5	3.4	87.6

[89] The county codes and names are as follows.
00: SHIP REGISTER, 01: ØSTFOLD, 02: AKERSHUS, 03: OSLO, 04: HEDMARK, 05: OPPLAND,
06: BUSKERUD, 07: VESTFOLD, 08: TELEMARK, 09: AUST-AGDER, 10: VEST-AGDER,
11: ROGALAND, 12: HORDALAND, 13: BERGEN, 14: SOGN OG FJORDANE, 15: MØRE OG ROMSDAL,
16: SØR-TRØNDELAG, 17: NORD-TRØNDELAG, 18: NORDLAND, 19: TROMS, 20: FINNMARK

Table 4.17. Percentage of inhabitants who migrated from each county in 1875 to each county in 1900 in the population linked by the extended feature based model (unique + best matches).

1900	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
1875																					
01	0	97.2	0.5	1.6	0.1	0	0.1	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0
02	0	0.5	92.8	5.9	0.1	0.1	0.2	0.2	0.1	0	0	0	0	0	0	0	0	0	0	0	0
03	0	0.2	1.5	96.7	0.1	0.1	0.3	0.3	0.1	0.1	0.1	0.1	0	0.1	0	0.1	0.1	0	0	0	0
04	0	0.1	0.1	0.6	98.9	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
05	0	0	0.1	0.4	0.1	99.1	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
06	0	0.1	0.2	1.2	0	0.1	97.8	0.4	0.1	0	0	0	0	0	0	0	0	0	0	0	0
07	0.1	0.2	0.2	2.2	0	0	0.4	96.1	0.3	0	0.1	0.1	0	0.1	0	0	0.1	0	0	0	0
08	0.1	0.1	0.1	0.8	0.1	0	0.1	0.3	98.2	0.1	0.1	0	0	0	0	0	0	0	0	0	0
09	0.1	0.1	0.1	1.1	0	0	0	0.1	0.3	95.6	2.3	0.1	0	0.1	0	0	0	0	0	0	0
10	0	0	0.1	0.8	0	0	0.1	0.1	0.1	0.3	98	0.4	0	0.1	0	0	0	0	0	0	0
11	0	0	0	0.3	0	0	0	0	0	0.1	0.1	98.9	0.2	0.1	0	0	0	0	0	0	0
12	0	0	0	0.1	0	0	0	0	0	0	0	0.2	98.4	1.1	0	0	0	0	0	0	0
13	0	0	0.1	0.8	0	0	0	0.1	0.1	0.1	0	0.1	3.4	94.5	0.4	0.2	0.1	0	0.1	0	0
14	0	0	0	0.1	0	0	0.1	0	0	0	0	0	0.1	0.4	99	0.1	0	0	0	0	0
15	0	0	0	0.3	0	0	0	0	0	0	0	0	0	0.1	0.1	98.8	0.5	0	0.1	0	0
16	0	0	0.1	0.5	0.1	0	0	0	0	0	0	0	0	0	0	0.2	98.5	0.3	0.1	0	0
17	0	0.1	0.1	0.3	0	0	0	0	0	0	0	0	0	0	0	0.1	0.9	98.1	0.3	0	0
18	0	0	0	0.3	0	0	0	0	0	0	0	0	0	0.1	0	0	0.1	0.1	98.9	0.3	0
19	0	0	0	0.4	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0.1	0.1	1.1	97.6	0.4
20	0.1	0.1	0	0.6	0.1	0	0	0.1	0	0	0	0.1	0	0.1	0.1	0.1	0.4	0.1	0.4	0.7	97.1

Table 4.18. Percentage of inhabitants who migrated from each county in 1875 to each county in 1900 in the population linked by the rule based model (ABE-JW).

1900	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
1875																					
01	0.5	79.8	3.5	11.5	0.4	0.2	0.9	1.8	0.4	0.2	0.1	0.1	0.1	0.1	0	0.1	0.2	0.1	0.1	0.1	0
02	0.1	3.3	63.4	26.4	1.2	1	1.9	1.1	0.3	0.2	0.2	0.1	0	0.1	0.1	0	0.2	0.1	0.1	0.1	0
03	0.2	3.1	7.9	76.7	1.5	0.9	2.5	2.1	0.7	0.5	0.7	0.4	0.2	0.7	0.1	0.3	0.8	0.2	0.2	0.1	0.2
04	0.1	0.7	3.3	7.4	83.8	2.1	0.5	0.4	0.1	0.1	0.1	0.1	0.1	0	0	0.1	0.4	0.1	0.2	0.3	0.1
05	0	0.4	1.3	3.6	1.8	89.8	0.9	0.3	0.1	0.1	0.1	0	0.2	0.1	0.1	0.3	0.3	0.2	0.4	0.1	0
06	0.3	0.9	2	8.9	0.3	1	81.6	2.9	0.8	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0
07	1.2	1.2	1.1	8.8	0.2	0.2	2.8	80.9	1.7	0.4	0.4	0.2	0	0.2	0	0.1	0.2	0.1	0.1	0.1	0
08	0.7	0.5	0.6	4	0.2	0.2	0.8	1.7	88.7	1.4	0.5	0.2	0.1	0.1	0.1	0.1	0.1	0	0.1	0	0
09	1.6	0.6	0.4	4.3	0.1	0.1	0.3	0.6	1.6	84.9	4.5	0.3	0.1	0.2	0	0.1	0.1	0	0.1	0	0.1
10	0.7	0.3	0.3	3.1	0	0.1	0.2	0.4	0.4	2.6	90.1	1.1	0.1	0.2	0	0.1	0.1	0	0.1	0	0
11	0.8	0.2	0.2	1.4	0.1	0.1	0.2	0.2	0.2	0.4	0.9	92.6	1.2	0.8	0.1	0.2	0.1	0	0.2	0.1	0
12	0.5	0.1	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.2	0.2	1.5	91	4.6	0.5	0.2	0.1	0.1	0.3	0.1	0
13	1.4	0.3	0.6	4.4	0.1	0.1	0.1	0.5	0.2	0.4	0.4	1.1	7.5	77.5	2.2	1.2	0.7	0.2	0.9	0.1	0.1
14	0.2	0.1	0.2	0.7	0.1	0.2	0.2	0.1	0	0.1	0.1	0.2	2.9	6.1	87.5	0.2	0.2	0.1	0.6	0.1	0
15	0.2	0.1	0.1	1.3	0.1	0.2	0.1	0.1	0	0.1	0.1	0.2	0.2	0.7	0.5	92.8	1.8	0.2	0.8	0.2	0.1
16	0.2	0.2	0.3	2.7	0.5	0.4	0.1	0.2	0.1	0.1	0.1	0.2	0.2	0.4	0.1	1.1	88.6	2.2	1.5	0.5	0.3
17	0.1	0.1	0.1	1.1	0.1	0.1	0.1	0.1	0.1	0	0	0.1	0.1	0.1	0	0.3	4.3	90.4	2.2	0.2	0.2
18	0.1	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0	0.1	0.1	0.2	0.2	0.4	0.1	0.5	0.7	0.7	93.9	1.2	0.4
19	0.1	0.1	0.1	1	0.1	0.2	0	0.1	0	0.1	0.1	0.1	0.1	0.4	0.1	0.3	0.8	0.2	3.3	90.5	2.3
20	0.3	0.2	0.2	1.8	0.1	0.1	0.1	0.3	0.1	0.1	0	0.2	0.1	0.4	0	0.4	1.2	0.2	2.6	3.1	88.4

In linked populations by the time-invariant feature based model and by the rule-based method, the flow of people moving from all counties to Oslo is seen. Other than that, it can

be seen people moving mostly to neighboring counties. However, in the linked population by the extended feature based model, this trend is less obvious than in other linked populations. This may be related to the high proportion of people living in the place of birth or rural areas with less migration in the population linked by the extended feature based model. However, since the performance (precision) of the extended feature based model is higher, the results of Tables 4.9 and 4.11 also need to be critically reviewed. Because it could be the result of incorrect matches having an effect. Since moving-related attributes are considerably related to the representativeness issues, it seems that considerable caution is needed in estimating moving in the population at the time using the linked population.

Changes in marital status between 1875-1900 in linked populations are shown in Tables 4.19, 4.20 and 4.21.

Table 4.19. Percentage of changes from each marital status[90] in 1875 to marital statuses in 1900 in the population linked by the time-invariant feature based model (Unique + Best matches). Each row sums to 100%

1900	Unknown	Unmarried	Married	Widow(er)	Divorced	Seperated
1875						
Unknown (n=120,049)	0.6	37.8	59.1	2.4	0.1	0.1
Unmarried (n=196,095)	0.5	30.5	64.3	4.5	0.1	0.1
Married (n=143,996)	0.4	1.7	71.7	26.2	0.1	0.1
Widow(er) (n=9,187)	0.7	3.2	23.9	72.1	0.1	0.1
Divorced (n=215)	0.9	5.6	33.5	51.2	6.5	2.3

Table 4.20. Percentage of changes from each marital status in 1875 to marital statuses in 1900 in the population linked by the extended feature based model (Unique + Best matches).

1900	Unknown	Unmarried	Married	Widow(er)	Divorced	Seperated
1875						
Unknown (n=121,975)	0.4	41.8	55	2.6	0.1	0.1
Unmarried (n=195,010)	0.5	32.5	62.3	4.6	0.1	0.1
Married (n=187,376)	0.3	0.2	74.5	24.9	0	0.1
Widow(er) (n=10,141)	0.9	0.7	22.4	75.9	0.1	0.1
Divorced (n=227)	1.8	0.4	31.3	58.1	6.2	2.2

[90] Since less than 10 cases were recorded as 'separated' in 1875, I excluded them from the table.

Table 4.21. Percentage of changes from each marital status in 1875 to marital statuses in 1900 in the population linked by the rule based model (ABE-JW).

1900	Unknown	Unmarried	Married	Widow(er)	Divorced	Seperated
1875						
Unknown (n=68,871)	0.6	39.3	57.6	2.4	0.1	0.1
Unmarried (n=114,440)	0.5	31.9	63.2	4.2	0.1	0.1
Married (n=91,959)	0.4	1.3	72.4	25.8	0	0.1
Widow(er) (n=5,997)	0.7	2.6	23.9	72.6	0.1	0.1
Divorced (n=141)	0.7	5	31.9	55.3	6.4	0.7

It is difficult to estimate the exact rate of change due to the high rate of missing marital status in 1875, but overall there is no significant difference between the linked populations. About 30-33% of unmarried people remained single, 71-75% of married people remained married, and 22-24% of widow(er)s remarried. In the population linked by the extended feature based model, the rate of maintaining the same marital status is slightly higher, and the rate of non-logical change[91] is lower because changes in marital status and non-logical changes in marital status are also used as features for linking.

Changes in family position (relation to household head) between 1875-1900 in linked populations are shown in Tables 4.22, 4.23 and 4.24.

Table 4.22. Percentage of changes from each family position in 1875 to family positions in 1900 in the population linked by the time-invariant feature based ML model (Unique + Best matches). Each row sums to 100%

1900	01	02	03	04	05	06	07	08	09	10	11	12	13	99
1875														
01. Householder	83.9	0.3	0.1	0	2.5	2	0.1	0	0	0.5	0.4	9.3	0	0.9
02. Spouse	0.8	76	0.2	0	4.3	3	0.2	0	0	0.8	0.7	13.2	0	0.7
03. Child	36	27.7	18	0.5	0.1	0.1	1.1	0.1	0	0.1	0.7	13.8	0.1	1.9
04. Child-in-law	65.6	23.3	1.3	0.3	1.5	0.7	0.3	0	0	0.1	0.4	5.6	0	0.8
05. Parent	23.8	24	0.2	0	10	7.6	0.2	0	0	3	0.6	28.9	0.1	1.6
06. Parent-in-law	13.6	18.4	0	0	8.9	13.6	0.3	0.3	0	3.5	0.3	39	0	2.2
07. Sibling	29.1	30.4	1.5	0	1	1.2	5.8	0.1	0	1.6	0.7	27.1	0.1	1.4
08. Sibling-in-law	48.3	13.9	1.7	0	0	2.3	1.7	0.6	0	2.9	1.7	25	0	1.7
09. Grandchild	30.3	30.6	15.6	0.8	0	0.1	0.6	0.1	0.3	0.3	0.8	19	0.1	1.5
10. Other relatives	31	31	5.1	0	0.7	0.3	1.5	0	0	1.6	0.8	25.5	0.2	2.3
11. Partner/Friend	52.6	23.4	4.3	0.1	0.6	0.5	0.8	0.1	0	0.3	1.3	14.1	0	1.8
12. Boarder	35.5	37.3	2.5	0.2	0.5	0.3	0.9	0.1	0	0.3	0.6	20.6	0.1	1.2
13. Inst. inmate	32.2	22.9	2.6	0	1.9	0	1.1	0.4	0	0.4	0.4	30	1.9	6.4
99. Unknown	41.4	27.1	4.9	0.2	1	0.8	1.2	0.1	0.1	0.3	0.6	19.3	0.1	2.7

[91] Marital status changes from married, widow(er) and divorced to unmarried.

Table 4.23. Percentage of changes from each family position in 1875 to family positions in 1900 in the population linked by the extended feature based ML model (Unique + Best matches).

1900	01	02	03	04	05	06	07	08	09	10	11	12	13	99
1875														
01. Householder	86.2	0.2	0.1	0	2.5	2.1	0.1	0	0	0.5	0.2	7.7	0	0.5
02. Spouse	0.7	79.4	0.1	0	4.3	2.9	0.1	0	0	0.7	0.3	10.8	0	0.6
03. Child	35.4	25.9	24.5	0.5	0.1	0.1	1.4	0.1	0	0.1	0.6	10.9	0	0.6
04. Child-in-law	69.2	23.6	0.9	0.3	1.8	0.5	0.1	0	0	0.2	0.5	2.8	0	0.1
05. Parent	23.1	22.7	0.2	0	11.9	9.3	0.1	0	0	4.1	0.3	26.7	0.1	1.5
06. Parent-in-law	14.1	15.8	0.2	0	7.4	16.2	0.4	0.4	0	3.6	0	40.2	0	1.7
07. Sibling	26.9	27.5	1.4	0.1	0.8	1.6	7.7	0.1	0	1.8	0.8	30	0	1.5
08. Sibling-in-law	46.7	13.3	1.2	0	0	2.4	2.4	1.2	0	4.2	1.2	26.7	0	0.6
09. Grandchild	28.6	31.1	20.2	1.3	0	0	0.7	0.1	0.5	0.4	0.5	16.1	0.1	0.7
10. Other relatives	33.2	28.5	5.8	0	0.4	0.6	1.5	0	0	1.5	0.6	26.6	0	1.3
11. Partner/Friend	49	28.5	4.1	0.1	0.5	0.3	0.9	0.1	0	0.2	2.1	13.3	0	0.9
12. Boarder	34.5	38.3	2	0.1	0.5	0.4	0.8	0.1	0	0.3	0.6	21.5	0	0.9
13. Inst. inmate	36.3	23.2	1.1	0	1.1	0	0	0	0	0.5	0.5	19.5	2.1	15.8
99. Unknown	39.2	28.4	6.2	0.2	1	0.9	1.4	0.1	0.1	0.3	0.5	19.3	0.1	2.3

Table 4.24. Percentage of changes from each family position in 1875 to family positions in 1900 in the population linked by the rule based model (ABE-JW).

1900	01	02	03	04	05	06	07	08	09	10	11	12	13	99
1875														
01. Householder	84.4	0.3	0.1	0	2.5	1.9	0.1	0	0	0.5	0.3	9	0	0.8
02. Spouse	0.9	76.9	0.2	0	4	2.8	0.2	0	0	0.7	0.6	13	0	0.8
03. Child	39.4	23.7	19.1	0.5	0.1	0.1	1.1	0.1	0	0.1	0.5	13.5	0.1	1.7
04. Child-in-law	70.7	20.2	1	0.2	1.2	0.5	0.5	0	0	0	0.5	4.3	0	1
05. Parent	26.5	23.4	0.3	0	10	8.1	0.1	0	0	3.2	0.5	26	0.3	1.6
06. Parent-in-law	13.2	17.1	0	0	8.3	13.6	0.4	0.4	0	2.6	0	40.8	0	3.5
07. Sibling	30.6	26	1.4	0.1	0.7	1.1	7.1	0.2	0	1.8	0.4	28.9	0	1.7
08. Sibling-in-law	50.5	7.3	1.8	0	0.9	3.7	2.8	0.9	0	3.7	1.8	25.7	0	0.9
09. Grandchild	35.8	25.4	16.1	1	0	0	0.6	0	0.6	0.2	0.5	18.4	0.1	1.2
10. Other relatives	34.6	24.8	4.7	0	0.9	0.6	1.6	0	0	2.5	0.9	27	0.3	1.9
11. Partner/Friend	55	20.6	4.4	0.1	0.5	0.1	1.1	0.1	0	0.2	1	15.3	0	1.6
12. Boarder	38.7	33.9	2.4	0.1	0.5	0.3	0.9	0.1	0	0.3	0.5	20.9	0.1	1.3
13. Inst. inmate	32.4	24.3	1.4	0	0.7	0	0	0	0	0	0	31.8	2.7	6.8
99. Unknown	43.7	23.3	5.2	0.1	1.1	0.9	1.6	0.1	0.1	0.3	0.6	20	0.1	2.7

In linked populations, the proportion of household heads and spouses maintaining their family positions is higher than 80% and 75%, respectively. Other family positions also have a high rate of changes to the position of household heads or spouses. This may be because the number of households increased and the number of household members decreased between 1875-1900.[92] Household members who are not the children of the householder also have

[92] The number of households: 326,289 in the 1875 census, 459,752 in the 1900 census

quite a high rate of becoming boarders. Although the proportion of maintaining the same family position and the proportion of institutional inmates linked to unknown in the population linked by the extended feature based model are slightly higher, there is no significant difference between the linked populations. However, as shown in Table 4.13, as the linked populations have a higher proportion of children and a lower proportion of boarders in common compared to the census, it seems that there are some limitations in estimating changes in family position using these results.

Average number of household members: 7.22 (std 6.50) in the 1875 census, 6.67 (std 4.43) in the 1900 census (excluding institutional inmates)

Chapter 5

Discussion

I begin by reviewing the results of the study along with the research questions raised in the introduction of the study.

5.1. Does linking by the machine learning approach improve the linking rate of HPR?

The results of linking the censuses by a machine learning approach show that it is possible to improve the linking rate between the 1875-1900 censuses, which is the bottleneck of the current HPR. Although both traditional rule-based linking and linking by machine learning led to an increase in the linking rate, the machine learning results were generally better in performance. When linking using different feature sets (time-invariant feature set and extended feature set) and match selections (taking only unique matches and taking both unique and best matches) to see the impact of different machine learning methods, the model's performance was better for the extended feature set and for taking both unique and best matches. When the 1875-1900 censuses are linked nationwide by using the extended feature set and taking both unique and best matches, which had the best performance, the linking rate is improved by 34%[93] compared to the existing HPR. In regards to the link quality, when using the extended feature set and taking both unique and best matches, the F1 score for the test set from the training data is 0.77, and the F1 score for the test set provided by the NHDC is 0.94.

Comparing the rule-based linking with the machine learning linking with time-invariant features and only unique matches, which is the closest to the rules in the rule-based method in this study, the rule based linking rate is 22% and the machine learning linking rate is 21.2%, which is almost similar. Both are improved by about 15-16% compared to the existing

[93] Percentage of population aged 20 and over in 1900

HPR link rate (6.2%). However, in terms of performance, the rule-based method's F1 score is 0.57 (precision 0.56, recall 0.58) / 0.73 (precision 0.97, recall 0.59), and the machine learning linking's F1 score is 0.67 (precision 0.67, recall 0.67) / 0.77 (precision 0.98, recall 0.63), which is higher for machine learning. The HPR links used as training data for machine learning models are highly accurate because they are generated through manual reviews and additional sources such as church books. Machine learning models trained on these HPR links show high performance even in linking using only the census without using these additional sources. This demonstrates the strength of machine learning approaches that learns on its own the match and non-match patterns underlying in training data.

One of the issues discussed in previous record linkage studies is the selection of variables to use for linking [27][28][24]. This issue, which can be summarized as a tradeoff between link rate (match size) / accuracy and representativeness, led to a distinction between linking using only time-invariant variables and linking using extended variables including mutable variables. The results of linking using these two feature sets in this study showed that the result using the extended feature set had higher link rate (larger match size), higher performance, and lower representativeness, which is similar to those discussed in previous studies. Regarding match size and performance (representativeness is covered in section 5.2), both precision and recall are improved when using an extended feature set rather than a time-invariant feature set, but in particular, the recall is greatly improved. It seems that using extended features not only reduces false matches, but also contributes greatly to finding matches that could not be determined by only time-invariant features because more information is available.

Although it has not been emphasized much in previous studies, the results of the linking show that the manner in which multiple matches are managed (i.e., cases in which multiple candidates are classified as matches for one person) has a significant impact on performance. Ways of handling multiple matches includes discarding all multiple matches to minimize false matches [23][14], randomly selecting a candidate from among multiple matches [39], or taking the best candidate when certain conditions are met (i.e., a sufficiently certain answer) among multiple matches [21][24]. Discarding all multiple matches increases precision by reducing the possibility of incorrect matches, but lowers recall by eliminating correct matches as well. Taking the best matches among multiple matches as final matches increases the recall significantly at the cost of slightly lowering precision, thereby improving overall performance.

5.2. Is the linked population representative of the entire population?

The results of the study show that the linked populations are generally not representative of the original census population. This is already discussed in several previous studies [22][24][26][28][38][40][41]. In the linked population, people with more easily linkable traits are overrepresented, such as older people, men, householder's children, people living in their birthplace, and people living in rural areas. This is the same whether linked by machine learning or by the rule based method. However, it is also true that there are differences in characteristics between linked populations depending on the algorithm used for linking. For example, the population linked by the rule based method is particularly less representative in gender as men's names are better maintained, and the population linked by the extended feature based machine learning model is less representative in age, family relations (children and boarders), living in urban/rural areas, and living in birthplace as the model used the family and residence information. Overall, the difference in representation between linked populations and the census population is larger than the difference between the linked populations.

Comparing the time-invariant feature based linked population and the extended feature based linked population, the latter is generally less representative. However, for some attributes such as gender, birthplace, and municipalities, which are variables related to names, birth year, and birthplace that play a major role in time-invariant features, the time-invariant feature based linked population was less representative than the extended feature based one. Depending on which features play an important role in algorithms, certain characteristics appear to be overrepresented and to varying degrees.

In comparison between taking only unique matches and taking both unique and best matches, the former is generally less representative. As shown in Table 4.13, the representativeness of the population in which only unique matches are taken is lower for most characteristics (There are some differences depending on the feature set). Taking the best candidate among multiple match candidates increases the size of the linked population by not discarding candidates with high match potential. And it seems to contribute to alleviating selection bias by including people who are not very certain, i.e., people who do not have easily linking characteristics, in matches.

Looking at changes in characteristics over time in linked populations, it is implied that these differences in representativeness may affect demographic estimates. For example, the

rule-based linked population and time invariant feature-based linked population differ from the extended feature based linked population by about 10% in the proportion of people living in their birthplace[94]. This results in much lower moving rates in the extended feature-based linked population in county changes of 1875-1900, as shown in Tables 4.15-17. Therefore, when making demographic estimates using linked data, it will be necessary to closely examine the representativeness of the characteristics related to the purpose of the study and to investigate whether it does not affect the study.

One strategy that can alleviate the problem of representativeness of linked populations is weighting. Although not covered in this study, the effect of mitigating the bias of linked samples by weighting has been discussed in previous studies [40][41][28]. As an application of inverse probability weighting [41], dividing the linked data into sub-cells according to observed characteristics, and generating a pseudo-population by weighting using the inverse of the probability for each cell and the representation for the cell in the census can be one of the useful strategies to improve representativeness. When weighting, sufficient cell sizes are also important so that special traits of some in the sub-cells are not overrepresented. Thus, large match-sized linked data based on an extended feature set and by taking both unique and best matches can be better for weighting [28].

[94] Census: 55%, Rule based linked data: 64%, Time-invariant feature based linked data: 63-64%, Extended feature-based linked data: 72-75% (See Table 4.13)

5.3. Findings of the study

In addition to the answers to the research questions, I describe additional findings obtained during this study.

For accurate and reliable census linking, it is necessary to understand the historical and cultural context of the census data prior to methods such as algorithm and feature selection. Although the census form is similar worldwide and the census linking process can be generalized, its application to actual census data requires appropriate processing based on consideration of the background and context of the census data.

In this study, the understanding of the historical and cultural context of the Norwegian census data affected the entire research process, from linking to interpretation. For example, Norwegian census data is based on the municipality as a regional unit and it is advantageous for linking because the municipality has a relatively small range. However, there have often been changes in the boundaries of municipalities in history, so if this is not considered in the linking process, the error rate can be increased. Also, in the late 19th and early 20th centuries, there was a change in the convention of surnames in Norway. This has resulted in some people changing their surnames over time, a variable which is generally not expected to change. If this is not taken into account in the linking process, some true matches may be lost. Moreover, in the interpretation of the linking results, the overrepresentation of people born/lived in the northern regions compared to the original census in the linked data can lead to misinterpretation unless it is taken into account that the rural ratio in the northern regions is higher and the census collection in the rural area is more accurate.

Although this study tried to illustrate the unique characteristics and challenges of the Norwegian census data, and to deal with them in the linking process, there may still be limitations. The importance of domain knowledge in census linking reminds us that census linking is a collaborative work in multiple fields. This implies that cooperation with domain experts is necessary not only for the linking process such as designing derived variables and selecting features suitable for the target census, but also for interpretation of linking results and improvement of linking models in the future.

Next, this study showed that the size of the dataset can influence the determination of match selection parameters (absolute probability cutoff α[95] and relative probability cutoff β[96]) related to the performance of the model. As the size of the dataset increases, the possibility of multiple matches also increases, so how multiple matches are dealt with has a significant impact on performance. One way to reduce multiple matches is to raise the absolute cutoff or the relative cutoff so that fewer candidates meet the criteria. However, this results in lowering the recall, and the overall performance is lowered because the decrease in recall is larger than the increase in precision. If the recall is too low, the match size (linking rate) also decreases, and as shown in Table 4.12, that is not completely free from the issue of representativeness. Another way to deal with multiple matches is to select the candidate with the best probability in multiple matches as a match. Then, the number of incorrect predictions increases, but the number of correct predictions also increases, and the increase in recall is greater than the decrease in precision, which increases overall performance. When the F1-score was used as a performance metric in this study, the match selection parameters that can achieve the best performance were determined at lower values[97] as the dataset became larger.

However, excessive compromise of precision is also problematic. If the precision is too low and the number of incorrect matches increases, it leads to errors in demographic estimates. Since it is much more difficult to correct the False Positive (FP, Type I error) than the False Negative (FN, Type II error) [22], a large loss in precision is undesirable. Therefore, when determining match selection parameters based on precision and recall, minimum baselines may have to be established. In determining the feature set and match classification hyperparameters, this study is based on F1-score with equal weight on precision and recall, but if necessary, matching results suitable for the purpose can be obtained by varying weights to precision and recall or setting minimum baselines. This kind of refined model tuning can be made according to the researcher's needs. Alternatively, it will be possible to build and provide several types of linked sample sets, such as a sample set with maximum representativeness of the entire population[98], a sample set with as many individuals as

[95] When the predicted probability of a match candidate exceeds this value, it is classified as a match.

[96] When the ratio of the predicted probability of the best match candidate and the prediction probability of the second best match candidate exceeds this value, it is classified as a match.

[97] Other studies [21][24] in which these hyperparameter values are mentioned show similar results. α: 0.14, β: 1.375 In [21], α: 0.27, β: 1 in [24]. Helgertz used Matthew's Correlation Coefficient (MCC) as a metric for measuring performance. That places more weight on precision in asymmetric classes like record linking. When MCC was used instead of F1 score as a performance metric for this study, α increased to 0.2.

[98] IPUMS Linked Representative Samples (IPUMS-LRS) is an example [42].

possible linked, or a sample set with the highest precision, by the NHDC through evaluation by experts.

Another way to deal with multiple matches is to reduce ambiguity in multiple match candidates by increasing the available information itself. That is why the performance of extended feature based models is good. While this approach improves both linking rates and performance, there has been a representativeness issue with people with certain characteristics being overrepresented in linked data. However, if representativeness issues are inevitable in linked data and weighting can alleviate the representativeness problem [41], and if a large match size also helps in weighting by increasing the size of subcells [28], then using the extended features to improve accuracy and match size at the expense of representativeness and correcting the biases through weighting is also an approach worth considering.

5.4. Limitations of the study

Since machine learning learns the match and non-match patterns from the training data, it is important that training data should be as close as possible to the census data, which is the final target. In this study, the high accuracy data from municipalities manually reviewed and linked with additional sources such as parish records were used as training data, without constructing new training data. This has the advantage of being able to easily improve the link rate in the current situation without the effort to build a new training dataset, but it also has the disadvantage that the bias in training data is not completely controlled. Since most of the training data are from the municipalities of Troms county, the rate of birth/residence in the northern region is high, and there are differences from the census in sex, marital status, and whether they live in their birthplace. Similar biases appear in the linked results of this study, but it is difficult to ascertain whether this is due to linking or training data.[99] If the training data is sampled to be representative of the census, and the machine learning model is trained on this, it will be possible to separate the bias caused by the training data and the bias inherent in linking. For example, if samples are evenly extracted in consideration of a variety of characteristics from the national census, and they are linked through manual review and with additional sources such as parish records or genealogy platforms, less biased and high accuracy training data can be built.[100] If the training data consists of samples that are more representative of the census, the performance of the models trained on them will also be improved.

When linking two censuses, the linking rate is generally less than 100% because there are people who die or migrate during the period between them. The population available for linking in the 1900 census[101] is smaller than the population available for linking in the 1875 census[102] because people who died or emigrated between 1875 and 1900 are excluded. Since this study only included people living in Norway at the time of the census, there is a limitation in that the migrant population to/from outside Norway was not included in the study. Between 1875 and 1900 was a period of large-scale migration (and return) from

[99] Although there are similar biases in the linked population, it is difficult to affirm that it is the effect of the training data, as it is the same in the result linked by the rule-based method that does not use the training data. In addition, the training data is somewhat similar to the census in age, so the characteristics of the training data are not directly followed by the population linked by machine learning.

[100] Link-Lives project in Denmark focuses on the quality of training data for machine learning models and builds high-quality training data involving domain experts. [29][43]

[101] Population aged 20 and over in the 1900 Census (The population under the age of 25 in 1900 would not have existed in 1875, but in this study, 20 is used instead of 25 to account for errors in the records.)

[102] Total Population in the 1875 Census

Norway to North America. This will be one of the factors that should be considered both in linking the censuses and in constructing the life histories of Norwegians in this period.[103] If Norway's migration records are integrated into the census records later, it will not only reduce the difference in link rates and enable more accurate linking, but also support various studies related to the migrant population. Also, although mortalities between 1875 and 1900 were not considered in this study, there are funeral records of individuals in parish registers at the time. Just as the baptism and marriage records of parish registers were used for linking censuses, if these funeral records are linked with the census to improve data quality, both the match rate and performance of linking will be improved.

In addition, in the process of building machine learning models for linking, there are limitations in that the missing matches in indexing (about 10% of true matches) were not taken into account, and the tested machine learning algorithms were not carefully tuned with hyperparameters. If these limitations are supplemented in the future, the performance of the models can be improved.

Also, there is a limitation in not presenting more clearly by quantitatively measuring the representativeness of linked samples and the possibility of errors in demographic estimation. If this is supplemented, it will be able to support more convincingly that there should be an examination of representativeness when using linked data for research, and that representativeness should be improved through post-processing if necessary.

[103] For studies analyzing the migrants from Norway to North America during the age of mass migration (1850-1913), see [16], [17].

Chapter 6

Conclusion and Future work

This study began with the motivation to improve the linking rate between the 1875-1900 Norwegian censuses of the Historical Population Registers (HPR), the integrated microdata of the Norwegian population since the 19th century. To this end, I applied machine learning approaches to the record linking process to link the two censuses, and examined the results and their implications. I will conclude the study by summarizing the findings of the study and future work.

The historical record of a country reflects the culture of that society. Although the census forms are mostly similar internationally and the process of linking them can be generalized, the culture and characteristics of the country must be taken into account for its interpretation and application. The unique characteristics of the Norwegian census, such as small regional units (municipalities), changes in regional boundaries, changes in last names, and urban/rural regional distributions, influenced the entire process of the study from variable creation, feature selection, model performance, and interpretation of results. For linking historical records, it is important to understand the cultural context to which they belong, and apply this knowledge to the linking process properly.

Regarding the primary motivation of the study, I showed that the machine learning approach can be an effective strategy for improving the linking rate of HPR. The machine learning model trained using the high-quality data included in the HPR as training data improved the linking rate by up to 34% nationwide with high performance (F1 score 0.77 / 0.94). This was also higher performance compared to the traditional rule-based linking method. The approach attempted in this study is an improvement that can be applied to the current HPR without the input of additional resources.

In addition, I tested various machine learning algorithms, feature sets, and match selection options in this study and presented the results. By demonstrating that data linked by these

different models can vary in performance and representativeness, I emphasized the flexibility and potential of the machine learning approach. These options and hyperparameters may be adjusted for linking according to the researcher's needs or expert evaluation.

Finally, I brought up some considerations to keep in mind when using the linked populations for research by examining the representativeness of the linked data for the entire population. Linked data are generally not representative of the entire population, and the characteristics or degrees of bias vary depending on the linking method used. By displaying that these differences may affect demographic estimates, I pointed out that the use of linked data requires careful examination and, if necessary, further processing depending on the purpose of the study.

Because this study focused on linking, it did not cover mitigating bias and improving the representativeness of the linked data, and applying it to analysis on specific topics. I leave them up to future studies. In addition, occupational variables in Norway's historical census data, which can be indicators of people's economic and social status, have not yet been standardized. Standardizing them is also one of the future works necessary for the wide utilization and in-depth analysis of data. The reliable life history dataset of Norwegians since the 19th century, which HPR pursues and this study aims to contribute, will be a resource that can inspire significant studies in a variety of fields.

Works cited

1. Asher J, Resnick D, Brite J, Brackbill R, Cone J. An Introduction to Probabilistic Record Linkage with a Focus on Linkage Processing for WTC Registries. Int J Environ Res Public Health. 2020;17. doi:10.3390/ijerph17186937

2. Christen P. Data Matching. 2012. doi:10.1007/978-3-642-31164-2

3. Sommerseth HL, Thorvaldsen G. The Impact of Microdata in Norwegian Historiography 1970 to 2020. Historical Life Course Studies. 2022. pp. 18–41. doi:10.51964/hlcs11675

4. Thorvaldsen G. Using NAPP Census Data to Construct the Historical Population Register for Norway. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2011. pp. 37–47. doi:10.1080/01615440.2010.517470

5. Hammel EA. Demographic Techniques: Family Reconstitution. 2001.

6. Thorvaldsen G. Fra folketellinger og kirkebøker til norsk befolkningsregister. Heimen. 2008;45: 341–359.

7. Thorvaldsen G, Andersen T, Sommerseth HL. Record Linkage in the Historical Population Register for Norway. Population Reconstruction. 2015. pp. 155–171. doi:10.1007/978-3-319-19884-2_8

8. Abramitzky R, Boustan L, Eriksson K, Feigenbaum J, Pérez S. Automated Linking of Historical Data. J Econ Lit. 2021;59: 865–918.

9. Modalsli J. Intergenerational Mobility in Norway, 1865–2011. The Scandinavian Journal of Economics. 2017. pp. 34–71. doi:10.1111/sjoe.12196

10. Sadinle M, Fienberg SE. A Generalized Fellegi–Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems. Journal of the American Statistical Association. 2013. pp. 385–397. doi:10.1080/01621459.2012.757231

11. Fu Z, Boot HM, Christen P, Zhou J. Automatic Record Linkage of Individuals and Households in Historical Census Data. International Journal of Humanities and Arts Computing. 2014;8: 204–225.

12. Goiser K, Christen P. Towards automated record linkage. AusDM. 2006;6: 23–31.

13. Sobek M, Ruggles S. The IPUMS Project: An Update. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 1999;32: 102–110.

14. Ferrie JP. A new sample of males linked from the public use microdata sample of the 1850 U.s. federal census of population to the 1860 U.s. federal census manuscript schedules. Hist Methods. 1996;29: 141–156.

15. Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J. Big Data and Social Science: Data Science Methods and Tools for Research and Practice. CRC Press; 2020.

16. Abramitzky R, Boustan LP, Eriksson K. Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. Am Econ Rev. 2012;102: 1832–1856.

17. Abramitzky R, Boustan LP, Eriksson K. Have the poor always been less likely to migrate? Evidence from inheritance practices during the age of mass migration. J Dev Econ. 2013;102: 2–14.

18. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969;64: 1183–1210.

19. Winkler WE. Overview of record linkage and current research directions. BUREAU OF THE CENSUS. 2006. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.1519

20. Abramitzky R, Mill R, Pérez S. Linking individuals across historical sources: A fully automated approach. Hist Methods. 2020;53: 94–111.

21. Feigenbaum JJ. Automated census record linking: A machine learning approach. 2016. Available: https://ranabr.people.stanford.edu/sites/g/files/sbiybj5391/f/machine_learning_approach.pdf

22. Bailey M, Cole C, Henderson M, Massey C. How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data. J Econ Lit. 2020;58: 997–1044.

23. Goeken R, Huynh L, Lenius T, Vick R. New Methods of Census Record Linking. Hist Methods. 2011;44: 7–14.

24. Helgertz J, Price J, Wellington J, Thompson KJ, Ruggles S, Fitch CA. A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel. Hist Methods. 2021; 1–18.

25. Rijpma A, Cilliers J, Fourie J. Record linkage in the Cape of Good Hope Panel. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2020;53: 112–129.

26. Price J, Buckles K, Van Leeuwen J, Riley I. Combining family history and machine learning to link historical records: The Census Tree data set. Explor Econ Hist. 2021;80: 101391.

27. Ruggles S. Linking Historical Censuses: a New Approach. History and Computing. 2002;14: 213–224.

28. Antonie L, Inwood K, Minns C, Summerfield F. Selection Bias Encountered in the Systematic Linking of Historical Census Records. Social Science History. 2020. pp. 555–570. doi:10.1017/ssh.2020.15

29. Løkke A. Link-Lives, Historical Big Data: reconstructing millions of life courses from archival records using domain ex-perts and machine learning. 2021. Available: http://ceur-ws.org/Vol-3019/LinkedArchives_2021_paper_9.pdf

30. Wisselgren MJ, Edvinsson S, Berggren M, Larsson M. Testing Methods of Record Linkage on Swedish Censuses. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2014;47: 138–151.

31. Edvinsson, Engberg. A Database for the Future: Major Contributions from 47 Years of Database Development and Research at the Demographic Data Base. Life Course Studies. 2020. Available: https://testplatform.openjournals.nl/hlcs/article/view/9305

32. Gjelseth M. Relasjonsdatabaser som verktøy i en historisk-demografisk studie. unpublished dissertation, University of Oslo. 2000.

33. Holden L, Thorvaldsen G, Bråthen TR. Historisk befolkningsregister og DNF 1814. Heimen. 2012. pp. 399–414. doi:10.18261/issn1894-3195-2012-04-04

34. Vick R, Huynh L. The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2011;44: 15–24.

35. Alhaug G. 10 001 navn Norsk fornavnleksikon. Cappelen Damm; 2011.

36. Sogner S. Folkevekst og flytting: en historisk-demografisk studie i 1700-årenes Øst-Norge. 1979.

37. Abramitzky R, Boustan LP, Eriksson K. A Nation of Immigrants: Assimilation and

Economic Outcomes in the Age of Mass Migration. J Polit Econ. 2014;122: 467–506.

38. Antonie L, Inwood K, Lizotte DJ, Andrew Ross J. Tracking people over time in 19th century Canada for longitudinal analysis. Machine Learning. 2014. pp. 129–146. doi:10.1007/s10994-013-5421-0

39. Nix E, Qian N, National Bureau of Economic Research. The Fluidity of Race: "Passing" in the United States, 1880-1940. 2015.

40. Ferrie. Longitudinal Data for the Analysis of Mobility in the US, 1850-1910. Work Pap, Dep Econ, Northwestern Univ, Evanston. 2004. Available: https://www.mcgill.ca/economics/files/economics/ferrie.pdf

41. Bailey M, Cole C, Massey C. Simple strategies for improving inference with linked data: a case study of the 1850–1930 IPUMS linked representative historical samples. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2020;53: 80–93.

42. Ruggles S, Fitch C, Roberts E. Historical Census Record Linkage. Annu Rev Sociol. 2018;44: 19–37.

43. Revuelta-Eugercios. LINK-LIVES: BUILDING HISTORICAL BIG DATA FROM ARCHIVAL RECORDS FOR USE BY RESEARCHERS AND THE DANISH PUBLIC. digitaltreasures.eu. Available: https://www.digitaltreasures.eu/wp-content/uploads/2021/12/LinkLives-project-Revuelta-Eugercios.pdf

44. Mill, Roy. 2013. "Chapter 4. Linking Records across Historical Sources." PhD diss. 'Inequality and Discrimination in Historical and Modern Labor Markets', Stanford University. https://stacks.stanford.edu/file/druid:br608gp5134/thesis_toplevel-augmented.pdf

45. Abramitzky, Ran, Boustan, Leah, Eriksson, Katherine, Feigenbaum, James, and Pérez, Santiago . Data and Code for: Automated Linking of Historical Data. Nashville, TN: American Economic Association [publisher], 2021. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-09-24. https://doi.org/10.3886/E133781V1

46. Jonas Helgertz, Steven Ruggles, John Robert Warren, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Joseph P. Price, Evan Roberts, and Matthew Sobek. IPUMS Multigenerational Longitudinal Panel: Version 1.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D016.V1.0

47. Steven Ruggles, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Evan Roberts, Megan Schouweiler, and Matthew Sobek. IPUMS Ancestry Full Count Data: Version 3.0 [dataset]. Minneapolis, MN: IPUMS, 2021.

48. Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.3 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D020.V7.3

49. Price, Joseph , and Buckles, Kasey. Data and code from Price, Buckles, Van Leeuwen, & Riley (Explorations in Economic History). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-02-10. https://doi.org/10.3886/E130961V4