HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Selection into Migration: American Finns and the Karelian Fever

Kaarle Myllyneva
Master's Thesis
Master's Programme in Economics
Faculty of Social Sciences
University of Helsinki
May 2025

Abstract

Faculty: Faculty of Social Sciences

Degree programme: Master's Programme in Economics

Author: Kaarle Myllyneva

Title: Selection into Migration: American Finns and the Karelian Fever

Level: Master's Thesis

Month and year: May 2025

Number of pages: 50 + 11

Keywords: Historical Record Linking; Census Linking; Labor Migration

Supervisors: Torsten Santavirta, Andrei Markevich

Additional information: -

Abstract: In this master's thesis, I study economic and occupational selection into migration during the Karelian Fever movement, when thousands of Finnish Americans and Canadians migrated to Soviet Karelia in the early 1930s. I use digitized Soviet Karelian immigration records, which I link to individual-level data from the 1930 United States Census and the 1931 Canadian Census. The linking process builds on state-of-the-art historical record linkage methods, and I develop an algorithm to link cross-national and multilingual datasets. The resulting linked dataset enables a detailed analysis of migrant characteristics and how they differed from those who remained in North America.

To study migration selection, I apply both descriptive methods and regression analysis. I compare occupational and geographic patterns of migrants and stayers, and estimate whether migrants were systematically selected based on income or occupational status. I use occupation-based income variables for the U.S. sample and individual earnings data for the Canadian sample. The results show no clear selection among U.S. migrants, but indicate that Canadian migrants had significantly lower earnings than non-migrants, suggesting negative economic selection. I also find that Finnish-origin individuals in the 1930 U.S. Census had systematically lower occupational income scores than the general population – including other Nordic immigrant groups. Additionally, I show that traditional occupation-based income scores may underestimate income variation. In addition to economic factors, I discuss ideological motivations and the role of organized recruitment in creating migration decisions.

By combining historical data, cross-national linkage methods, and economic analysis, this thesis contributes to research on historical record linkage, labor migration and the economics of migration.

Contents

1 Introduction

The emigration of American Finnish workers to Soviet Karelia is a unique and underexplored phenomenon. While migration from Europe to North America and the corresponding return migration have been extensively studied (Abramitzky et al., 2012, 2019), this particular case – in which subsequent migration was directed to a different country – remains less well understood. The American Finnish migration movement, often referred to as Karelian Fever, was driven by a combination of economic setbacks, ideological motivations, and systematically organized recruitment efforts by Soviet Karelian authorities (Kangaspuro, 2012). Understanding selection into migration provides valuable insights into the economic and socioeconomic determinants of migration, as well as the role of ideology in creating migration decisions.

In the early twentieth century, the Finnish population in North America was employed in blue-collar industries such as mining, lumbering, and construction (Korkiasaari, 1989). Employment in these industries deteriorated significantly with the onset of the Great Depression, which began in the United States and quickly spread to Canada. At the same time, the Soviet Union set ambitious objectives under the first Five-Year Plan. In response, Soviet Karelian authorities launched an active recruitment campaign targeting North American workers, promising stable employment and better living conditions. The most in-demand occupations were forestry workers, carpenters, and construction workers. Despite the generous promises, many migrants faced harsh conditions, political repression, and eventual purges under Stalin's regime. (Golubev and Takala, 2014).

This study aims to answer the following key research questions:

- How can state-of-the-art historical record linkage methods be applied to cross-national and multilingual datasets, and how effective are they in the case of American Finnish migration to Soviet Karelia?
- To what extent were American Finnish migrants to Soviet Karelia selected based on economic status, and how do occupation-based income measures affect the detection of selection patterns?

- What new insights does the first individual-level analysis of the Karelian Fever contribute to the broader understanding of labor migration?

By addressing these questions, this thesis contributes to the broader literature on historical record linkage, labor migration and the economics of migration.

The existing literature has examined the emigration of the American Finnish population to Soviet Karelia (Kangaspuro, 2012; Kero, 1983; Golubev and Takala, 2014), but this is the first study to analyze migration selection using detailed individual-level linked data. Previous research indicates that both economic hardship and ideological factors influenced migration decisions. However, quantitative methods have not yet been applied to assess whether there was a systematic pattern in the selection into migration. Linking Soviet Karelian immigration records with U.S. and Canadian census data introduces a novel empirical approach to studying migration selection and occupational outcomes.

This study also contributes to the literature by providing evidence that occupation-based income scores, such as *OCCSCORE* (Ruggles et al., 2024), may not accurately capture income variation within occupations. These types of occupation-based income scores are widely used in historical income analysis and labor economics (Inwood et al., 2019). I address these concerns by comparing *OCCSCORE*-based analysis with individual-level earnings data from the 1931 Canadian Census.

The empirical analysis is based on digitized historical records from Soviet Karelia, the United States, and Canada. The primary sources include (i) Soviet Karelian immigration records collected from various archival sources in the Republic of Karelia; (ii) 1930 U.S. Census data; and (iii) 1931 Canadian Census data.

The use of these datasets involves historical record linkage techniques to link individuals across sources. Using automated linking algorithms, I identify migrants and stayers, enabling a comparative analysis of their occupational and socioeconomic characteristics. The main empirical methods include descriptive statistics, regression models estimating selection into migration, and robustness checks addressing potential biases in the occupational income score measure.

The key findings of this study are as follows. American Finnish emigrants to Soviet Karelia largely reflected the overall occupational distribution of American Finns in North America. The results indicate negative selection into migration among Canadian migrants, as those who left for Karelia had, on average, lower earnings than those who stayed. This pattern, however, is not observed among U.S. migrants, possibly due to the absence of personal income data for the U.S. emigrants. Nevertheless, it cannot be ruled out that, in the United States, there was no selection into migration based on economic status. The results suggest that analyzing within-occupation earnings variation is critical for understanding migration selection.

The analysis highlights significant variation in occupational categorization between North American census records and Soviet Karelian immigration records. Although *forestry worker* was the most common occupation recorded in Soviet Karelian immigration documents, its prevalence was notably lower in the North American census records. The results indicate a fundamental shift in labor classification under the Soviet planned economy.

The structure of the thesis is the following. Section 2 provides a historical background on American Finnish emigration to Soviet Karelia and introduces the labor migration framework. Section 3 describes the data sources and the measures of economic status in census records. Section 4 outlines the historical record linkage techniques. Section 5 presents the research design and empirical strategies used to analyze migration selection. Section 6 reports the main findings on selection into migration. It also discusses the measurement issues. Section 7 discusses potential mechanisms underlying migration decisions, including economic conditions, ideological motivations, and labor market dynamics. Section 8 concludes the study by summarizing the main findings and suggesting directions for future research.

During the writing process, I have used artificial intelligence tools to assist with grammar correction and language refinement. All research design, analysis, and substantive content are my own.

2 Background

In this section, I provide historical context for American Finnish emigration to Soviet Karelia and discuss labor migration theory. First, I examine the emigration of Finns to North America, briefly describing the economic and social factors that influenced migration decisions and outlining the characteristics of Finnish migrants. Second, I go through the phenomenon known as Karelian Fever, which led thousands of Finnish Americans and Canadians to migrate to Soviet Karelia during the decade of the 1930s. I outline the economic and ideological motivations identified in the existing literature behind this movement and emphasize the organized recruitment efforts that facilitated it. Finally, I present the labor migration framework, focusing on models of migrant self-selection based on relative economic returns.

2.1 Historical Background

2.1.1 The Finnish Population in North America

One of the largest migration episodes in modern history occurred between 1850 and 1913, a period known as the Age of Mass Migration. During this time, nearly 30 million people immigrated to the United States. The phenomenon was mostly driven by European immigrants (Abramitzky et al., 2014). In line with this international trend, one of the most significant waves of emigration in Finland's history occurred between 1860 and 1930. By the onset of the Second World War, an estimated 380,000 to 400,000 Finns had emigrated to North America, with approximately 80 percent settling in the United States and the remainder in Canada. The largest influx of Finnish immigrants to the United States occurred between 1899 and 1913. After the United States imposed immigration restrictions in the 1920s, the flow of Finnish immigrants shifted to Canada. This second wave, however, was substantially smaller than the previous one to the United States. (Korkiasaari, 1989).

The general reasons behind the migration were primarily economic. The push factors included rapid population growth and an increase in the landless population. North America was considered attractive due to its need for workers and its chances of a better life (Ko-

rkiasaari, 1989). The socioeconomic status of emigrants was generally low rather than high. Between 1893 and 1910, over 85 percent of the emigrants belonged to the tenant class or were workers (Korkiasaari, 1989). For example, in 1905, 89% of all emigrants were farmers, crofters, cottagers, workers, or their children (Kero, 1974). Most Finnish immigrants settled in areas near the U.S.-Canada border. East Coast states, such as New York and Massachusetts, along with the Great Lakes region – particularly Michigan and Minnesota – were the primary centers of Finnish settlement. In Canada, the Finnish population spread along the border from east to west. (Korkiasaari, 1989).

The otherwise fairly cohesive Finnish American community was politically divided into two groups: the Reds, representing the radical labor movement, and the Whites, who identified as patriotic loyalists (Kangaspuro, 2012). The Finnish leftists played a significant role in the labor movements of the U.S. and Canada. In many regions, the Finns were the largest foreign nationality involved in the Communist movement. Nearly 45% of members of the Workers Party of America were Finnish in 1923. The party was a legal party of American Communists from 1921 to 1929. Facilitated by interactions within this socialist movement, Finnish American radical leftists established connections with Finnish refugee radicals in Soviet Russia. (Kostiainen, 1978).

2.1.2 Karelian Fever: Migration from North America to Soviet Karelia

Following their initial migration to North America in search of better economic opportunities, many Finnish immigrants found themselves facing hardship during the 1930s. The Great Depression (1929–1939) was a severe global economic downturn triggered by the 1929 stock market crash, leading to widespread unemployment and economic hardship across the United States. As Canada was an important export partner of the United States, the effects of the recession were substantial there as well, driven by falling global market prices (Kero, 1983). The recession had a significant impact on the Finnish-American population. A majority of Finnish Americans worked in industries vulnerable to economic fluctuations, such as construction, mining, lumbering, and farming. The risk of unemployment was particularly high for the most recent first-generation immigrants. (Kangaspuro, 2012).

Meanwhile, the Soviet Union was recovering from the civil war (1917–1922) and sought both political and economic support from the international labor movement. In 1928, the central authorities launched the First Five-Year Plan as part of a broader effort to industrialize the Soviet economy. The ambitious targets assigned to Soviet Karelia under the plan were unattainable without the importation of additional labor. The most needed occupational groups were forestry workers, carpenters, and construction workers, as Soviet Karelia had significant forest reserves and its role in the Soviet economy was to produce wood. (Kangaspuro, 2012).

In the early 1920s, a group of Finnish Americans established the commune of *Kylväjä* (*The Sower*) in the southern Soviet Union, followed by the founding of the *Säde* commune in Soviet Karelia. The American Finnish leftist press, particularly the newspapers *Työmies* (United States) and *Vapaus* (Canada), promoted the Karelian communes and actively participated in fundraising efforts for Soviet Karelia. In the summer of 1930, the Communist Party of the Soviet Union called for the organized recruitment of labor immigrants to Soviet Karelia. Finnish American Communists were an obvious target of this campaign. As early as 1921, some radical leftist activists in the United States had established the Soviet Karelia Relief Committee, which raised funds and sent equipment to Soviet Karelia. The committee's financial supervisor, Matti Tenhunen, was now assigned a key role in organizing the emigration of Finns from the United States to Karelia. He began working as a representative of the People's Commissariat of Labor of Soviet Karelia in New York. (Golubev and Takala, 2014).

To organize the propagandist recruitment campaign, the Karelian Technical Aid Committee (KTAC) was founded in New York in May 1931. The committee had its main office in New York and a second office in Toronto, Canada. It also had representatives in several U.S. states that had large Finnish populations. About half a year later, the Resettlement Administration of the Council of People's Commissars of Soviet Karelia was established in Petrozavodsk to oversee immigration to Soviet Karelia. It was responsible for recruitment, labor reception, and the registration of new arrivals. KTAC employees in North America officially worked for the Resettlement Administration. The KTAC played a significant role in

supplying equipment and other resources to Soviet Karelian industries. The majority of its funding came from the immigrants themselves, many of whom planned to leave America permanently and sold their property before emigrating. Purchased equipment was transported to Soviet Karelia mainly alongside the immigrants, and these deliveries were not trivial for the Karelian industries. Many immigrants also brought their own tools and equipment with them. (Golubev and Takala, 2014).

The recruitment process was systematically organized by the Soviet Karelian authorities and carried out in several stages. Prospective emigrants from North America were required to submit an application to a representative of the Karelian Technical Aid Committee to obtain a letter of recommendation. Information about immigration opportunities spread through leftist newspapers such as *Työmies* or by word of mouth within Finnish communities. Once approved by the KTAC, applications were sent to Soviet Karelia, where local authorities sent them to Moscow, where the entry visas were issued. (Golubev and Takala, 2014).

The importation of foreign labor did not occur without conflicts among political institutions in the Soviet Union. The autonomous regional administration in Karelia, led by Finnish politician Edward Gylling, strongly supported Finnish labor migration. However, the central government in Moscow sought to reduce the power of regional autonomies and was concerned about the growing influence of nationalist ideologies among the Finnish population in Karelia (Kangaspuro, 2012). At the same time, the Communist Parties in the United States and Canada feared that mass emigration of the American Finnish population to Soviet Karelia would weaken their ranks, as Finns constituted a significant portion of party membership in many areas. Despite this resistance, the KTAC's efforts successfully generated enthusiasm for Soviet Karelia among American Finnish communities. (Golubev and Takala, 2014).

This enthusiasm and the migration to Soviet Karelia are known in historiography as *Karelian Fever*. Organized, propaganda-driven enthusiasm for Karelia was undoubtedly a significant factor behind the migration. However, there are several other factors, such as the weak economic conditions in America and the availability of secure jobs in Soviet Karelia. Additionally, Finnish communities in America were often somewhat separated from the broader society, which prevented the Finnish identity, culture, and language from vanishing even

in the long term. Soviet Karelia, being close to Finland and sharing common characteristics such as climate, language, and societal practices, was seen as a potential place to live. (Kangaspuro, 2012).

During the Karelian Fever, an estimated 6,000 to 6,500 American Finns migrated to Karelia (Kero, 1983). Figure 1 shows the annual number of arrivals from the U.S. and Canada. The data were collected from Soviet Karelian archival sources and are described in detail in the next section. As illustrated in the figure, recruitment was most intense in 1931 and 1932, after which the number of arrivals declined rapidly in 1933 and 1934. The last group of Finns from America migrated to Karelia in the summer of 1935 (Kero, 1983).

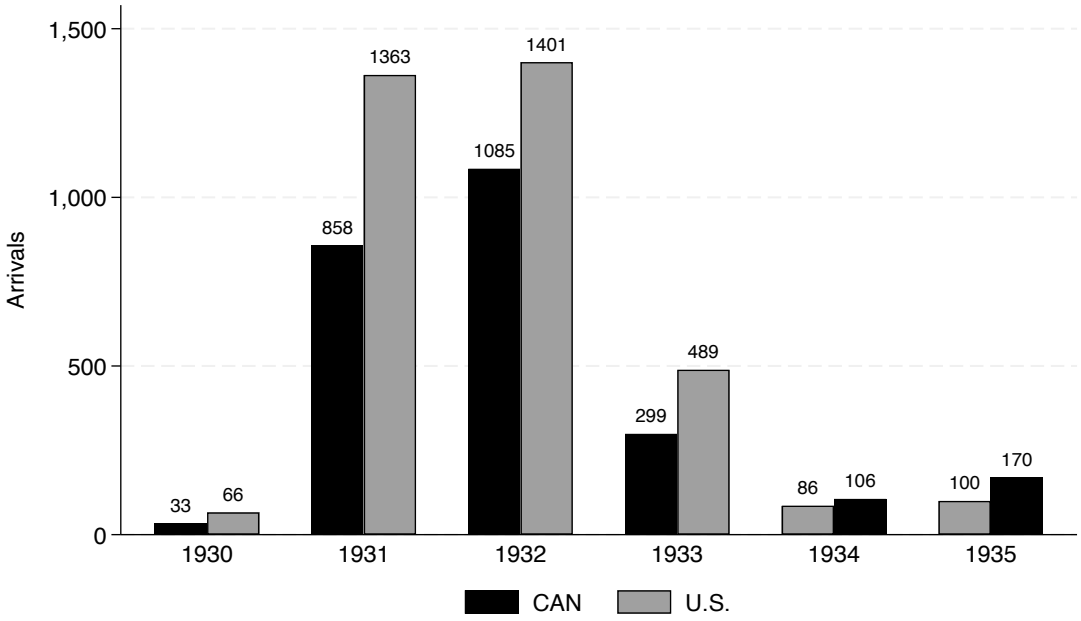

Figure 1: Annual Arrivals of Karelian Fever Migrants by Country (1930–1935)
Notes: Based on the author's own calculations. Anomalous years excluded from the figure due to their likely inaccuracy.

The conditions at the Soviet Karelia were likely a surprise for the American Finns. They were accustomed to the high standard of living in North America. Housing, food, and work conditions were harsh and significantly different from what the arrivals had expected. The propaganda had painted a vastly different picture of life in the destination. Cultural clashes were not avoided. The local supervisors and management took a negative view of the

immigrants' more efficient working methods. Hierarchical management structures and strict discipline hindered the adoption of best practices in forestry and other industries. However, in some areas, production of wood in Karelia increased significantly with the arrival of American Finns. (Kangaspuro, 2012).

As the political climate in the Soviet Union grew increasingly hostile toward foreign nationalities in the 1930s, conditions became difficult for American Finns. Many of the workplaces where they were employed became targets of the most severe purges during Stalin's Great Terror (1937–1938). The campaign against foreign nationalities resulted in executions, imprisonments, and deportations. An example of this is the fate of the four prominent leaders of KTAC: three of them were executed, and only one survived, never visiting Soviet Karelia (Golubev and Takala, 2014). By 1950, only 405 Finns from the United States and 205 from Canada were recorded in Karelia's official statistics (Kangaspuro, 2012).

2.2 Labor Migration Framework

Understanding the economic logic behind migration decisions is crucial for analyzing selection into migration. The theoretical framework in the labor migration literature often builds on the work of Borjas (1987), which itself is based on the Roy model (Roy, 1951). This model frames migration as a form of self-selection driven by relative returns to skill. Individuals choose to migrate if the expected gain in lifetime earnings exceeds the cost of migration. More specifically, when skills are assumed to be transferable between the origin and destination countries, it is the relative return to skill – not absolute income levels – that determines who migrates. According to the model, if the destination country offers a higher return to skill than the origin country (indicating greater income inequality), migrants will tend to come from the upper end of the skill distribution. This phenomenon is referred to as positive selection. Conversely, if the origin country rewards skill more than the destination country (reflecting a more equal income distribution in the destination), migrants will tend to come from the lower end of the skill distribution, referred to as negative selection. (Borjas, 2016).

Empirical studies generally support the model. For example, research has found a negative cross-sectional relationship between immigrants' earnings in the United States and the

level of income inequality in their countries of origin (Cobb-Clark, 1993; Bratsberg, 1995). However, there are also contradictory findings. Chiquiar and Hanson (2005), for instance, show that Mexican migrants to the United States were drawn from the middle of the income distribution, not the lower end, as the Borjas model would suggest.

A key assumption of the model is that migration is voluntary and driven purely by economic incentives. The case of Karelian Fever is more complex. Migration was not only the result of individual decisions but was also shaped by a coordinated recruitment campaign organized by Soviet authorities. These complexities suggest that selection may have occurred within, rather than between, occupational groups – which is not easily captured by the standard Borjas-Roy framework. Rather than comparing income distributions between North America and Soviet Karelia, Finnish workers may have evaluated their prospects within their own professions.

3 Data

To examine the factors driving the migration of American Finns to Soviet Karelia, I link individual-level data from Soviet Karelia immigration records with U.S. and Canadian census records. The censuses provide pre-migration characteristics of the immigrants, and linking them with the Soviet Karelian immigration records enables me to use statistical methods to analyze the underlying factors of migration. In this section, I introduce the data sources and their properties. Additionally, I discuss the measures of economic status in census records.

3.1 Soviet Karelian Immigration Records

Data on American Finns who migrated to Soviet Karelia are collected from the National Archives of the Republic of Karelia (Karjalan tasavallan Kansallisarkisto, 2022). The archive collection consists of documents about Finns who migrated to the republic of Karelia from the United States and Canada. Most of these documents are part of the Emigration Administration (*Siirtolaishallinto*) collection (R-685) under the Council of People's Commissars of

the Karelian Autonomous Soviet Socialist Republic (Karelian ASSR). Additionally, documents come from Council of People's Commissars of Karelian ASSR, the Presidium of the Supreme Council of the Karelian ASSR, the Communist Party of Karelian ASSR and several Karelian enterprises where North American Finns were employed. The language of the documents varies between Russian, Finnish and English.

The archival records have later been digitized and contain a significant number of duplicates, as they often list the same individual multiple times. For example, the same person may appear both in a specific passenger record and in a comprehensive summary document. The records often vary considerably depending on the source, and the digitization process has resulted in inconsistent treatment of variables.

By combining these archival sources, I create a definitive dataset of American Finns who migrated to Soviet Karelia in 1930–1935. Data cleaning, deduplication, and improvements are described in detail in Appendix B. The resulting dataset contains 6,663 individuals and includes variables such as name, age, gender, and date of arrival. For some individuals, information is also available on their occupational title in Soviet Karelia, the work place, and the municipality where they worked. The descriptive statistics are shown in Table 1. The statistics are similar for both United States and Canadian immigrants: approximately two-thirds are men, and the average age is about 30 years.

Table 1: Descriptive Statistics of Soviet Karelian Immigration Records

	U.S.	Canada	Missing
Share of men (%)	64.35	66.84	71.86
Year of birth, mean	1899.45	1903.51	1897.23
	(14.69)	(13.02)	(12.98)
Year of arrival, mean	1931.84	1932.04	1932.21
	(1.00)	(1.13)	(1.99)

Notes: Standard deviations are reported in parentheses. The column *Missing* includes individuals with no recorded country variable.

3.2 North American Census Data

3.2.1 United States Census Records

The United States Census data come from the IPUMS database, which provides a complete-count microdata sample of the 1930 U.S. Census (Ruggles et al., 2024). For the linking process with the Soviet Karelian immigration records, I use a subsample consisting of all records where Finland is registered as the individual's birthplace, mother's birthplace, or father's birthplace. This subsample contains 336,773 individuals.

3.2.2 Canadian Census Records

The Canadian Census data are retrieved from Statistics Canada (2023) and originate from the 1931 Census. The sample includes only individuals whose ethnicity is recorded as "Finn" or "Finnish". Additionally, individuals whose ethnicity is recorded as "Swedish" but who were born in Finland are included, as they are likely Swedish-speaking Finns. The resulting dataset contains 38,113 individuals. According to Korkiasaari (1989), there were 43,585 Canadian Finns, including both first- and second-generation individuals. Compared to this number, it is likely that my sample does not fully capture the entire population of Canadian Finns. This shortcoming is likely due to data consistency issues, such as misspelled ethnicity variables in the digitized census. It is also likely that I am unable to capture every second-generation Finn. However, this should not pose major issues, as most of the Karelian Fever immigrants included in the later analysis were first generation immigrants.

3.2.3 Measures of Economic Status in Census Records

The variables measuring economic status in census records are crucial for migration analysis, as they are essential for explaining selection in economic terms. The U.S. Census data do not include individual income information, as earnings were not recorded until the 1940 Census (Inwood et al., 2019). Because of this, the analysis must be conducted using an occupational-

level income variable. The U.S. Census data contain an *OCCSCORE*[1] variable, which is the key variable in the empirical analysis in the Section 6. I also merge *HISCO*[2] occupation variable for specific analysis purposes.

Contrary to the 1930 U.S. Census, the 1931 Canadian Census includes an individual earnings variable, which is used in the Canadian-specific analysis. Since this earnings variable is more specific than the occupation-level *OCCSCORE* and *HISCO* variables, it provides more precise information about the socioeconomic characteristics.

Inwood et al. (2019) show that occupation-based income scores are not particularly useful, especially when comparing income level differences between groups. Their findings indicate that using occupational income scores tends to underestimate earnings differences between native-born and foreign-born individuals. They also demonstrate that occupation-based income scores even moved in the opposite direction of individual-level earnings between the 1921 and 1931 Canadian censuses. This discrepancy arises because immigrants experienced a decline in earnings while their occupational profiles remained approximately the same. Abramitzky et al. (2014) note that occupation score is a proxy for labor market earnings that varies between occupations but not within occupations. Saavedra and Twinam (2020) find out that using *OCCSCORE* biases the race and gender earnings gap toward zero and can even result in estimated gaps with the wrong sign.

On the other hand, the Canadian Census lacks a pre-classified occupation variable and instead contains a digitized descriptive profession variable. This variable is translated into a categorical variable based on the HISCO classification. Due to the significant variation in the handwritten forms of the occupation variable, it is not possible to assign a HISCO category to all observations. This *HISCO* variable is again used to merge the *OCCSCORE* variable to the Canadian census sample. This allows me to run pooled regressions for both

[1]"*OCCSCORE* is a constructed 2-digit numeric variable that assigns occupational income scores to each occupation in all years which represents the median total income (in hundreds of 1950 dollars)", (Ruggles et al., 2024).

[2]The HISCO classification (Historical International Standard Classification of Occupations) is a systematic framework by Leeuwen et al. (2002) used to categorize and classify historical occupational titles. It provides a standardized way to compare and analyze occupational data across time periods, regions, and datasets. Developed in the 1990s and 2000s, HISCO is based on the structure of the International Labour Organization's International Standard Classification of Occupations (ISCO) system but adapted for historical contexts.

samples. However, in the Canada-specific analysis, I primarily use the individual earnings variable.

4 Linking

During the past decade, the digitization of historical census records has created significant opportunities for research in the social sciences and economics. These often extensive datasets contain detailed personal data, enabling analyses that were previously very challenging. One prominent application of these datasets is historical record linkage, which allows researchers to compare individuals' characteristics either across different observation years (Abramitzky et al., 2014) or between national censuses of different countries (Abramitzky et al., 2019). In this section, I first present the most commonly used methods for linking historical records in the literature. Then, I describe how Soviet Karelian immigration records are linked with North American census data. Finally, I outline my methodological contribution to historical record linking.

4.1 Overview of Historical Record Linking

The availability of complete count censuses has driven the development of automated linking algorithms. However, census data typically lack personal identification numbers, making it challenging to definitively link individuals between datasets. Most algorithms rely on reported names and ages as key variables to establish matches. This approach faces a significant challenge: how to handle cases where multiple individuals share identical names and ages. Historical record data are often inconsistent, with name spellings differing substantially between records. The age variable is also prone to inaccuracies, as individuals may report rounded ages. The timing of census data collection can affect age variables, introducing at least a one-year discrepancy in the results.

According to Abramitzky et al. (2021), an optimal algorithm fulfills four primary objectives: First, the algorithm needs to be correct with very few wrong matches (reduce type I errors).

Second, it must work well by finding most of the correct matches (reduce type II errors). Third, it should create linked samples that are similar to the group being studied. Finally, it should be easy for most researchers to use, considering the current limits of computers and resources.

State-of-the-art automated linking algorithms fall into two categories: machine learning algorithms and probabilistic linking algorithms. An example of the former is the algorithm developed by Feigenbaum (2016), which utilizes machine learning techniques. An example of the latter is the fully automated probabilistic algorithm created by Abramitzky et al. (2020), which "uses the expectation maximization (EM) algorithm to combine age and name distances into a single score reflecting the probability that each potential pair of records is a true match." Another commonly used approach is the more conservative ABE linking algorithm, developed by (Abramitzky et al., 2012, 2014, 2019), which builds on the work of Ferrie (1996). Next, I will provide a more detailed explanation of the ABE linking algorithm according to Abramitzky et al. (2021).

Typically, the ABE linking algorithm involves two datasets, A and B, with the objective of finding a match in dataset B for each record in dataset A. The linking variables are assumed to be first name, last name, and birth year. First, the name variables in both datasets are standardized through cleaning and unification. Common nicknames are replaced with their formal equivalents to ensure consistency. Dataset A is filtered to include only unique observations. For each record in dataset A, the algorithm identifies any identical observations in dataset B based on the name and birth year. If an exact match is found and it is unique, the pair is considered a match. If multiple potential matches are found, the observation is excluded from the process to avoid ambiguity. If no exact match is identified, the algorithm expands the search to include records in dataset B with a birth year within ± 1 year of the record in dataset A. If no matches are still found, the scope is further widened to ± 2 years. If a match cannot be identified even after this step, the observation is dropped from the process and treated as a unique. (Abramitzky et al., 2021).

This is the basic process of the ABE linking algorithm. However, modifications are often applied to improve its accuracy. One common enhancement is the standardization of name

variables using phonetic algorithms, such as the NYSIIS or Soundex transformations. Another approach involves calculating the Jaro-Winkler string similarity for the name variables of potential matches, which provides a measure of similarity between strings. This method is described in detail in Abramitzky et al. (2019).

4.2 Linking the Soviet Karelian Immigration Records to the U.S. and Canadian Censuses

The linkage between the Soviet Karelian data and the census records is conducted using the ABE linking algorithm with Jaro-Winkler string comparison. Typically, the algorithm is applied to datasets in the same language. However, in this case, the names have been transliterated across different languages, thus contributing to existing research on historical record linkage.

The linking process begins by filtering census records to include only individuals with Finnish heritage. Names are then cleaned and standardized before linking. The linking is conducted in two stages: first, using both name similarity (measured by Jaro-Winkler string distances) and birth year, and second, for unmatched cases, using only name similarity. Minor age discrepancies of up to two years are permitted. Only unique matches are retained to improve match quality. Conducting the linking process using a two-stage approach balances accuracy and maximizes matches while minimizing false matches. Since the linkage is conducted between immigration records and census data, allowing matches without known birth year information does not pose the same risk of false matches as it does in census-to-census linkage. The linking algorithm is described in detail in Appendix C.

The resulting match rates are similar to those reported in previous literature. The match rates are presented in Table 2. For example, Abramitzky et al. (2014) achieves a forward (linking individuals from an earlier census to a later one) match rate of 16% for native-born men between the 1900 U.S. Census and 1910 and 1920 censuses. The rate for foreign-born men in Abramitzky et al. (2014) is lower, averaging 12%. One possible explanation for the lower match rate for foreign-borns is return migration to the country of origin. However, the

return migration does not affect my matching rates, as I do not perform census-to-census linking. Ferrie (1996) reports a linkage rate of 19%, which is notably higher than the rate achieved in this study. Abramitzky et al. (2012) achieves a forward match rate of 29% (and a corresponding backward match rate of 23%), while Long and Ferrie (2013) attain a rate of 22% in forward linking process.

Table 2: Match Rates by Country

Country	Observations	Number matched	Match rate
U.S.	4,108	649	15.8%
Canada	3,108	364	11.7%

Notes: The total number of observations is 6,663. The U.S. and Canadian samples in this table overlap because some records lack information on the country of origin. These records are first included in the U.S. linking process and, if no match is found, subsequently included in the Canadian linking process. Therefore, they appear in observations from both the U.S. and Canada.

Several factors may lower match rate in this linking process. Linking is typically conducted through census-to-census linkage, where records tend to be more consistent (Abramitzky et al., 2014). In contrast, my study links records from various sources, such as travel documents and censuses. Another possible explanation is the variation in the spelling of Finnish American names across different sources (see Figure 2). The names were originally Finnish, but upon migration to North America, individuals either anglicized their names for integration or retained their original forms. During the recruitment process to Soviet Karelia, the names were recorded using the Cyrillic alphabet. In the digitization of the Karelian archives, these names were subsequently transliterated back into Finnish and, in some cases, further harmonized. This process introduces variation in name spellings, adding noise to the linking process and potentially lowering the match rates. Additionally, I include all individuals listed in the Soviet Karelian immigration records, including women and children. This may lower the match rates, as historical records tend to be more complete for men (Folbre and Abel, 1989).

However, certain factors are likely to increase the match rate. I do not require matches to be

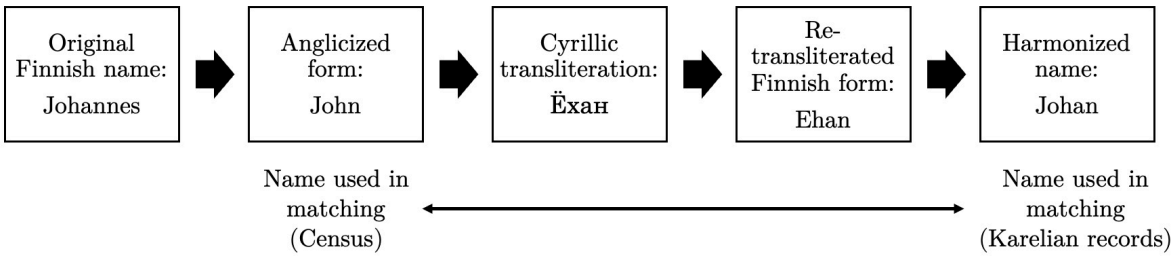

Figure 2: Example of the Name Transformation Process Between Records

unique within a five-year band. In the second stage of linking, I include observations from the Karelian records with no recorded birth year. Although this raises the match rate, the effect is marginal (18 additional matches in the U.S. sample and 22 in the Canadian sample). Since the censuses were recorded in 1930 and 1931, and the main sources for the Soviet Karelian immigration records originate from the years immediately following, the temporal proximity may also further improve match rates.

I assume that the linking process does not introduce bias that would systematically mismatch or exclude certain observations. Table 3 presents the comparison between the Karelian immigration records and the linked sample across the following variables: sex, year of birth, occupational score, and country of origin. The proportion of males in the original dataset is slightly higher than in the linked sample, but the difference is not statistically significant. The mean of the birth year variable is identical in both samples. The *OCCSCORE* variable is slightly higher in the Karelian records, but the difference is marginally significant only at the 10% significance level. The distribution of the country variable appears to differ between the two samples. The share of U.S.-origin immigrants is over six percentage points higher in the linked sample, suggesting that the linking algorithm performs better when linking to the U.S. Census. This may be due to the lower quality of the name variables in the Canadian Census. It is also possible that the original Soviet Karelian immigration archives contain records with misreported country variables (e.g., individuals recorded as being from Canada when they should have been listed as from the United States).

However, since most of the analyses are conducted separately for the U.S. and Canadian

censuses, this difference in proportions should not cause significant issues. Overall, Table 3 suggests that there is no significant bias in the linking process, apart from country-related discrepancies.

Table 3: Comparison of Soviet Karelian Immigration Records and Linked Sample

Variable	Soviet Karelian Immigration Records	Linked	p-value
Sex (% male)	65.9%	64.0%	0.22 (Chi-square)
Year of birth, mean	1901	1901	0.95 (t-test)
OCCSCORE, mean	18.4	17.9	0.09 (t-test)
Country (% USA)	57.9%	64.1%	0.00 (Chi-square)

Notes: The p-values indicate statistical significance levels for comparisons between the Soviet Karelian records and the linked sample. The t-test is used for continuous variables (Year of birth, *OCCSCORE*) to compare means, while the Chi-square test is used for categorical variables (Sex, Country) to assess differences in proportions. The sample sizes are 6,663 for the Soviet Karelian data and 1,013 for the linked dataset.

5 Research Design

This section explains the research framework used to study whether American Finnish migrants to Soviet Karelia were economically selected. The main part of the analysis is based on regression models that compare occupational status and income levels between migrants and non-migrants, using individual-level data linked from census and immigration records. The analysis begins with summary statistics, followed by more detailed descriptive analysis of the distribution of occupations and geographic locations among those who migrated and those who stayed. This gives a first look at the background of the migrants and shows if they were more likely to come from certain jobs or areas. The descriptive results also help the interpretation of the main results.

Before the analysis of selection into migration, I first examine how American Finns were positioned in the North American labor market compared to the general population. This provides important context for understanding their economic conditions prior to migration. Utilizing the complete U.S. Census of 1930, I assess whether American Finn men had sys-

tematically lower occupational income scores than other men in the population. I start with a baseline OLS model, in which the main explanatory variable is *American Finn*$_i$, a dummy indicating whether the individual has Finnish origins. Additional control variables are added to account for demographic and socioeconomic characteristics, along with regional-specific effects. The baseline equation is

$$\ln(OCCSCORE_i) = \alpha + \beta_1 \cdot American\ Finn_i + \beta_2 \cdot age_i + \mathbf{X}_i\boldsymbol{\gamma} + \epsilon_i \tag{1}$$

where $\ln(OCCSCORE_i)$ is the natural logarithm of the occupation score (measured in 1950 dollars) for individual i, *American Finn*$_i$ is a dummy variable equal to one if individual i has Finnish origins, age_i is the age of individual i and $\mathbf{X}_i$ is a vector of additional control variables, such as squared age, family size, urban/rural status, years in the United States, literacy, and English proficiency, with $\boldsymbol{\gamma}$ representing the corresponding coefficient vector. β_1 and β_2 are the coefficients to be estimated, and ϵ_i is the error term.

This analysis is limited to the U.S. Census because the Canadian dataset includes only individuals of Finnish origin, making it impossible to compare Finnish Canadians to the general population. A key limitation of this approach is that *OCCSCORE* does not measure individual-level earnings but instead provides a proxy based on occupational income rankings. As a result, the findings shed more light on the occupational distribution of American Finns rather than their direct earnings.

I then turn to the main analysis of selection into migration. The objective is to examine whether individuals who migrated to Soviet Karelia differed systematically from those who remained, particularly with respect to occupational status and income. To investigate this, I employ a similar OLS regression framework to estimate selection into migration, using regressions of the following form:

$$\ln(OCCSCORE_i) = \alpha + \beta_1 \cdot recruited_i + \beta_2 \cdot age_i + \mathbf{X}_i\boldsymbol{\gamma} + \epsilon_i \tag{2}$$

where $\ln(OCCSCORE_i)$ is the natural logarithm of the occupation score (measured in 1950 dollars) for individual i, *recruited*$_i$ is a dummy variable equal to one if individual i was

recruited to migrate to Soviet Karelia, age_i is the age of individual i and $\mathbf{X}_i$ is a vector of additional control variables (such as squared age, English proficiency, and year of immigration), with $\boldsymbol{\gamma}$ representing the corresponding coefficient vector. β_1 and β_2 are the coefficients to be estimated, and ϵ_i is the error term. Using this approach, I measure the effect of income on migration with β_1, which captures the difference in earnings between migrants and stayers.

The Canadian census data include individual-level earnings information, enabling a more direct estimation of income differences. Therefore, I replicate the analysis using the Canadian earnings variable instead of *OCCSCORE*. This approach is particularly valuable for assessing whether the *OCCSCORE* variable effectively captures income differences.

The Great Depression led to rising unemployment, and American Finns were in a vulnerable position in the labor market (Kero, 1983). To test whether unemployment increased the likelihood of migration from Canada to Soviet Karelia, I run a naïve OLS regression of the following form:

$$recruited_i = \alpha + \beta_1 \cdot unemployment_i + \epsilon_i \tag{3}$$

where $recruited_i$ is a dummy variable equal to one if individual i was recruited to migrate from Canada to Soviet Karelia, $unemployment_i$ is a dummy variable if the individual i was unemployed, β_1 is the coefficient to be estimated and ϵ_i is the error term.

My recruited sample is relatively small compared to total number of Finnish-origin individuals in North America. This can result in large standard errors and difficulties in obtaining statistically significant effects. The statistical power may be weak, making it challenging to detect meaningful effects of occupational income on migration.

6 Results

This section presents the empirical findings of the thesis and is organized into the following subsections: the occupational status of Finnish-origin individuals in the United States in 1930; the overall characteristics of the Finnish-origin population in North America; the

occupational and geographical characteristics of those recruited to Soviet Karelia; and the economic selection into migration. The main analysis investigates whether migrants were positively or negatively selected based on occupational income or individual earnings. The section concludes with a discussion of measurement issues, emphasizing the limitations of using occupation-based income variables, such as *OCCSCORE*, in income analysis.

6.1 Occupation-Based Income of American Finns in the 1930 U.S. Census

To provide background on the economic position of American Finns before migration, I begin by examining their occupational status in the 1930 U.S. Census. The analysis compares Finnish-origin individuals to the broader male population using occupational income scores. Since direct earnings data are not available, the results reflect differences in occupational income levels rather than within-occupation wage variation.

Table 4 presents the baseline estimates from equation (1) on the effect of being an American Finn on the occupation-based income variable. Across all specifications, being an American Finn is associated with a statistically significant penalty on income. The coefficient is negative in all models and takes values between –0.094 and –0.156. Since the dependent variable is in logarithmic form, it can be interpreted as a percentage change.

Table 4: Occupation-Based Income in the 1930 U.S. Census

	Dependent variable: ln(*OCCSCORE*)				
	(1)	(2)	(3)	(4)	(5)
American Finn	-0.116***	-0.116***	-0.094***	-0.156***	-0.148***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Age		0.007***	0.048***	0.059***	0.056***
		(0.000)	(0.000)	(0.000)	(0.000)
Age^2			-0.001***	-0.001***	-0.001***
			(0.000)	(0.000)	(0.000)
Number of family members			-0.012***		
			(0.000)		
Urban			0.464***		
			(0.000)		
Years in the United States				0.003***	-0.000***
				(0.000)	(0.000)
Literate				0.320***	
				(0.000)	
Speaks English				0.115***	0.240***
				(0.001)	(0.001)
Region fixed effects	✗	✗	✗	✗	✓
Adj. R-squared	0.000	0.030	0.287	0.099	0.147
Observations	34,498,319	34,498,319	34,498,319	32,873,993	32,873,993

Notes: Standard errors in parentheses. Region fixed effects included in 5. Only working-age (15–64 years) men included. The dependent variable is the logarithm of $100 \times$ *OCCSCORE*.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results suggest that, even after controlling for regional differences, literacy, and English proficiency, Finnish-origin individuals in the U.S. in 1930 earned over 10% less than the rest of the population recorded in the census.

I also run the regression for the Scandinavian subsample of the 1930 U.S. Census. The results differ only slightly from those in Table 4 and are presented in Appendix A, Table A.4. The income gap remains above 10%, indicating that American Finns had lower-paying

occupations even compared to Swedes, Norwegians, Danes, and Icelanders.

There are several possible explanations for this earnings gap. American Finns were likely overrepresented in low-paying industries such as forestry, mining, and construction. In addition, they may have faced labor market discrimination due to their leftist ideological affiliations. Limited English proficiency might also have contributed to this disparity, as Finnish communities tended to be relatively isolated from broader society. These hypotheses could potentially be tested using Canadian data, including occupation fixed effects, if the dataset included other Nordic nationalities. Unfortunately, this is not the case, and the analysis must be left to future research.

Once again, individual-level earnings data are not available, and the findings reflect occupational status rather than income differences within occupational groups. The results may help explain why the phenomenon of Karelian Fever emerged: Finnish immigrants were not as well positioned in the labor market as the average American or other Nordic nationalities at the time. The prospect of an alternative future – or a new beginning – in Soviet Karelia may thus have appeared particularly attractive to them.

6.2 Descriptive Analysis

Understanding the occupational characteristics, demographic profiles, and geographical composition of American Finns who migrated to Soviet Karelia provides valuable insight into both the broader migrant population and the recruitment policies that shaped this migration. By comparing occupational distributions across the Soviet Karelian immigration records, the recruited sample (previously referred to as the linked sample), and census data, it is possible to assess whether recruitment targeted specific skill groups. In this subsection, I compare the summary statistics of recruited individuals with those of stayers. I also examine the most common occupations across the datasets, particularly the representation of forestry workers, and analyze the geographical distribution of American Finns in the census records to assess whether migration was geographically selective.

Table 5 presents summary statistics for Finnish-origin men aged 15 to 64 in the 1930 U.S.

Census and the 1931 Canadian Census, categorized by migration status. Only men are included, as the subsequent analysis relies on these male subsamples.

Table 5: Summary Statistics by Country and Migration Status

Variable	U.S., Recruited	U.S., Stayers	Canada, Recruited	Canada, Stayers
Age, mean	37.0	34.8	32.4	33.8
Year of immigration, mean	1910.3	1906.3	1921.4	1918.9
Unemployed (%)	23.2%	25.2%	25.8%	23.9%
Speaks English (%)	92.0%	92.6%	72.7%	78.8%
Earnings, mean	—	—	465.9	523.3

Notes: The table presents summary statistics for Finnish-origin men aged 15–64 in the 1930 U.S. Census and 1931 Canadian Census, divided by whether they were recruited to migrate to Soviet Karelia or not. Percentages indicate the share unemployed or English-speaking. U.S. earnings are not available. The sample sizes are 131,215 for the U.S. and 17,667 for the Canadian sample. Due to missing data, not all individuals are included in the calculation of each variable.

The mean age of stayers is similar in both the United States and Canada, at approximately 34–35 years. However, recruited individuals in the United States are older than the stayers, whereas in Canada, the opposite pattern is observed – migrants are younger than the stayers. The average year of immigration to North America differs notably between the two countries, which is explained by U.S. immigration restrictions introduced in the 1920s that redirected the flow of Finnish immigrants toward Canada (Korkiasaari, 1989). The shorter period spent in North America may help explain why the Soviet Karelian immigration records include relatively more Canadians than U.S.-origin migrants, as recent arrivals are more likely to undertake onward migration (Borjas, 2016). This hypothesis is further supported by the finding that the recruited individuals had arrived in North America approximately three to four years later than the stayers, as indicated in the table.

Unemployment rates range from 23 to 26 percent across all groups. In the United States, unemployment is higher among stayers, whereas in Canada, the unemployment rate is higher among migrants. English proficiency is substantially higher in the U.S. sample. This may partly reflect data limitations: 16 percent of observations for the English proficiency variable

are missing in the U.S. census, compared to near-complete reporting in the Canadian data. It is likely that non-English speakers are overrepresented among the missing cases in the U.S. sample. Within the U.S. data, English proficiency does not vary significantly between migrants and stayers. In contrast, the Canadian data show a notable gap – English proficiency is six percentage points higher among stayers – suggesting that language skills may have influenced migration decisions.

Finally, the table reports mean earnings for the Canadian subsample. The average earnings of migrants are more than 10 percent lower than those of the stayers. This disparity suggests the possibility of negative selection into migration within the Canadian subsample. However, descriptive statistics alone cannot fully explain the observed selection patterns. To more precisely assess these patterns, Section 6.3 presents a regression analysis that evaluates the economic selection into migration to Soviet Karelia.

6.2.1 Occupational Statistics

The most common occupations in the Soviet Karelian immigration records, recruited sample and the pooled censuses are shown in Table 6. The top ten occupations are very similar in both the recruited and census samples, with no occupations appearing to be significantly over- or underrepresented in the recruited sample. However, most of the recruited individuals have a different occupation listed in the Soviet Karelian immigration records. The most notable difference is that approximately 45% of working-age men are reported as forestry workers in the immigration records, whereas the share of forestry workers is only about 5% in the recruited and census samples. Additionally, there are virtually no recorded farmers in the Soviet Karelian immigration records, but they still represent the second-largest occupational group in the recruited and census samples. The occupational change from pre-migration census data to Soviet Karelian records suggests a process of occupational reclassification, either administratively or through real job changes post-migration. It is important to note that recorded census occupations do not provide information about an individual's work history or professional skills. The census occupation captures professional characteristics only at a specific point in time.

Table 6: The 10 Most Common Occupations for American Finns: Soviet Karelian Immigration Records, Recruited and Censuses

Soviet Karelian Immigration Records				**Recruited Sample**				**Censuses**			
Occupation	Freq.	Percent	HISCLASS 5	Occupation	Freq.	Percent	HISCLASS 5	Occupation	Freq.	Percent	HISCLASS 5
Forestry worker	1,193	44.66	5	General worker	118	24.58	5	General worker	21,157	17.01	5
Carpenter	516	19.32	4	Farmer	79	16.46	3	Farmer	17,556	14.11	3
General worker	234	8.76	5	Carpenter	55	11.46	4	Farm worker	12,327	9.91	5
Driver	176	6.59	4	Miner	38	7.92	4	Miner	11,053	8.88	4
Sawyer	92	3.44	4	Farm worker	31	6.46	5	Carpenter	10,275	8.26	4
Mechanics	83	3.11	4	Forestry worker	24	5.00	5	Forestry worker	6,954	5.59	5
Metal worker	56	2.10	4	Other prod worker	16	3.33	5	Other prod worker	5,257	4.23	5
Miner	50	1.87	4	Fisherman	13	2.71	5	Mechanic	4,079	3.28	4
Farm worker	48	1.80	5	Blacksmith	9	1.88	4	Manager	3,668	2.95	2
Machinist	36	1.35	4	Manager	8	1.67	2	Driver	3,497	2.81	4
Total (top 10)	2,484	93.00		Total (top 10)	391	81.47		Total (top 10)	95,823	77.02	
Outside top 10	187	7.00		Outside top 10	89	18.53		Outside top 10	28,584	22.98	
Total	**2,671**	**100.00**		**Total**	**480**	**100.00**		**Total**	**124,407**	**100.00**	

Notes: Table consists of men aged 15 to 64, with age calculated using the 1930 U.S. Census and the 1931 Canadian Census. Occupations are categorized based on the HISCLASS 5 classification (Kok and Mandemakers, 2009). HISCLASS 1–2 = 1 (Elite), HISCLASS 3–7 = 2 (Lower middle class), HISCLASS 8 = 3 (Self-employed farmers), HISCLASS 7 and 9 = 4 (Skilled workers), HISCLASS 10–12 = 5 (Unskilled workers and farm workers).

A similar table for the recruited sample, based on occupational titles recorded in Soviet Karelia (see Table A.3 in Appendix A), yields an occupational distribution that closely follows the Soviet Karelian immigration records presented in the first column of Table 6. This similarity suggests that the recruited sample accurately represents the broader Soviet Karelian immigration data and supports the validity of linking individual-level U.S. and Canadian census records with Soviet archival material.

The separate tables for the U.S. and Canadian samples are presented in Appendix A (Tables A.1 and A.2). The share of forestry workers among U.S. individuals in the Karelian immigration records is 38.4%, while it is 54.4% for Canadians. Overall, the occupational distribution is very similar between U.S. and Canadian individuals in the Soviet Karelian data. The most significant difference between these two countries in the census and recruited samples is that, in the Canadian Census, most occupations fall into the "general worker" category. No further information is available regarding the specific occupational distribution within this category. However, some inferences can be drawn from Table 7, which lists the ten most common Soviet Karelian occupations among recruited pre-migration general workers.

To assess whether a specific pattern existed in the recruitment policy for Soviet Karelia, Table 7 presents (i) the most common occupations classified under the occupation of *forestry worker* in the Soviet Karelia and (ii) the most common occupations classified under the occupation of *general worker* in the censuses.

Table 7: Occupation Transition Table

Census Occupations of Karelian *Forestry Workers*			**Karelian Occupations of Census *General Workers***		
Occupation	Freq.	Percent	Occupation	Freq.	Percent
General worker	51	28.49	Forestry worker	51	53.68
Farmer	32	17.88	Construction worker	10	10.53
Farm worker	17	9.50	Carpenter	9	9.47
Forestry worker	14	7.82	Sawyer	7	7.37
Miner	14	7.82	Driver	4	4.21
Carpenter	8	4.47	Electrician	2	2.11
Blacksmith	7	3.91	Miner	2	2.11
Other prod worker	4	2.23	Roofer	2	2.11
Manager	3	1.68	Blacksmith	1	1.05
Housekeeper	3	1.68	Concreting	1	1.05
Total (top 10)	153	85.47	Total (top 10)	89	93.68
Total	**179**	**100.00**	**Total**	**95**	**100.00**

Notes: Table consists of men aged 15 to 64, with age calculated using the 1930 U.S. Census and the 1931 Canadian Census. Only the ten most common occupations are displayed under each variable.

The first column of the table presents the pre-migration occupations of immigrants who were recorded as forestry workers in the Soviet Karelian immigration records. The most common occupations are general worker, farmer, farm worker, and forestry worker. While their proportions are slightly higher than in the overall census sample, the differences remain minor. The share of forestry workers in the census sample is 5.6 percent, while 7.8 percent of the Karelian forestry workers were recorded as forestry workers in the census – a slightly higher proportion. This suggests that many individuals employed in forestry in Soviet Karelia were not necessarily working in that sector prior to migration.

However, this does not necessarily mean they lacked relevant experience. Many Finnish migrants initially worked in the timber and mining industries and later transitioned to farming, often establishing their own small farms. The mean age of census-recorded forestry workers is 38, compared to 45 for farmers, supporting the hypothesis of a career progression from manual labor to farming. The complete age structures of forestry workers and farmers are presented in Appendix A, Figure A.1. Moreover, American Finnish farmers often cleared

their own land, indicating that they possessed practical logging skills despite being recorded as farmers in the census. This underscores the limitation of a single census snapshot in capturing the full scope of an individual's work experience.

Further evidence is provided in the second column of Table 7, which shows that individuals recorded as general workers prior to migration were later dispersed across a wide range of occupations, with no single post-migration job dominating relative to the overall occupational distribution in Soviet Karelia. Many transitioned into forestry work, suggesting that general worker category likely included individuals with logging experience.

Among the 24 recruited individuals who were forestry workers before migration, 20 have post-migration occupations recorded, and 14 of these worked in forestry in Soviet Karelia. This means that about 70% of pre-migration forestry workers continued in occupation consistent with their North American occupation. This finding further supports the idea that pre-migration occupations influenced job placements in Soviet Karelia.

The occupational distributions reported here are consistent with Efremkin (2011), who found that 48% of immigrants were forestry workers, 18% carpenters, 8% construction workers, 5% drivers, and 4% sawyers. While these shares do not align neatly with the pre-migration census data, it is likely that many immigrants possessed relevant skills, whether or not officially recorded. Recruiters may have prioritized individuals with experience in forestry-related work, including farmers who had engaged in land clearing and manual labor.

Some of the mismatch between census and post-migration occupations also reflects structural differences in labor allocation. In contrast to the self-reported, market-driven North American labor system, Soviet Karelia followed a centrally planned model. Individuals were often assigned jobs based on state needs rather than personal preference or background. Although farmers were the second-largest group in the census data, they may have been redirected into forestry due to the elimination of private farming under collectivization and the campaign's specific focus on recruiting forestry workers.

Overall, the apparent mismatch between census occupations and Soviet employment records likely reflects a combination of selective recruitment practices – favoring those with practical skills – and centralized labor assignments in the Soviet system. Farmers, general workers,

and others not formally classified as forestry workers may nonetheless have been seen as suitable for forestry work based on their experience.

6.2.2 Geographic Distribution

Following the examination of occupational patterns, I analyze the geographical distribution of both the recruited individuals and the stayers in North America. This comparison helps assess whether the migrants came from specific regions. The geographical distribution of Finnish Americans in the 1930 U.S. Census and the recruited sample is presented in Table 8.

Table 8: Geographical Distribution of American Finns in the 1930 U.S. Census and in the Recruited Sample

1930 U.S. Census			**Recruited Sample**		
State	Freq.	Percent	State	Freq.	Percent
Michigan, USA	76,869	22.83	Michigan, USA	135	20.80
Minnesota, USA	62,966	18.70	Minnesota, USA	79	12.17
New York, USA	29,488	8.76	New York, USA	64	9.86
Massachusetts, USA	27,986	8.31	Massachusetts, USA	61	9.40
Washington, USA	23,707	7.04	Ohio, USA	48	7.40
California, USA	17,594	5.23	Washington, USA	42	6.47
Wisconsin, USA	15,236	4.53	Illinois, USA	34	5.24
Ohio, USA	13,157	3.91	Wisconsin, USA	28	4.31
Oregon, USA	12,814	3.81	Pennsylvania, USA	25	3.85
Illinois, USA	10,385	3.08	New Hampshire, USA	20	3.08
Other states	46,484	13.81	Other states	113	17.41
Total	**336,686**	**100.00**	**Total**	**649**	**100.00**

Most Finns in the United States were concentrated in Michigan (22.83%) and Minnesota (18.70%). These states were key industrial, mining, and forestry regions that attracted Finnish immigrants due to employment opportunities (Kero, 1996). Michigan and Minnesota are followed by New York (8.76%) and Massachusetts (8.31%). The pattern in the recruited sample is very similar: Michigan (20.80%), Minnesota (12.17%), New York (9.86%), and Massachusetts (9.40%) are the most common home states for migrants. These results suggest that that Karelian Fever migrants were not geographically selected into migration: no state is significantly over- or underrepresented in the recruited sample.

The geographical distribution of Finns in the 1931 Canadian Census and the recruited sample is presented in Table 9. The results are very similar to those in the U.S. census and recruited sample.

Table 9: Geographical Distribution of American Finns in the 1931 Canadian Census and in the Recruited Sample

1931 Canadian Census			**Recruited Sample**		
Province	Freq.	Percent	Province	Freq.	Percent
Ontario, CA	24,315	63.76	Ontario, CA	235	64.56
British Columbia, CA	6,136	16.09	British Columbia, CA	75	20.60
Alberta, CA	2,983	7.82	Alberta, CA	26	7.14
Saskatchewan, CA	2,132	5.59	Saskatchewan, CA	14	3.85
Quebec, CA	1,435	3.76	Quebec, CA	10	2.75
Manitoba, CA	921	2.42	Manitoba, CA	2	0.55
New Brunswick, CA	101	0.26	New Brunswick, CA	2	0.55
Nova Scotia, CA	69	0.18			
Yukon, CA	34	0.09			
Northwest Territories, CA	4	0.01			
Other provinces	3	0.01	Other provinces	0	0.00
Total	**38,133**	**100.00**	**Total**	**364**	**100.00**

In Canada, Ontario overwhelmingly dominates (63.76%), followed by British Columbia (16.09%) and Alberta (7.82%). The proportions in the recruited sample are very similar, with Ontario at 64.56%, British Columbia at 20.60%, and Alberta at 7.14%. In Ontario, Finns primarily settled near mining areas, construction sites, and centers of the railroad and forestry industries. As in the United States, the results provide evidence that there was no geographical pattern in migration from Canada to Soviet Karelia.

6.3 Main Analysis: Selection into Migration

The main analysis of this study focuses on economic selection into migration. This subsection examines the migration from North America to Soviet Karelia by analyzing the relationship between pre-migration occupational and income status and the likelihood of migration. The analysis explores whether migrants differed systematically from stayers in terms of occupa-

tional income and other background characteristics, and whether those who left for Soviet Karelia were positively or negatively selected based on their economic standing.

Table 10 presents my baseline estimates from equation (2) on the effect of pre-migration occupational income on selection into migration. Specifications (1) and (4) include only age as a control variable, while specifications (2) and (5) also control for a quadratic in age, English proficiency, and year of immigration to the country of census. Specifications (3) and (6) exclude the year of immigration but include fixed effects for region.

Table 10: Selection into Migration from North America to Soviet Karelia

	U.S. Sample			Canadian Sample		
	Dependent variable: ln(*OCCSCORE*)			Dependent variable: ln(*OCCSCORE*)		
	(1)	(2)	(3)	(4)	(5)	(6)
Recruited	0.030	-0.013	-0.043*	0.029	0.017	0.013
	(0.023)	(0.021)	(0.020)	(0.029)	(0.028)	(0.028)
Age	0.006***	0.065***	0.059***	0.001	0.019***	0.025**
	(0.001)	(0.009)	(0.010)	(0.001)	(0.003)	(0.005)
Age^2		-0.001***	-0.001***		-0.000***	-0.000**
		(0.000)	(0.000)		(0.000)	(0.000)
Speaks English		0.175***	0.155***		0.030	0.035**
		(0.027)	(0.025)		(0.031)	(0.010)
Year of immigration		-0.001			0.003	
		(0.001)			(0.002)	
Region fixed effects	✗	✗	✓	✗	✗	✓
Adj. R-squared	0.020	0.075	0.163	0.000	0.019	0.118
Observations	110,776	106,653	106,653	13,631	11,301	13,630

Notes: Standard errors in parentheses. The samples include working-age (15–64 years) men for the 1930 U.S. Census and 1931 Canadian Census. The dependent variable in the U.S. sample is the logarithm of $100 \times OCCSCORE$. The dependent variable in the Canadian sample is the logarithm of $100 \times OCCSCORE$. U.S. standard errors clustered by state and Canadian by province. Fixed effects for region included in specifications (3) and (6).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In the U.S. sample, the coefficient on the *recruited* dummy variable is close to zero and

statistically insignificant in models (1) and (2). In model (3), the coefficient is slightly higher but remains relatively small, although it is statistically significant at the 5% level. Overall, the effect is minor and does not provide strong evidence that migrants differed from stayers in terms of occupational status in the U.S. sample. As expected, age has a positive effect on the dependent variable. English proficiency is associated with significantly higher income.

In the Canadian sample, the coefficient for *recruited* remains relatively small (between 0.013 and 0.029) across all specifications. It is also statistically insignificant, indicating that the occupational status of migrants was not significantly different from that of stayers. Overall, the effects of the control variables are relatively small, suggesting that this model does not effectively explain occupational income levels.

However, the Canadian Census data include an individual-level earnings variable. Since it is not determined solely by occupational status, it potentially contains greater variation, providing an opportunity to deeper analyze the effect of income on selection into migration within the Canadian sample. I run a regression identical to that presented in Table 10, but with the logarithm of earnings as the dependent variable instead of the logarithm of the occupational income score. The results of this regression are shown in Table 11. For comparability, the results from Table 10 for the U.S. sample are included.

Table 11: Selection into Migration from North America to Soviet Karelia, Individual Earnings Variable in the Canadian Sample

	U.S. Sample			Canadian Sample		
	Dependent variable: ln(*OCCSCORE*)			Dependent variable: ln(*earnings*)		
	(1)	(2)	(3)	(4)	(5)	(6)
Recruited	0.030	-0.013	-0.043*	-0.197***	-0.151**	-0.193**
	(0.023)	(0.021)	(0.020)	(0.040)	(0.043)	(0.044)
Age	0.006***	0.065***	0.059***	0.007***	0.038***	0.038***
	(0.001)	(0.009)	(0.010)	(0.001)	(0.006)	(0.003)
Age^2		-0.001***	-0.001***		-0.000***	-0.000***
		(0.000)	(0.000)		(0.000)	(0.000)
Speaks English		0.175***	0.155***		0.377***	0.406***
		(0.027)	(0.025)		(0.012)	(0.028)
Year of immigration		-0.001			-0.009***	
		(0.001)			(0.001)	
Region fixed effects	✗	✗	✓	✗	✗	✓
Adj. R-squared	0.020	0.075	0.163	0.005	0.039	0.043
Observations	110,776	106,653	106,653	10,429	8,966	10,428

Notes: Standard errors in parentheses. The samples include working-age (15–64 years) men for the 1930 U.S. Census and 1931 Canadian Census. The dependent variable in the U.S. sample is the logarithm of $100 \times$ *OCCSCORE*. The dependent variable in the Canadian sample is the logarithm of earnings. U.S. standard errors clustered by state and Canadian by province. Fixed effects for region included in specifications (3) and (6).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Now, the results for the Canadian sample differ significantly from those obtained when the dependent variable is the occupational income score. The coefficient for the *recruited* variable is consistently negative and statistically significant across all specifications, ranging from –0.151 to –0.197. This indicates that migrants from Canada had 15–20% lower pre-migration earnings compared to non-migrants, suggesting negative selection in Canada. English proficiency appears to increase earnings by nearly 40%, which is a notably large effect. However, the adjusted R-squared values indicate that these models explain relatively little variation in earnings. The results for the U.S. sample have been described previously.

The samples used for the regressions with the occupational income score and the earnings variable are not identical, as some Canadian census records include occupational information but lack earnings data. This discrepancy is reflected in the sample size: the Canadian sample in Table 11 is over 20 percent smaller than in Table 10. This difference in sample composition poses a potential threat to the comparability of the results. To address this concern, I test whether the results remain consistent when restricting the analysis to individuals included in Table 10. The results are presented in Appendix A Table A.5. The regression coefficients are similar to those in the unrestricted sample and statistically significant at the five percent level only in specifications (1) and (3). These results still provide evidence that in Canada, the recruited migrants were negatively selected into migration.

According to Efremkin (2011), unemployment rates in Canada reached approximately 32 percent of the population in 1932. He notes that unemployment among Karelian Fever migrants from Canada exceeded 30 percent. To test whether unemployment influenced the decision to migrate, I estimate equation (3). The results are presented in Table 12.

Table 12: The Effect of Unemployment on Migration

	Dependent variable: *recruited*	
	(1)	(2)
Unemployment	-0.001	-0.001
	(0.001)	(0.001)
Age	-0.000	0.001
	(0.000)	(0.000)
Speaks English	-0.002*	-0.000
	(0.001)	(0.001)
Age^2		-0.000
		(0.000)
Constant	0.018***	0.005
	(0.003)	(0.009)
Region and industry fixed effects	✗	✓
Adj. R-squared	0.000	0.003
Observations	10,409	10,409

Notes: Standard errors in parentheses. The samples include working-age (15–64 years) men in the 1931 Canadian Census. The dependent variable is a dummy *recruited* indicating if the individual migrated from Canada to Soviet Karelia. Standard errors clustered by province. Fixed effects for region and industry included in specification (3).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results in Table 12 do not support the hypothesis that unemployment was a major determinant of migration to Soviet Karelia. In both specifications, the coefficient on unemployment is small and statistically insignificant. Although no direct relationship between unemployment and migration is found, it is possible that migrants had lower earnings as a consequence of being unemployed. To investigate this, I estimate a model in which unemployment is included as a control variable in the earnings regression. The results, presented in Table 13, show that even after accounting for unemployment, recruited individuals earned over 15 percent less than stayers. The coefficient on the unemployment variable is substantial, suggesting that unemployment reduced earnings by roughly 50 percent.

Table 13: Selection into Migration, Controlling for Unemployment

	Dependent variable: ln(*earnings*)			
	(1)	(2)	(3)	(4)
Recruited	-0.197**	-0.191**	-0.177***	-0.185**
	(0.041)	(0.043)	(0.031)	(0.041)
Age	0.007***	0.039***	0.034***	0.032***
	(0.001)	(0.005)	(0.004)	(0.005)
Age^2		-0.000***	-0.000***	-0.000***
		(0.000)	(0.000)	(0.000)
Speaks English		0.417***	0.309***	0.273***
		(0.019)	(0.016)	(0.023)
Unemployment				-0.538***
				(0.029)
Region and industry fixed effects	✗	✗	✗	✓
Adj. R-squared	0.005	0.034	0.098	0.153
Observations	10,409	10,409	10,409	10,409

Notes: Standard errors in parentheses. The samples include working-age (15–64 years) men in the 1931 Canadian Census. The dependent variable is the logarithm of earnings. Standard errors clustered by province. Fixed effects for region and industry included in specification (4).
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Taken together, the results presented in Table 12 do not support the hypothesis that unemployment was a major determinant of migration to Soviet Karelia. In both specifications, the coefficient on unemployment is small and statistically insignificant, indicating no relationship between unemployment and the probability of being recruited. This result does not necessarily imply that economic challenges were irrelevant to the migration decision. It may rather reflect the presence of liquidity constraints: individuals who were unemployed might have lacked the financial resources or social support needed to undertake an international migration, even if they had strong incentives to leave.

This interpretation aligns with broader findings in the migration literature, which emphasize that migration requires a minimum level of liquidity. Individuals whose liquid assets fall below the cost of migration may benefit from migrating but are unable to finance it (Chernina

et al., 2014; McKenzie and Rapoport, 2010). In this context, the poorest individuals may have been effectively trapped in place.

This interpretation is further supported by the earnings regression in Table 13. While unemployment is not associated with a higher likelihood of migration, it is strongly linked to substantially lower earnings – reducing occupational score-based earnings by over 50 percent. This suggests that although unemployment had severe economic consequences, these may have acted as a constraint on mobility rather than a catalyst for it.

6.4 Measurement Issues

While the main analysis provides estimates of economic selection into migration, the findings, particularly in the Canadian sample, vary depending on the income variable used. As shown in Section 6.3, no significant relationship is found when using the occupational income score (*OCCSCORE*), whereas the individual-level earnings variable indicates a negative income-based selection into migration. This discrepancy raises important questions about the measurement properties and reliability of the income proxies used in the analysis.

To assess which results are more reliable, I construct a new occupational income variable – referred to as the Income Index – calculated as the mean of observed earnings within each OCC1950-classified occupation (Ruggles et al., 2024). Using this earnings-based occupational measure, I evaluate whether the results remain consistent. Separate regressions are conducted for each income variable: individual earnings, *OCCSCORE*, and the *Income Index*. The set of independent variables remains constant across all estimations: *recruited*, age, and English proficiency. The results are presented in Table 14.

Table 14: Selection into Migration, Different Income Variables

	Dependent variable:		
	ln(*earnings)*	ln(*OCCSCORE*)	ln(*Income Index*)
	(1)	(2)	(3)
Recruited	-0.137*	-0.019	-0.003
	(0.049)	(0.037)	(0.022)
Age	0.006***	0.002*	0.001
	(0.001)	(0.001)	(0.001)
Speaks English	0.398***	0.033	0.116***
	(0.032)	(0.035)	(0.019)
Constant	5.331***	7.493***	6.043***
	(0.044)	(0.044)	(0.034)
Adj. R-squared	0.031	0.005	0.027
Observations	8,224	8,224	8,224

Notes: Income Index is a manually created variable that represents average earnings at the industry level. It functions similarly to the *OCCSCORE* variable but is constructed within the 1931 Canadian Census Finnish sample. Standard errors in parentheses. Standard errors clustered by province.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results suggest that the occupation-based earnings variables do not adequately explain selection into migration. Using occupation-based income proxies – both the widely used *OCCSCORE* and the newly constructed *Income Index* – no significant income selection is observed. This suggests that variation in income occurs primarily within occupations rather than between them.

That being said, the U.S. census estimations do obviously not capture internal wage variation. Finns were employed in industries with lower average wages, but it is impossible to determine whether they earned lower wages within those industries. Therefore, the results presented in Table 10, for example, must be interpreted with caution. As demonstrated in the Canadian case, the occupation-based income variable can produce misleading results. These findings align well with Inwood et al. (2019).

While occupation-based income variable remains a standard measure in historical income

analysis (see for instance Abramitzky et al. (2012), Abramitzky et al. (2014) and Collins and Wanamaker (2014)), my results indicate that it does not fully capture income differences within occupations. The significant variance observed suggests that individuals within the same occupational category had notably different income levels, raising concerns about the precision of occupation-based income estimates.

7 Mechanisms

In the previous results section, I show that Karelian Fever migrants may have been economically selected into migration. Although no clear patterns emerge in terms of occupations or regions of origin, the analysis reveals that the recruited individuals in Canada earned approximately 15–20 percent less than non-migrants. This finding, which is based only on the Canadian sample, suggests negative selection: individuals with fewer financial resources may have been more likely to migrate.

Applying the standard Roy model of migration (Borjas, 1987) to these results yields a mixed picture. Negative selection suggests that migrants believed that the income distribution in Soviet Karelia was more equal than in North America, offering higher relative returns to lower-skilled workers. However, Finnish migrants often outperformed local Soviet workers and even received higher wages (Kero, 1983), which challenges the explanatory power of a purely economic model in the context of Karelian Fever. Since there are no statistics on the return to skill for occupations such as logging in either North America or Soviet Karelia, it is difficult to assess how migrants evaluated the economic returns to migration. It is possible, for example, that lower-skilled individuals perceived the income distribution in Soviet Karelia to be flatter than it actually was.

As not all economic indicators support this pattern, it appears that financial hardship alone does not fully explain the observed migration behavior. The analysis has focused on the economic factors underlying Karelian Fever. However, the existing literature identifies several additional motivations behind the migration decisions. Lee (1966) presents that migration behavior is influenced by four different types of factors: push factors in the origin area, pull

factors in the destination area, intervening factors, such as economic realities, and individual factors, such as migrant's own ideology.

Push factors refer to conditions in the home country that drive individuals to leave, making life there unfavorable. Examples include job loss, lack of opportunities, unfavorable political conditions, or discrimination. (Lee, 1966).

The Great Depression caused severe economic distress in North America, serving as a major push factor for migration. American Finnish communities were particularly affected by the collapse of key industries such as mining, lumber, and farming, leading to widespread unemployment. Simpson and Swan (1937), for example, show that lumber production in the United States declined by nearly 70% between 1925 and 1933. During the same period, the gross income of the lumber industry decreased by approximately 66.5%. The decline of the industries that had previously provided employment made the North American labor market increasingly unfavorable for Finnish workers, thereby shaping their migration decisions.

According to Kangaspuro (2012), the Finnish population in North America was not fully integrated into society. They were viewed as radical due to their political leaning toward communism. Suspicion toward communism, which intensified during the economic recession, was particularly directed at the Finnish population and can be considered an ideological push factor. Additionally, Finnish immigrants faced challenges in labor market integration due to limited language proficiency. These conditions represent clear push factors contributing to the formation of Karelian Fever.

Pull factors refer to conditions that attract people to a particular destination, such as better job opportunities, favorable political conditions, or higher living standards (Lee, 1966). The Karelian Technical Aid Committee promoted an image of a worker's utopia with jobs for all, an appealing prospect for American Finns suffering from the effects of the Great Depression. Employment was highly valued, and the promise of job security served as a major pull factor. Moreover, American Finns had already emigrated once for work, making them potentially more open to relocating again compared to the general population (Kangaspuro, 2012).

In addition to economic pull factors, Soviet Karelia offered significant cultural incentives. The political environment there was highly favorable to many Finns living in North America.

The presence of previously migrated Finns and the idea of a Finnish-led government in the Republic of Karelia led many recruits to believe that the region could become a "New Finland". Furthermore, its proximity to the Finnish border and its familiar climate can also be considered pull factors in this context (Kangaspuro, 2012).

Intervening factors are obstacles that migrants must overcome before migration can take place (Lee, 1966). These include distance and transportation. The Karelian Technical Aid Committee significantly reduced these obstacles by providing support throughout the entire migration process. By organizing transatlantic travel, assisting with documentation, and coordinating employment and housing in Soviet Karelia, the committee lowered the effective cost of migration, potentially also altering the usual economic self-selection dynamics predicted by the Borjas-Roy model (Borjas, 1987).

Research on this topic has also focused on the ideological reasons behind the Karelian Fever. Lee (1966) identifies individual factors as the fourth necessary condition for migration. This condition applied to at least a fraction of American Finns, as many in both the U.S. and Canada leaned strongly to the left politically. To examine whether ideological reasons influenced migration decisions, I compare the distribution of the home states of Karelian Fever migrants with historical Finnish socialist party membership rates. Figure 3 illustrates the share of Karelian Fever migrants by U.S. state, as well as the membership rates of the Finnish Socialist Federation in 1917 and the Finnish Federation of the Workers Party in 1923[3].

There appears to be no clear relationship between migration and party membership rates. The states with the highest number of Karelian migrants appear on the left side of the chart, yet these states do not correspond to those with the highest party membership rates. Data are collected 10–20 years before the Karelian fever migration, making it difficult to draw conclusions based solely on these data. However, it is reasonable to assume that membership rates, to some extent, reflect the ideological characteristics of the American Finn population across states, even a few years after the data were collected.

Ideological factors do not appear to have been the primary drivers of migration, but they

[3]The Finnish Socialist Federation split in 1921, and afterward, the majority of its members joined the Communist Workers Party of America in 1922. In 1924, it was renamed the Finnish Federation of the Workers Party of America. (Kostiainen, 1978)

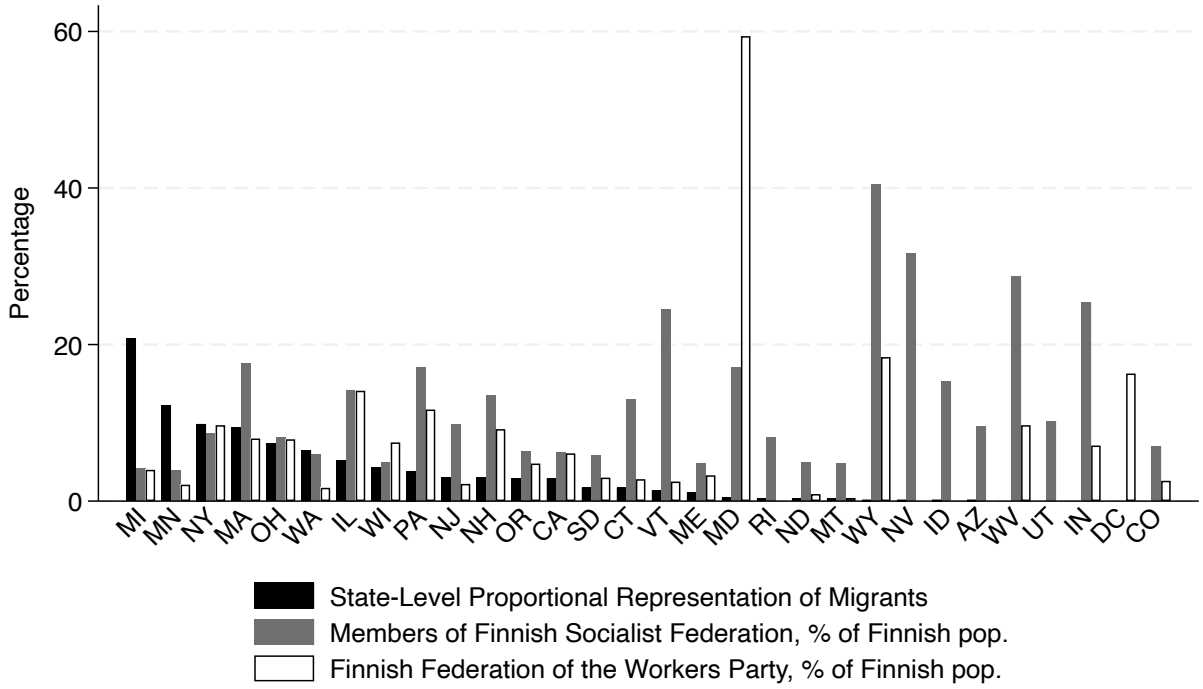

Figure 3: Share of Karelian Migrants by State, Membership of the Finnish Socialist Federation in 1917, and Membership of the Finnish Federation of the Workers' Party in 1923

Notes: The data on party membership rates come from Kostiainen (1978). The figure includes only the states presented in his research.

certainly played a role. The four conditions for migration outlined by Lee (1966) are met in the case of the Karelian Fever: selection into migration likely resulted from a combination of multiple factors acting simultaneously. Economic and cultural push-and-pull factors, combined with an effective recruitment campaign and an ideologically receptive environment, made Soviet Karelia appear to be an attractive place to build a new future.

8 Conclusions

During the Age of Mass Migration, millions of Europeans migrated to the United States and Canada. This migration movement was often followed by return migration, as many did not find the promising future they had expected in North America. The case of Karelian Fever

presents a fascinating anomaly within the broader migration pattern, as it involved substantial repeat migration to Soviet Karelia. This study helps to fill a gap in the existing literature by utilizing detailed historical individual-level data to examine the factors underlying the Karelian Fever.

The results of the migration analysis provide insights into the economic differences between migrants and stayers. In the Canadian sample, using individual-level earnings data from the 1931 Census, individuals who migrated to Soviet Karelia earned approximately 15–20% less than those who remained. This negative selection suggests that individuals facing greater economic hardship were more likely to migrate. However, it is worth noting that, when using the occupational-level income variable, the selection effect is either statistically insignificant or small in size. Additional analysis shows that unemployment status alone does not predict migration to Soviet Karelia in the Canadian sample. After controlling for unemployment, migrants still have significantly lower earnings than stayers.

The examination of the U.S. sample provides no consistent evidence of selection into migration. This analysis is conducted using *OCCSCORE* as the dependent variable. The results may reflect either a true absence of economic selection or limitations associated with using occupational-level variables as proxies for income. However, I find that Finnish-origin individuals in the 1930 U.S. Census had systematically lower occupational income scores than the general population, including other Nordic immigrant groups. Since the U.S. census data do not include individual-level earnings, it is not possible to make a direct comparison with the Canadian case. These findings underline the importance of using direct earnings measures over occupation-based proxies, which often fail to capture income variation within occupations.

In addition to its findings on labor migration and the economics of migration, this thesis makes a methodological contribution by advancing the use of historical record linkage across different archival sources. It presents a successful application of an automated historical linking algorithm that utilizes string comparison (the ABE-JW algorithm) to link Soviet Karelian immigration records with North American census data. As far as the author is aware, this is the first systematic linkage of Soviet-origin immigration records with North

American census data. The challenges of transliteration, duplicate records, and varying name conventions are addressed, contributing to broader methodological developments in historical data linkage.

This study shows that although economic drivers played a role in the migration of Finnish individuals to Soviet Karelia, it was not just about jobs or money. The Karelian Fever was driven by a combination of different factors: poor economic conditions in North America, strong political convictions, and organized recruitment efforts by Soviet Karelian authorities. These migrants were not only seeking better employment opportunities – they were also motivated by the idea of building a new kind of society. Although some questions remain open, this study provides a foundation for future research on migration decisions using large-scale individual-level datasets.

References

Abramitzky, R., Boustan, L., and Eriksson, K. (2012). Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. *American Economic Review*, 102(5):1832–1856. 1, 4.1, 4.2, 6.4, C

Abramitzky, R., Boustan, L., and Eriksson, K. (2014). A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. *Journal of Political Economy*, 122(3):467–506. Publisher: The University of Chicago Press. 2.1.1, 3.2.3, 4, 4.1, 4.2, 4.2, 6.4, C

Abramitzky, R., Boustan, L., and Eriksson, K. (2019). To the New World and Back Again: Return Migrants in the Age of Mass Migration. *ILR Review*, 72(2):300–322. Publisher: SAGE Publications Inc. 1, 4, 4.1, C

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., and Pérez, S. (2021). Automated Linking of Historical Data. *Journal of Economic Literature*, 59(3):865–918. 4.1

Abramitzky, R., Mill, R., and Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach*. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):94–111. Publisher: Routledge _eprint: https://doi.org/10.1080/01615440.2018.1543034. 4.1

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *The American Economic Review*, 77(4):531–553. 2.2, 7

Borjas, G. J. (2016). *Labor Economics*. McGraw Hill, New York, 7th edition. 2.2, 6.2

Bratsberg, B. (1995). The incidence of non-return among foreign students in the united states. *Economics of Education Review*, 14(4):373–384. 2.2

Chernina, E., Castañeda Dower, P., and Markevich, A. (2014). Property rights, land liquidity, and internal migration. *Journal of Development Economics*, 110:191–215. 6.3

Chiquiar, D. and Hanson, G. (2005). International migration, self-selection, and the distribution of wages: Evidence from mexico and the united states. *Journal of Political Economy*, 113(2):239–281. 2.2

Cobb-Clark, D. (1993). Immigrant selectivity and wages: The evidence for women. *American Economic Review*, 83(4):986–93. 2.2

Collins, W. J. and Wanamaker, M. H. (2014). Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data. *American Economic Journal: Applied Economics*, 6(1):220–252. 6.4

Efremkin, E. (2011). Recruitment in North America: An analysis of emigrants to Soviet Karelia, 1931–1934. In Kangaspuro, M. and Saramo, S., editors, *Victims and Survivors of Karelia: A Special Double Issue of the Journal of Finnish Studies*, pages 103–125. Journal of Finnish Studies, Huntsville, Tex. 6.2.1, 6.3

Feigenbaum, J. J. (2016). Automated census record linking: A machine learning approach. `https://open.bu.edu/handle/2144/27526`. 4.1

Ferrie, J. P. (1996). 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. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29(4):141–156. 4.1, 4.2, C

Folbre, N. and Abel, M. (1989). Women's Work and Women's Households: Gender Bias in the U.S. Census. *Social Research*, 56(3):545–569. Publisher: The New School. 4.2

Golubev, A. and Takala, I. (2014). To Karelia! In *The Search for a Socialist El Dorado*, Finnish Immigration to Soviet Karelia from the United States and Canada in the 1930s, pages 27–50. Michigan State University Press. 1, 2.1.2, 2.1.2

Inwood, K., Minns, C., and Summerfield, F. (2019). Occupational income scores and immigrant assimilation. Evidence from the Canadian census. *Explorations in Economic History*, 72:114–122. 1, 3.2.3, 6.4

Kangaspuro, M. (2012). Amerikansuomalaisten tie Neuvosto-Karjalaan: Leipää, poliittista aktivismia ja murskattuja unelmia. In Hänninen, K. and Saaritsa, S., editors, *Työväki maahanmuuttajana*, Väki voimakas, pages 16–45. Työväen historian ja perinteen tutkimuksen seura, Helsinki. 1, 2.1.1, 2.1.2, 2.1.2, 7

Karjalan tasavallan Kansallisarkisto (2022). Pohjois-Amerikan suomalaiset Neuvosto-Karjalassa 1920-1930-luvuilla. Karjalan tasavallan Kansallisarkisto. 3.1

Kero, R. (1974). *Migration from Finland to North America in the years between the United States civil war and the First World War / by Reino Kero.* PhD thesis, Institute for Migration, Turku, Finland. Edition: Reprint. ISBN: 951-9266-00-3. 2.1.1

Kero, R. (1983). *Neuvosto-Karjalaa rakentamassa : Pohjois-Amerikan suomalaiset tekniikan tuojina 1930-luvun Neuvosto-Karjalassa / Reino Kero.* Historiallisia tutkimuksia, 122. Suomen historiallinen seura, Helsinki. 1, 2.1.2, 5, 7

Kero, R. (1996). *Suureen länteen: siirtolaisuus Suomesta Yhdysvaltoihin ja Kanadaan.* Siirtolaisuusinstituutti, Turku, Finland. 6.2.2

Kok, J. and Mandemakers, K. (2009). "Je zoudt maar last van mij hebben". Verwanten in het Nederlandse huishouden, 1860-1940. *TSEG - The Low Countries Journal of Social and Economic History*, 6(4):139–165. Number: 4. 6, A.1, A.2

Korkiasaari, J. (1989). *Suomalaiset maailmalla : Suomen siirtolaisuus ja ulkosuomalaiset entisajoista tähän päivään / Jouni Korkiasaari.* Siirtolaisuusinstituutti, Turku. 1, 2.1.1, 3.2.2, 6.2

Kostiainen, A. (1978). *The forging of Finnish-American communism, 1917-1924 : a study in ethnic radicalism.* Migration studies. Migration Institute, Turku. 2.1.1, 3, 3

Lee, E. S. (1966). A theory of migration. *Demography*, 3(1):47–57. 7, 7

Leeuwen, M. H. D. v., Maas, I., and Miles, A. (2002). *HISCO: Historical International Standard Classification of Occupations.* Leuven University Press. 2

Long, J. and Ferrie, J. (2013). Intergenerational Occupational Mobility in Great Britain and the United States since 1850. *American Economic Review*, 103(4):1109–1137. 4.2

McKenzie, D. and Rapoport, H. (2010). Self-selection patterns in mexico-u.s. migration: The role of migration networks. *The Review of Economics and Statistics*, 92(4):811–821. 6.3

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146. 2.2

Ruggles, S., Flood, S., Sobek, M., Backman, D., Chen, A., Cooper, G., Richards, S., Rodgers, R., and Schouweiler, M. (2024). IPUMS USA: Version 15.0 [U.S. Census 1930]. 1, 3.2.1, 1, 6.4

Saavedra, M. and Twinam, T. (2020). A machine learning approach to improving occupational income scores. *Explorations in Economic History*, 75:101304. 3.2.3

Simpson, J. and Swan, E. L. (1937). Improvements in the Lumber Industry. *The ANNALS of the American Academy of Political and Social Science*, 193(1):110–119. Publisher: SAGE Publications Inc. 7

Statistics Canada (2023). Census of Canada 1931. Retrieved October 8, 2024 from Statistics Canada. 3.2.2

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 354–359. American Statistical Association. 4

Appendix For Online Publication

A Additional Tables and Figures

Table A.1: Common Occupations in the 1930 U.S. Census

Soviet Karelian Immigration Records				**Recruited Sample**				**Census**			
Occupation	Freq.	Percent	HISCLASS 5	Occupation	Freq.	Percent	HISCLASS 5	Occupation	Freq.	Percent	HISCLASS 5
Forestry worker	539	38.42	5	Farmer	55	18.03	3	General worker	15,944	14.40	5
Carpenter	353	25.16	4	General worker	50	16.39	5	Farmer	15,062	13.60	3
General worker	111	7.91	5	Carpenter	43	14.10	4	Farm worker	11,304	10.21	5
Driver	98	6.99	4	Miner	23	7.54	4	Miner	10,392	9.38	4
Mechanics	63	4.49	4	Farm worker	21	6.89	5	Carpenter	9,515	8.59	4
Metal worker	37	2.64	4	Other prod worker	16	5.25	5	Forestry worker	5,974	5.39	5
Sawyer	36	2.57	4	Forestry worker	13	4.26	5	Other prod worker	5,256	4.75	5
Farm worker	27	1.92	5	Manager	8	2.62	2	Mechanic	3,967	3.58	4
Miner	23	1.64	4	Blacksmith	8	2.62	4	Manager	3,577	3.23	2
Machinist	19	1.35	4	Mechanic	7	2.30	4	Driver	3,343	3.02	4
Total (top 10)	1,306	93.09		Total (top 10)	244	80.00		Total (top 10)	84,334	76.14	
Outside top 10	97	6.91		Outside top 10	61	20.00		Outside top 10	26,441	23.87	
Total	**1,403**	**100.00**		**Total**	**305**	**100.00**		**Total**	**110,775**	**100.00**	

Notes: Table consists of men aged 15 to 64, with age calculated using the 1930 U.S. Census. Occupations are categorized based on the HISCLASS 5 classification (Kok and Mandemakers, 2009). HISCLASS 1–2 = 1 (Elite), HISCLASS 3–7 = 2 (Lower middle class), HISCLASS 8 = 3 (Self-employed farmers), HISCLASS 7 and 9 = 4 (Skilled workers), HISCLASS 10–12 = 5 (Unskilled workers and farm workers).

Table A.2: Common Occupations in the 1931 Canadian Census

Soviet Karelian Immigration Records				**Recruited Sample**				**Census**			
Occupation	Freq.	Percent	HISCLASS 5	Occupation	Freq.	Percent	HISCLASS 5	Occupation	Freq.	Percent	HISCLASS 5
Forestry worker	613	54.44	5	General worker	68	38.86	5	General worker	5,213	38.24	5
Carpenter	125	11.10	4	Farmer	24	13.71	3	Farmer	2,494	18.30	3
General worker	111	9.86	5	Miner	15	8.57	4	Farm worker	1,023	7.50	5
Driver	63	5.60	4	Carpenter	12	6.86	4	Forestry worker	980	7.19	5
Sawyer	50	4.44	4	Forestry worker	11	6.29	5	Carpenter	760	5.58	4
Farm worker	21	1.87	5	Farm worker	10	5.71	5	Miner	661	4.85	4
Fisherman	21	1.87	5	Fisherman	9	5.14	5	Railroad worker	270	1.98	5
Miner	21	1.87	4	Machinist	4	2.29	4	Fisherman	267	1.96	5
Metal worker	17	1.51	4	Food processor	3	1.71	4	Machinist	253	1.86	4
Machinist	17	1.51	4	Housekeeper	3	1.71	4	Housekeeper	192	1.41	3
Total (top 10)	1,059	94.05		Total (top 10)	156	89.14		Total (top 10)	12,113	88.86	
Outside top 10	67	5.95		Outside top 10	16	9.14		Outside top 10	1,519	11.14	
Total	**1,126**	**100.00**		**Total**	**175**	**100.00**		**Total**	**13,632**	**100.00**	

Notes: Table consists of men aged 15 to 64, with age calculated using the the 1931 Canadian Census. Occupations are categorized based on the HISCLASS 5 classification (Kok and Mandemakers, 2009). HISCLASS 1–2 = 1 (Elite), HISCLASS 3–7 = 2 (Lower middle class), HISCLASS 8 = 3 (Self-employed farmers), HISCLASS 7 and 9 = 4 (Skilled workers), HISCLASS 10–12 = 5 (Unskilled workers and farm workers).

Table A.3: The Most Common Occupations in Soviet Karelian Immigration Records, Recruited Sample

Occupation	Freq.	Percent
Forestry worker	179	45.55
Carpenter	60	15.27
Construction worker	34	8.65
Driver	18	4.58
Sawyer	14	3.56
Farm worker	11	2.80
Fisherman	10	2.54
Mechanics	8	2.04
Electrician	7	1.78
Stonemason	7	1.78
Total (top 10)	348	88.55
Outside top 10	45	11.45
Total	**393**	**100.00**

Notes: Table consists of men aged 15 to 64, with age calculated using the 1930 U.S. Census and the 1931 Canadian Census. The data on occupations are more detailed than that in Table 6; therefore, the occupational categories are not fully comparable.

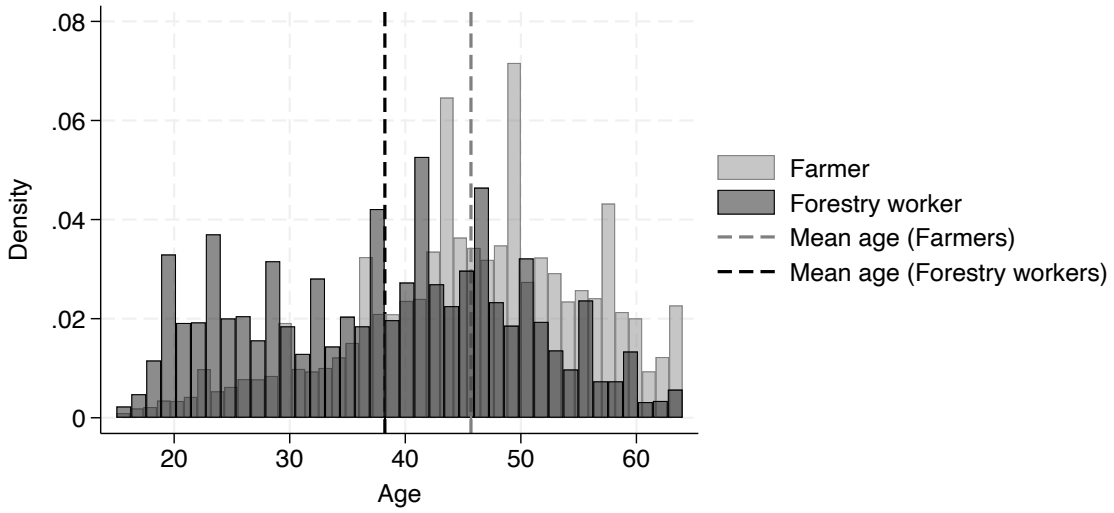

Figure A.1: Age Distribution of Farmers and Forestry Workers in North American Census Data

Notes: Figure consists of men aged 15 to 64, with age calculated using the 1930 U.S. Census and the 1931 Canadian Census.

Table A.4: Occupation-Based Income in the 1930 U.S. Census, Scandinavian Subsample

	(1)	(2)	(3)	(4)	(5)
American Finn	-0.116***	-0.108***	-0.076***	-0.102***	-0.141***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Age		0.005***	0.045***	0.053***	0.052***
		(0.000)	(0.000)	(0.000)	(0.000)
Age^2			-0.001***	-0.001***	-0.001***
			(0.000)	(0.000)	(0.000)
Number of family members			-0.005***		
			(0.000)		
Urban			0.515***		
			(0.001)		
Years in the United States				0.001***	-0.000***
				(0.000)	(0.000)
Literate				0.124***	
				(0.005)	
Speaks English				0.119***	0.162***
				(0.006)	(0.005)
Region fixed effects	✗	✗	✗	✗	✓
Adj. R-squared	0.004	0.017	0.313	0.053	0.129
Observations	1,284,256	1,284,256	1,284,256	1,228,786	1,228,786

Notes: Standard errors in parentheses. Region fixed effects included in 5. Only working-age (15–64 years) Scandinavian-origin (Finland, Sweden, Norway, Denmark and Iceland) men included. The dependent variable is the logarithm of $100 \times OCCSCORE$.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.5: Selection into Migration from Canada to Soviet Karelia, Restricted Sample

	(1)	(2)	(3)
Recruited	-0.147*	-0.097	-0.144*
	(0.046)	(0.050)	(0.048)
Age	0.007***	0.036***	0.044***
	(0.001)	(0.004)	(0.007)
Age^2		-0.000***	-0.001***
		(0.000)	(0.000)
Speaks English		0.385***	0.406***
		(0.030)	(0.030)
Year of immigration		-0.008***	
		(0.001)	
Region fixed effects	✗	✗	✓
Adj. R-squared	0.005	0.038	0.034
Observations	8,224	7,159	8,223

Notes: Standard errors in parentheses. The samples include working-age (15–64 years) in the 1931 Canadian Census. The dependent variable is the logarithm of earnings. Standard errors clustered by province. Fixed effects for region included in specification (3).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B Construction of the Soviet Karelian Immigration Records Dataset

The first sources are two extensive datasets of American Finns who emigrated to Soviet Karelia between 1931 and 1935: the first dataset contains Canadian individuals (*Dataset A*), and the second contains U.S. individuals (*Dataset B*). The Canadian dataset consists of 1,840 observations, while the U.S. dataset includes 2,585 observations. These datasets are not well-structured and require cleaning and formatting before use. They contain individual-level variables such as first name, last name, birth year, occupation, arrival date in the Soviet Union, and workplace in Karelia. However, many of these variables are missing, making it difficult to identify the same individuals across different datasets. The data are imprecise and contain a significant number of duplicates.

The original language of the records is Russian. Some first and last names have been transliterated and translated into Finnish, while others have not. To address this inconsistency, missing transliterations and Finnish forms are filled in using known existing transliterations and expressions. The two datasets are then combined, resulting in a dataset (*Dataset AB*) containing 4,648 observations. However, this dataset includes 106 duplicate records based on first name, last name, country of origin, and birth year. These duplicates are merged vertically, meaning that missing values in the uppermost observation are replaced by values from duplicate entries. Since the ordering is practically arbitrary, it is possible that in some cases, all duplicate entries contain information for a given variable. As a result, some information is lost during the vertical merging process, as only one observation per duplicate group is retained.

Next, another dataset of Soviet Karelian immigration records is cleaned and formatted following the same procedures. This dataset (*Dataset C*) contains 7,562 individuals, of whom 4,135 migrated from the U.S., 2,885 from Canada, and 542 have missing home country information. However, only 5,855 observations are unique in terms of name, birth year, and country, indicating a significant number of duplicates. This duplication arises because the dataset includes 624 distinct archival sources, many of which overlap. The two largest sources contain 2,488 and 1,716 individuals, respectively, with country variables listed as U.S. and Canada. These sources are essentially the same as in *Datasets A* and *B*. Since they are better digitized in those datasets, they are dropped from *Dataset C* to avoid a laborious deduplication process later.

Next, selected variables from another dataset (*Dataset D*), containing 2,915 individuals, are merged into *Dataset C*. *Dataset D* shares archival source IDs with *Dataset C* which makes the merging process reliable. However, the data in it are not usable on its own. After deduplication, the resulting dataset *Dataset CD* contains 3,464 observations. Information from the dropped observations is merged to the retained records to preserve as much data as possible.

The final Soviet Karelian immigration records dataset is compiled from *Datasets AB* and *CD*. *Dataset AB* serves as the master dataset, with information from *Dataset CD* merged into it iteratively. First, observations are linked using the ABE linking algorithm based on name and birth year. Next, cases involving unique names are processed, followed by cases where names are not unique but differ in the country variable (after the ABE linking, all of

the remaining observations have missing birth year data). The resulting dataset consists of 7,244 observations.

The last step is to remove remaining duplicates based on name, birth year, and country. After dropping identified duplicates, the dataset contains 7,158 observations. Again, information from removed entries is merged vertically into the retained observations. Despite appearing to contain only unique records at this stage, one final round of deduplication is necessary. Previous deduplication steps identified only exact duplicates based on name variables; they did not account for minor variations such as "Vihtori" and "Wihtori", which are effectively the same name. Given the variability in name quality, this issue is addressed using Jaro-Winkler string similarity.

Individuals are sorted by name, and each n:th observation is compared to the n-1:th observation. If the Jaro-Winkler score exceeds 0.79 for both the first and last names, and the observations have matching country and sex variables, as well as birth years within a two-year range (or missing), the records are considered duplicates and removed. The final Soviet Karelian immigration records dataset contains 6,663 observations. The size of the dataset aligns with estimates in the existing literature on the number of Karelian Fever migrants.

C Data Linking Algorithm

The Sovier Karelian immigration records are linked to the 1930 U.S. Census and 1931 Canadian Census using the following algorithm:

The untouched IPUMS-origin U.S. Census dataset contains over 122 million observations. The census data is first filtered to retain only records where Finland is listed as the individual's, mother's, or father's place of birth. This filtering process results in a dataset of 336,773 observations, consisting of Finnish immigrants and their descendants. Next, the name variables are cleaned by removing non-alphanumeric characters and suffixes such as "Mr." or "Mrs.". Middle name initial variables are created, and nicknames are standardized by replacing them with their formal equivalents. The Soviet Karelian immigration records are prepared in a similar manner. The final Karelian dataset contains 6,663 observations, of which 3,523 list the United States as the country of origin, 2,557 list Canada, and 583 do not include information about the country of origin. Observations with missing country of origin are included in the U.S. Census linking process. Unmatched observations are subsequently

saved and included in the Canadian Census linking process. The Soviet Karelian dataset used in the U.S. Census linking thus contains 4,106 observations.

The linking algorithm follows the fully automated ABE approach developed by Abramitzky, Boustan, and Eriksson. This method was originally introduced by Ferrie (1996) and later adapted for computational use by Abramitzky et al. (2012, 2014, 2019). The linking process is conducted in two parts. In the first stage, only observations with information on birth year are included. The algorithm begins by creating blocks of observations based on the initials of the first and last names. This blocking step is necessary to reduce computational demands by limiting comparisons to records with matching initial letters in both first and last names. However, this approach may exclude potential matches when initials differ due to spelling variations (e.g., Vihtori vs. Wihtori).

Within these blocks, the algorithm compares birth years and calculates Jaro-Winkler string distances for both first and last names of each possible pair[4]. Thresholds for the distance scores are set at 0.12 for last names and 0.15 for first names[5]. If a pair exceeds both thresholds, their birth years are compared. If multiple matches share the same name and birth year, the match is considered non-unique and is rejected. If only one match with the same name and birth year exists, it is accepted. In cases with no exact match, the closest match within an age difference of ± 2 years is retained, provided it is unique. Only matches with the same sex are accepted. Note that the sex variable in the Soviet Karelian immigration records is inferred based on first name variables, which may introduce errors. This first stage yields 631 matches.

The identified matches are set aside, and the second stage begins with all remaining observations, regardless of whether the birth year variable is present. The process is otherwise similar to the first part, but birth years are not compared. Linking is based solely on name variables, with the same Jaro-Winkler thresholds as in the first stage. Because birth year is

[4]Winkler (1990) defines string similarity as follows: If $c > 0$, the string similarity distance is given by $\Phi = 1/3 \cdot (c/d + c/r + (c - \tau)/c)$, where d is the length of the first string, r the length of the second string, τ number of transpositions of characters, and c number of characters in common in pair of strings. If $c = 0$, then $\Phi = 0$. Two characters are matching only if they are same and not farther than $m/2 - 1$ characters apart, where $m = \max(d, r)$. The number of transpositions means the number of matching characters that are not in the right order divided by two.

[5]The thresholds for Jaro-Winkler scores are determined empirically by examining and refining the results. Since Jaro-Winkler similarity gives higher weight to the initial characters of a string and does not heavily penalize differences in string length, only matches differing by a maximum of two characters are permitted. This adjustment is particularly important for Finnish names, as many Finnish surnames are formed by adding suffixes. For instance, Mäki and Mäkinen are distinct names, but their Jaro-Winkler similarity score is 0.086, which means the algorithm interprets them as the same name using the conventional thresholds.

not used, some resulting matches may have mismatching birth years; matches with an age difference of three years or more are excluded. Matches are accepted when one or both birth year values are missing. This second stage results in 18 additional matches.

The two-stage linking approach is necessary because 22.5% of the Soviet Karelian immigration records lack birth year information. Combining the results yields a dataset of 649 matched individuals, corresponding to a match rate of 15.8%. The linking procedure for the 1931 Canadian Census closely follows the method used for the U.S. Census. In this case, the Soviet Karelian dataset, containing only observations with Canada or missing as the country of origin, includes 3,108 records. A total of 364 matches are identified, resulting in a match rate of 11.7%.

D Code Availability

The full code used in this thesis is available here.