Little School on the Prairie: A Push for Structural Change

Devin Bissky Dziadyk *†

March 19, 2025

Abstract

The one room school house looms large in North American history, but we have little information on the role these rural schools played in economic development. Using linked full count census records from the early 20th century Canadian prairies combined with precise geolocations for individual households and schools, I examine the effect of the roll out of one room schoolhouses on children's educational participation and later life outcomes. I find that geographic proximity plays an important roll in education participation, and educational participation correlates with adult income and decreases the likelihood of working in agriculture. Expanding to a difference in differences strategy exploiting the timing of school rollout, I find that a treated child who has a school constructed nearby is 3.92pp less likely to be found in agriculture as an adult, and lives 16.8% further from their childhood home. This suggests that education access can act as a push factor for structural change in rural areas.

"And although many farmers would have preferred that their children remain on the land, most knew it would prove impossible. The best they could do was to endow them with education to be mobile."

Goldin and Katz (2000)

Introduction

The one room school house is a quintessential part of the development of the American West. These single room schools provided widespread primary education in rural areas, allowed for feasible commuting for children, were decentralized to adapt to local communities (Dippel & Ottinger, 2020), and helped establish a rural education advantage that acted as an input to the later high school movement (Goldin & Katz, 2008). Despite their widespread

*University of Toronto, devin.bisskydziadyk@mail.utoronto.ca. Supported in part by funding from the Social Sciences and Humanities Research Council. I would like to thank Stephan Heblich, Shari Eli, and Tianyi Wang for their guidance and support. Thanks as well to Irisa Zhou, Abdelrahman Amer, Chenyue Liu, Ben Coulliard and many others. In memory of Jojo Zhang.

†JEL Codes: I25, R12, N31

availability, one room schoolhouses where criticized for providing poor quality education and had lower returns to education than other forms of schooling (Lachanski, 2024; Goldin & Katz, 2000). What then was the impact of one room schools on the children and communities that constructed them?

This paper answers this question by analyzing the rollout of one room schools in the western Canadian prairies at the start of the 20th century. The Canadian prairies are highly geographically and culturally similar to the American Midwest and were dotted by one room schoolhouses which provided primary education to local school children under the supervision of a single teacher. As in the United States these schools were primarily driven by local demand, and came under criticism for having poor standards educational standards, but being flexible to local needs and demands (Noonan et al., 2006; Baergen, 2005). As a result, the benefits of constructing one room schools are highly uncertain.

Using the Canadian prairies as a setting is ideal because of the uncoordinated rollout of schools across Manitoba, Saskatchewan, and Alberta and the widespread use of legal land descriptions (survey boundaries) as addresses. As a result it is possible to geolocate both historical school locations and household locations to land section (1 mile square parcel). This allows for estimation of the effects of school construction on individuals. A comparable strategy is impossible in the United States because rural address were not recorded as precisely in census records, often only to the township (6 mile block) level.

Estimating these effects I find that there are significant distance effects on education, with children adjacent to a school attending approximately 3/4 of a month longer that those 5km away. Exploiting the dynamic rollout of schools I find that children who experienced a school opening near them are less likely to be working in agriculture as adults, and have moved further away from home. Taken together this suggests that the expansion of one room schools can be an important force in the reallocation of labor out of agriculture and towards other sectors.

This paper is most closely related to Goldin and Katz (2000) which estimates the returns to education in Iowa starting in 1915. Goldin and Katz shows that the returns for rural education were significant, with the return for an extra year of high school increasing male earnings 11%, while the returns to a year of common school are lower at 4%. Importantly they estimate that the returns to education as significant for both agricultural and non-agricultural occupations. Relative to Goldin and Katz (2000) I estimate the causal impact of rural education on later life outcomes by exploiting the rollout pattern of common (one room) schools.

There is growing literature on the spatial effects of education policy. Hsiao (2023) examines the case of Indonesia, finding that school construction increases inequality between rural and urban areas, with large gains for people who relocate but small gains for areas that engage in school construction. Ghose (2023) shows that the endogenous response of education is responsible for two thirds of the gains from an export shock, and that education reduces regional inequality. Finally Eckert and Kleineberg (2021) shows that in spatial equilibrium local education policies become less effective due to household sorting. Relative to the models in the spatial education literature this paper adds an agricultural sector with declining returns to scale, and allows for external increasing returns in the non-agricultural sector.

This paper also speaks to a literature on the effects of education on structural change and

development. Caselli and Coleman II (2001) argues that the regional convergence between the North and South United States can be attributed to declining education costs, leading successive generations to choose non-agricultural occupations. This idea is then shown with microdata in Porzio et al. (2022) which shows that the reallocation out of agriculture is due to successive generations becoming progressively more educated. Additionally, this paper speaks to the complementarity between skills and different production technology as in Goldin (1998)

Although this paper focuses on the development of the Canadian west, the institutional features for education are similar to that of the United States. Dippel et al. (2020) explores the role of federalist school provision and preference heterogeneity for schooling in the United States. Schaede (2021) utilizes variation in land endowments for school funding to show that areas with poorly funded schools are slower to transition away from agriculture. French (2022) examines the adoption of the tractor in the United States, showing that areas that adopted the tractor earlier saw a faster transition out of agriculture and more intergenerational mobility, proposing education as the main mechanism.

Finally this paper relates to the settlement and land institutions of the American and Canadian west. Smith (2022), Mattheis and Raz (2019), and Leonard and Kogelmann (2022) make use of newly digitized homestead records to examine the effect of the Homestead Act in the United States. In yet to be released work, Bagagli (2023) explores similar data for Canada. Nagy (2023) explores the relationship between farm hinterlands and city formation in the United States. The roll of surveying systems has been explored in Libecap and Lueck (2011) and Hornbeck (2010). There is additionally later work on the dust bowl, dryfarming, and farm size as in Hansen and Libecap (2004), Libecap and Hansen (2000), Hornbeck (2012), and Hornbeck (2022)

Historical Background

Western Expansion

After the end of the American Civil War and the acquisition of Alaska, the Canadian Government also desired to expand territory westward (Lambrecht, 1991). To accomplish this the Canadian government acquired Rupert's Land and the North-West Territories in 1870 from the Hudson's Bay Company. To further expand, British Columbia entered confederation with Canada in 1871, under the condition that a railway be constructed to the west coast from the east coast no later than 1881 (completed in 1885). The railway would simultaneously link the opposite sides of the dominion, while allowing settlers into the fertile prairies.

To enable settlement existing claims on prairie land has to be resolved. The Canadian government signed seven major treaties between 1871 and 1877 across the prairies to extinguish indigenous land title (Lambrecht, 1991). The exact contents of these "Numbered Treaties" varied, but throughout they established reserves for indigenous people, administered by the Indian Act. The second group of claimants was the Métis, people who were the result of intermarriage between indigenous and european parents, who were granted 1.4 million acres under the Manitoba Act of 1870 and later lands in the North West Territories

in 1885. The settling of Métis land claims was a source of continuous error, with repeated issuance of scrip which could be redeemed for land as late as 1925 (Lambrecht, 1991). The overwhelming motivation behind the settling of indigenous and Métis land claims was to extinguish any claims that would impede settlement and risk the establishment of Canadian soveriegnty [1].

Non reserve land would then be administered by the Dominion Lands Act in 1872 (Lambrecht, 1991). The act laid out a survey for lands in Manitoba, the North-West Territories (to largely become Alberta and Saskatchewan in 1905), and later the railway belt and peace river block in British Columbia. Based on the Public Land Survey System in the United States, the Dominion Land Survey would both measure and allocate land across western Canada, and remains largely in place as of writing as the dominant way to describe parcels of land.

Surveying

The Dominion Land Survey uses a regular hierarchical pattern to survey land. The base unit of land is a township, consisting of an area 6 miles by 6 miles. The location of each township is defined by three values, the township number, range number, and meridian. Township numbers give the north-south position, start at the southern border of Canada on the 49th parallel, and count up as one proceeds north. Range numbers start at the Prime Meridian (approximately 97° 27'), and increase as one moves west. Range numbers reset as after crossing one of six meridians. Any position in the east-west direction is therefore defined by a range number, and the position of the range relative to the meridians. For example, a township could be located by 5-23-W2, or the 5th township, 23rd range, west of the second meridian. Figure 1 show the extent of the Dominion Land System in western Canada.

Each 6 mile by 6 mile township is further subdivided in 36 sections, which is further split into four quarter sections. Each section is numbered 1 through 36, starting in the lower right hand corner. Figure 2 shows an example of how sections are laid out within townships. Between each section is a road allowance of 90 feet, although this was decreased to 66 feet and only between every second section horizontally after 1881. The road allowances is an important reason why grid of land parcels remains largely in place today (Lambrecht, 1991).

Within a township land was largely allocated between four parties, and although land policy changed over time the main pattern was largely established by the 1881 Dominion Lands Regulations Lambrecht (1991). Odd numbered sections, excluding 11 and 29, were reserved for land grants for railway companies to incentivize railway construction. In addition a small number of colonization companies could purchase and then resell land for the purposes of settlement. Even numbered sections, with the exception of 8 and most of 26 were sold as homesteads. Homesteaders were granted entry to a quarter section for a minimal fee, and after three years of cultivation and residence would be granted the title to the land. Sections 11 and 29 were reserved for the purpose of funding local schools, although the funds generated were pooled at the federal level and then redistributed to the provincal

[1] Prime Minister John A. MacDonald would defend his role as minister of the interior with regards to Métis land claims stating "Whether they had any right to those lands or not was not so much the question as it was a question of policy to make an arrangement with the inhabitants of that Province ... in order to introduce law and order there and assert the soverieignty of the Dominion" (Lambrecht, 1991)

Figure 1: Survey boundaries for townships under the Dominion Land System. Each square consists of a 6 mile by 6 mile township. The red lines give the meridians. Each township can be uniquely identified by a township number, giving the north-south direction, and a range number plus position relative to a meridian, giving the east-west direction.



Map: Index to townships in Manitoba, Saskatchewan, Alberta and British Columbia showing the townships for which official and preliminary plans have been issued to January 1st, 1929. Topographical Survey of Canada, Department of the Interior. Compiled, drawn and printed by the Topographical Survey of Canada, Ottawa. [cartographic material]
Source: Library and Archives Canada/Maps, plans and charts/NMC43265

Figure 2: Survey boundaries for sections within each township under the Dominion Land System. Even numbered sections, with the exception of 8 and 26 were sold as homesteads. Odd numbered sections with the exception of 11 and 29 were reserved for railways. Sections 11 and 29 were sold reserved to fund schools. Section 8, and three-quarters of section 26 (and the entirety of section 26 every fifth township) where given to the Hudson's Bay Company in exchange for Rupert's Land and the North West Territories.

31 Railway	32 Homestead	33 Railway	34 Homestead	35 Railway	36 Homestead
30 Homestead	29 School	28 Homestead	27 Railway	26 HBC	25 Railway
19 Railway	20 Homestead	21 Railway	22 Homestead	23 Railway	24 Homestead
18 Homestead	17 Railway	16 Homestead	15 Railway	14 Homestead	13 Railway
7 Railway	8 HBC	9 Railway	10 Homestead	11 School	12 Homestead
6 Homestead	5 Railway	4 Homestead	3 Railway	2 Homestead	1 Railway

and territorial governments. Sections 8 and three quarters of section 26, and the entirety of section 26 every fifth township in the agricultural parts of the west were granted to the Hudson's Bay Company in exchange for their surrender of Rupert's Land and the North-West Territories in 1870. Details on the allocation process of land can be found in section 0.2 of the appendix.

Schools

Throughout the early 1900's the expansion of schooling was rapid, giving rise to the motto "A New School Everyday for Twenty Years" in 1900 (One Room School Project, 2023). The development of the education system in Saskatchewan and Alberta was significantly later than Manitoba, reflecting the slower settlement and further west location of the provinces. Prior to Saskatchewan and Alberta becoming provinces in 1905, they were part of the Northwest Territories, and shared education policy.

Separate Schools

The Northwest Territories Act of 1875 gave the Territorial Council the right to pass ordinances regarding education, on the condition that the religious minority of either Protestants or Catholics could establish their own school and ratepayers would only be liable for the school of their religion (Moir, 1934; Noonan et al., 2006). The first major ordinance, The North-West Territory School Ordinance of 1884, created a system of school districts, with every district designated as either Protestant or Catholic. The original intention was to create a "Dual" system similar to Quebec and Manitoba, with separate school systems for Protestants and Catholics (Noonan et al., 2006). Protestant and Catholic schools were overseen by a Catholic and Protestant sections, overseen by the Board of Education (Noonan et al., 2006). Each section separately determined subjects, textbooks, and teachers for the schools of their own faith (Moir, 1934).

The "Dual" system didn't last, and was changed to a "minority" system similar to that of Ontario in 1885 (Noonan et al., 2006). The minority system allowed two schools to operate simultaneously in a given area. The first school, either Protestant or Catholic, would be designated a "public" school. The minority in an area could then vote and create a "separate" school district, either Protestant or Catholic, for people of that faith. In several instances changing population of Catholics and Protestants led to a separate school, which was supported by the majority of tax payers (Noonan et al., 2006). Initially the boundaries of a separate school district did not need to overlap with the public school district, but this was changed so that both school districts had the same boundaries in 1892 (Noonan et al., 2006; Moir, 1934).

Although the minority system remained in place in the Northwest Territories, the differences between the Protestant and Catholic sections, and their influence, decreased over time. In 1885 the Board of Education was reduced to five members, two Catholic, two Protestant, and the lieutenant governor (Noonan et al., 2006). In 1887 the Board of Education took over the power to license teachers, meaning that Protestants and Catholics both jointly determine the teachers for each other's schools, despite differences in the materials being taught (Moir, 1934). In 1891 and 1892 most of the power of the sections was removed with the

lieutenant-governor-in-council taking over licensing of teachers, appointment of inspectors, the use of uniform textbooks, and uniform examination standards (Moir, 1934). Finally in 1901 a commissioner of education was created to administer schools, to be advised by an Educational Council consisting of three Protestants and two Catholics.

The differences in instruction across schools were also reduced over time. Religious education was limited to the last half hour of the day after 1892, with the religion determined by the local school board (Noonan et al., 2006). Following the Riel Rebellion in 1885 and a subsequent rise in anti-French sentiment, English was mandated as the language of instruction in 1901 (Noonan et al., 2006), with provisions for teaching some primary courses in French (Moir, 1934).

In practice, the majority of separate schools were Catholic, although a small number of separate Protestant schools were established in Catholic majority areas. To fund schools, Protestants and Catholics each paid taxes to the school of their respective faith, and after 1886 no owner of property was able to escape taxation all together (Moir, 1934). Prior to 1909 residents could determine which school board they would like to fund, but after a legal case it was determined that the assessment of taxes would be on the basis of religious faith, not personal preference (Noonan et al., 2006). Further legal opinions clarified that for the purposes of establishing a separate school, a Catholic was someone who believed that religious authority rests with the pope[2] (Noonan et al., 2006).

Details on how separate schools are handled in the data are discussed in section 0.5.1.

Data

Census Records

The primary data source for this paper comes from the Canadian complete count census records for 1901, 1911, 1921, and 1931, and the census of prairie provinces in 1906, 1916, and 1926 [3]. The project will focus on individuals who lived in Manitoba, and the portion of the Northwest Territories which became Alberta and Saskatchewan, starting in 1901. Census records for 1901, 1911, 1921, and 1931 can be linked to Census Subdivision (CSD) level polygons which have been digitized (HGIS Lab, 2023). In addition some variables of interest like months of schooling and income are available are only available in 1901, 1911, 1921, and 1931, while other variables the section, township, range, and meridian necessary to geolocate households are only available in some years. A full description of which variables are available in which years can be found in section 0.3 of the appendix.

Linking

Census records contain considerable personal information about each individual in a given wave of the census. Unfortunately census records do not often include identifiers that link

[2] As of 2018 this definition is still used in Alberta. This confusingly leads people who are Eastern Orthodox or Ukrainian Orthodox to be considered Protestants for the purposes of establishing a separate school board (Government of Alberta, 2018).

[3] Data access provide courtesy of The Canadian Peoples (2023)

individuals across censuses, posing a challenge for researchers to track individuals over time. To overcome this challenge researchers must link individuals across census waves using information contained in the census records. Early attempts to do this involved hand linked samples by researchers, which proved prohibitively costly. Currently machine linked samples have proved to be an efficient way to link large numbers of individuals across census waves.

The earliest attempts at automated linking utilized the methodology of Ferrie (1996) to match individuals across census waves using phonetic classification of names with other time invariant characteristics. This is notable in the first full count matching of census records in Abramitzky et al. (2012) and subsequent papers Abramitzky et al. (2014) and Abramitzky et al. (2019). Importantly these algorithms provide a completely replicable and transparent process, at the cost of excluding potentially valuable time-varying information that may be informative of a match.

More recent algorithms have been developed that utilize a broader set of characteristics to determine matches, at a cost of being less transparent to the researcher and requiring training data to generate matches. This includes Feigenbaum (2016) which implements a regression based approach for linking, Price et al. (2021) which utilizes FamilySearch data to train a model for linking, Buckles et al. (2023) which utilizes the FamilySearch data to link all US census records, and Helgertz et al. (2022) which links US census records using a two stage probabilistic approach (hereafter MLP) and a hand linked training sample. The method used in this paper will follow the MLP method of Helgertz et al. (2022).

The MLP method is a two stage procedure. In the first stage males are linked across census waves, restricting matches to those that have a high degree of certainty. In the second stage, the first stage links are used to infer families across census waves, which then allow family members of individuals in the first stage to be linked. The second stage allows for both males and females to be matched. The matching process is implemented using the Hlink package (Wellington et al., Apr. 27, 2022/2024).

In the first stage a list of all potential matches is generated. To ensure computational tractability, only matches which satisfy blocking characteristics are included. Blocking is done on sex, age within a three year window, the first letter of the surname, and the recorded birthplace (aggregated to the province, United States, United Kingdom, Ireland, or continent). The list of potential matches is then scored with a logit estimator to predict the likelihood of two records being a match. In order for a match to be accepted, the probability of a match being correct must be greater than a threshold $\alpha = 0.25$, and the score for a proposed match must better by some margin then the next most likely match $\beta = \frac{\alpha_1}{\alpha_2} = 1.1$. This ensures that only the most likely links are retained.

In the second stage the links in the first stage are used to infer family links across census waves. A second round of linking is then completed after blocking at the family level. A second logit model is used to score linking within the family, with the same thresholds for α and β applied. Linking within family also allows any blocking variables to be dropped, allowing for matches where some characteristics may have changed dramatically either due to enumerator error or deliberate choice on the part of the person being recorded.

In both the first and second stage of the matching procedure matches are determined probabilistically using a model trained on a set of known matches. Obtaining a training dataset is typically one of the most difficult pieces in a probabilistic matching algorithm. To overcome that lack of training data available for matching the Canadian census I utilize

Table 1: Preliminary Match Rates for Canadian Census

	Total	Matched	Children	Adults	Male	Female
1906-1911	100.00	22.80	26.69	21.43	26.75	17.76
1906-1921	100.00	19.73	20.59	20.12	25.62	12.02
1906-1931	100.00	16.20	16.27	16.91	23.02	7.18
1911-1921	100.00	19.95	23.06	18.25	24.69	14.67
1911-1931	100.00	13.37	13.92	13.19	18.86	7.21

the training data in the replication package of Helgertz et al. (2022) which links US census records. In order to make the training data relevant to the Canadian context I restrict the variables used in the training data to those also available in the Canadian census. This ensures that any variable used to train either of the logit models in steps 1 and 2 can be faithfully replicated in the Canadian data. Details on the variables used can be found in section 0.4.1 of the appendix.

Preliminary results on the match rates between Canadian census waves can be found in table 1. I'm able to match 30.16% of people between the 1906 and 1911 census waves, which declines to 26.43% across the 1906 and 1921 census waves when the difference in time between waves is larger. These results are broadly consistent with the literature, including Antonie et al. (2014) who claim a match rate of 24.22% when matching the 1871 to 1881 Canadian census.

One possible concern with the MLP linking method is that it may introduce bias from selection into matching. It is important to note that no automated linking method produces unbiased samples (Bailey et al., 2017). In particular there is concern that including time varying variables like location will bias the results towards matches where individuals don't move. This comes at the benefit of being able to distinguish otherwise identical matches without this information. In section 0.4.2 of the appendix I compare the results of the matching with the ABE method, which uses on time invariant factors, and find that changes in the match rate by distance are consistent with large improvements in match rate with little cost in increased distance bias.

Schools

Records of school locations[4] and opening dates come from three main sources. Alberta's schools are documented in Baergen, 2005, and later digitized by the University of Calgary (University of Calgary, 2019). Saskatchewan's schools were scraped from Saskatchewan Gen Web, 2023, a genealogy website. Manitoba's schools came from Manitoba Historical Society, 2019, which are largely based on the extensive archival work of Perfect, 1978. In addition I supplement these records with annual reports for the Saskatchewan department of education

[4]I use the term "schools" and "school districts" interchangeably in this paper. Most school districts consisted of a single school, while a small percentage could consist of two schools. In a small number of cases a school district could be formed and then contract with a nearby school district to convey the students elsewhere. In 1913 there were 3,230 school districts in existence with 2,747 having a school in operation (Saskatchewan Department of Education, 1906, 1907, 1908, 1910, 1911, 1912, 1914), although some of these would reflect new districts that had not yet opened their school.

Table 2: Geolocation rates by census year for Alberta, Saskatchewan, and Manitoba. Exact geolocations are at the section level or in CSDs that are less than 1.5 miles in area. General geolocations are those that are at the township level or finer. CSD geolocations are at the CSD level. Splitting the geolocation rate by urban and rural households reveals that the majority of geolocations in both census waves are for the rural households.

	Total	**Exact**	**General**	**CSD**	**Exact (Rural)**	**Exact (Urban)**
1906	100.00	55.57	84.41	100.00	71.94	0.00
1921	100.00	65.97	69.61	100.00	83.89	36.35
1931	100.00	42.76	86.57	100.00	70.81	0.00

(Saskatchewan Department of Education, 1906, 1907, 1908, 1910, 1911, 1912, 1914).

In many cases records may be missing the date of opening or location. I infer missing opening dates from the fact that within Alberta and Saskatchewan, and separately from Manitoba, school districts were numbered sequentially. Details on the process of inferring opening dates can be found in section 0.5 I additionally assume that no schools closed during my period of interest. This is a reasonable assumption as the population of the Canadian prairies was rapidly expanding due to immigration, and the school consolidation movement had not yet come into effect. In cases where the school location is not known exactly, I utilize the township centroid or the centroid of townships whose children attended the school. Details on school locations can be found in section 0.6.3 of the appendix.

Geolocation

A unique feature of the Canadian prairies is the widespread use of land surveys and the legal land descriptions as addresses. In census records and colloquial use the legal land description was used as an the address of rural households, schools, and other places. The combination of section, township, meridian, and range uniquely identifies a single location at a one mile resolution, and down to a half mile resolution if the quarter section number is available as well. The legal land descriptions continue to be in use and unchanged today, allowing for precise geolocation over one hundred and fifty years after the Dominion Land Survey was completed.

Geolocation of rural households and schools requires the legal land description or a location and the corresponding parcel location. To get the parcel locations I merge shapefiles for parcels in Alberta (Government of Alberta, 2020), Saskatchewan (Information Services Corporation (ISC), 2020), and Manitoba (Manitoba Government, 2020). For households I locate to the nearest section, and identify potentially erroneous locations based on the location of neighbors and plausible paths that enumerators would have taken. For schools I adopt a sequential approach, first locating at the nearest quarter section, then section, then township. Detail on the geolocation procedure for households and schools can be found in section 0.6 of the appendix.

Table 2 gives the rate of geolocation across census waves. In total I'm able to geolocate 55.57% of individuals in 1906 and 65.97% of individuals in 1921 to the exact land parcel, with the vast majority coming from rural areas. In rural areas, where the majority of geolocation occurs, the rates are 71.94% and 83.89% respectively.

Table 23 in the appendix gives the counts of geolocated schools. As of 1911 I'm able to geolocate 73.7% of schools to the exact land parcels, and 97.4% of schools to the township level. Details on school geolocation can be found in section 0.6 in the appendix.

Other Data

Railways formed the most common mode of transportation in the early 20th century, and their location determined broad settlement patterns. To account for this I use a dataset of the location and year of construction of Canadian railways (Cartography office, 2022).

Agricultural productivity is also a key variable in determining the viability of an area. I make use of the FAO GAEZ v4 database FAO and IIASA (2024), utilizing the expected output density with the oldest possible time period (1971-2000), the CRUTS32 climate model, historical representative concentration pathway (RCP), rainfed water supply, low input level, and without CO_2 fertilization. I include data for expected yields for wheat, oats, barley, flax, which were the most common crops of the time, and reed canary grass as a proxy for livestock suitability.

Reduced Form Results

Despite the significant anecdotal evidence that geographic distance reduced school attendance (including quotes in section 0.1 in the appendix), there is little quantitative evidence that the distance to school affected attendance or later life outcomes. To show this I divide the work into three parts. I first link together the 1906 and 1911 censuses to show that schooling decreased on the intensive and extensive margin with distance. I then link together the 1911 and 1921 census records to show that differences in school attendance as a child had an impact on later life outcomes. The first two parts rely strictly on cross sectional differences.

In the final part I link the 1906 and 1931 censuses and exploit variation in the time and location of school openings across the prairies in a difference-in-differences framework. To estimate the effect of school openings to follow the methodology of Duflo (2001), estimating the effect of school openings on a younger cohort of children who are able to benefit from openings, relative to an older cohort who are too old to benefit.

Across all estimation strategies I find a consistent empirical pattern that lower school distance increases schooling, that higher schooling decreases the likelihood of being in agriculture as an adult, and that taken together the process of new school construction increases the likelihood of leaving agriculture for other opportunities.

1906-1911

I first examine the impact of school distance on the intensive and extensive margin of schooling. I link children in 1906, where I observe exact geolocations, to 1911, where I observe school attendance in that year. The sample is restricted to those children whose Census Subdivisions (CSD) are within 200km of each other between 1906 and 1911. I do this to capture children who have likely not moved in the intervening five year period, as the CSD

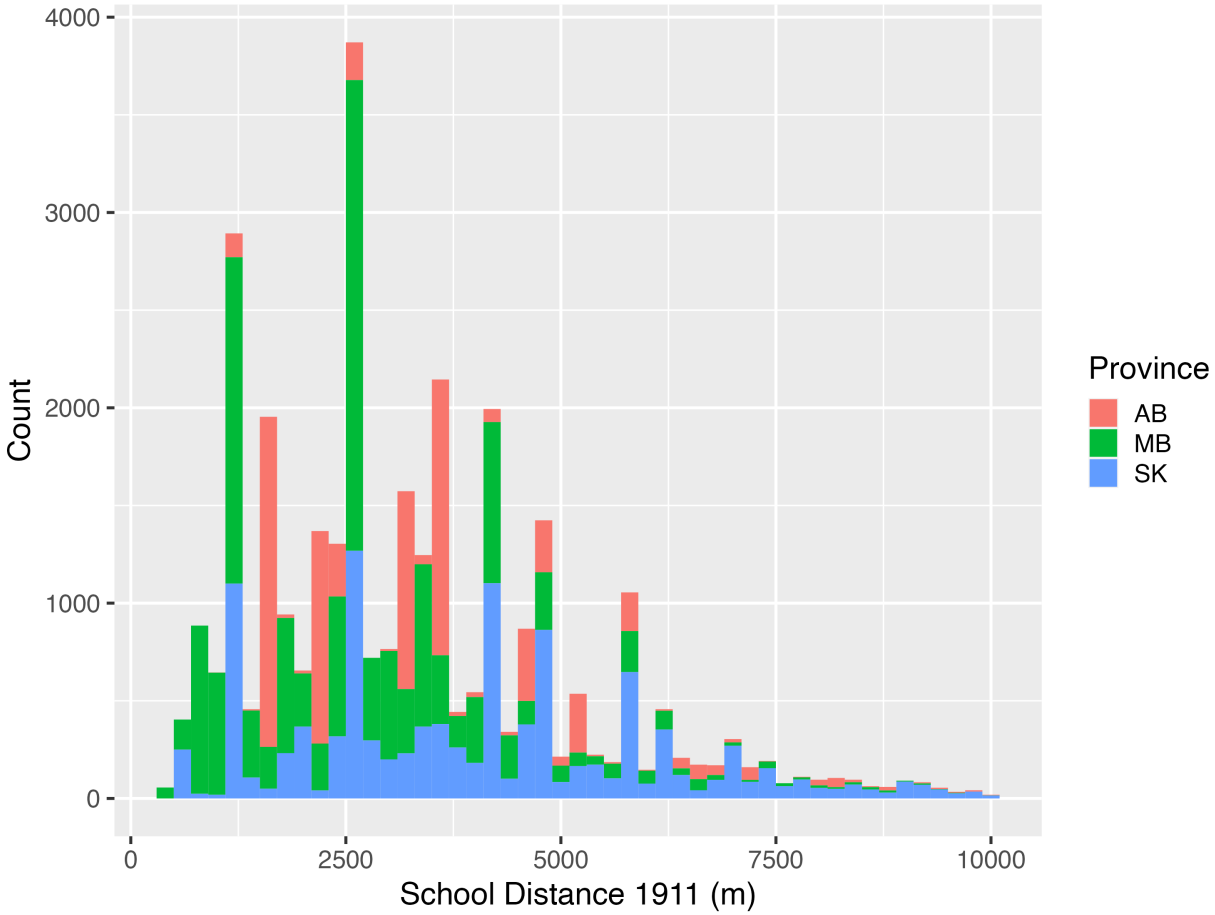


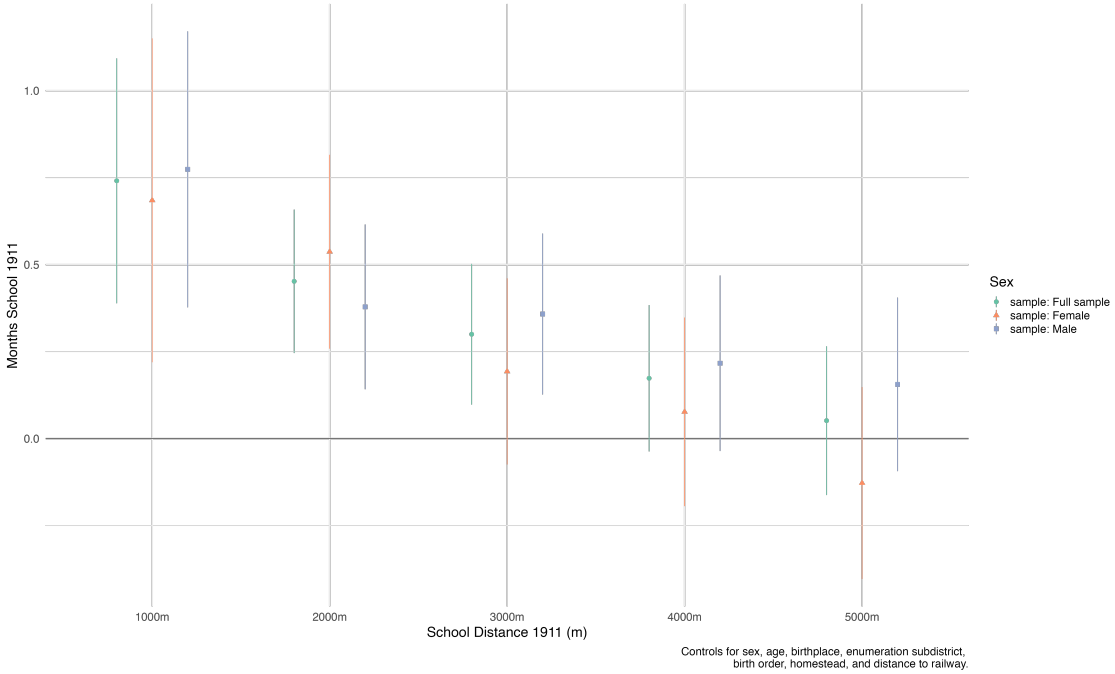
Figure 3: Distance to school for children in 1911, based on 1906 locations and restricting the sample to those that have not moved. Note the distribution has distinctive peaks including those at approximately 1250m and 2500m. This reflects the precision of the geolocation, with 0.75 miles being 1207m, and 1.5 miles being 2.41km.

boundaries between the two years are not consistent. Holding the 1906 locations constant and assuming the child has not moved, I then compute the distance to school in 1911 with the set of schools open in 1911. The median school distance in 1911 is approximately 2.9km, with a significant spread as seen in figure 3.

The sample is further restricted to those that are at least five years old in 1911 (so they are guaranteed to be alive in 1906) and are less than eighteen years old in 1911, or the time that modern secondary schooling ends. In addition I restrict the sample to those that are listed as a descendant of the head of household in 1911. This restricts the sample to children who are living with their parents, and drops those who may have already moved away for work or higher schooling in 1911.

I begin by estimating the effect of distance on schooling in 1911. I estimate the following equation:

Figure 4: Months of Schooling versus Distance



Months of schooling by distance in 1911, in 1km bins. All coefficients are expressed relative to the furthest bin of 5km or more. Includes controls for sex, age, birthplace, birth order, homestead, and distance to the nearest railway.

$$y_{idg} = \sum_{k=1}^{5} \beta_k \mathbb{1}\{d \in k_{1km}\} + X_i + \alpha_g + \epsilon_{idg} \tag{1}$$

where y_i is the number of months of school that the child attended in 1911, or whether or not a child attended school in 1911. β_k is the distance to the nearest school in 1911 in $1km$ rings, and X_i is a series of controls including sex, age, birthplace, and birth order, and α_g is an enumeration subdistrict fixed effect. In addition I include an indicator for whether the family lives on a homestead parcel of land, as this is a good proxy for family wealth in 1906 when it is not observed. Finally I include the distance to the nearest railway line as a control.

Figure 4 shows the estimated total effect of distance on the number of months of schooling in 1911. All coefficients being relative to the children who are five kilometers or further from the school in 1911. I find that there are significant decreases in the months of school attended in 1911, with children immediately adjacent to the school attending for approximately 0.75 months more than children who are 5km or further away. The distance effects are not statistically different for boys versus girls.

There is a monotonic decline in the estimated effect size with distance, with the effect being statistically insignificant from zero for the full sample when children are in the 3-4km bin. These estimates align with the recommended school district size, which was for 16-20

square miles (sections of land), with the longest distance being no more than five miles in the province of Alberta (Noonan et al., 2006). Assuming a circular district, this gives a radius of approximately 2.5 miles or 4km at the largest suggesting that policy makers of the time accurately judged the largest possible distance that a child would be able to travel to school.

I can further break down these estimates into an intensive and extensive margin effect of distance. The average child who is within 1km of a school is approximately 5 percentage points more likely to have attended school in 1911 relative to a child who is further than 5km from a school. The mean school attendance for all children is 61.5%, so a decline of 6 percentage points is a significant determinant of school attendance. The intensive margin gives a difference of approximately 0.55 months of schooling for a child who is within 1km of a school compared to a child who is more than 5km away. The intensive margin appears to decay faster in space relative to the main effect, with the effects rapidly approaching zero in the 2-3km bin. Figure 14 in the appendix shows these results.

The coefficients reported thus far should not be taken as the causal estimate of distance on schooling, as the distance to school in a single year (in this case, 1911) is likely to be endogenous. There are three possible sources of endogeneity that could bias OLS results of the effect of school distance on schooling.

The primary source of endogeneity is that parents who are more likely to value education highly and send their children to school are more successful at constructing a school near them. This is a particularly strong concern given that school districts were primarily formed at the request of local communities. As a result we would expect that OLS estimates would be biased larger, as the parents who are more likely to send their children to school would be more likely to organize a school near themselves.

One possible way to address this threat is to note that although communities had the power to create schools, individual parents did not. After a proposal for a school district was submitted, the government provided a list of suitable sites for the school, with a preference for choosing the centroid of the district to minimize walking time for children. As a result individual parents could have influence to have a school constructed in their area, but the exact distance to school was largely out of their controls unless they were successful in centering the school district around their home location. Nonetheless, this source of endogeneity remains a concern to be addressed.

A second source of endogeneity is that parents may sort to be closer to the school. If this is the case, we would expect that OLS estimates would be biased larger. I can address this by relying on the fact that homesteaders were required to live and farm on their land for three years following entry to a homestead 0.2, restricting their mobility. Farmers who purchased land from a railway, colonization company, the Hudson's Bay Company, or in a sale of school lands did not face this requirement, but homesteaders were the majority of settlers at the time, making it unlikely that parents sorting closer to schools is driving the results. Figure 8 in the appendix shows the population by section number in 1906, finding that the majority of people were settled on homestead parcels in 1906 and would be subject to these constraints.

The final source of endogeneity is that school and household locations are measured with error. Error could come from the use of centroid of parcels as locations, while in practice households and schools would be located near the corner or edge of a parcel to get access to the road. This pattern in seen in figure 3. This is especially likely to be a problem in

the Saskatchewan data, where just under one half of schools can be located to the one mile resolution as seen in table 23 in the appendix. For these schools I locate them to the centroid of the nearest township (or group of townships), which gives a resolution of approximately 3 miles.

Another possibility in terms of measurement error is that the actual locations of schools may be mis-measured in a given year. In some cases schools closed, which is not reflected in the data and only captures school openings. In addition, schools sometimes changed their locations which I cannot capture. If this were a significant problem we would expect the actual number of schools operating to be lower than the number of schools I report. Instead I find slightly fewer schools operating than I would expect as per table 23 in the appendix. Regardless of the source of measurement error, we would expect the coefficients obtained from OLS to be biased towards zero.

To address these concerns I adopt an instrumental variable strategy. As an instrument I use the market access of each household to the nearest ten schools, while dropping the nearest closest two. The instrument works by instrumenting for the supply of schools, as schools were generally rolled out in a consistent pattern as seen in figure 13, while removing any local variation in school construction from the closest two schools. In addition if there are concerns about measurement error, the second closest school may actually be the closest, which I would also want to instrument for. The school market access for household i is constructed by:

$$MA_i = \Sigma_{i=3}^{10} \frac{1}{log(dist_i)}. \tag{2}$$

I can then estimate equation 1 using the log distance to school as the distance measure. Column 1 of table 3 gives the estimate of the extensive margin of school attendance on the distance to school. The estimated coefficient is negative and statistically significant, indicating that a 100% increase in the distance to school decreases the likelihood of attending school by approximately 3 percentage points. Notably this is a stronger effect that the distance to the nearest railway line, which has been emphasized in the literature.

Column 2 of table 3 gives the first stage if the market access instrument. I find a strong first stage of 1,148.6, with higher values of market access corresponding to lower distances to school in 1911.

Column 3 of table 3 gives the instrumental variable estimate. I find that the expected effect on distance increases significantly to -0.1713, indicating that a 100% increase in distance decreases the likelihood of school attendance by approximately 17 percentage points. The IV estimate is larger than the OLS estimate, suggesting that of the sources of bias measurement error, which biases OLS towards zero, is the most prominent.

I then switch the dependent variable to be the number of months attending school, to estimate the intensive margin response of distance to school. The results of these estimations can be found in table 4. Consistent with figure 15a I find a negative OLS estimate, with an elasticity of -0.369. This estimate decreases significantly to -0.2696 when instrumenting, consistent with OLS being biased towards zero as in table 3.

Taken together these results suggest that the distance to school was a significant determinant of a child's attendance on both the intensive and extensive margins.

Table 3: Attendance vs Distance

Dependent Variables:	Attends School 1911	Log School Distance	Attends School 1911
Model:	(1)	(2)	(3)
	OLS	First Stage	IV
Variables			
Log School Distance	-0.0280***		-0.1713***
	(0.0052)		(0.0446)
Log RR Distance	-0.0107***	0.0407***	-0.0021
	(0.0028)	(0.0058)	(0.0039)
school_MA		-0.0776***	
		(0.0059)	
Fixed-effects			
Birthplace	Yes	Yes	Yes
Enumeration Subdistrict	Yes	Yes	Yes
Homestead	Yes	Yes	Yes
Family Agriculture 1911		Yes	Yes
Sex/Age/BO Interactions	Yes	Yes	Yes
Fit statistics			
F-test (1st stage)		1,148.6	
F-test (1st stage), Log School Distance			1,148.6
Observations	40,926	40,926	40,926

Clustered (Enumeration Subdistrict) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Estimates of the effect of distance to school on school attendance in percentage points. The first column gives the OLS estimate, the second column gives the first stage estimate of distance to school on the market access for schooling, and the third column gives the IV estimate for the effect of school distance on school attendance. Controls are included for birthplace, enumeration subdistrict (geography), whether the family is living on a homestead, whether the family is employed in agriculture, and all possible interaction terms between sex, age, and birth order.

Table 4: Months of School vs Distance

Dependent Variables:	Log Months School 1911	Log School Distance	Log Months Schoo
Model:	(1)	(2)	(3)
	OLS	First Stage	IV
Variables			
Log School Distance	-0.0369***		-0.2696***
	(0.0062)		(0.0591)
Log RR Distance	-0.0140***	0.0418***	-0.0006
	(0.0033)	(0.0071)	(0.0048)
school_MA		-0.0773***	
		(0.0071)	
Fixed-effects			
Birthplace	Yes	Yes	Yes
Enumeration Subdistrict	Yes	Yes	Yes
Homestead	Yes	Yes	Yes
Sex/Age/BO Interactions	Yes	Yes	Yes
Fit statistics			
F-test (1st stage)		670.83	
F-test (1st stage), Log School Distance			670.83
Observations	24,967	24,967	24,967

Clustered (Enumeration Subdistrict) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Estimates of the effect of distance to school on months in school in 1911, expressed in logs. The first column gives the OLS estimate, the second column gives the first stage estimate of distance to school on the market access for schooling, and the third column gives the IV estimate for the effect of school distance on school attendance. Controls are included for birthplace, enumeration subdistrict (geography), whether the family is living on a homestead, whether the family is employed in agriculture, and all possible interaction terms between sex, age, and birth order.

Sex

A descriptive feature emphasized in the school board reports in section 0.1 is that boys stay in school only until they have reached puberty and their labor becomes more valuable. To test if this is quantitatively meaningful I examine differences in school attendance for boys versus girls before and after the age of 13, when puberty typically occurs. I first estimate the difference in school attendance between girls and boys with the following specification:

$$y_{ig} = \beta sex_i \times age_i + sex_i + age_i + X_i\alpha_g + \epsilon_{ig} \tag{3}$$

where sex_i is a binary variable with 0 for female and 1 for male, age_i is an age fixed effect for all ages between 6 and 17, and X_i is a series of other controls including the distance to the nearest railway line in 1911, the distance to the nearest school, birthplace, birth order, and whether or not they family lives on homestead land, and α_g is the enumeration subdistrict. As in the previous section I use both the whether a child attended school in 1911 and the number of months they attended school in 1911 as outcome variables.

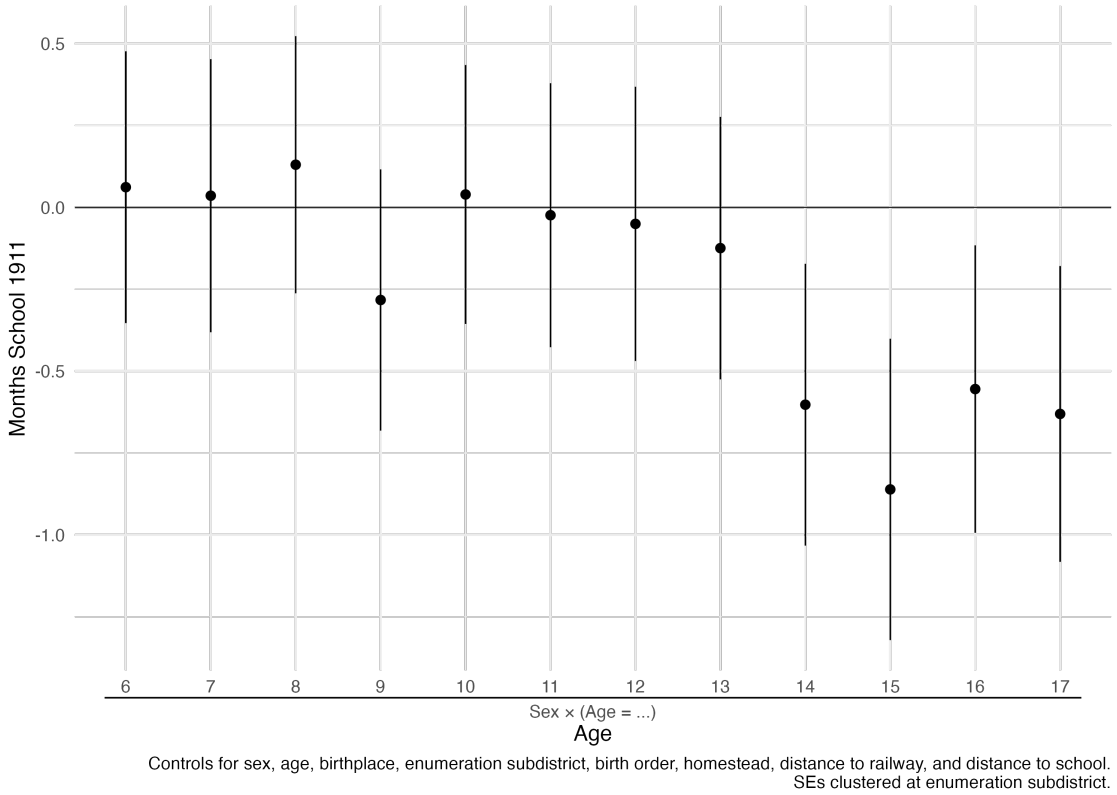
Figure 5 shows the coefficients of the interaction terms between sex_i and age_i from ages 6 to 17. As described in the school board reports there is a clear decline in male school attendance relative to females starting at age 13. The difference becomes significant and large by the late teens, amounting to approximately 0.75 months of school in a year, or three weeks. The average school attendance for children between the ages of 13 to 17 is 4.2 months, so a decline of 0.75 months represents a 17% decline in the amount of schooling attended.

Figure 15 in the appendix shows that this result is primarily driven on the intensive margin. Boys decrease their likelihood of going to school only slightly relative to girls, but conditional on going to school decrease the amount that they go. This is consistent with boys remaining attached to the school, but being more selective about which days of school are attended as the value of their farm labor increases.

I then repeat the estimation in equation 3, but use an indicator for being over the age of 12 (13-17) to capture the effect on boys of puberty. The main result is in column 3, of table 5 where I find a coefficient of -0.6580 on the months of school for boys when over the age of 12, in line with the results in figure 5. I then vary the fixed effects used in the estimation. Column 2 drops the enumeration subdistrict fixed effect, and column 1 drops the homestead fixed effect. In each case I find negligible changes in the interaction effect between sex and being over 12. Column 4 does the opposite, and adds a family fixed effect. The interpretation of this regression is that any changes in male school attendance over the age of 12 are relative to their siblings. Comfortingly I find coefficients that are nearly identical.

These results are unsurprising, and mirror what is report in the Iowa state census in 1915 as reported in Goldin and Katz (2000), that the total stock of education for women was higher than men. A related result from goldin2000 is that the returns to education for women were only significant until the point they were qualified to become a teacher or clerk as adults, and essentially zero aftewords. This suggests that the causal impact of education access may be to drive women towards those occupations, but this remains an area to be further researched.

Figure 5: Males vs Females



Coefficients of the interaction term of sex and age regressed on total months at school. Controls for sex, age, birth order, birthplace, enumeration subdistrict (geography), and whether the family is living on a homestead. The decline in the coefficients represents females attending more school relative to males.

Table 5: Months of School vs Age by Sex

Dependent Variable:	Months School 1911			
Model:	(1)	(2)	(3)	(4)
Variables				
Sex	-0.0212	-0.0187	0.0434	0.0745
	(0.0619)	(0.0619)	(0.0593)	(0.0547)
Sex × Older 12	-0.6048***	-0.6065***	-0.6311***	-0.6615***
	(0.1046)	(0.1048)	(0.0999)	(0.1066)
Older 12	-0.6773***	-0.6801***	-0.7262***	-1.817***
	(0.0929)	(0.0925)	(0.0889)	(0.1032)
Log RR Distance	-0.2468***	-0.2366***	-0.1579***	
	(0.0337)	(0.0340)	(0.0300)	
Log School Distance	-0.3326***	-0.3183***	-0.3357***	
	(0.0612)	(0.0610)	(0.0585)	
Fixed-effects				
Birthplace	Yes	Yes	Yes	Yes
Birth Order	Yes	Yes	Yes	Yes
Homestead		Yes	Yes	
Enumeration Subdistrict			Yes	
Family				Yes
Fit statistics				
Observations	32,293	32,293	32,293	32,293
R^2	0.04889	0.05063	0.15177	0.69390

Clustered (Enumeration Subdistrict) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

People 5-17 in 1911. Lived in Alberta, Saskatchewan, or Manitoba in 1911.
Months of school in 1911 versus sex interacted with age. Column 1 reports the regression without any family characteristics. Column 2 include whether the family lives on a homestead parcel. Column 3 includes enumeration subdistrict fixed effects. Column 4 includes family fixed effects. In each case the coefficient on sex interacted with being older than 12 is nearly identical, indicating that boys decrease their schooling after the age of 12.

1911-1921

In the previous section I established that the distance to school is an important factor in a child's schooling. Next, I want to establish if schooling obtained in 1911 correlates with any later life outcomes. Note that this relies on the months of school that is observable in 1911, which is a flow of schooling in a particular year, rather than the stock of total schooling a person observes. Nonetheless, a positive correlation between schooling in 1911 and later life outcomes is plausible if we believe that schooling either enhances skill or acts as a signalling mechanism for individuals to signal their quality to employers.

I do this by linking the 1911 census records to the 1921 census records, and restricting the sample to people who were between the ages of 6 and 18 in 1911. As both the 1911 census and 1921 census were national censuses, I further restrict the sample to those living in Alberta, Saskatchewan, and Manitoba in 1911, but place no restriction on their location in 1921. The sample is then further restricted to people living in rural areas in 1911, although there is no requirement for precise geolocation.

Income

The first outcome of interest is whether schooling in 1911 has any measurable impact on income as an adult. I restrict the sample to those children who have moved away from home as adults, and estimate the following specification:

$$y_{ig} = \beta s_i + \gamma y_{i-1,g} + X_i + \alpha_g + \epsilon_{ig} \tag{4}$$

where y_i is the log income of the child in 1921, and y_{i-1} is the log family income in 1911 when the child lives at home. s_i is the log months of school in 1911. X_i is a set of controls for individual characteristics, and family characteristics, and α_g is an enumeration subdistrict fixed effect.

The results of this regression can be found in table 6. I find that the elasticity of income with respect to months of schooling is 0.1411 in the preferred specification in column 3. This figure should be treated with some caution however, as many farmers did not report income in both the 1911 and 1921 census waves. The sample size is therefore much smaller than the total possible sample, and is likely biased with only the highest income and most successful farmers in 1911 reporting any income and therefore being their children being included in the sample. Nonetheless, this figure suggests that schooling as a child did have a positive impact on reported incomes as an adult.

Agriculture

The sample of children being analyzed is limited to those in rural areas, where the primary employment is agriculture. One possibility is that access to education may act as a "push" factor, encouraging people to leave agricultural occupations, as suggested by Caselli and Coleman II (2001) and Porzio et al. (2022). To show this I modify equation 5 to get:

$$y_{ig} = \beta s_i + X_i + \alpha_g \epsilon_{ig} \tag{5}$$

Table 6: Log Income versus Schooling

Dependent Variable:	Log Income (1921)		
Model:	(1)	(2)	(3)
Variables			
Log Months School 1911	0.1694***	0.1404***	0.1562**
	(0.0512)	(0.0512)	(0.0681)
Family Log Income (1911)	0.0401**	0.0287	0.0466**
	(0.0167)	(0.0188)	(0.0214)
Fixed-effects			
Individual Characteristics	Yes	Yes	Yes
Family in Agriculture 1911		Yes	Yes
Family Size (1911)		Yes	Yes
Enumeration Sub District (1911)			Yes
Fit statistics			
Observations	1,935	1,935	1,935
R^2	0.14199	0.15704	0.37152

Clustered (Enumeration Sub District (1911)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

People 6-18 in 1911 who are not dependents in 1921. Lived in Alberta, Saskatchewan, or Manitoba in 1911.

Log income in 1921 as an adult, versus log months of schooling in 1911, controlling for family income. Controls for individual characteristics include sex, age, birth order, and birthplace. Column 1 includes only controls for individual characteristics, Column 2 adds controls for whether the family is working in agriculture and family size, and Column 3 adds enumeration subdistrict fixed effects.

Table 7: Likelihood of Being in Agriculture versus Schooling

Dependent Variable:	Family in Agriculture 1921				
Sex x Birth Order	Full sample	First Daughter	First Son	Later Daughter	Later So
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Log Months School 1911	-0.0578***	-0.0641	-0.0276	-0.0660***	-0.0663**
	(0.0123)	(0.0455)	(0.0358)	(0.0250)	(0.0188)
Family in Agriculture 1911	0.1400***	0.0914*	0.0990**	0.0828***	0.1982**
	(0.0157)	(0.0479)	(0.0432)	(0.0256)	(0.0230)
Fixed-effects					
Individual Characteristics	Yes	Yes	Yes	Yes	Yes
Family Size (1911)	Yes	Yes	Yes	Yes	Yes
Enumeration Sub District (1911)	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	10,326	1,100	1,475	3,566	4,185
R^2	0.16626	0.48376	0.39485	0.24398	0.24348

Clustered (Enumeration Sub District (1911)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
People 6-18 in 1911 who are not dependents in 1921. Lived in Alberta, Saskatchewan, or Manitoba in 1911.
Likelihood of a person's family working in agriculture in 1921 as an adult, versus log months of schooling in 1911, controlling for whether the family worked in agriculture as a child. Controls for individual characteristics include sex, age, birth order, and birthplace. Column 1 includes the entire sample, while columns 2 through 5 split the sample by birth order and sex. Firstborn sons and daughters are the firstborn children in their families, while later sons and daughters include all children who are second or later in the birth order.

where y_{ig} is the probability of the family working in agriculture in 1921, and s_i is the months of schooling in 1911. I also include controls for whether the person's family was working in agriculture in 1911, as working in agriculture is likely to be persistent between generations.

Column 1 of table 7 shows that the coefficient on log months of school and being in agriculture is -0.0550, that is for every doubling of months in school as a child, the likelihood of remaining in agriculture as an adult decreases by approximately 5 percentage points. There is significant heterogeneity by birth order and sex. Columns 2 through 5 break out each child into whether they are the firstborn son or daughter, or a later child. While all coefficients are negative, the estimates for firstborn sons are significantly smaller in magnitude. This likely reflects the idea that the firstborn son should inherit the family land, so that any additional schooling does not affect their likelihood of leaving agriculture as an adult.

Difference in Differences

Until now the entirety of the results have been cross sectional, relying on variation within a single wave of the census on later outcomes. I now extend the analysis by utilizing the rollout of schools over time to provide variation in school exposure.

The 1906 census provides precise locations for rural households, while the assembled school data covers school openings in the period from 1871 to 1912. Assuming that movement is limited over short time periods, as in the previous section, then distances to school will be most accurate around the year 1906. As a result I utilize the changes in school distance between 1900 and 1912. Additionally I link the 1906 census to the 1931 census to get the adult outcomes of children. I restrict the sample to men only, as most women who are children in 1906 are likely to have married by 1931, making matching unreliable due to changing last names.

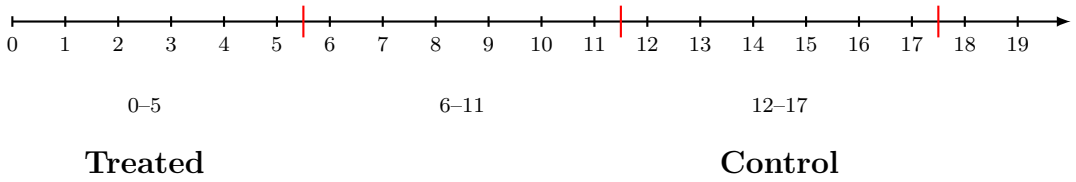


Figure 6: Ages in 1906

To identify the effect of school opening on later life outcomes I follow the design of Duflo (2001). I take a cohort of children as my treated cohort if they are between the ages of 0-5 in 1906, and control as those between the ages of 12-17 in 1906 as in figure 6. Treatment intensity is then if the distance to school between 1901 and 1906 decreases significantly. Within the treated and control cohort there are then treated individuals who saw significant decreases in their distance to school, and control individuals who did not see significant decreases.

To identify the effect of school construction we assume that school construction benefits the 0-5 cohort, but does not benefit the 12-17 cohort who are too old to take advantage of the access to schooling. This is a realistic assumption as children at the time typically decreased school participation around age 12, especially boys as shown in table 5. Using an exposed versus unexposed cohort helps assuage concerns that parents may have influenced school construction locations. If parents were able to influence construction locations there would be no concern as long as they could not manipulate the timing of school construction significantly.

This design also controls for other possible shocks we might worry could confound the results. Including an older control cohort allows me to control for other aggregate shocks which affects both cohorts equally, for example bad harvests that drive families out of agriculture. Shocks to one cohort are also not a concern so long as they do not correlate with treatment status. This is a particular concern with World War I participation which would affect the 12-17 cohort, who will be 20-25 and therefore prime military age in 1914 at the start of the war. Thankfully the design accounts for this, so long as participation in the war is not correlated with the likelihood of being treated by school construction.

Formally, I construct the estimation equation as follows. The comparison is between those who are in the 0-5 cohort ($t_c = \mathbb{1}\{age_{1906} < 6\}$) in 1906 and those who are 12-17 in 1906. An individual is considered treated if the distance to school decreased by 5km or more between 1901 and 1906 ($s_i = \mathbb{1}\{\Delta dist_{i,1901,1906} > 5km\}$). Unlike Duflo (2001), my treatment is defined at the individual, not region level. The estimation equation is therefore:

$$y_{icg} = \gamma s_i t_c + s_i + t_c + \beta X_i + \alpha_g + \epsilon_{icg} \tag{6}$$

where y_{icg} is the outcome of interest in 1931, s_i is the treatment, t_c is whether a person is in the treated cohort, X_i is a series of controls, and α_g is a geographic fixed effect for the enumeration subdistrict.

I first examine if any member of the household in 1931 works in an agricultural occupation. Table 8 shows the results. Aligning with the results in the previous section I find that having a school constructed nearby decreases the likelihood of being in agriculture by 4.69pp in the baseline specification in column 1. This value decreases slightly with the inclusion of birthplace controls for own birthplace, fathers birthplace, and mothers birthplace, increasing again with the inclusion of homestead, livestock, and agricultural controls.

Column 5 of table 8 includes a control for the log distance to railway in 1906, and is the preferred specification. Railways were the primary mode of transportation at the time, providing both outside goods and a location to sell grain. I find that the likelihood of remaining in agriculture is positively associated with distance to the railroad in 1906, consistent with more remote regions being more likely to be agricultural. This however understates the true impact of the railway, as rail construction continued rapidly throughout this period.

Column 6 of table 8 takes a more dynamic approach to estimating the impact of the railway. I construct the same treatment measure as with schooling, treating an observation as treated if the distance to the railway decreases by 5km or more in the period between 1901 and 1906. The estimated effect in column 6 of the railway construction is positive (although only weakly significant), suggesting that rail construction induces people to stay in agriculture. Importantly including this additional treatment increases the magnitude of the effect of school opening to -4.75pp. The opposite movement of these two forces shows that while education can act as a push out of agriculture, increases in market access may cause people to specialize further in agriculture.

A important caveat of this observation is that the assumption that the treatment effect reflects the effect on the treated cohort only is likely violated. Railroad construction is likely to affect both the 0-5 cohort and the 12-17 cohort. As a result we should not interpret the coefficient as the treatment effect on the treated cohort.

To ensure the robustness of the results I vary the definitions of the treated and control cohorts, and definition of the treatment intensity, the and the cutoff value of 5km. These results can be found in tables 25 through 26 in the appendix.

As a further robustness check I check an alternate definition of working in agriculture by focussing only on the individual who I observe in 1906. I code them as working in agriculture of they are listed in an agricultural occupation, or if they are listed as a labourer working on a farm. Table 9 shows the results. A with table 8 I vary the set of contols, with the preferred specification in column 5, giving a -3.25pp decrease in the likelihood of being a farmer as an

Table 8: Family in Agriculture in 1931

Dependent Variable:	Family in Ag				
Model:	(1)	(2)	(3)	(4)	(5)
School Constr. × Treated Cohort	-0.0478***	-0.0374**	-0.0417**	-0.0413**	-0.0413**
	(0.0161)	(0.0167)	(0.0165)	(0.0164)	(0.0164)
School Constr.	0.0480***	0.0382**	0.0428**	0.0413**	0.0367**
	(0.0180)	(0.0182)	(0.0181)	(0.0184)	(0.0186)
Treated Cohort	0.0364***	0.0431***	0.0637***	0.0632***	0.0623***
	(0.0100)	(0.0102)	(0.0100)	(0.0100)	(0.0100)
Log RR Distance (1906)					0.0196***
					(0.0044)
Match Prob.	0.4508***	0.4403***	0.4279***	0.4279***	0.4291***
	(0.0250)	(0.0252)	(0.0251)	(0.0253)	(0.0252)
Enumeration Subdistrict (1906)	Yes	Yes	Yes	Yes	Yes
Fam. Bpl.		Yes	Yes	Yes	Yes
Homestead			Yes	Yes	Yes
Livestock			Yes	Yes	Yes
Ag. Productive.				Yes	Yes
Observations	14,747	14,747	14,747	14,747	14,747
R^2	0.07750	0.08399	0.09483	0.09733	0.09896

Clustered (Enumeration Subdistrict (1906)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Likelihood of being having at least one family member in agriculture as an adult in 1931. School construction is defined as 1 if the distance to school decreased by more than 5km between 1901 and 1906. The treated cohort is those who are 0-5 in 1906, as compared to those who are 12-17. Railway construction is defined as 1 if the distance to the nearest railway line decreases by more than 5km between 1901 and 1906. Homestead is a control for the type of land parcel a family is on in 1906. Family birthplace are controls for own, father's and mother's birthplace. Livestock controls for family holdings of cattle, hogs, horses, milk cows, and sheep in 1906 as taken from the census. Agricultural conditions are for the agricultural productivity in wheat, oats, barley, flax, and reed canary grass.

adult.

Finally I turn to another measure of economic mobility, the distance that one lives from home. Table 10 shows the results. I find that having a school constructed nearby significantly increases the likelihood of moving further away across specifications. Many individuals are still located on the same parcel of land that they were as children. The results hold regardless of how these individuals are treated, as seen between column 2 which uses the inverse hyperbolic arcsine transformation, and column 3 which uses a log transformation. Column 4 performs a separate estimation checking if an individual is located on their family farm from 1906. As expected I find that school construction decreases the likelihood to be found on the family farm.

Table 9: Farmer in 1931

Dependent Variable:	Farmer				
Model:	(1)	(2)	(3)	(4)	(5)
School Constr. × Treated Cohort	-0.0417**	-0.0322*	-0.0364**	-0.0359**	-0.0359**
	(0.0172)	(0.0175)	(0.0173)	(0.0173)	(0.0173)
School Constr.	0.0534***	0.0436**	0.0487**	0.0477**	0.0430**
	(0.0185)	(0.0189)	(0.0188)	(0.0191)	(0.0192)
Treated Cohort	-0.0184*	-0.0117	0.0100	0.0098	0.0088
	(0.0102)	(0.0106)	(0.0106)	(0.0106)	(0.0106)
Log RR Distance (1906)					0.0200***
					(0.0045)
Match Prob.	0.4762***	0.4639***	0.4517***	0.4523***	0.4535***
	(0.0251)	(0.0255)	(0.0252)	(0.0253)	(0.0253)
Enumeration Subdistrict (1906)	Yes	Yes	Yes	Yes	Yes
Fam. Bpl.		Yes	Yes	Yes	Yes
Homestead			Yes	Yes	Yes
Livestock			Yes	Yes	Yes
Ag. Productive.				Yes	Yes
Observations	14,747	14,747	14,747	14,747	14,747
R^2	0.07873	0.08476	0.09590	0.09747	0.09906

Clustered (Enumeration Subdistrict (1906)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Likelihood of working as a farmer or farm labourer in 1931. School construction is defined as 1 if the distance to school decreased by more than 5km between 1901 and 1906. The treated cohort is those who are 0-5 in 1906, as compared to those who are 12-17. Railway construction is defined as 1 if the distance to the nearest railway line decreases by more than 5km between 1901 and 1906. Homestead is a control for the type of land parcel a family is on in 1906. Family birthplace are controls for own, father's and mother's birthplace. Livestock controls for family holdings of cattle, hogs, horses, milk cows, and sheep in 1906 as taken from the census. Agricultural conditions are for the agricultural productivity in wheat, oats, barley, flax, and reed canary grass.

Table 10: Migration Distance between 1906 and 1931

Dependent Variables:	Migration Dist. (km)	Asinh Migration. Dist.	Log Migration Dist.	On Family Farm
Model:	(1)	(2)	(3)	(4)
School Constr. × Treated Cohort	30.94***	0.2310**	0.1737**	-0.0321*
	(9.814)	(0.0897)	(0.0768)	(0.0173)
School Constr.	-15.29	-0.1218	-0.0638	0.0213
	(9.757)	(0.1009)	(0.0914)	(0.0185)
Treated Cohort	-7.760	-0.1904***	-0.0050	0.0451***
	(6.615)	(0.0517)	(0.0469)	(0.0100)
Log RR Distance (1906)	-4.688*	-0.0399*	-0.0166	0.0117***
	(2.415)	(0.0219)	(0.0193)	(0.0038)
Match Prob.	-975.7***	-5.547***	-4.734***	0.5049***
	(27.88)	(0.1231)	(0.1035)	(0.0189)
Enumeration Subdistrict (1906)	Yes	Yes	Yes	Yes
Fam. Bpl. + Homestead	Yes	Yes	Yes	Yes
Livestock + Ag.	Yes	Yes	Yes	Yes
Observations	14,747	14,747	13,025	14,747
R^2	0.28121	0.21798	0.23677	0.12190

Clustered (Enumeration Subdistrict (1906)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Distance from childhood home in 1906 versus adult home in 1931. Column 1 reports the effect on linear distance, which are highly skewed. Column 2 resolves this and reports the estimated effect taking the Inverse Hyperbolic Arcsine of the distances. To check if the presence of zeros is driving the results, column 3 reports the log migration distance. Finally Column 4 reports a binary outcome for if an individual is living on the same farm as an adult that they were as a child. School construction is defined as 1 if the distance to school decreased by more than 5km between 1901 and 1906. The treated cohort is those who are 0-5 in 1906, as compared to those who are 12-17. Railway construction is defined as 1 if the distance to the nearest railway line decreases by more than 5km between 1901 and 1906. Homestead is a control for the type of land parcel a family is on in 1906. Family birthplace are controls for own, father's and mother's birthplace. Livestock controls for family holdings of cattle, hogs, horses, milk cows, and sheep in 1906 as taken from the census. Agricultural conditions are for the agricultural productivity in wheat, oats, barley, flax, and reed canary grass.

Conclusion

The role of the one room school house in the development of the Canadian (and American) prairies is widely reported, but not extensively studied. This paper uses geolocated households and schools to show that these rural schools both encouraged school participation, that school participation gave children new opportunities outside of agriculture, and that the estimated causal impact of school construction was to push children out of agricultural occupations. This is consistent with education acting as a push factor for structural change out of agriculture.

Further work is still needed to fully explore the effect of one room schools. The formation of local schools was often a chaotic process, with schools opening based on the demands of the local community (Noonan et al., 2006; Baergen, 2005). It's unclear what factors would have lead schools to open earlier or later. In addition there was widespread pressure to consolidate rural schools starting in the 1910's, with concerns that rural schools were too small to be effective and had low teaching standards (Noonan et al., 2006; Baergen, 2005). It is unclear whether consolidation of rural schools would have been beneficial, given the significant geographic effects found by this paper.

The results presented here also hint at a relationship with rail construction and market access, which pushed in the opposite direction of school construction. If this the case it suggests that the exact infrastructure that governments choose to prioritize has an important impact in rural-urban migration and structural change.

References

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson (Mar. 2019). "To the New World and Back Again: Return Migrants in the Age of Mass Migration". In: *ILR Review* 72.2, pp. 300–322. ISSN: 0019-7939, 2162-271X. DOI: 10.1177/0019793917726981. URL: http://journals.sagepub.com/doi/10.1177/0019793917726981 (visited on 09/04/2024).

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (Aug. 1, 2012). "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration". In: *American Economic Review* 102.5, pp. 1832–1856. ISSN: 0002-8282. DOI: 10.1257/aer.102.5.1832. URL: https://pubs.aeaweb.org/doi/10.1257/aer.102.5.1832 (visited on 09/04/2024).

— (June 2014). "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration". In: *Journal of Political Economy* 122.3, pp. 467–506. ISSN: 0022-3808, 1537-534X. DOI: 10.1086/675805. URL: https://www.journals.uchicago.edu/doi/10.1086/675805 (visited on 01/22/2021).

Abramitzky, Ran et al. (Sept. 1, 2021). "Automated Linking of Historical Data". In: *Journal of Economic Literature* 59.3, pp. 865–918. ISSN: 0022-0515. DOI: 10.1257/jel.20201599. URL: https://pubs.aeaweb.org/doi/10.1257/jel.20201599 (visited on 01/22/2024).

Antonie, Luiza et al. (Apr. 1, 2014). "Tracking People over Time in 19th Century Canada for Longitudinal Analysis". In: *Machine Learning* 95.1, pp. 129–146. ISSN: 1573-0565. DOI: 10.1007/s10994-013-5421-0. URL: https://doi.org/10.1007/s10994-013-5421-0 (visited on 06/20/2024).

Baergen, William Peter (2005). *Pioneering with a Piece of Chalk: The One-Room Country Schools of Alberta, 1885-1982*. Stettler, AB: W.P. Baergen. 622 pp. ISBN: 978-0-9739491-0-0.

Bagagli, Sara (2023). *Once Upon a Time in the (Canadian) West: Railroads, Location Fundamentals, and the Growth of Cities*. Sara Bagagli. URL: https://sarabagagli.github.io/research/publication/ (visited on 03/21/2023).

Bailey, Martha et al. (Nov. 2017). *How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data*. DOI: 10.3386/w24019. URL: https://www.nber.org/papers/w24019 (visited on 03/14/2023). Pre-published.

Buckles, Kasey et al. (Sept. 2023). *Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project*. DOI: 10.3386/w31671. URL: https://www.nber.org/papers/w31671 (visited on 09/14/2023). Pre-published.

Canada, Library and Archives (Mar. 14, 2023a). *1931 Census of Canada to Be Released on June 1, 2023*. URL: https://www.canada.ca/en/library-archives/news/2023/03/1931-census-of-canada-to-be-released-on-june-1-2023.html (visited on 09/12/2024).

— (May 24, 2023b). *Prairie Provinces Census, 1870-1926*. URL: https://library-archives.canada.ca/eng/collection/research-help/genealogy-family-history/censuses/pages/prairie-provinces.aspx (visited on 09/12/2024).

Cartography office, Geography Department (June 2, 2022). *Historical Canadian Railroads*. Version 2. Borealis. DOI: 10.5683/SP2/UCCFVQ. URL: https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP2/UCCFVQ (visited on 10/24/2024).

Caselli, Francesco and Wilbur John Coleman II (2001). "The U.S. Structural Transformation and Regional Convergence: A Reinterpretation". In: *Journal of Political Economy* 109.3, pp. 584–616. ISSN: 0022-3808. DOI: 10.1086/321015. URL: https://www.jstor.org/stable/10.1086/321015 (visited on 11/20/2023).

Dippel, Christian, Dustin Frye, and Bryan Leonard (2020). *Property Rights without Transfer Rights: A Study of Indian Land Allotment*. DOI: 10.3386/w27479. URL: https://www.nber.org/papers/w27479 (visited on 09/18/2022). Pre-published.

Dippel, Christian and Sebastian Ottinger (2020). "The "Educated American" in the Age of Mass Migration: School Decentralization and Preference Heterogeneity". In.

Duflo, Esther (Sept. 1, 2001). "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment". In: *American Economic Review* 91.4, pp. 795–813. ISSN: 0002-8282. DOI: 10.1257/aer.91.4.795. URL: https://pubs.aeaweb.org/doi/10.1257/aer.91.4.795 (visited on 06/27/2023).

Eckert, Fabian and Tatjana Kleineberg (2021). "Saving the American Dream? Education Policies in Spatial General Equilibrium". In.

Enamorado, Ted, Ben Fifield, and Kosuke Imai (Nov. 17, 2023). *fastLink: Fast Probabilistic Record Linkage with Missing Data*. Version 0.6.1. URL: https://cran.r-project.org/web/packages/fastLink/index.html (visited on 09/30/2024).

FAO and IIASA (2024). *Global Agro Ecological Zones Version 4 (GAEZ V4)*. URL: http://www.fao.org/gaez/.

Feigenbaum, James J (2016). "A Machine Learning Approach to Census Record Linking". In.

Ferrie, Joseph P. (Oct. 1, 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". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 29.4, pp. 141–156. ISSN: 0161-5440. DOI: 10.1080/01615440.1996.10112735. URL: https://doi.org/10.1080/01615440.1996.10112735 (visited on 09/04/2024).

French, Jacob (2022). "Technological Change, Inequality, and Intergenerational Mobility: The Case of Early 20th Century Agriculture". In: p. 61.

Ghose, Devaki (2023). "Trade, Internal Migration, and Human Capital: Who Gains from India's IT Boom?" In.

Goldin, Claudia (1998). "America's Graduation from High School: The Evolution and Spread of Secondary Schooling in the Twentieth Century". In: *The Journal of Economic History* 58.2, pp. 345–374. ISSN: 0022-0507. URL: https://www.jstor.org/stable/2566738 (visited on 11/22/2023).

Goldin, Claudia and Lawrence F. Katz (Sept. 2000). "Education and Income in the Early Twentieth Century: Evidence from the Prairies". In: *The Journal of Economic History* 60.3, pp. 782–818. ISSN: 1471-6372, 0022-0507. DOI: 10.1017/S0022050700025766. URL: https://www.cambridge.org/core/journals/journal-of-economic-history/article/education-and-income-in-the-early-twentieth-century-evidence-from-the-prairies/0F5357717F81B36ADFDB9B8DC0C0A848 (visited on 12/09/2023).

— (2008). *The Race Between Education and Technology*. Cambridge, UNITED STATES: Harvard University Press. ISBN: 978-0-674-03773-1. URL: http://ebookcentral.proquest.com/lib/utoronto/detail.action?docID=3300082 (visited on 11/14/2024).

Government of Alberta (1912). "Annual Report of the Department of Education of the Province of Alberta." In: *Annual report of the Department of Education of the Province of Alberta.*

— (2018). *Establishing a Separate School District Information Package.* URL: https://education.alberta.ca/media/3772257/2018-separate-school-establishment-manual.pdf (visited on 09/27/2024).

— (2020). *ATS v4.1 Polygons - Township Index - Open Government.* URL: https://open.alberta.ca/opendata/a2f3c476882041a8b7e0def645122a1a#detailed (visited on 12/15/2023).

Government of Canada (1906). *Order-in-Council Issue of Proclamation for Taking the Census of Manitoba, Saskatchewan, and Alberta on 1906/06/24.* URL: https://www.bac-lac.gc.ca/eng/census/1906/Documents/Order-in-Council-1906-0887.pdf (visited on 10/08/2024).

Government of Manitoba (1911, 1912). *Report of the Department of Education.*

Hamilton, Beckey (n.d.). "Explaining the Absence and the Characteristics of the Migrants: French Canadians in Southwestern Saskatchewan at the Turn of the 20th Century". In: 13 ().

Hansen, Zeynep K and Gary D Libecap (Apr. 2004). "The Allocation of Property Rights to Land: US Land Policy and Farm Failure in the Northern Great Plains". In: *Explorations in Economic History* 41.2, pp. 103–129. ISSN: 00144983. DOI: 10.1016/j.eeh.2003.10.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0014498303000640 (visited on 11/02/2022).

Helgertz, Jonas et al. (Jan. 2, 2022). "A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel". In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55.1, pp. 12–29. ISSN: 0161-5440. DOI: 10.1080/01615440.2021.1985027. URL: https://doi.org/10.1080/01615440.2021.1985027 (visited on 05/24/2024).

HGIS Lab (2023). *The Canadian Historical GIS — Canadian Historical Geographic Information System.* URL: https://hgiscanada.usask.ca/download (visited on 12/15/2023).

Hornbeck, Richard (May 1, 2010). "Barbed Wire: Property Rights and Agricultural Development*". In: *The Quarterly Journal of Economics* 125.2, pp. 767–810. ISSN: 0033-5533. DOI: 10.1162/qjec.2010.125.2.767. URL: https://doi.org/10.1162/qjec.2010.125.2.767 (visited on 02/02/2023).

— (June 2012). "The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe". In: *American Economic Review* 102.4, pp. 1477–1507. ISSN: 0002-8282. DOI: 10.1257/aer.102.4.1477. URL: https://www.aeaweb.org/articles?id=10.1257/aer.102.4.1477 (visited on 03/10/2023).

— (2022). "Dust Bowl Migrants: Environmental Refugees and Economic Adaptation". In.

Hsiao, Allan (2023). "Educational Investment in Spatial Equilibrium". In.

Information Services Corporation (ISC) (2020). *Cadastre, Saskatchewan - Overview.* URL: https://www.arcgis.com/home/item.html?id=724e988767cd4872a8ca14ce4ca15647 (visited on 12/15/2023).

Lachanski, Michael S. (2024). *The Elusive Rural U.S. Historical Educational Advantage: New Evidence for the 1881 to 1914 Cohorts.*

Lambrecht, Kirk N. (1991). *The Administration of Dominion Lands, 1870-1930*. 1st ed. Occasional Papers 3. Regina, Sask: Canadian Plains Research Center, University of Regina. 405 pp. ISBN: 978-0-88977-049-2.

Leonard, Bryan and Brian Kogelmann (2022). "Does Equality Persist? Evidence from the Homestead Act". In: *Journal of Political Institutions and Political Economy* 3.2, pp. 215–241. ISSN: 2689-4823, 2689-4815. DOI: 10.1561/113.00000059. URL: http://www.nowpublishers.com/article/Details/PIP-0059 (visited on 02/02/2023).

Libecap, Gary and Zeynep Kocabiyik Hansen (June 2000). *"Rain Follows the Plow" and Dryfarming Doctrine: The Climate Information Problem and Homestead Failure in the Upper Great Plains, 1890-1925*. h0127. Cambridge, MA: National Bureau of Economic Research, h0127. DOI: 10.3386/h0127. URL: http://www.nber.org/papers/h0127.pdf (visited on 11/02/2022).

Libecap, Gary D. and Dean Lueck (June 2011). "The Demarcation of Land and the Role of Coordinating Property Institutions". In: *Journal of Political Economy* 119.3, pp. 426–467. ISSN: 0022-3808, 1537-534X. DOI: 10.1086/660842. URL: https://www.journals.uchicago.edu/doi/10.1086/660842 (visited on 09/18/2022).

Manitoba Government (2020). *Manitoba Original Survey Legal Descriptions*. URL: https://geoportal.gov.mb.ca/ (visited on 12/15/2023).

Manitoba Historical Society (2019). *MHS Archives: Historic Sites of Manitoba*. URL: https://www.mhs.mb.ca/docs/sites/index.shtml (visited on 12/18/2023).

Mattheis, Ross and Itzchak Tzachi Raz (2019). "There's No Such Thing As Free Land: The Homestead Act and Economic Development". In: p. 91.

Moir, George (1934). *The Separate School Question in Canada*. The Ryerson Press, Toronto.

Nagy, Dávidkrisztián (Jan. 31, 2023). "Hinterlands, City Formation and Growth: Evidence from the U.S. Westward Expansion". In: *The Review of Economic Studies*, rdad008. ISSN: 0034-6527, 1467-937X. DOI: 10.1093/restud/rdad008. URL: https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdad008/7017614 (visited on 03/20/2023).

Noonan, Brian, Diane Hallman, and Murray Scharf (2006). *A History of Education in Saskatchewan: Selected Readings*. ISBN: 0-88977-190-1.

One Room School Project (2023). *Saskatchewan Gen Web - One Room School Project - Evolution*. URL: https://sites.rootsweb.com/~cansk/school/Evolution.htm (visited on 04/29/2023).

Perfect, Mary B. (1978). *One Hundred Years in the History of the Rural Schools of Manitoba: Their Formation, Reorganization and Dissolution (1871-1971)*.

Porzio, Tommaso, Federico Rossi, and Gabriella Santangelo (Aug. 2022). "The Human Side of Structural Transformation". In: *American Economic Review* 112.8, pp. 2774–2814. ISSN: 0002-8282. DOI: 10.1257/aer.20201157. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20201157 (visited on 11/20/2023).

Price, Joseph et al. (Apr. 2021). "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set". In: *Explorations in Economic History* 80, p. 101391. ISSN: 00144983. DOI: 10.1016/j.eeh.2021.101391. URL: https://linkinghub.elsevier.com/retrieve/pii/S0014498321000024 (visited on 05/28/2024).

Saskatchewan Department of Education (1906, 1907, 1908, 1910, 1911, 1912, 1914). *Annual Report of the Saskatchewan Department of Education.*

Saskatchewan Gen Web (2023). *One Room School Project*. URL: https://saskgenweb.ca/cansk/school/schooldistricts.html (visited on 12/18/2023).

Schaede, Ursina (2021). *The Long Run Effects of Funding for Public Education: Evidence from Land Grants*. Ursina Schaede. URL: https://ursinaschaede.github.io/research/ (visited on 03/22/2023).

Sharp, Paul F. (1947). "The American Farmer and the "Last Best West"". In: *Agricultural History* 21.2, pp. 65–75. ISSN: 0002-1482. URL: https://www.jstor.org/stable/3739391 (visited on 11/14/2024).

Smith, Cory (2022). "Land Concentration and Long-run Development in the Frontier United States". In.

The Canadian Peoples (2023). *The Canadian Peoples*. The Canadian Peoples. URL: https://thecanadianpeoples.com/ (visited on 04/30/2023).

University of Calgary (2019). *Glenbow Archives School District Database, Courtesy of Libraries and Cultural Resources at the University of Calgary*.

Wellington, J, R Harper, and KJ Thompson (Aug. 28, 2024). *Hlink*. Institute for Social Research and Data Innovation, University of Minnesota. URL: https://github.com/ipums/hlink (visited on 09/04/2024).

Appendix

0.1 Selected Quotes

There is significant evidence from the department of education reports that school attendance in rural areas was significantly lower that desired, and that boys in particular were leaving school at a faster rate than girls. The following quotes provide some accounts of the challenges in rural schools from the saskatchewan department of education (Saskatchewan Department of Education, 1906, 1907, 1908, 1910, 1911, 1912, 1914).

"Parents who keep children out of school because of distance or danger, appear to have no hesitation in sending the boys to market" - A. Kennedy, Inspector of Schools, 1910

"The attendance in rural districts is not so regular as one would expect or could hope for. Undoubtably, it is in part due to the scarcity of farm help and the long distances some children have to go; but I fear it is also caused by a lack of interest on the part of many parents", H.H. Smith, Inspector of Schools, Saskatoon, December 31, 1909

"When a boy is old enough to work an outfit on a farm, his school days are over", John S. Huff, Davidson, Saskatchewan., March 1st, 1912

0.2 Land Allocation

Dominion land policy was marked by frequent changes between the acquisition of Rupert's Land and the North-West Territories and later settlement. I rely on Lambrecht (1991) as the primary source of both narrative and legislative background. As described in section the primary landholders after settlement were railways, homesteaders, the Hudson's Bay Company, and land reserved for the funding of schools. Despite the relative simplicity of this formula, there was significant variation and modification of land policy that is not captured by it.

To fund the construction of railways, approximately 32 million acres was given away as grants to encourage construction between 1881 and 1894. The system was initiated to subsidize the construction of the Canadian Pacific Railway (CPR), but also applied to a multitude of smaller railways as well. Like Canadian land policy in general, there was significant variation in policy over time before the pattern seen in figure 1 became firmly established.

The 1870's featured a variety of attempts at establishing land grants and highlighted conflicting goals in Canadian policy. The first grant system in 1871 allowed three full townships (approximately 18 miles) on either side of the railway to be withdrawn as a grant, superseded in 1872 by the Dominion Lands Act which allowed land to be withdrawn up to 20 miles on either side. Reserving these lands for the railway had the unintended consequence of limiting settlement, another main goal of the government. To overcome this the land grant policy was modified again in 1879 to allow belts of land at different distances to have different checkerboard patterns with homesteads and railway land. Finally in 1881 the Canadian Pacific Railway Act authorized the current system, with alternate sections within 24 miles of the railway given to the CPR for a total of 25 million acres.

Other railways constructed across the prairies had similar arrangements with the government. Individual contracts typically specified 6,400 acres of land for mile of line, the

equivalent of every other section within a 20 mile band around the railway. In all of these contracts, including the CPR, was an indemnity provision which allowed the railway to not take a particular parcel of land if it was not "fairly fit for settlement". This provision meant that railways would often select lands that were far from their railway line, especially as rail lines started to cross. Although land grants were removed in 1894, the indemnity provision led to a delay in railways selecting their parcels, which was not resolve until 1908.

After the railway land grant system ended there were still significant supports for railway construction through subsidized corporate debt. The Railway Act allowed a rail company to pass through the subsidy lands of another, encouraging more railway construction. Railways began to consolidate as well, with Canadian Northern forming from absorbing smaller railways with land grants, and later merging with Grand Trunk Pacific to form Canadian National.

Railway construction was primarily devised to support settlement, which was enabled through homesteads. The homestead system was designed to compete with the United States, including advertisements at which referred to Canada as the "Last Best West" (Sharp, 1947). As a result the homestead system in Canada resembles that of the United States, but with additional features that would make it more appealing to would be settlers. As with railway land there was some variance in policy over time.

At the core of the system was a 160 acre quarter section, which would occupy one quarter of a section of land. The majority of homesteads would be on even numbered sections, except for 8 and 26, with some railway sections also being sold as homesteads if the railway used the indemnity provision to not take possession of a particular parcel. Obtaining a homestead first required "entry" by filing for a homestead at a Dominion Lands office, followed by a three year period of homestead duties which required living on and cultivating that parcel. After fufilling homestead duties, the homesteader could apply for letters patent, or "perfecting" the homestead, and become the owner of the parcel. Some homesteaders purchased their land in less than three years.

While applying for entry, homesteaders were able to apply for pre-emption, or a second parcel of land which was contiguous with the homestead. The homesteader could cultivate this land immediately, and could be purchased after fulfilling homestead duties. The combination of pre-emption and railway indemnity slowled the rate of settlement by restricting the amount of land available. As a result the right of pre-emption was removed in 1890, then reintroduced in 1908 after the railway indemnities were resolved. More than half of homestead entries would include a pre-emption.

The eligibility for a homestead was intentionally broad. In 1874 the required age was reduced from 21 to 18, and in 1876 women were excluded unless they were the head of the household. There were no nationality requirements until 1908, when it was required that a homesteader had to be or intend to be a British subject. After 1916 at the start of World War I, entry was restricted to British subjects.

There were some other mechanisms that could be used to obtain homestead land. From 1883 to 1886 it was possible to have a second homestead, and this was again possible after 1908. Widespread crop failures after WWI prompted the government to allow a second homestead if the first had failed. In 1881 there were also colonization companies, which would purchase land and receive rebates depending on the success of their sales. Approximately 3 million acres were allocated to these companies, but they failed after the completion of the

initial boom of construction with the CPR in 1885.

The final method for land allocation was as a reward for military service. Grants and scrip were given for service in the militia and North-West Mounted Police starting in 1871, with repeated extensions for entry until 1900. A similar program was established for soldier and nurses who participated in the Boer War, and involved the transfer of approximately 2 million acres in Saskatchewan and 1 million acres in Alberta. This transfer was criticized for not imposing any domicile conditions unlike regular homesteads. Finally soldiers returning from WWI were able to make an entry for an additional 160 acres in addition to a regular homestead, with subsided loans to strengthen agricultural life and "make men settlers".

The Hudson's Bay Company (HBC) was granted 5% of all available land in the fertile belt in exchange for Rupert's Land and the North-West Territories as part of the deed of surrender. This land amounted to section 8 in every township, the entirety of section 26 in every fifth township, and three quarter of section 26 in four fifths of the townships [5]. The fertile belt was defined as the land bounded by the United States border to the south, the North Saskatchewan river to the north, the Rocky Mountains to the west, and the Laek Winnipeg-Lake of the Woods water system to the east. In addition the HBC was given title to any land that it had fur trading posts on and adjacent parcels not to exceed 50,000 acres. There were some negotiations to allow the government to fufil other objectives if section 8 or 26 were needed, but by 1924 a total of 6,639,059 acres had been given to the HBC.

The HBC, recognizing the value of the land that it had been granted, held onto it's portfolio of land strategically and only sold parcels when the surrounding area had been settled to maximize the value (Hamilton, n.d.). This strategy was achieved by first allowing settlement to occur in the surrounding land, including the development of railways. Land closer to railways would sell faster (Hamilton, n.d.). The HBC continued to sell land throughout the twentieth century, and by 1961 formed the Rupert's Land Trading Company to administer the remaining parcles. By 1984 the final 5,100 acres in Saskatchewan were donated to the Saskatchewan Wild Life Federation (Hamilton, n.d.).

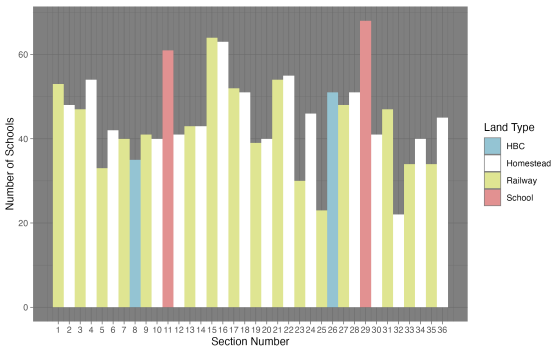
Sections 11 and 29 were explicitly reserved to fund school boards. Unlike the United States (Schaede, 2021) these "school sections" were not sold or rented to fund local school boards. Instead revenue from sales and rentals was collection by the federal government and placed in the School Endowment Fund. These funds were invested in Dominion securities, with the revenues distributed to the provinces.

As with the HBC land the government held off on selling land until the surrounding land was settled to maximize the value. There was a bias towards leasing land rather than selling, to maintain a steady flow of income.

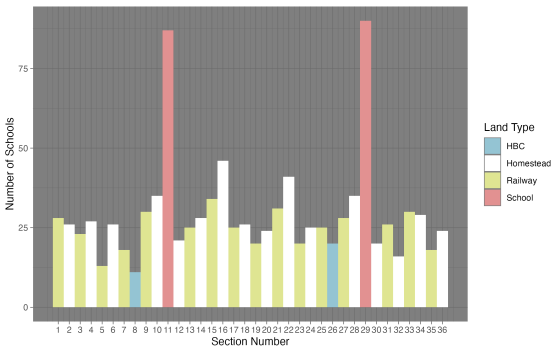
There is some evidence that schools were more likely to locate on school sections relative to other parcels. This appears largely true in Saskatchewan, less so in Alberta, and not at all in Manitoba. This could reflect the lower transaction costs to construct schools on land explicitly designated for schools. Figure 13 shows the distribution of schools across sections.

Figure 8 shows the distribution of population by section in 1906 and 1921, two of the census years where the section number is available. School and HBC sections were the least populated in both years, consistent with those sections being sold off strategically once the surrounding areas had been settled. The highest population sections are the homesteads,

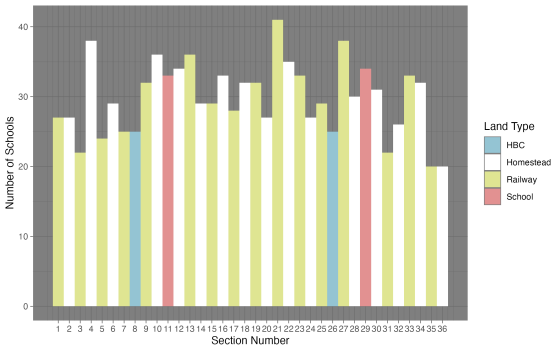
[5] $\frac{1}{36} + \frac{1}{36} * \frac{1}{5} + \frac{3/4}{36} * \frac{4}{5} = 1/20$



(a) Alberta



(b) Saskatchewan



(c) Manitoba

Figure 7: Distribution of schools by section type by province.

although this gap narrows in 1921. This is consistent with railway lands being slower to settle due to the indemnity provision in railway contracts.

0.3 Census Records

The Canadian full count census records are available courtesy of The Canadian People's project (The Canadian Peoples, 2023) and a partnership with Ancestry.com. The records used in this project consist of waves starting in 1901 and ending 1931, the most recent wave released[6]. I utilize both census waves that cover the entirety of Canada in 1901, 1911, 1921, and 1931, and census waves that cover the prairie provinces only in 1906, 1916, and 1926. These additional waves, taken to track the growth of the west after the creation of Alberta and Saskatchewan in 1905 (Canada, 2023a), mean that the relevant areas have census waves every five years.

Although the census was being conducted frequently, different years of the census contain different information. The availability of different variables is recorded in table 11. One important advantage that the Canadian census has is that income is available every ten years starting in 1901, nearly forty years prior to the United States first asking about income in 1940.

0.4 Linking

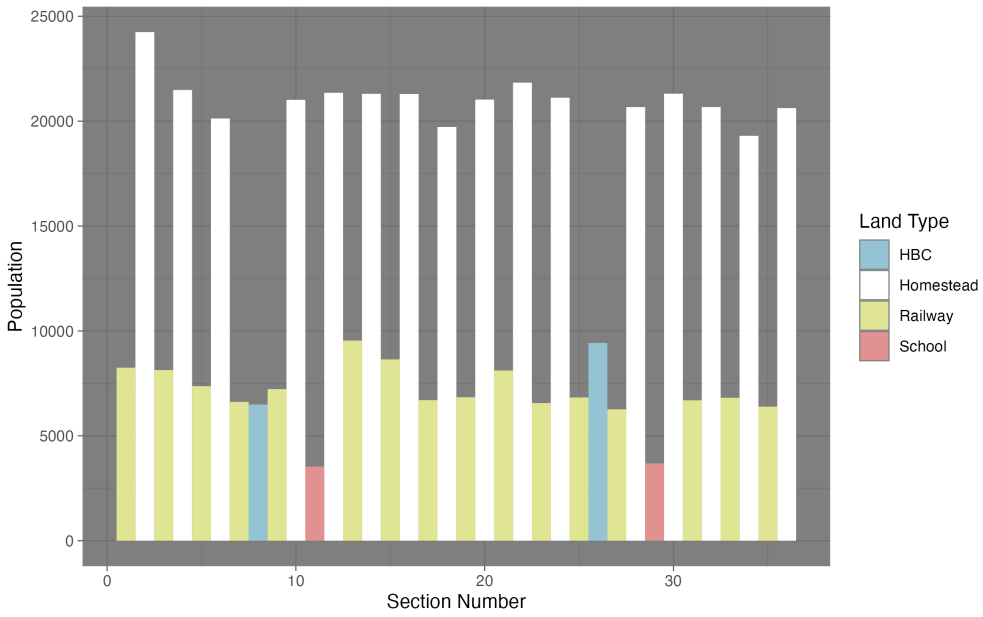
0.4.1 Linking Variables

In order to score potential matches I use the training data from Helgertz et al. (2022) which links US census records. The variables used in stage 1 of linking between individuals can be found in table 12. The names of variables created in the Canadian data are identical to those in the training data for easy reproducibility.

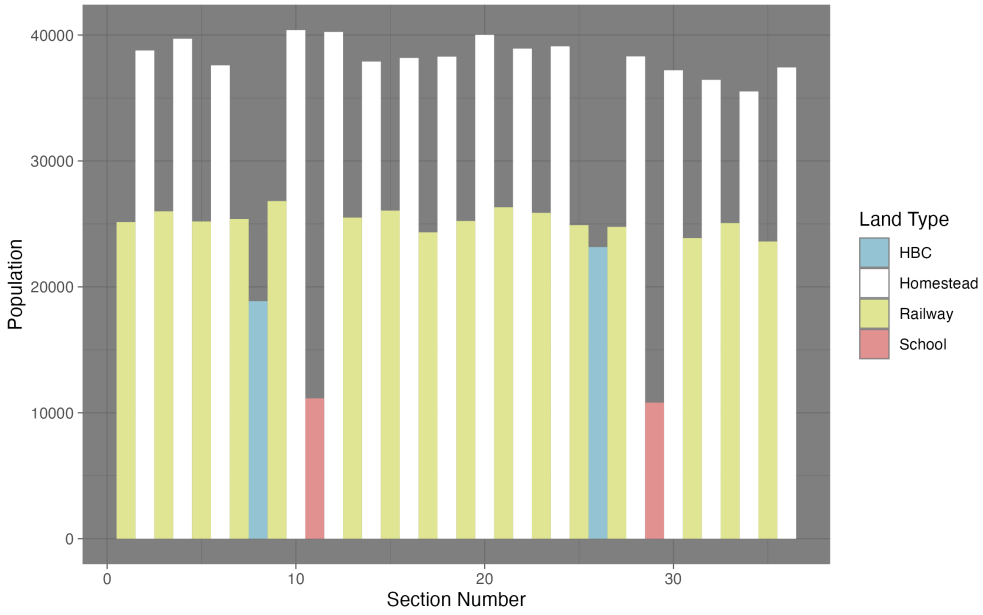
To compute the distances between census records I utilize the link between the census records from The Canadian Peoples, 2023 and the census subdivision (CSD) files computed by HGIS Lab, 2023. I take the centroid of each CSD for the purpose of computing the distance, and compute pairwise distances between all CSDs as a lookup table. CSD polygons are only available in complete census years (e.g. 1901, 1911, 1921, etc.), and not for years where a census of the prairie provinces was taken (e.g. 1906, 1916, 1926, etc.). For years where I do not observe the CSD, I create a synthetic CSD based on the exact locations I can observe combined with the enumeration subdistricts or enumeration districts. For details on geolocation, see section 0.6 of the appendix.

The variables used in stage 2 of linking can be found in table 13. Note that the number of variables and type is lower than in stage 1. This reflects the fact that the stage 2 matching occurs within households based on the stage 1 link, thus requiring fewer variables to reach a similar level of precision. Unlike stage 1 no blocking is required, instead variables are included for sex, age, name initials, and birthplace codes that were previously used for blocking. This allows second stage matches between entries that might have differences across these variables. For example and individual whose sex was misrecorded in one year would not be matched in stage 1 due to the blocking on sex, but could be matched in stage 2

[6]The Canadian government operates on a 92 year lag for releasing census records (Canada, 2023b)



(a) 1906



(b) 1921

Figure 8: Population distribution between land parcel types in 1906 and 1921. Early in the settlment process the majority of settlers lived on even numbered sections which were sold as homesteads. Over time more settlement occurred on railway, school, and Hudson's Bay Company lands, as evidenced by the figure from 1921.

Table 11: Variables available in the Canadian census by census wave. Rural locations include the Section, Township, Range, and Meridian allowing for precise geolocation of households. 1911 excludes this information so location to the nearest township is the best possible. Some years include the exact address for urban households, although 1916 only includes the street. Income is consistently available in full census years, but missing in prairie province census years.

Year	**Name**	**Months in School**	**Location (Rural)**	**Location (Urban)**	**Income**	**Other**
1901	Yes	Yes	Not Digitized	Not Digitized	Yes	Locations contained in a separate schedule, which has not been digitized.
1906	Yes	No	Yes	No	No	Contains information on livestock holdings by each individual.
1911	Yes	Yes	No	No	Yes	Contains information on insurance holdings by individual.
1916	Yes	No	Township	Street	No	Rural addresses at the township level only. Contains information on insurance holdings by individual.
1921	Yes	Yes	Yes	Yes	Yes	Contains information on house materials.
1926	Yes	Not Digitized	Not Digitized	Not Digitized	No	
1931	Yes	Yes	Yes	Yes	Yes	Contains estimated value of home.

Table 12: Linking variables in Stage 1. Code refers to the name of the variable in both the Canadian data and Helgertz et al. (2022). Type gives the data type of the variable. Description offers a brief description of the variable and how it is created.

Variable	**Code**	**Type**	**Description**
Last Name JW	namelast_jw	0-1	Jaro-Winkler string distance between last names.
First Name JW	namefrst_jw	0-1	Jaro-Winkler string distance between first names.
Second Generation	sgen	Binary	Is the individual a second generation immigrant?
County Distance	distance	Continuous (m)	Distance between counties in each census wave, in meters.
Distance Squared	distance2	Continuous (m^2)	County Distance squared.
Birth Year	byrdif	Continuous	Difference between recorded birth years.
Middle Initial	mi	0, 1, or 2	0 if middle initial is empty, 1 if middle initial matches any possible middle initial, 2 if there is no possible match.
First Soundex	fsoundex	Boolean	1 if soundex matches, 0 if otherwise.
Last Soundex	lsoundex	Boolean	1 if soundex matches, 0 if otherwise.
Multiple Exact Matches	exact_mult	Boolean	1 if there are multiple exact matches with identical first and last name, 0 if otherwise
Immigration Year	immyear_caution	0, 1, 2	0 if there is no difference in immigration year or <= 5 years, 1 if the difference is <= 10 years, 2 if the difference is > 10 years. 0 if either record is missing the immigration year
Hits	hits	Integer	The number of potential matches for a given individual, looking from the first dataset
Hits Squared	hits2	Integer	Hits squared.

Table 13: Linking variables in Stage 2. Code refers to the name of the variable in both the Canadian data and Helgertz et al. (2022). Type gives the data type of the variable. Description offers a brief description of the variable and how it is created.

Variable	Code	Type	Description
Last Name JW	namelast_jw	0-1	Jaro-Winkler string distance between last names.
First Name JW	namefrst_jw	0-1	Jaro-Winkler string distance between first names.
Sex	sexmatch	Boolean	Indicates if two records have matching sex.
Relation Type	idtype	1, 2, or 3	1 if both entries are biologically related family, 2 if both entries are non biologically related famiy, 3 otherwise.
Birthplace	bplmatch	Boolean	1 if birthplace code matches, 0 if otherwise.
Birth Year	byrdifcat	Continuous	Difference between recorded birth years. No restriction on maximum possible difference like in stage 1.
First Initial Match	f1_match	1 or 2	1 if the first initial of the first first name matches any of the potential first initials, 2 otherwise.
Middle Initial Match	f2_match	0, 1, or 2	0 if middle initial is empty, 1 if middle initial matches any possible middle initial, 2 if there is no possible match. Analogous to middle initial match in stage 1.

if another family member was matched and enough other characteristics are similar between the two census waves.

For details on the Hlink package used the implement the linking, please see the Github page for Hlink (Wellington et al., Apr. 27, 2022/2024).

0.4.2 Linking Comparison

The linking method used in this paper from Helgertz et al. (2022) (MLP), is a relative newcomer to the linking literature. The more established method of Abramitzky et al. (2021) (ABE) offers greater transparency in the linking process, at the cost of excluding some potentially useful information that might allow for larger match rates and more precise matches.

To compare the performance of these two methods I focus on matches between 1906 and 1911 and compute matches using both the ABE and MLP methods. Table 14 shows the number of matches using both the MLP and ABE methods for the target sample of children. Both methods have a significant overlap in the matched population of 41359, but

Table 14: Links by ABE vs MLP between the 1906 and 1911 Canadian Census. Links are generated for children who are between the ages of 0 and 17 in 1906 and who live in Alberta, Saskatchewan, or Manitoba. MLP links are generated using the procedure document in section 0.4.1. ABE links are generated by first blocking on sex, place of birth, and fathers place of birth. No NYSIIS name cleaning is applied, and the loosest requirements for an exact match are used, requiring uniqueness only on the exact age to be considered a match.

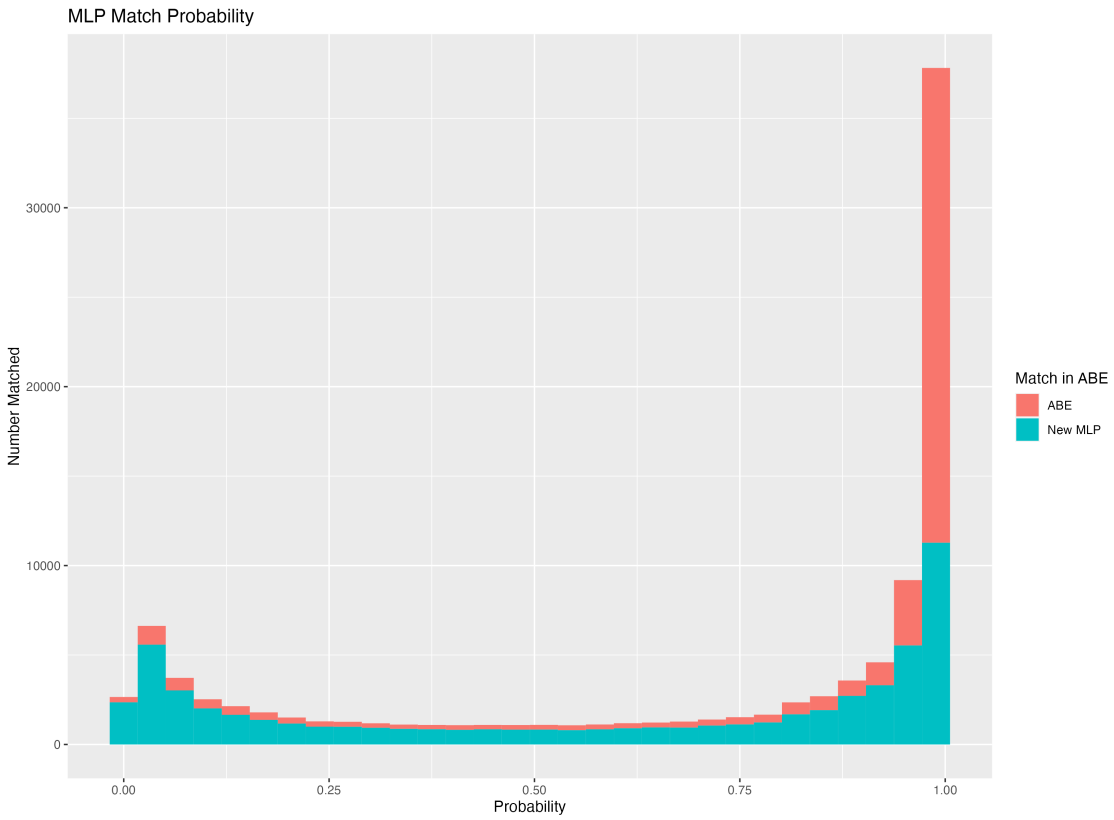
Method	In MLP	Not in MLP	Total
In ABE	32867	23094	55961
Not in ABE	47303	197091	244394
Total	80170	220185	300355

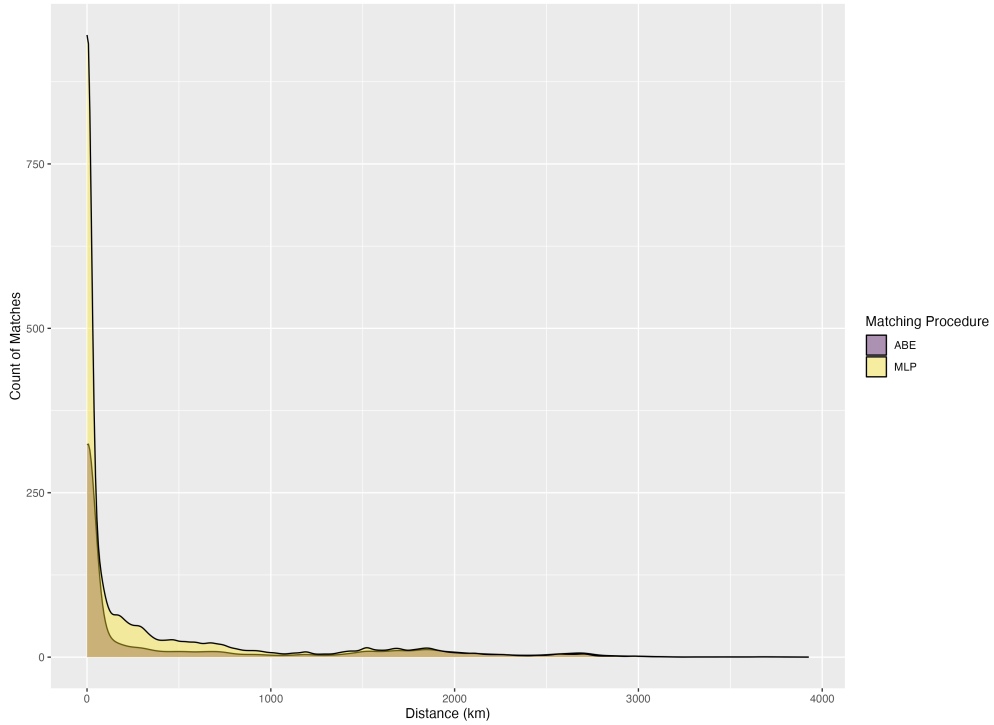
MLP recognizes a much larger set off additional links of 64664, relative to ABE at 14602. This is consistent with MLP using more information to disambiguate links, leading to a higher match rate.

To understand where additional matches are coming from, I plot the likelihood scores for each match in figure 9, and decompose the matches into whether they are an ABE match or are a new MLP match. These likelihoods can only be computed for matches that are completed by MLP, thus the matches are the equivalent of those in column 1 of table 14. I find that the ABE matches have a large mass near 1, with a tail that extends all the way to zero. This suggests that ABE is picking up the highest quality matches, but that by failing to take into account additional information some very low quality links are created. MLP in contrast picks up the high quality matches that ABE does, but also many others whose likelihood is lower. Importantly by creating an explicit score, the researcher is able to select only those matches which have a high probability of being a correct match. Note that this method does not score any ABE matches which are not captured at all by MLP, the "In ABE" and "Not in MLP" group in table 14. If these matches were to be included, there would be a heavier tail of low probability matches attributed to ABE. The results suggest that MLP is able to tell apart matches that may be present, but are not as obvious.

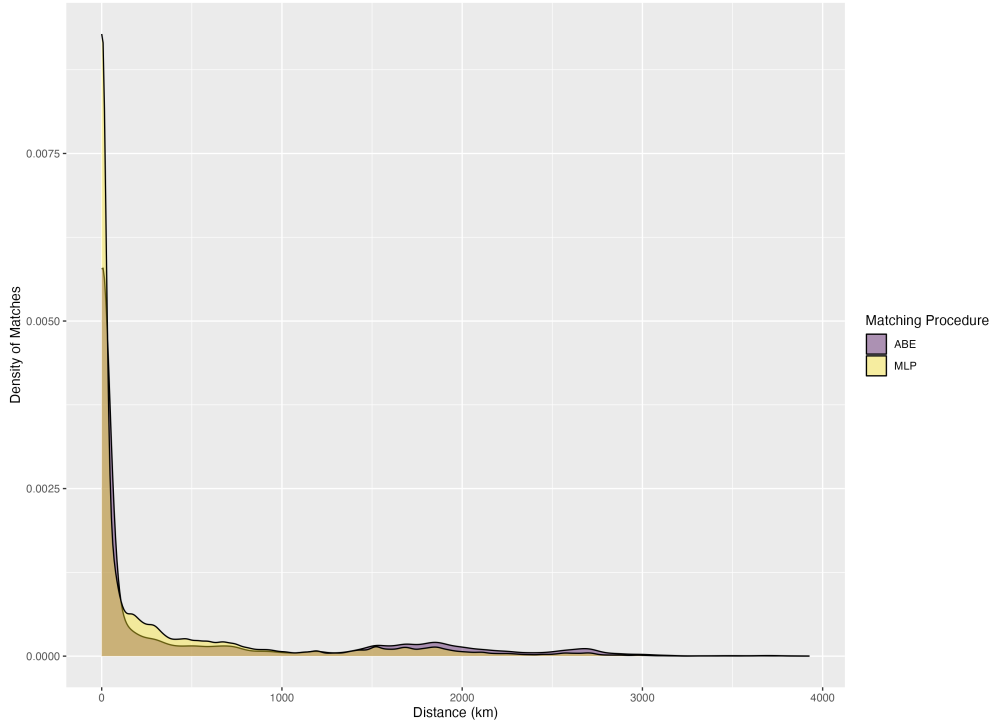
The ability to distinguish between two otherwise identical records in MLP comes at a cost of potentially introducing bias in the matching procedure. One of the most likely sources of bias is the inclusion of distance as a linking variable. If the locations of the individual between two census waves are closer, then a higher probability of a link being true will be predicted. The consequence of this is that any predicted mobility estimates will be biased towards zero, as the match probability decreases with distance. To get a sense of the magnitude of this problem, we can compare the density distribution between matches made by ABE versus those made by MLP in figure 9. I find that MLP dramatically increases the number of matches for individuals in close proximity, but that there are gains in the count of matches across the entire distribution of distances. The distribution of counts can be seen in figure 10a. When examining the density in figure 10b I find that the densities are remarkably similar, although there is slightly more mass at lower distances. Taken together this suggests that although including distance as a linking variable may introduce some bias, it is unlikely to dramatically change any results.

Figure 9: MLP Match Probability vs ABE Links. This figure includes all links that are captured by MLP. Links that are not captured by MLP do not have a score, and cannot be included.





(a) Count of matches.



(b) Density of matches.

Figure 10: Distribution of distance between matches in 1906-1911, for children in Alberta, Saskatchewan, and Manitoba who are between the ages of 0 and 17.

0.4.3 Linking Analytics

All linking proceedures are biased and are likely to be unrepresentative of the population that we want to link (Helgertz et al., 2022). It is therefore important to understand which populations are likely to be linked and if this may bias any of the results. To do this, I estimate the likelihood of being linked conditional on being in the first of a pair of census waves. This choice is made so that I can utilize the 1906 sample as the first wave, which only include the prairie provinces, at the cost of being an understatement of the likelihood of being linked as some people will have emmigrated or died and will be impossible to link.

To understand why I take the the likelihood of being matched conditional on appearing in the first census wave, consider the opposite. To evaluate the effect of the match between 1906 and 1911, we would take the later wave, and remove anyone born after 1906 and who immigrated after 1906. The fraction of people remaining who are linked would be the estimated linkage rate. The 1911 census contains the entirety of Canada, so the total match rate would be too low, as only those in 1906 who were in Alberta, Saskatchewan, and Manitoba would be linked, relative to the entire population of Canada in 1911. This would not be an accurate measure of the likelihood of being linked. An alternate approach would be to restrict the sample to those living in Alberta, Saskatchewan, and Manitoba in 1911 to get the denominator correct, but this would remove the possibility of people moving out of the prairies. Instead I take the earlier year, being concious that some individual will not be matched.

For each pair of census waves I regress the likelihood of being matched of a set of characteristics of the individual including age, birthplace, relation to the head of household, enumeration subdistrict, religion, whether the household lives on a homestead, family holdings of cows and horses, and agricultural productivity of wheat. All effect of estimated jointly. Tables 15, 16, 17, 18, 19, and 20 give the results.

0.5 Schools

After collecting the original school data from each of the three sources, several cleaning steps are required to obtain the opening date and location of each school.

For the Saskatchewan data there can be multiple locations reported for each school district number. This can reflect different original sources for a school's location, or can reflect if a school district spans a geographic area with the exact location of the school being unknown. Table 21 shows an example of the data. To assign unique identifiers I use the fastLink R package (Enamorado et al., 2023) to deduplicate records and assign unique identifiers.

To remove duplicates I take the most precise location as the location of the school. In the case where two schools have the same level of precision (i.e. both have locations at the section level), I keep both and assign them separate identifiers. This ensures that I am likely to be over counting schools rather than under counting. I keep a single entry for any schools that don't have a location, to keep the total number of schools as accurate as possible.

Many entries in the Saskatchewan data are missing the school opening dates. In cases where the opening date is missing I infer it from the school district number. Within the Northwest Territories (later split to form Alberta and Saskatchewan in 1905) school districts were numbered sequentially. As a result if the school district numbers are known for each

Table 15: Linking rates by Age

Dependent Variable:	Matched				
Model:	(1)	(2)	(3)	(4)	(5)
	1906-1911	1906-1921	1906-1931	1911-1921	1911-1931
Age = 1	0.0730***	0.0786***	0.0673***	0.0686***	0.0367***
Age = 2	0.0074*	0.0155***	0.0583***	-0.0178***	0.0132***
Age = 3	-0.0355***	0.0100***	0.0521***	-0.0302***	0.0088***
Age = 4	-0.0083**	0.0209***	0.0443***	0.0052***	0.0201***
Age = 6	-0.0268***	-0.0569***	-0.0872***	-0.0325***	-0.0582***
Age = 7	-0.0458***	-0.1299***	-0.1473***	-0.0806***	-0.1143***
Age = 8	-0.0642***	-0.1663***	-0.1502***	-0.1197***	-0.1272***
Age = 9	-0.1267***	-0.1751***	-0.1359***	-0.1434***	-0.1211***
Age = 10	-0.1999***	-0.1869***	-0.1474***	-0.1499***	-0.1134***
Age = 11				-0.1073***	-0.0706***
Age = 12				-0.1558***	-0.1033***
Sex	0.1307***	0.1465***	0.1467***	0.1197***	0.1084***
Family Size	0.0041***	0.0010***	0.0001	-5.8×10^{-5}	-7.95×10^{-5}***
Birthplace	Yes	Yes	Yes	Yes	Yes
Relation	Yes	Yes	Yes	Yes	Yes
Enum. Subdist.	Yes	Yes	Yes	Yes	Yes
Homestead	Yes	Yes	Yes		
Milk Cows	Yes	Yes	Yes		
Horses	Yes	Yes	Yes		
Wheat Prod.	Yes	Yes	Yes		
Religion				Yes	Yes
Observations	430,101	430,101	430,101	1,191,247	1,191,247
R^2	0.08367	0.08690	0.10244	0.05913	0.06569

Clustered (Enum. Subdist.) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Linking rates by age for different decades of life. The reference group is those in the 5th age group, those between the ages of 40-49. Each column gives the estimated effect for different pairs of matched years. Columns (1), (2) and (3) focus on the rural population who are able to be geolocated and are the population of interest. Columns (4) and (5) are census waves which include all of Canada, but are restricted to people in Alberta, Manitoba, and Saskatchewan.

Table 16: Linking rates by Birthplace

Dependent Variable:	Matched				
Model:	(1)	(2)	(3)	(4)	(5)
	1906-1911	1906-1921	1906-1931	1911-1921	1911-1931
Sex	0.1307***	0.1465***	0.1467***	0.1197***	0.1084***
Family Size	0.0041***	0.0010***	0.0001	-5.8×10^{-5}	-7.95×10^{-5}***
Birthplace = UnitedStates	-0.0347***	-0.0533***	-0.0462***	-0.0260***	-0.0283***
Birthplace = Canada+BC	-0.0491***	-0.0425***	-0.0391***	-0.0131***	-0.0101***
Birthplace = MB	-0.0378***	-0.0389***	-0.0417***	0.0113***	0.0062***
Birthplace = Maritimes	-0.0322***	-0.0213***	-0.0155**	-0.0111***	-0.0132***
Birthplace = QC	-0.0563***	-0.0268***	-0.0199***	0.0067	0.0032
Birthplace = Caribbean	-0.1479***	-0.0886***	-0.1132***	-0.0441**	-0.0571***
Birthplace = Nordic	-0.0760***	-0.0547***	-0.0334***	-0.0327***	-0.0133***
Birthplace = UnitedKingdom	0.0319***	0.0322***	0.0358***	0.0003	0.0090***
Birthplace = Ireland	-0.0157*	-0.0178**	-0.0030	-0.0009	0.0065*
Birthplace = RestofEurope	-0.1094***	-0.0683***	-0.0469***	-0.0306***	-0.0115***
Birthplace = ROW	-0.0722***	-0.1175***	-0.0977***	-0.0556***	-0.0498***
Age	Yes	Yes	Yes	Yes	Yes
Relation	Yes	Yes	Yes	Yes	Yes
Enum. Subdist.	Yes	Yes	Yes	Yes	Yes
Homestead	Yes	Yes	Yes		
Milk Cows	Yes	Yes	Yes		
Horses	Yes	Yes	Yes		
Wheat Prod.	Yes	Yes	Yes		
Religion				Yes	Yes
Observations	430,101	430,101	430,101	1,191,247	1,191,247
R^2	0.08367	0.08690	0.10244	0.05913	0.06569

Clustered (Enum. Subdist.) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Linking rates for different birthplaces. The reference group is those born in Ontario. Each column gives the estimated effect for different pairs of matched years. Columns (1), (2) and (3) focus on the rural population who are able to be geolocated and are the population of interest. Columns (4) and (5) are census waves which include all of Canada, but are restricted to people in Alberta, Manitoba, and Saskatchewan.

Table 17: Linking rates by Land Type

Dependent Variable:	Matched				
Model:	(1)	(2)	(3)	(4)	(5)
	1906-1911	1906-1921	1906-1931	1911-1921	1911-1931
Sex	0.1307***	0.1465***	0.1467***	0.1197***	0.1084***
Family Size	0.0041***	0.0010***	0.0001	-5.8×10^{-5}	-7.95×10^{-5}***
Birthplace = UnitedStates	-0.0347***	-0.0533***	-0.0462***	-0.0260***	-0.0283***
Birthplace = Canada+BC	-0.0491***	-0.0425***	-0.0391***	-0.0131***	-0.0101***
Birthplace = MB	-0.0378***	-0.0389***	-0.0417***	0.0113***	0.0062***
Birthplace = Maritimes	-0.0322***	-0.0213***	-0.0155**	-0.0111***	-0.0132***
Birthplace = QC	-0.0563***	-0.0268***	-0.0199***	0.0067	0.0032
Birthplace = Caribbean	-0.1479***	-0.0886***	-0.1132***	-0.0441**	-0.0571***
Birthplace = Nordic	-0.0760***	-0.0547***	-0.0334***	-0.0327***	-0.0133***
Birthplace = UnitedKingdom	0.0319***	0.0322***	0.0358***	0.0003	0.0090***
Birthplace = Ireland	-0.0157*	-0.0178**	-0.0030	-0.0009	0.0065*
Birthplace = RestofEurope	-0.1094***	-0.0683***	-0.0469***	-0.0306***	-0.0115***
Birthplace = ROW	-0.0722***	-0.1175***	-0.0977***	-0.0556***	-0.0498***
Age	Yes	Yes	Yes	Yes	Yes
Relation	Yes	Yes	Yes	Yes	Yes
Enum. Subdist.	Yes	Yes	Yes	Yes	Yes
Homestead	Yes	Yes	Yes		
Milk Cows	Yes	Yes	Yes		
Horses	Yes	Yes	Yes		
Wheat Prod.	Yes	Yes	Yes		
Religion				Yes	Yes
Observations	430,101	430,101	430,101	1,191,247	1,191,247
R^2	0.08367	0.08690	0.10244	0.05913	0.06569

Clustered (Enum. Subdist.) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Linking rates for different land types. The reference group is those living on a homestead section. Each column gives the estimated effect for different pairs of matched years. Columns (1), (2) and (3) focus on the rural population who are able to be geolocated and are the population of interest. Columns (4) and (5) are census waves which include all of Canada, but are restricted to people in Alberta, Manitoba, and Saskatchewan.

Table 18: Linking rates by Livestock Holdings

Dependent Variable:	Matched		
Model:	(1)	(2)	(3)
	1906-1911	1906-1921	1906-1931
Sex	0.1313***	0.1467***	0.1466***
Family Size	0.0041***	0.0011***	9.13×10^{-5}
Milk Cows = 1-4	0.0266***	0.0266***	0.0199***
Milk Cows = 5-9	0.0537***	0.0475***	0.0372***
Milk Cows = 10-49	0.0618***	0.0495***	0.0403***
Milk Cows = 50+	0.0173	0.0080	-0.0086
Horses = 1-4	0.0311***	0.0280***	0.0205***
Horses = 5-9	0.0547***	0.0499***	0.0385***
Horses = 10-49	0.0634***	0.0559***	0.0436***
Horses = 50+	0.0406**	0.0231*	0.0205**
Age	Yes	Yes	Yes
Birthplace	Yes	Yes	Yes
Relation	Yes	Yes	Yes
Province	Yes	Yes	Yes
Homestead	Yes	Yes	Yes
Wheat Prod.	Yes	Yes	Yes
Observations	430,101	430,101	430,101
R^2	0.05674	0.06955	0.08882

Clustered (Enum. Subdist.) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Linking rates by livestock holdings. The reference group is those holding no cows or horses. Each column gives the estimated effect for different pairs of matched years. Columns (1), (2) and (3) focus on the rural population who are able to be geolocated and are the population of interest. Columns (4) and (5) are census waves which include all of Canada, but are restricted to people in Alberta, Manitoba, and Saskatchewan.

Table 19: Linking rates by Relation to Head

Dependent Variable:	Matched				
Model:	(1)	(2)	(3)	(4)	(5)
	1906-1911	1906-1921	1906-1931	1911-1921	1911-1931
Family Size	0.0043***	0.0013***	0.0004	-4.14×10^{-5}	-6.45×10^{-5}***
Relation = Boarder	-0.1487***	-0.1326***	-0.1166***	-0.1034***	-0.0951***
Relation = Daughter	-0.1340***	-0.2163***	-0.2192***	-0.1301***	-0.1627***
Relation = Lodger	-0.1398***	-0.1533***	-0.1187***	-0.1049***	-0.0944***
Relation = Other	-0.1284***	-0.1156***	-0.0941***	-0.0972***	-0.0842***
Relation = Sister	-0.1942***	-0.2063***	-0.1794***	-0.1756***	-0.1637***
Relation = Son	0.0110***	0.0118***	0.0425***	0.0319***	0.0435***
Relation = Wife	-0.0379***	-0.0605***	-0.0819***	-0.0285***	-0.0458***
Age	Yes	Yes	Yes	Yes	Yes
Birthplace	Yes	Yes	Yes	Yes	Yes
Enum. Subdist.	Yes	Yes	Yes	Yes	Yes
Homestead	Yes	Yes	Yes		
Milk Cows	Yes	Yes	Yes		
Horses	Yes	Yes	Yes		
Wheat Prod.	Yes	Yes	Yes		
Religion				Yes	Yes
Observations	430,101	430,101	430,101	1,191,247	1,191,247
R^2	0.08044	0.08238	0.09728	0.05538	0.06190

Clustered (Enum. Subdist.) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Linking rates by relation to the head of household. The reference group is the head of household. Each column gives the estimated effect for different pairs of matched years. Columns (1), (2) and (3) focus on the rural population who are able to be geolocated and are the population of interest. Columns (4) and (5) are census waves which include all of Canada, but are restricted to people in Alberta, Manitoba, and Saskatchewan.

Table 20: Linking rates by Religion

Dependent Variable:	Matched	
Model:	(1)	(2)
	1911-1921	1911-1931
Sex	0.1197***	0.1084***
Family Size	-5.8×10^{-5}	-7.95×10^{-5}***
Religion = Orthodox	-0.0393***	-0.0289***
Religion = Anglican	0.0653***	0.0438***
Religion = Presbyterian	0.0681***	0.0441***
Religion = Methodist	0.0670***	0.0425***
Religion = Unitarian	0.0625***	0.0405***
Religion = Lutheran	0.0402***	0.0292***
Religion = Baptist	0.0630***	0.0425***
Religion = Mennonite	0.0105	0.0008
Religion = Pentecostal	0.0389	0.0661
Religion = Evangelical	0.0583***	0.0583***
Religion = Fundamentalist	0.0321	0.0247
Religion = Adventist	0.0461***	0.0233***
Religion = Mormon	0.0671***	0.0310***
Religion = Holiness	0.0512***	0.0319***
Religion = Spiritualist	0.0476**	0.0152
Religion = Christian	0.0283***	0.0144***
Religion = Muslim	-0.0350	-0.0278
Religion = Jewish	0.0203***	-0.0094*
Religion = Dharmic	0.0365*	0.0319
Religion = Polytheist	-0.0700***	-0.0650***
Religion = Bahai'i	-0.1391***	0.9297***
Religion = Ethical	0.0169	0.0188
Religion = Other	-0.0984***	-0.0716***
Age	Yes	Yes
Birthplace	Yes	Yes
Relation	Yes	Yes
Enum. Subdist.	Yes	Yes
Observations	1,191,247	1,191,247
R^2	0.05913	0.06569

Clustered (Enum. Subdist.) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Linking rates by religion. The reference group is Catholics. Each column gives the estimated effect for different pairs of matched years. Columns (1), (2) and (3) focus on the rural population who are able to be geolocated and are the population of interest. Columns (4) and (5) are census waves which include all of Canada, but are restricted to people in Alberta, Manitoba, and Saskatchewan.

Table 21: An example of assigning unique identifiers in the Saskatchewan data. I use the fastLink package (Enamorado et al., 2023) to deduplicate records, taking into account the school name, the school district number, and the township, range, and meridian. To resolve duplicates I first take the most precise locations with section numbers (or in some cases quarter section numbers), then the less precise locations with township range and meridian locations only.

Unique ID	**School Name**	**School District Number**	**Section**	**Township**	**Range**	**Meridian**
186	Alexander	1029	NA	20	29	W3
186	Alexander Plain	1029	1	20	29	W3
200	Alexandra	1908	NA	27	19	W2
200	Alexandra	1908	35	27	20	W2
200	Alexandria	1908	NA	27	19	W2
200	Alexandria	1908	NA	27	20	W2
200	Alexandria	1908	NA	28	19	W2
200	Alexandria	1908	NA	28	20	W2
200	Alexandria	1908	NA	27	19	W2

school, and school opening dates are known for Alberta, school opening dates can be inferred for Saskatchewan. Table 22 shows an example of how this process works in practice. When a school could have two possible opening dates, I take the earlier of the two.

After the formation of Alberta and Saskatchewan in 1905 school districts in Alberta continued to be numbered following the same order, while Saskatchewan continued with the same numbering scheme until 1911, when the government went back and filled in numbers occupied by Alberta schools (Baergen, 2005; Saskatchewan Department of Education, 1906, 1907, 1908, 1910, 1911, 1912, 1914). I utilize reports from the Saskatchewan department of education to get the year of school opening until the end of 1912.

I utilize the school district numbers to infer the year of opening for any missing year. For Saskatchewan, no information is available on the date of school opening, but Alberta and Saskatchewan shared the same school numbering system until 1905 when the provinces were created (Baergen, 2005). Using the known dates for Alberta, I then create a joint list of school districts, and interpolate between the known Alberta dates to determine an estimate of the opening date for each Saskatchewan school.

One possible concern with this approach is that some schools could be missing in the records for Alberta or Saskatchewan. After Alberta and Saskatchewan were formed Alberta continued with the same numbering scheme (Baergen, 2005), while Saskatchewan went back and filled in new school districts with the numbers previously allocated to schools now in Alberta (Saskatchewan Department of Education, 1906, 1907, 1908, 1910, 1911, 1912, 1914). As a result any schools missing from the Alberta records and present in the Saskatchewan records before the numbers previously used in Alberta were filled up would be falsely attributed to have been constructed earlier than they actually were. To resolve this I manually record which school districts were opened from the Saskatchewan Department of Education annual reports from 1906 through 1912, which record which school district numbers were

Table 22: Inference procedure for schools with missing establishment years. If an entry for Saskatchewan has a corresponding entry in Alberta, then it remains NA, as the school number was originally allocated to a school in Alberta when both provinces were part of the Northwest Territories. If there is no corresponding Alberta school then the date of opening can be inferred from the earlier of the two dates that surround it. Instead I turn to the annual reports of the Saskatchewan Department of Education for the overlapped values, giving a year of 1911 for Clunie.

Province	School District Name	School District Number	Year (Data)	Year (Inferred)	Year (Archive)
Alberta	Anderson	434	1896	1896	1896
Saskatchewan	Clunie	434	NA	NA	1911
Saskatchewan	New Finland	435	NA	1896	1896
Saskatchewan	Glenmorris	436	NA	1896	1896
Alberta	Bon Accord	438	1897	1897	1897

reassigned (Saskatchewan Department of Education, 1906, 1907, 1908, 1910, 1911, 1912, 1914).

For Manitoba I scrape Manitoba Historical Society, 2019. Each school has a single page describing the location of the school and some brief history. To get the opening dates of each school I utilize the ChatGPT API to read the body of each page and return the year of construction of the school. The total school openings in each year from 1870 to 1911 can be seen in figure 11.

There are two caveats for utilizing the school data. The first is that I do not observe closing dates for any schools, and therefore assume that any school that opens remains open until 1911. This is a reasonable assumption as settlement ensured that there would always be demand for schools. The second caveat is that I do not observe if schools moved locations, and in some cases will have a second location in the dataset with way to tell when a move occurred. In these cases I take both locations as school locations, with separate school identifiers.

Table 23 shows the totals split by province of the number of schools in the database and compares this to the number of schools listed in each provinces department of education reports. I find very comparable numbers for Alberta with 1784 schools listed as being organized in 1911, versus 1750 in the data.

0.5.1 Separate Schools

As discussed in section , the question of separate schools for Catholics and Protestants was an important feature of the late nineteenth and early twentieth century prairies. In the data I therefore need to be careful how different school boards are handled.

In Manitoba separate schools were eliminated in 1890 (Noonan et al., 2006), and the data reflect the entire set of possible schools.

In Alberta the data only contain public schools, either Protestant or Catholic, and do not contain any information on separate schools established by the minority faith. Checking the reports from the Department of Education I find that of the 1,784 schools in existence

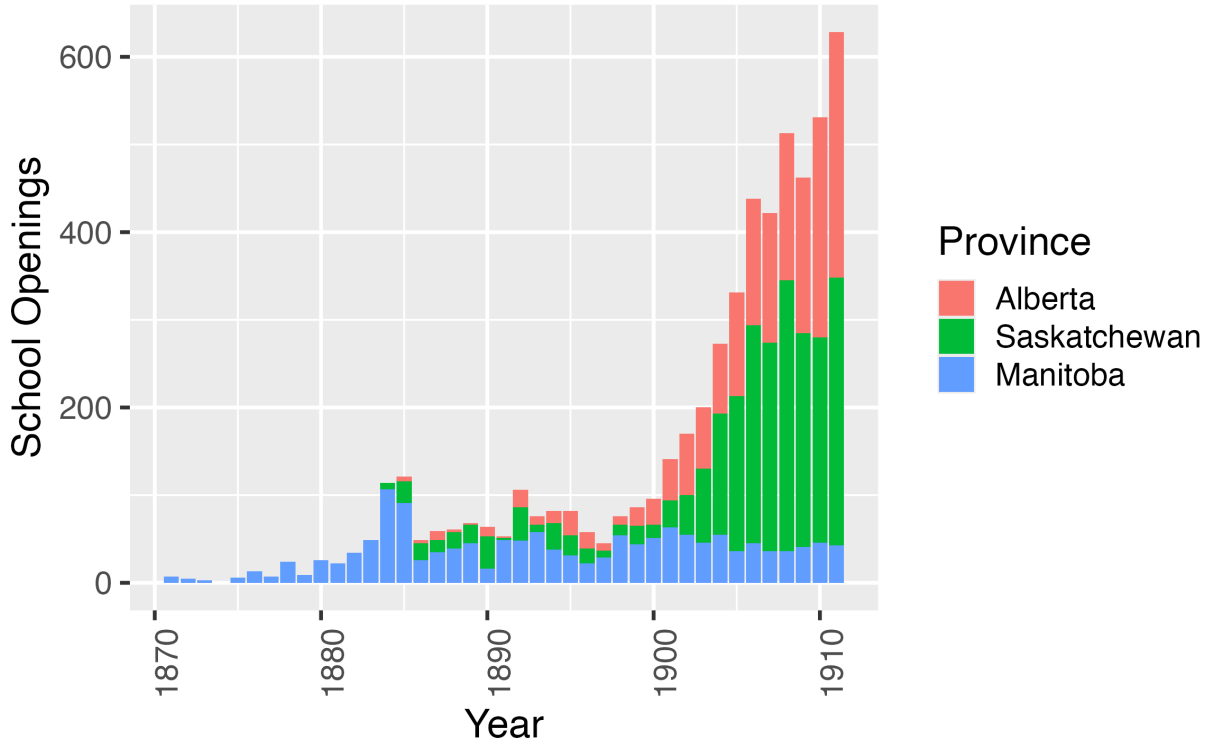


Figure 11: School openings for Alberta, Saskatchewan, and Manitoba from 1870 to 1911. Prior to 1870 only a handful of school existed in Manitoba. The last year that I have data of school openings for is 1912.

Table 23: Summary statistics by province for the number of schools in 1911. The first row lists the number of schools reports in the 1911 report from the department of education for Alberta (Government of Alberta, 1912), Saskatchewan (Saskatchewan Department of Education, 1906, 1907, 1908, 1910, 1911, 1912, 1914), and Manitoba (Government of Manitoba, 1911, 1912). This figure includes both public and separate schools in Alberta and Saskatchewan while the data contain only public schools, so is a slight overestimate. The second row, "Observed" lists the number of schools I find in the data in 1911, which requires that they have an opening date before 1911. Any schools with an unknown opening date are dropped. "Geolocated" reports the number of schools, of those which are observed, which I am able to geolocate. "Geolocated (Section)" reports the number of schools where the geolocation is at the section level (1 mile square) or finer. Details on the geolocation proceedure can be found in section 0.6 of the appendix.

Type	**Total**	**Alberta**	**Saskatchewan**	**Manitoba**
Actual	5955	1784	2573	1598
Observed	5618	1750	2370	1498
Geolocated	5472	1750	2227	1495
Geolocated (Section)	4141	1619	1027	1495

at the end of 1911, only 14 were separate schools and 5 were private schools (Government of Alberta, 1912). In addition 9 of these schools were located in cities or towns, and are likely to be located near other schools as they are separate school districts. It is therefore unlikely that they are affecting my distance calculations.

In Saskatchewan there are 13 of 2,516 total school districts which were separate in 1911 (Saskatchewan Department of Education, 1906, 1907, 1908, 1910, 1911, 1912, 1914). Although I can observe separate schools in the data, the numbering system for separate schools is separate for that of public schools, making it impossible to infer the opening dates from school numbers alone. Although the numbering system is shared with Alberta prior to 1905, I also don't know the school numbers for separate schools in Alberta as described the paragraph above. In addition some of the numbers of separate schools were back filled as with public schools in section 0.5, so that even a small number of opening dates is insufficient to determine the others. As a result archival research is therefore required to determine the opening date for these schools. I therefore drop all separate schools from the Saskatchewan data to be consistent with Alberta.

0.6 Geolocation

After merging shapefiles for parcels in Alberta (Government of Alberta, 2020), Saskatchewan (Information Services Corporation (ISC), 2020), and Manitoba (Manitoba Government, 2020), I drop parcels which do not conform to the Section-Township-Range-Meridian system. This consists primarily of river lots located in Manitoba, near Winnipeg, with 13,258 dropped out of a total of 196,864 in Manitoba. For each parcel I take the centroid as the location of the parcel. I create three different levels of aggregation at the quarter section, section, and township level.

0.6.1 Geolocating Households

To geolocate households I first split households into rural and urban households. Rural households are located via the section, township, range, and meridian, while urban households are located using CSD polygons from HGIS Lab (2023). In most cases the rural geolocation is more precise than the urban geolocation, as each section is only a single mile square.

To locate rural households in 1906 and 1921 I first match each household using the section, township, range and meridian to the known coordinates for that section. This gives an initial set of geolocations for each household in rural areas.

One unique feature of census records is that a single enumerator would have to travel between households in order enumerate each household. In addition the instructions to enumerators consistently advise across census years to start at one end of a district, then to move systematically to the other end[7]. As a result the path of an enumerator is informative

[7]For example the enumerators instructions in 1906 advise *In a township, parish or other country district, where the houses are scattered, it is advisable to start on a road or highway at the border line of the district and visit in succession every occupied house until the other side of the district is reached, when the next road may be taken in the same way, and so on until the whole area assigned to the enumerator is covered, taking care to finish the census of one farm or lot before proceeding to the next.* (Government of Canada, 1906)

if the geolocation of a household is correct, and significant deviations from a regular path are likely to be errors in geolocation.

To utilize this feature I first take the raw locations in 1906, and group them by enumeration subdistrict. Any observations which are greater than four standard deviations away from the mean longitude and latitude of the enumeration subdistrict are removed. Figure 12 shows an example of this process. In 1921, where I have CSD polygons, I remove any geolocations which are greater than 3 kilometers from the edge of their CSD polygon. The remaining addresses I consider to be the exact location, as they correspond to a single section of land.

For any addresses that remain unlocated I do a second round of locating to the nearest township, then a third round of locating to the nearest city. In 1921 there a several small towns which have an area close to that of a section. If the area of the CSD is less than 1.5 square miles (relative to a section which has an area of 1 square mile), then I consider the location to also be exact.

0.6.2 Creating Synthetic CSDs

In complete census years (1901, 1911, and 1921) CSD boundaries and centroids have been digitized by the HGIS lab at the University of Saskatchewan (HGIS Lab, 2023). For census of the prairie provinces and the most recent complete census year (1906, 1916, 1926, and 1931) CSDs have not been digitized. To overcome this challenge I create synthetic CSDs using the enumeration subdistricts and the location of known households.

I first take the known locations from geolocated households, and then compute the centroid of those locations. I use this centroid if there are at least (number of households) located in that enumeration subdistrict. This gives the centroid of the synthetic CSD when the CSD is unknown for (A number) of enumeration districts out of a total of (Another number). For the remaining (A number) enumeration subdistricts I use the centroid of the enumeration district as the centroid of the enumeration subdistrict. This leaves only two enumeration subdistricts in northern Alberta and Saskatchewan which I geocode to an approximate location.

0.6.3 Geolocating Schools

I implement two proceedures to geolocate schools.

For Alberta and Saskatchewan, I know the legal land description for each school, and use this to get the parcel of land that corresponds to that description. I start with the finest possible geography, the Quarter Section, and try to match each school. The remaining schools I attempt to match at the section level, and any remaining after that I match at the township level. As shown in table 21 there are cases where a school only has location information at the township level, will multiple possible townships reflecting the entire area of the school division. To resolve this I take the centroid of a group of townships as the location of the school.

For Manitoba, where I already have longitude and latitude, I use the existing coordinates to retrieve the legal land description of the location of the school. I include this information in my school database for possible use in later regressions.

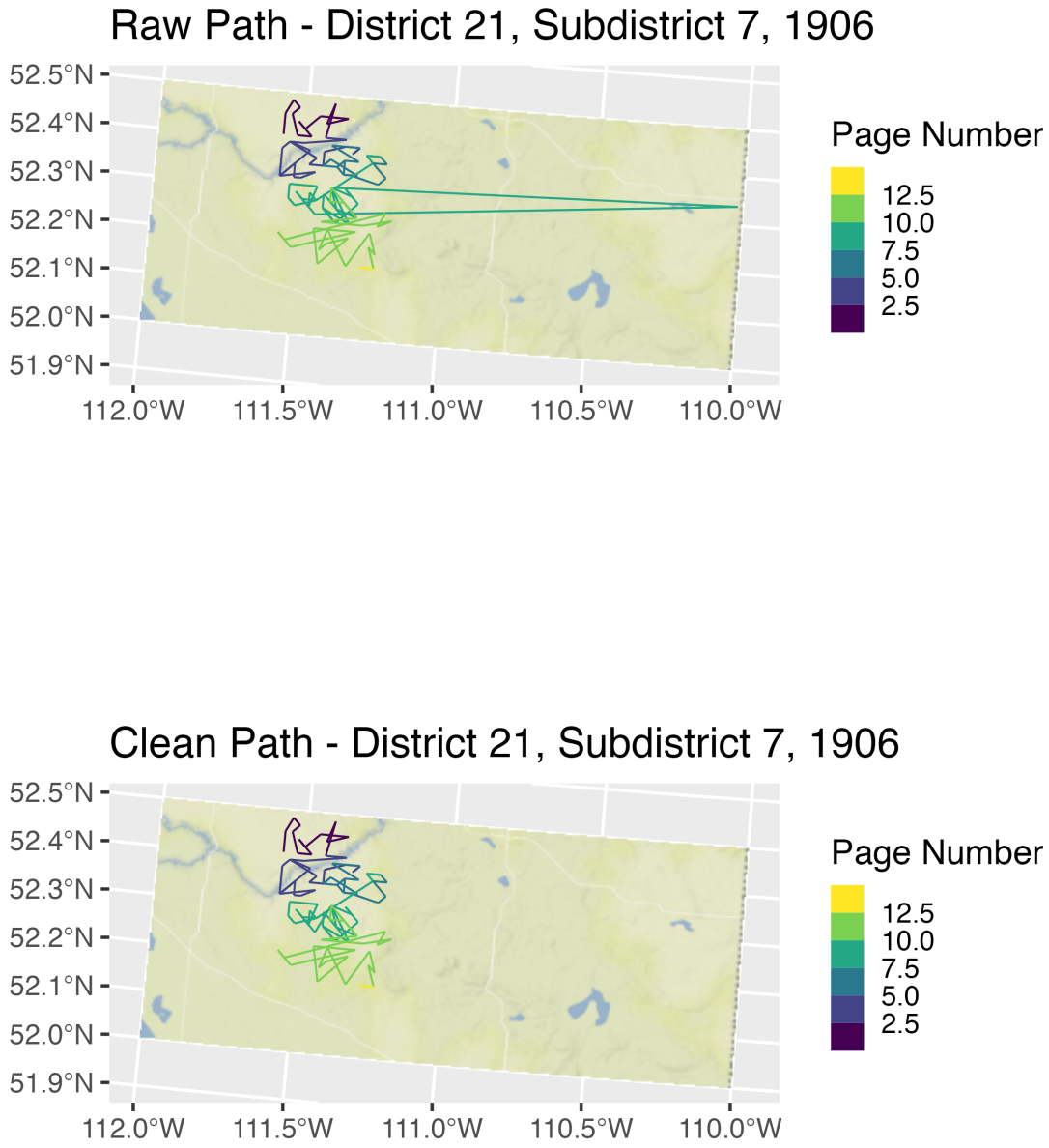
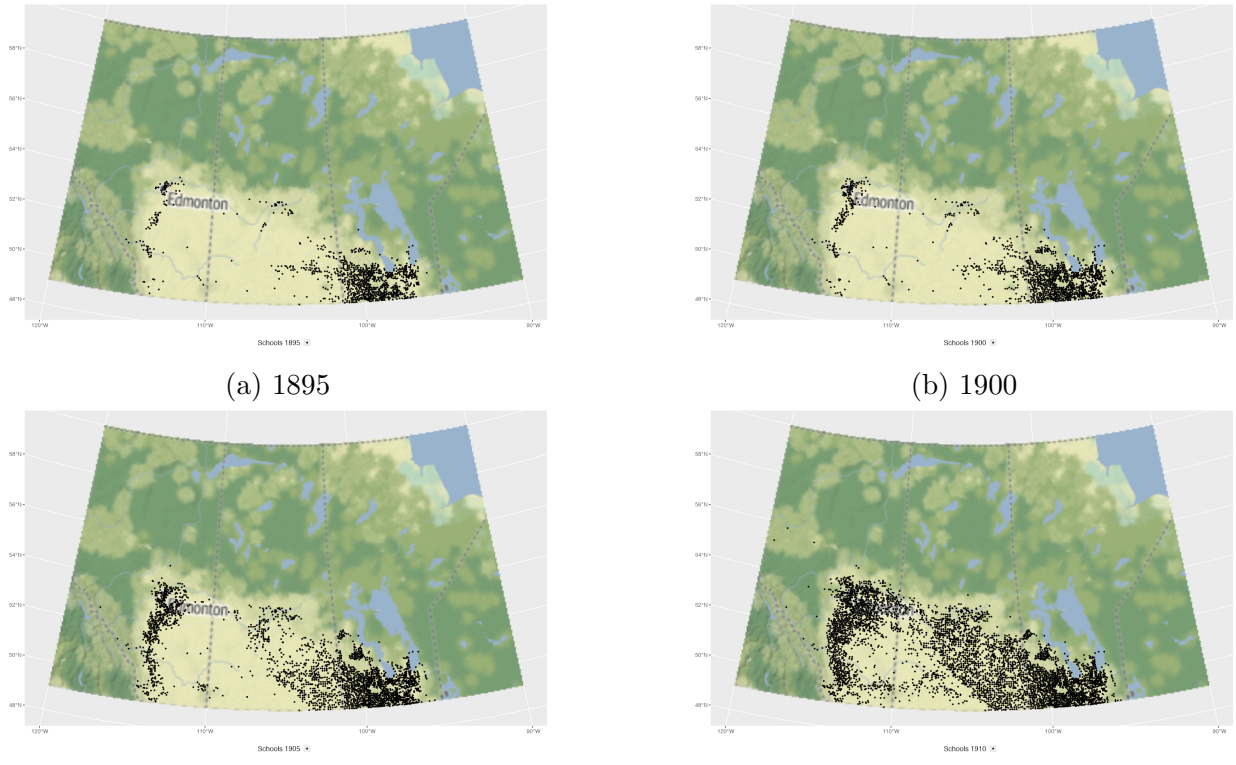


Figure 12: Enumeration path for a single enumeration subdistrict, before and after the cleaning proceedure is applied. Any location which is further than four standard deviations in either longitude or latitude from the centre of the enumeration subdistrict is removed.

The bottom rows of table 23 show the number of schools I am able to geolocate in 1911. For both Alberta and Manitoba I am able to geolocate nearly all schools reported, while in Saskatchewan I am only able to geolocate 93% of schools. When I focus on the schools I am able to geolocate most precisely, at the section level, I find I'm able to geolocate 93% of schools in Alberta, 99% in Manitoba, but only 43% in Saskatchewan. As a result I show that the results are robust regardless of the school locations used in Saskatchewan.



(a) 1895

(b) 1900

(c) 1905

(d) 1910

Figure 13: Expansion of schools in western Canada from 1895 to 1910.

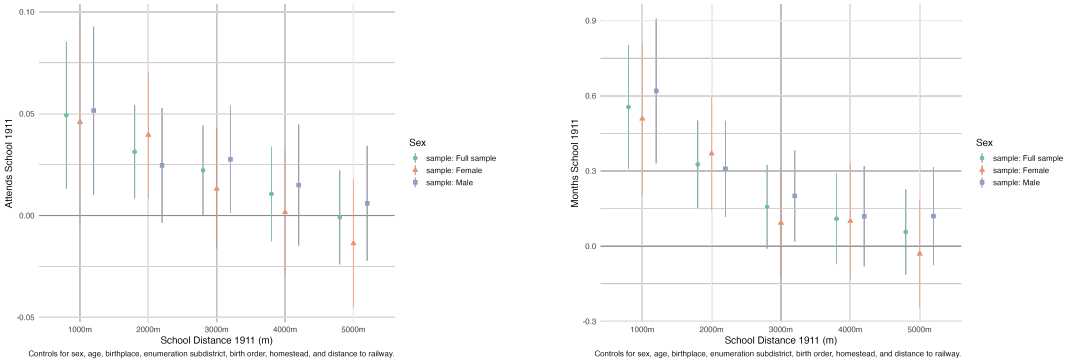
Figure 13 shows the expansion of school across western Canada from 1895 to 1910. Although schools were initially formed around Winnipeg in Manitoba, westward expansion led settlers to move further west, and construct schools. As a result we see the expansion of schools in Saskatchewan and Alberta (also seen in figure 11).

0.7 Additional Results

0.7.1 Intensive and Extensive Margin and Distance

Figure 15a gives the estimation results repeating the estimation in figure 4, but examining the extensive margin of school attendance only. Figure **??** also repeats the same estimation strategy, but dropping observations with zero attendance to examine intensive margin effects of school attendance. In both cases there is a clear decay pattern over distance, although the estimates are less precise than the main effect.

Figure 14: Intensive and Extensive Margin of School Attendance Over Distance



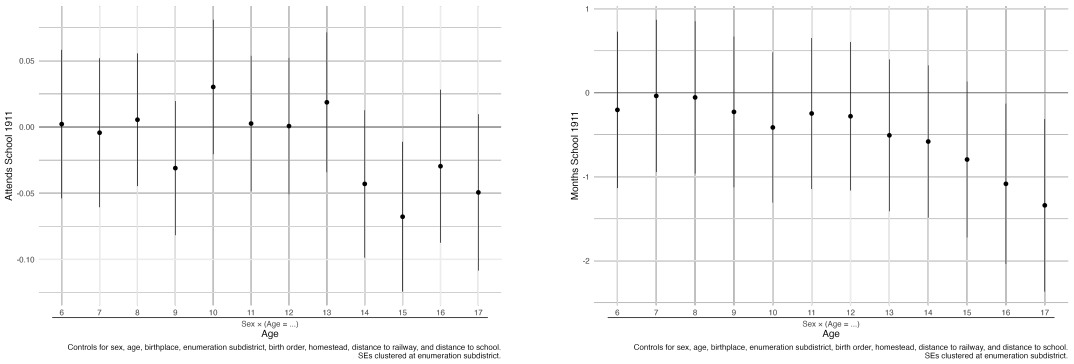
(a) School Attendance (Extensive Margin) (b) Months of Schooling (Intensive Margin)

Extensive and intensive margin effects of distance on schooling in 1911, in 1km bins. All coefficients are expressed relative to the furthest bin of 5km or more. Includes controls for sex, age, birthplace, birth order, homestead, and distance to the nearest railway.

0.7.2 Male vs Female Intensive and Extensive Margin

Difference in Differences Robustness

Figure 15: Intensive and Extensive Margin of School Attendance for Boys vs Girls



(a) School Attendance (Extensive Margin) (b) Months of Schooling (Intensive Margin)

Coefficients of the interaction term of sex and age regressed on whether or not a child attended school in 1911 and the months of school attended in 1911. Controls for sex, age, birth order, birthplace, enumeration subdistrict (geography), and whether the family is living on a homestead. The decline in the coefficients represents females attending more school relative to males.

Table 24: Treated Robustness for Family in Agriculture in 1931

Dependent Variable:	Family in Ag					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	0-5 vs 6-11	0-5 vs 12-17	0-5 vs 18-23	6-11 vs 12-17	6-11 vs 18-23	12-17 vs 18-23
Variables						
School Constr. × Treated Cohort	-0.0334**	-0.0413**	-0.0465***	-0.0205	-0.0096	0.0057
	(0.0157)	(0.0164)	(0.0166)	(0.0171)	(0.0164)	(0.0170)
School Constr.	0.0267	0.0367**	0.0058	0.0526***	0.0071	0.0041
	(0.0174)	(0.0186)	(0.0159)	(0.0176)	(0.0174)	(0.0159)
Treated Cohort	0.0501***	0.0623***	0.0429***	0.0120	-0.0153	-0.0148
	(0.0095)	(0.0100)	(0.0114)	(0.0096)	(0.0104)	(0.0100)
Log RR Distance (1906)	0.0223***	0.0196***	0.0227***	0.0126***	0.0163***	0.0137***
	(0.0044)	(0.0044)	(0.0043)	(0.0043)	(0.0044)	(0.0044)
Match Prob.	0.3957***	0.4291***	0.4162***	0.4170***	0.4128***	0.4393***
	(0.0255)	(0.0252)	(0.0228)	(0.0264)	(0.0238)	(0.0232)
Fixed-effects						
Enumeration Subdistrict (1906)	Yes	Yes	Yes	Yes	Yes	Yes
Fam. Bpl.	Yes	Yes	Yes	Yes	Yes	Yes
Homestead	Yes	Yes	Yes	Yes	Yes	Yes
Livestock	Yes	Yes	Yes	Yes	Yes	Yes
Ag. Productive.	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	15,602	14,747	15,718	14,219	15,190	14,335
R^2	0.09948	0.09896	0.09610	0.10221	0.09838	0.10133

Clustered (Enumeration Subdistrict (1906)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Likelihood of being having at least one family member in agriculture as an adult in 1931. School construction is defined as 1 if the distance to school decreased by more than 5km between 1901 and 1906. The treated cohort is varied between different groups. Railway construction is defined as 1 if the distance to the nearest railway line decreases by more than 5km between 1901 and 1906. Homestead is a control for the type of land parcel a family is on in 1906. Family birthplace are controls for own, father's and mother's birthplace. Livestock controls for family holdings of cattle, hogs, horses, milk cows, and sheep in 1906 as taken from the census. Agricultural conditions are for the agricultural productivity in wheat, oats, barley, flax, and reed canary grass.

Table 25: School Construction Robustness for Family in Agriculture in 1931

Dependent Variable:	Family in Ag		
Model:	(1)	(2)	(3)
	Improvement - Year	Improvement - Age	Cutoff
School Constr.	0.0367**	0.0261	0.0212
	(0.0186)	(0.0192)	(0.0256)
Treated Cohort	0.0623***	0.0574***	0.0596***
	(0.0100)	(0.0094)	(0.0110)
School Constr. × Treated Cohort	-0.0413**	-0.0371*	-0.0410
	(0.0164)	(0.0217)	(0.0284)
Log RR Distance (1906)	0.0196***	0.0198***	0.0190***
	(0.0044)	(0.0044)	(0.0051)
Match Prob.	0.4291***	0.4290***	0.4214***
	(0.0252)	(0.0252)	(0.0295)
Enumeration Subdistrict (1906)	Yes	Yes	Yes
Fam. Bpl. + Homestead	Yes	Yes	Yes
Livestock + Ag.	Yes	Yes	Yes
Observations	14,747	14,747	10,332
R^2	0.09896	0.09875	0.11463

Clustered (Enumeration Subdistrict (1906)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Likelihood of being having at least one family member in agriculture as an adult in 1931. Column 1 reports the baseline specification where school construction is defined as 1 if the distance to school decreased by more than 5km between 1901 and 1906. Column 2 reports an alternate specification where school construction is defined as 1 if the distance to school decreased by at least 5km between the ages of 0 to 5 in the treated cohort, and between the ages of 12 to 17 in the control cohort. Column 3 defines school construction as being 1 if the school started more than 10km away at age 0, and being less than 5km away at age 5 for the treated cohort, and being more than 10km away at age 12 and less than 5km away for the control cohort. The treated cohort is those who are 0-5 in 1906, as compared to those who are 12-17. Railway construction is defined as 1 if the distance to the nearest railway line decreases by more than 5km between 1901 and 1906. Homestead is a control for the type of land parcel a family is on in 1906. Family birthplace are controls for own, father's and mother's birthplace. Livestock controls for family holdings of cattle, hogs, horses, milk cows, and sheep in 1906 as taken from the census. Agricultural conditions are for the agricultural productivity in wheat, oats, barley, flax, and reed canary grass.

Table 26: Distance Robustness for Family in Agriculture in 1931

Dependent Variable:	Family in Ag				
Model:	(1)	(2)	(3)	(4)	(5)
	3km	4km	5km	6km	7km
School Constr. × Treated Cohort	-0.0214	-0.0299*	-0.0413**	-0.0345**	-0.0339**
	(0.0160)	(0.0164)	(0.0164)	(0.0168)	(0.0171)
School Constr.	0.0370**	0.0223	0.0367**	0.0286	0.0256
	(0.0173)	(0.0179)	(0.0186)	(0.0186)	(0.0200)
Treated Cohort	0.0576***	0.0599***	0.0623***	0.0598***	0.0591***
	(0.0103)	(0.0102)	(0.0100)	(0.0099)	(0.0098)
Log RR Distance (1906)	0.0194***	0.0198***	0.0196***	0.0197***	0.0197***
	(0.0044)	(0.0044)	(0.0044)	(0.0044)	(0.0044)
Match Prob.	0.4295***	0.4292***	0.4291***	0.4289***	0.4286***
	(0.0252)	(0.0252)	(0.0252)	(0.0252)	(0.0252)
Enumeration Subdistrict (1906)	Yes	Yes	Yes	Yes	Yes
Fam. Bpl. + Homestead	Yes	Yes	Yes	Yes	Yes
Livestock + Ag.	Yes	Yes	Yes	Yes	Yes
Observations	14,747	14,747	14,747	14,747	14,747
R^2	0.09889	0.09874	0.09896	0.09880	0.09877

Clustered (Enumeration Subdistrict (1906)) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Likelihood of being having at least one family member in agriculture as an adult in 1931. School construction is defined as 1 if the distance to school decreased by more than 5km between 1901 and 1906 in the baseline specification in column 3. All other columns alternate this value. The treated cohort is those who are 0-5 in 1906, as compared to those who are 12-17. Railway construction is defined as 1 if the distance to the nearest railway line decreases by more than 5km between 1901 and 1906. Homestead is a control for the type of land parcel a family is on in 1906. Family birthplace are controls for own, father's and mother's birthplace. Livestock controls for family holdings of cattle, hogs, horses, milk cows, and sheep in 1906 as taken from the census. Agricultural conditions are for the agricultural productivity in wheat, oats, barley, flax, and reed canary grass.