Finnish Municipalities 1880 to 1974:

Creating Harmonized Statistical Units Over a Century

Jonas Mueller Gastell*

August 25, 2020

1 Introduction

Finland, like other Nordic countries, has kept astonishingly detailed records of its citizens' lives over the last centuries. Many of these records were aggregated into well audited and maintained statistical tables immediately after collection. These aggregate records are now easily available to researchers, thanks to the scanning efforts of the National Archives and Statistics Finland (`https://www.doria.fi/`). However, the unit of statistical record keeping, to which individual surveys and administrative entries were aggregated, has been unstable. This instability has meant that research conducted with the aggregated data has been hampered by the need to engage in highly labor intense and *ad hoc* manual adjustments.

The procedure described in this paper is the first systematic approach to provide a harmonized definition of statistical units of Finland. Using detailed records of the size of population movements caused by statistical area changes combined with baseline population counts, I create cross-walks from any year between 1880 and 1973 to any later year between 1881 and 1974. I also provide automated code routines to enable researchers to easily "translate" a record created in a given year into the equivalent record on the basis of a latter year's statistical boundaries. The method developed could also be applied to other countries or time periods, given a sufficiently rich record of unit changes. Eckert et al. provide a related approaches to harmonize US counties based on *area changes*, while the

*This project would not be possible without the large number of dedicated academic contributors, sharing data, time, and expertise. I would in particular like to acknowledge the support from Sakari Saaritsa, who funded (thanks to an Academy of Finland grant) and oversaw the digitization of the original Census and Parish files from 1880 to 1938, in collaboration with Jarmo Peltola, for the urban population data from 1880 to 1938, see `https://doi.org/10.1080/1081602X.2019.1598462`, with Eero Simanainen, who undertook the digitization of the 1914 to 1938 parish data and did the heroic initial aggregation work of these parishes into kunta borders. In addition, I am deeply indebted to Matti Sarvimäki, who funded (thanks to an Academy of Finland grant) the digitization of the original Census tables from 1938 to 1974 and the municipality changes files from 1880 to 1974, see `http://www.aalto-econ.fi/sarvimaki/forced.pdf`, jointly with Maarit Olkkola, who contributed funding (thanks to a grant by the Yrjö Jahnsson Foundation), managed, and oversaw the digitization of the original Census tables from 1938 to 1969, with the help of Satu Koskinen, and Elina Laakso, who undertook the digitization of the population transfers and 1970-1974 original tables. I would further like to acknowledge funding for this project from the Shultz Foundation at SIEPR, the Economic History Association, and the GRO Fund at Stanford. I also want to thank Aalto Department of Economics for their administrative support and generous hospitality. Further, I am deeply grateful to Matti Mitrunen, Katariina Mueller-Gastell, and Joonas Tuhkuri who all supported this project with their advice.

approach taken here is to use actually recorded *population changes* (c.f. Eckert et al., 2020).[1]

The standard unit of aggregation in Finnish statistical publications is the "kunta," an administrative entity that provides many governmental services, such as schools and healthcare. Thus, the borders of this record keeping unit have shifted with the expediency of public administration and the changing currents of political preferences for more localized or more aggregated service provision. In 1880, the number of distinct "kuntas" is 497, reaches a high of 580 in the 1940s, and falls to 481 by 1974. In addition to the "'kunta" mergers and splits that cause these swings we also need to take account of similarly drastic and frequent re-drawings of "kunta"-borders. Figure 1 displays the evolution of "kunta" numbers and the population scale of administrative changes to "kuntas" from 1880 to 1974. As a "kunta" is an entity somewhere in between a county and a municipality, in this paper, I will use municipality for ease of reference going forward.

To understand the translation process proposed, consider a researcher who wants to regress the number of manufacturing plants in 1974 on the number of manufacturing plants in 1880, say. At their disposal are two data-frames. One containing the number of plants in each municipality in 1974, and one with the number of plants in the municipalities as of 1880. At a minimum, the researcher needs to 1) "add" the data of municipalities that no longer exist to the municipalities that incorporated them, and 2) account for where any newly formed municipalities have come from. But how to account for changes to borders of municipalities that exist in both 1880 and 1974?

The approach in this paper is to account for all changes to a municipality from 1880 to 1974 in terms of population changes: when municipality borders were adjusted, how many people were moved, statistically, from municipality A to municipality B? When municipality C was founded from parts of municipalities D and E, how many people were (statistically) re-homed? When municipalities F and G merged, how many people were affected? I then create a weighted representation of each municipality from these population movements. The weights represent the composition of a target year municipality in terms of the base year's municipalities, in absolute terms (municipality A is made up of $x\%$ of base year's municipality A and $y\%$ of base year's municipality B) and in relative terms ($a\%$ of municipality A come from base year's municipality A and $(100 - a)\%$ from base year's municipality B). These weights are created for adjacent years and are then chained together to create the full translation weights from any starting year to any target year.

This translation process creates a data-frame identical to what would had been collected in the baseline year had the target year's statistical boundaries been used, under either of two conditions: 1) if all variables are distributed *uniformly* in a given municipality's population *or* 2) that statistical population transfers lead to exactly representative transfers of population. In addition, we naturally need to hope that population totals and movements were recorded accurately. While these assumptions are almost certainly incorrect, the transparency, automatability, and reproducibility of the approach suggested herein will enable researchers to gauge the robustness of their analyses

[1] The two methods are complementary, each relying on a different set of assumptions and requiring different input data to construct. Ideally, researchers would have both types of translation tables available and could thus check the robustness of their results to changing from one harmonization procedure to the other.

to changing statistical areas more quickly and more efficiently than prior harmonization attempts. As an example application of the approach proposed, the paper presents time series of municipality demographics. In fact, and perhaps due to the difficulties associated with harmonizing the municipality data, such a consistent time series of the population of Finland's municipalities from 1880 to 1974 has not previously been collated.

The online data repository associated with this paper [2] contains the Python scripts to conduct the fully automated harmonization of data frames from a source to a target year. I also provide associated utilities to automatically match municipality names in statistical tables to the Statistics Finland municipality identification code and to audit the matching and translation outputs. Further included are the underlying population tables, the statistical population moves lists, and the construed cross-walks.

The next section describes the algorithm in detail, with an artificial example for illustrative proposes. I then describe the data collation steps undertaken in Section 3. In the Section 4, I further illustrate the algorithm using the population tables themselves. A detailed how-to for the Python module is found in the appendix in Section A, as well as an explanation how to use the crosswalks in Stata in Section B; further notes on the data construction steps and assumptions are detailed in the Appendix Section C.

2 Algorithm Description

The algorithm translates a data-frame construed from a given base year's statistical units to match the statistical units of a given target year. This translation seeks to approximate the data-frame that would have been observed if, in the base year, the data had been aggregated using the target year's statistical aggregation boundaries, instead of the base year's aggregation boundaries.

Three types of variables need to be considered in this translation exercise: 1) absolute scale, 2) relative scale, and 3) categorical variables. 'Absolute' variables are variables that a given statistical unit has a given number of, such as the number of manufacturing plants or the gross power consumption of a municipality. To translate an absolute scale variable, one needs to account for the total population in the target year's statistical unit that stems from each of the base year's statistical units. 'Relative' variables measure the magnitude of a variable relative to the population size, for example the percentage of the population that works in manufacturing.[3] To translate relative scale variables, one needs to describe each target year municipality as a population-weighted average of the baseline year's municipalities. Finally, categorical variables, such as which province a municipality is in, can be translated by treating "membership" x in each category i as an absolute variable, $x^i \in \{0, 1\}$.[4]

[2] `https://github.com/JonasMuellerGastell/FinlandKuntaHarmon`

[3] Many other variables are relative to a different denominator, for example, average manufacturing plant power consumption, which is relative to the number of plants. To translate these relative variables without access to the underlying items (number of plants, gross power consumption), one needs to make the (usually unpalatable) assumption that the denominator scales proportionally to population. In this example, that assumption would state that the number of manufacturing plants is proportional to population. In practice, one often has access to the variables that go into the composite measure and should translate both items independently as absolute variables and re-create the composite variable on the basis of these translated values.

[4] The resulting translated dummy variables will often be strictly inside the unit interval, and one would need to make a subject-

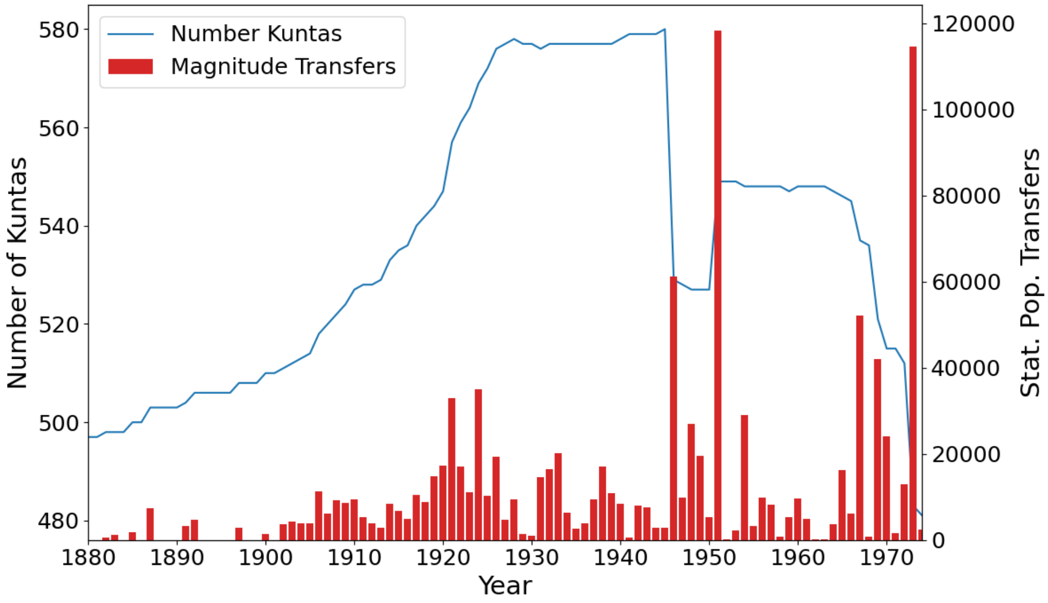


Figure 1: Number of "Kuntas" and the total number of people statistically re-distributed from 1880 to 1974. The left y-axis and the line in blue show the evolution of the raw number of municipalities in Finland. The upwards trend in the 19th century is driven by more market towns ("kauppala") being split off the rural countryside ("maalaiskunta"). The large drop in 1940 to 1944 corresponds to 1) the re-merging of market towns with the countryside, 2) the administrative formalization of the loss of Karelia, the Karelian isthmus islands, and Petsamo. The increase in 1950 corresponds to another splitting of rural municipalities from their urban cores. The declining trend starting in the 1960s is driven by the rationalization of smaller municipalities which yet again merges many smaller rural municipalities with their urban center. The right y-axis and the red bars show the sum of all statistical population transfers in a given year. This includes 100% of a municipality's population if the municipality is split off or merged with another municipality but only the number of people actually transferred between two existing municipalities in case of a simple border change.

The two "ingredients" to the algorithm are population counts in each year and a list of statistical population moves for each year. For the purposes of the algorithm, it does not matter whether a move is the result of a split (a new municipality is created by taking population from one or more existing municipalities), a merger (two or more municipalities combine with at least one ceasing to exist), a re-drawing of borders between existing municipalities, or a complex combination of these. The only relevant information for the algorithm is how many people who were counted under municipality i in a given year are counted under municipality j in the subsequent year.

To translate the statistical units of year k into the statistical units of year $k+1$, the algorithm creates weights that describe a unit i in year $k+1$ in terms of the populations carried forward from year k from i and the populations moved into i from units $j = 1, \ldots, N_{k+1}$, $i \neq j$, the N_{k+1} statistical units in year 1. For unit i, the weight for each unit j is the percentage of j's year k population that has been moved or carried forward into i. To turn these absolute weights (which answer the question 'how much of unit j is in unit i') into relative weights (i.e., to describe unit i as a weighted average of all units j who contributed population to i), we then sum over the populations moved into i and take the ratio of the magnitude of each move and the sum of all population moved or carried forward from k to $k+1$. The pseudo-code for the sub-algorithm is displayed in Algorithm 1.

The next step is to chain these weights across years. To go from $k-1$ to $k+1$, the algorithm reads the weights dictionaries that describe the units in a given year k in terms of year $k-1$ and propagates the weights forward using the translation dictionary from year k to $k+1$. Each entry in each dictionary is updated based on the in- and out-moves to and from the unit between years k and $k+1$ and the actual population counts of year k. The pseudo-code for this sub-algorithm is displayed in Algorithm 2.

The final piece of the algorithm is to use the weight dictionaries to translate a base year data-frame into a target year data-frame. Given base year municipalities $j \in \{1, \ldots, N_k\}$, denote the vector of stacked absolute weights corresponding to target municipality i, $\{w_{i,j}\}$, by w_i, and analogously for relative weights, r_i. To perform the translation for a given municipality i, the algorithm returns the dot product between the weight vector and the variable's column in vector form. This process is repeated for each municipality in the target year.

To gain intuition for how the algorithm works, consider the artificial population table, moves list, and resulting weights detailed in Table 1. The calculations are sketched in the table notes. The artificial example highlights two maintained assumptions that bear stating explicitly again. First, statistical population moves are attributed to all existing population sources proportionally. Thus, when municipality A gains population from municipality B, and then subsequently loses population to municipality C, municipality C is made up of population from both A and B. Second, changes in municipality populations between two years in excess of changes related to statistical unit changes are assumed to be attributable proportionally to all existing population sources. For example, when a municipality made up equally from two prior year statistical units sees its population grow by 20%, this growth will

expertise driven decision on how to treat these continuous approximations to the category membership. The simplest procedure would usually involve rounding to 0 or 1 for values sufficiently close to 0 and 1, and creating additional categories for highly mixed municipalities.

Algorithm 1 Create Translation Weights from Year k to Year $k+1$

Input: A list of statistical units in year $k+1$, Y_{k+1} with membership $i \in \{1, \ldots, N_{k+1}\}$ and in year k, Y_k, with membership $i \in \{1, \ldots, N_k\}$; a list containing populations p_i for all units i in year k; a list of moves between years k and $k+1$, with $s_{i,j}$ the population moved from i to j.

Output: A dictionary of absolute weights $w_{i,j}$ and relative weights $r_{i,j}$, for each unit j in year k to each unit i in year $k+1$

```
To create absolute weights $w$
for unit $i \in Y_{k+1}$ do
  if unit $i \in Y_k$ then
    $s_i = \sum_j^{N_{k+1}} s_{i,j}$
    $w_{i,i} = \frac{p_i - s_i}{p_i}$
  else
    $w_{i,i} = 0$
  end if
  for unit $j \in Y_k$ with $i \neq j$ do
    $w_{i,j} = \frac{s_{j,i}}{p_j}$
  end for
end for
```

```
To turn absolute weights into relative weights $r$
for unit $i \in Y_{k+1}$ do
  $r_{i,j} = \frac{p_j w_{i,j}}{\sum_j^{N_{k+1}} p_j w_{i,j}}$
end for
```

Algorithm 2 Chain Translation Weights from Year k to Year $k+n$

Input: A list of statistical units in each year l from k to $k+n$, Y_{k+n} with membership $i \in \{1, \ldots, N_l\}$; a list containing populations p_i^l for all units i in each year l; the dictionaries of absolute weights $w_{i,j}^l$ and relative weights $r_{i,j}^l$, for each unit-pair i, j, in each year l from k to $k+n$

Output: A dictionary of absolute weights $w_{i,j}^C$ and relative weights $r_{i,j}^C$, for each unit j in year k to each unit i in year $k+n$

```
Initialize for all $i, j$: $w_{i,j}^C = w_{i,j}^k$
for year $l \in \{k+1, \ldots, k+n\}$ do
  for unit $i \in Y_l$ do
    for unit $j \in Y_{l-1}$ (Note: including $i = j$) do
      for unit $q \in Y_k$ do
        $w_{i,q}^{C_{temp}} = w_{i,j}^l * w_{j,q}^C + w_{i,q}^{C_{temp}}$
      end for
    end for
  end for
  $\forall i, j: \quad w_{i,j}^C = w_{i,j}^{C_{temp}}$
  $\forall i, j: \quad w_{i,j}^{C_{temp}} = 0$
end for
```

be attributed equally to the two source units. Thus, the relative and absolute weights for both prior municipalities remain unchanged for the target year municipality

With the full translation cross-walks computed on the full list of moves and the complete population tables, we can examine the distribution of weights between specific base and target years, to gauge how much change has occurred in statistical unit borders. Figure 2 displays the evolution of the share of weights above 90% and below 19% (both for relative and for absolute weights) when translating the 1880 baseline forward to each year between 1881 and 1974. As expected from the evolution of municipality numbers and the magnitude of municipality changes, weights are mostly close to 1 for the pre 1910s. As the trend towards splitting municipalities accelerates in the early 20th century, the share of large weights goes down, but small weights do not increase much yet – most splits lead to more substantial weights. As more and more municipality mergers and divisions occur in the inter war and post-war period, the share of large weights further declines and very small weights become more common. The trends are very similar for absolute and relative weights. Figure 3 summaries the entire weight distribution when translating 1880 into 1974. We see that the distribution of medium sized weights is quite uniform between 0.1 and 0.9.

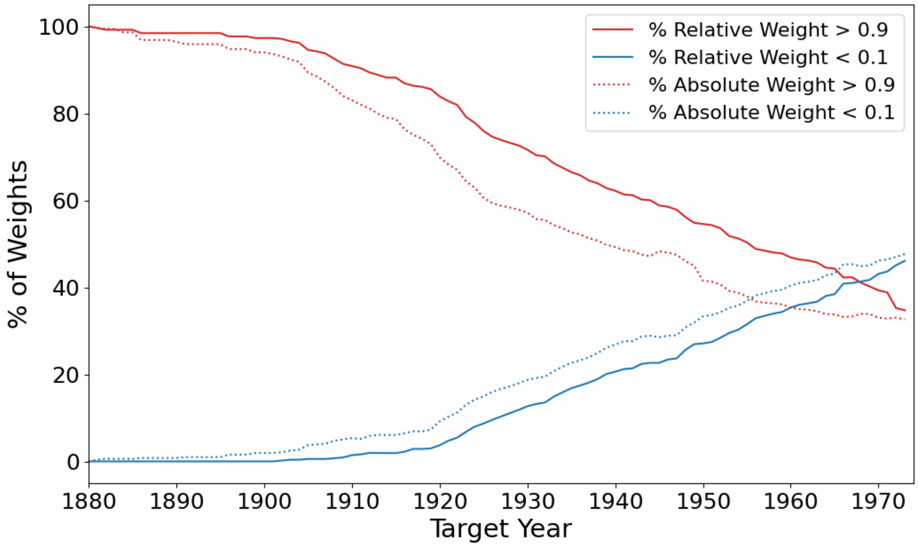


Figure 2: Evolution of absolute and relative weights when translating a data-frame on basis of 1880 to each year between 1881 and 1974.

3 Data Digitization and Description

The source for most data are the "Väkiluvun tilastoa," the yearly population statistics, which record the number of births, deaths, marriages, in- and out-migration of each statistical unit for a given year (Tilastokeskus (Statistics

	Population		
Municipality	Year 1	Year 2	Year 3
Helsinki	95,000	125,000	120,000
Espoo	23,000	25,000	40,000
Vantaa	10,000	5,000	7,000

(a) Population by Municipality and Year

Year	Source	Target	Pop. Moved
Year 1	Vantaa	Helsinki	5,000
Year 2	Helsinki	Espoo	12,500
Year 2	Espoo	Vantaa	2,500

(b) Population Moves

Target	Base	Absolute	Relative
Helsinki	Helsinki	1	0.95
—”—	Espoo	0	0
—”—	Vantaa	0.5	0.05
Espoo	Helsinki	0	0
—”—	Espoo	1	1
—”—	Vantaa	0	0
Vantaa	Helsinki	0	0
—”—	Espoo	0	0
—”—	Vantaa	0.5	1

(c) Weights Dictionary Year 1 to Year 2

Target	Base	Absolute	Relative
Helsinki	Helsinki	0.9	0.95
—”—	Espoo	0	0
—”—	Vantaa	0.45	0.05
Espoo	Helsinki	0.1	0.320
—”—	Espoo	0.9	0.667
—”—	Vantaa	0.05	0.013
Vantaa	Helsinki	0	0
—”—	Espoo	0.1	0.333
—”—	Vantaa	0.5	0.667

(d) Weights Dictionary Year 1 to Year 3

Table 1: Artificial example of population count tables, moves lists, and resulting population weights.
The details of the calculations are as follows. Between year 1 and year 2, Vantaa sees half its year 1 population moved into Helsinki. No other moves are recorded (but note that populations change in excess of the recorded moves due to organic migration, births, and deaths). In absolute terms, Vantaa in Year 2 is thus 50% of Vantaa in Year 1, while Year 2 Vantaa is still made up to 100% Vantaa Year 1 residents. In absolute terms, Helsinki's Year 2 population thus is 100% of Year 1 Helsinki and 50% of Year 1 Vantaa. In relative terms, the population comes to 5% from Vantaa and to 95% from Helsinki. Espoo sees no moves and thus its absolute and relative population weights are both 100% Espoo. Note throughout the maintained assumption that population changes unrelated to statistical moves are proportionally attributable to all population sources.
Between Year 2 and Year 3, Helsinki distributes 10% of its Year 2 population to Espoo ($12,500$ out of $125,000$), while Espoo loses 10% its own population to Vantaa ($2,500$ out of $25,000$). The chained weights from Year 1 to Year 3 are hence as follows: Helsinki loses 10% of each of its constitutent absolute weights but its relative weights remain unchanged. Espoo is made up out of 10% of Year 1 Helsinki, 5% of Year 1 Vantaa (via the 50% of Vantaa that moved into Helsinki between Year 1 and Year 2), and 90% of Year 1 Espoo (as Espoo lost 10% between Year 2 and Year 3). The relative weights for Espoo are given by the ratio of the inflow ($12,500$) to the Year 2 baseline ($25,000$), weighted by the original Year 1 source (95% Helsinki, 5% Vantaa). Vantaa gains $2,500$ from Espoo and hence is now made up of 10% of Year 1 Espoo ($2,500$ over $25,000$, as the population ratio is taken relative to Year 2 baseline) and still 50% of Year 1 Vantaa. The relative weights reflect the ratio of these inflows relative to Year 2 population.

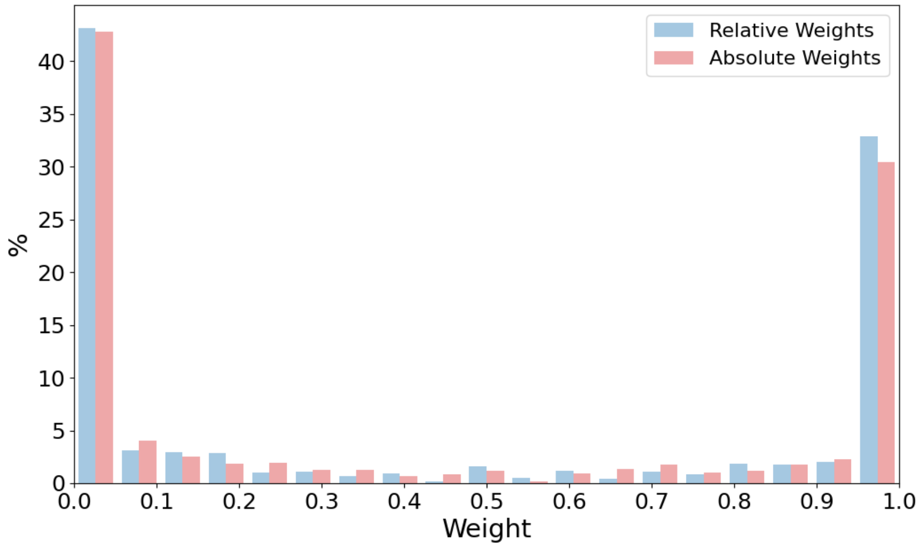


Figure 3: Histogram of absolute and relative weights when translating a data-frame on basis of 1880 to 1974.

Finland), 1880-1974).[5] In the 19th century and up to 1929, the record keeping and thus record aggregating statistical unit were church parishes and civil registries. Thereafter, "kuntas" themselves became the major statistical record keepers and thus the unit of statistical reporting. Already in the 19th century, however, full-count population censuses were conducted at the level of the "kunta," rather than the parish. As parish boundaries are defined by the intersection of denomination and geography, I have opted to aggregate parishes up to their "kunta" parents as accurately as possible and to report all time series and perform all harmonization at the level of the "kunta", not the parish. The decennial census documents, starting in 1880, thus provide the universe of names and identities of all statistical units for this project.[6]

The "Väkiluvun tilastoa" population tables are also the primary source for the information on population redistributions. Usually, a footnote in the original table would indicate whether a redrawing of municipality borders had occurred. The coverage of these footnotes is unclear – but given that moves of as few as one parish member were recorded, I believe that large scale redistributions that are not noted are rare. Nonetheless, the contributors of the data to this project, research assistants, and I have exhaustively examined the in- and out-migration statistics recorded to spot unusually large movements, which might have been associated with un-recorded statistical area changes. Similarly, we checked the raw time series of population trends to spot sudden drops or increases, and

[5] All of these files are available on `www.doria.fi`. E.g., `https://www.doria.fi/handle/10024/67320` is the population table for 1891.

[6] In the vast majority of cases, matching parishes to "kuntas" is straightforward, as parishes tend to be identical to most "kuntas" or strict subsets of larger "kuntas." The only problematic cases are non-Lutheran parishes. As Lutheranism is the religion of the vast majority of Finns, the Orthodox, Catholic, Baptist, and other parishes are very small and, to gain sufficient size, span the geographic area of multiple Lutheran parishes. Where a large overlap between a singular "kunta" and a non-Lutheran parish was possible, the match was performed one-to-one. Where no clear match was possible, the non-Lutheran parishes have been dropped. The total number of dropped parishes and the total population of these parishes is fortunately minute.

checked any suspect years against more detailed primary records of the municipalities population history.[7] In addition, research assistants have read the long-form descriptions of changes to "kunta" borders and definitions included in every year's statistical tables release and noted all population movement mentioned therein.

To construct the population tables from 1880 to 1974 through merges, splits, and name changes, and to match the recorded border changes to municipalities, one needs to have access to a time invariant municipality identifier. Fortunately, the Väestörekisterikeskus (Finnish Population Register Center) provides a three digit numeric identifier to all municipalities that have existed after 1945 and most that have been discontinued or ceded to the Soviet Union in or before 1944 Tilastokeskus (Statistics Finland).[8] I have relied on the historical information provided by Finnish Wikipedia[9] to trace any name changes not recorded in the original files. The original matching of municipality names to codes was performed manually. The script included in the module to match statistical table municipality names to codes has been tested on the manually performed matches and reconstructs the matches for the population series exactly.[10]

Finland performed a complete population count in each decennial census (i.e., complete counts are available for 1880, 1890, 1900, 1910, 1920, 1930, 1940, 1950, 1960, and 1970). Population counts between these years are estimated from the recorded net changes (births, deaths, in-migration, out-migration). Post 1950, the estimated counts are directly included in the published files. Pre-1950, estimated population counts are not directly included in the annual files. I hence impute the population counts by chaining the recorded net changes. I perform the calculation both going forward (i.e., adding net increases to a base decennial census year, say, from 1880 to 1881, to 1882, . . .) and going backward (i.e., subtracting net increases from, say, 1899 to 1900 from the 1900 figure to obtain the 1899 figure). The errors between both directions are likely different and thus I base the algorithm on the average of the two time series for each year. See the Appendix Section C.2 for more details.

4 Example: Harmonizing Population Tables

To see the harmonization process in action, I will present some descriptive analyses on two topics. First, I will look at the urban versus rural population of Finland. A stylized fact of Finnish history is that urbanization was late, with major urban in-migration still occurring in the 1960s. I will consider how the percentage of the urban population develops taking into account both the eventual definition of urban areas (as per 1974) and the original definition (i.e., as per 1880), keeping statistical units constant over time in both cases. Second, I will compare

[7] We did find a small number of such changes in populations that are very likely associated with un-recorded re-drawings of statistical boundaries in the pre 1939 data. Usually, these relate to how parishes were aggregated to "kuntas."

[8] For the small number of municipalities that had no existing three-digit code, I have manually created a unique code. Name changes are not always recorded in the population tables or long-form write-ups accompanying the tables.

[9] See, e.g., https://fi.wikipedia.org/wiki/Luettelo_Suomen_kunnista.

[10] The script automatically allows for name and spelling changes and correctly delineates between identically named municipalities in different provinces or identically named but distinct urban/market-towns and rural municipalities when sufficient information is provided.

mortality between urban and rural areas. I will again consider the original 1880 and the eventual 1974 definitions.

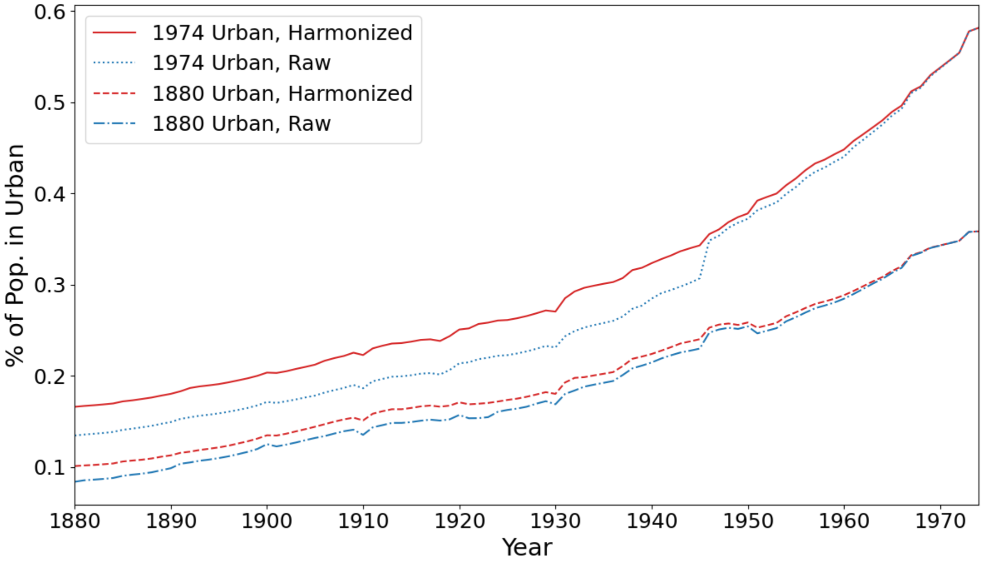


Figure 4: The percentage of total population that resides in urban areas, from 1880 to 1974. Solid and dotted lines represent the time series using the 1974 urban/rural classification system (retroactively applied to all municipalities). Dashed and dot-dashed lines display they time series using the 1880 classification system (pro-actively applied to municipalities going forward). In red are harmonized time series, in blue raw time series. Harmonization was performed with the scripts included in this software publication.

Figure 4 displays the evolution of the percentage of the total population that resides in urban areas, from 1880 and 1974. Solid and dotted lines represent the time series using the 1974 urban/rural classification system (retroactively applied to all municipalities). Dashed and dot-dashed lines display they time series using the 1880 classification system (pro-actively applied to municipalities going forward). In red are harmonized time series, in blue raw time series. We see the expected pattern of very low levels of urbanization in the late 19th century, when urbanization was as low as 10% in 1880. Even we are not using the urban 1880 based classification and retroactively declare rural areas that were urban in 1974 as urban, the percentage share of the urban population only climbs to 17%. Urbanization progresses linearly until about the second World War, when the pace of urbanization does seem to quicken.

Keeping statistical units constant and retroactively applying the urban area definition to areas that were not declared urban areas in the given year (partially) protects against over-estimating the amount of urban area growth that stems classification changes and border changes. Classification changes quite obviously fallaciously increase the urban population in years when a rural area that has grown quickly becomes a town. Border changes do so even more insidiously, as rural area's ceding land or being outright merged with urban areas is harder to spot when looking at a data series.

Considering both the 1880 and the 1974 definition of urban areas has two advantages. First, comparing growth based on the original urban areas to the eventual urban area classification tells us whether the urban population has increased mostly due to specific rural areas increasing and being re-classified as towns, or whether pre-existing urban areas were the main source of growth. Second, we are better equipped to spot the effects of the harmonization procedure when comparing both time series to an un-harmonized time series.

Comparing the four time series, two main patterns emerge: First, urbanization did seem to increase in speed from the 1930s to 1940s, as the slope of the red solid line (corresponding to the urban share based on harmonized 1974 classifications) increases. Second, some of this increased growth comes from newly urban-classified areas growing faster than other countryside areas. The red dashed line (corresponding to the urban share based on the harmonized 1880 classification) and the red solid line are remarkably parallel until the 1930s, naturally with a constant level difference. Then, in the 1940s and 1950s, the two lines start to diverge more strongly. This indicates that some of the towns that were newly classified as urban kept outgrowing the rural areas and other urban areas.

Figure 5 displays the mortality rates for urban (in red) and rural (in blue) areas in each year from 1880 to 1974. Again, urban and rural areas are defined both on the basis of 1974 and on the basis of 1880. In contrast to the urbanization rates displayed above, the decision which year's urban area definition to set as a baseline makes less of a difference. Mortality is defined as the number of deaths recorded in *all* urban (or all rural) areas over the total population recorded in those areas in the given year.

We see the overall precipitous decline in mortality in Finland from 1880 to 1974, punctuated by the drastic spikes that can be accounted for by the Spanish Flu (1918-1920) and the Winter and Continuation Wars with the Soviet Union (1939-1940, 1941-1944). Rural areas (in blue) consistently have higher mortality rates than urban areas, except for the Spanish Flu years. The difference seems to be increasing since the end of World War 2, which might be associated with the out-migration of younger people to the cities.

5 Conclusion

In this paper, I present the `FinlandKuntaHarmon` Python module, a set of Python scripts that automate the matching of municipality names in statistical publications to kunta-codes and the harmonization of the statistical record keeping units across years, for 1880 to 1974. I hope that, armed with the scripts and data provided in this paper and the associated software repository, more researchers can employ the rich, detailed universe of Finnish historic data more easily, accurately, and in more replicable manners. I plan to extend this time series to the present day in a future version of this paper and the Python module.

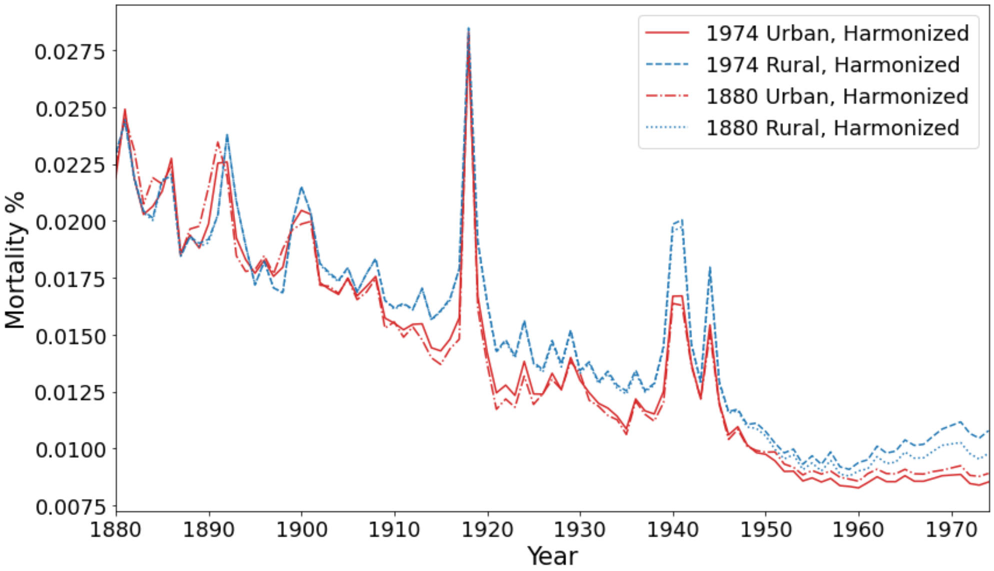


Figure 5: Mortality rates for the urban and rural population, from 1880 to 1974. Solid and dot-dashed red lines represent the mortality for urban areas, using the 1974 (solid) or 1880 (dot-dashed) classification (retroactively or pro-actively applied to all municipalities). Dashed and dotted blue lines display mortality for rural areas, using the 1974 (dashed) or 1880 (dotted) classification (retroactively or pro-actively applied to all municipalities). All displayed data series are harmonized using the scripts included in this software publication.

References

ECKERT, F., A. GVIRTZ, J. LIANG, AND M. PETERS (2020): "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790," *NBER Working Paper*.

TILASTOKESKUS (STATISTICS FINLAND) (1880-1974): *Suomen virallinen tilasto VI: Yleinen katsaus väkiluvunmuutoksiin Suomessa*, e.g., `https://www.doria.fi/handle/10024/67320` for 1891.

——— (1992): *Kuntanumerointi. Edelt. Kuntanumerointiluettelo. Tilastokeskus, kerran vuodessa 1975–1992. ISSN 0783-1404.*

A How to use the Python Module

To use the Python Module directly, go to the GitHub page,[11] select "Clone" or "Download ZIP", and unpack the downloaded folder into a directory on your Python path.[12] Below is a working code snippet that 1) reads a data-frame from 1950 and a data-frame from 1965 (stored in csv format), 2) matches the name column in both data-frames with the correct "kunta" identifiers, 3) translates the absolute scale variables, relative scale variables, and the categorical province identifier from 1950 into 1960, 4) merges the two data-frames, ready for running regressions! The script assumes the two data-frames live in some folder specified by the `path` variable, and the module is saved in the path referenced by the variable `moduleData`.The complete syntax with all options for each of the programs is explained in detail below.

[11] `https://github.com/JonasMuellerGastell/FinlandKuntaHarmon`

[12] If you do not know where your Python path points to or you do not wish to embed this module in your Python library, you will need to manually append the path to the module in any scripts you write.

```
path, moduleData = "[replace with your path to the data]",  "[replace with your path to the module]"

import pandas as pd
from harmonizer import harmonize, categoricalHarmonize
from code_linker import findCodes

# read the 1950 and 1965 data frame
df50 = pd.read_csv(path + "DataFrame1950.csv")
df65 = pd.read_csv(path + "DataFrame1965.csv")

# match the names (column called KuntaName) to kunta codes,
# and make the program print all accuracy checks
df50_withCodes, fixes50 = findCodes(df50, "KuntaName", yearvar = 1950, verbose=True,
                                    suggestFixes=True, path=moduleData)
df65_withCodes, fixes65 = findCodes(df50, "KuntaName", yearvar = 1965, verbose=True,
                                    suggestFixes=True, path=moduleData)
# inspect the fixes[yy] files to see  whether we need to make changes
print(fixes50, fixes65)

# translate the 1950 data frame to 1965 data frame.
# We want to translate the relative scale variable "AvgSalary", the categorical "Province")
# , and the other three absolute scale variables  ("HorsePowerTotal", "NumberPlants", "NumberWorkers")

df50_translated = harmonize(df50_withCodes, 1950, 1965,
                             absCols = ["HorsePowerTotal", "NumberPlants", "NumberWorkers"],
                             relCols = "AvgSalary", idCol="code", f=moduleData)
df50_categorical = categoricalHarmonize(df50_withCodes, 1950, 1965,
                                         "Province", idCol="code", f=moduleData)
df50_translated = df50_translated.merge(df50_categorical, on ="code")

# merge the 1950 translated data frame with the 1965 data-frame
df_final = df50_translated.merge(df65_withCodes, on="code")
```

A.1 Detailed Syntax Explanation

A.1.1 `code_linker.py`

This file provides a utility to match "kunta" names with the modern kunta-koodi system employed by Statistics Finland. The utility is called `findCodes`. To be able to use it, call `from code_linker import findCodes`

```
findCodes(df, names, province=False, urban=False,  yearvar = False,
          verbose=False, suggestFixes = False,
          path = '[your path]FinlandKuntaHarmon/Data/')
```

Output: *either* a data-frame with one column per absolute or relative scale column and one column per category per categorical variable, and an additional column 'code' containing kunta-coodi, *or* a tuple containing the above data-frame *and* a dictionary of suggestions (see below).

- `df`: the pre-loaded pandas data-frame in which you would like to link names, needs to contain a column with kunta-names
- `names`: string with the name of the column which contains the names, OR a list with strings with (in this order) the names of name-column, province-column, urban-column (province and urban indicator help the program deal with ambiguous names), OR a dictionary with 'name', 'province', 'urban' fields
- `province`: bool, should province names be used to help?
- `urban`: bool, should urban/rural indicator be used to help?
- `yearvar`: yearvar: bool, string, or integer; should the program attempt to check the matched list of names agains the population data file for the given year to identify kuntas who did not exist (in the population tables) in the given year? If so, include either the column name that contains the year or the year to be checked as an integer
- `suggestFixes`: bool, should the program make suggestions for what re-namings should occur to get non-anachronistic matches?
- `path`: string, path to the data necessary to run the module

A.1.2 `harmonizer.py`

This file provides utilities to translate a given data-frame from the kunta-boundaries of year A to the kunta-boundaries of year B (A¿B). The assumption maintained is that the variable of interest (an abslute value or a per-population rate or a categorical membership) is uniformly distributed within the population of the kunta.

Then, the utilities trace all (recorded) population transfers, kunta splits, merges, and foundations. By maintaining this assumption of uniformity and hence by using population weights, we create a best guess of what the data-frame would have looked like if the historical data was collected under the boundaries of the more recent kunta coding system. The two routines included are called `harmonize` and `categoricalHarmonize`. `harmonize(...)` harmonize a data frame with absolute or relative statistics from one year to another. `categoricalHarmonize(...)` harmonizes categorical variables (e.g., membership in a province) from one year to another. Import both using `from harominizer import *`.

```
harmonize(df, year1, year2, absCols = [], relCols = [], idCol = False,
              f= '[your path]/FinlandKuntaHarmon/')
```

Output: data frame with harmonized columns and an identifier columns (kunta-koodi).

- `df`: the pre-loaded pandas data-frame with a kunta-koodi id column and at least one data column to translate
- `year1`: int or string, the year on which the data-frame's kunta boundaries are based
- `year2`: int or string, the year to which you would like to translate the kunta boundaries
- `absCols`: string or list, the name(s) of the columns with abslute value (e.g., number of tractors) to translate
- `relCols`: string or list, the name(s) of the columns with relative values (e.g., casualty rate) to translate
- `idCol`: string, name of the kunta-koodi containing identifier column, if False assume idCol is one of ['code', 'id', 'kunta', 'grouper']
- `f`: string, path to where the kunta harmonization module lives

```
categoricalHarmonize(df, year1, year2, whichCol, idCol = False,
                         f= '[your path]/FinlandKuntaHarmon/')
```

Output: data frame with harmonized columns and an identifier columns (kunta-koodi). Note that categorical variables are translated into one column per level, with values continuously between 0 and 1, allowing the user to determine thresholds at which to assign 0 or 1 themselves

- `df`: the pre-loaded pandas data-frame with a kunta-koodi id column and at least one data column to translate
- `year1`: int or string, the year on which the data-frame's kunta boundaries are based
- `year2`: int or string, the year to which you would like to translate the kunta boundaries
- `whichCol`: string or list, the name(s) of the columns with categorical variables to translate

- `idCol`: string, name of the kunta-koodi containing identifier column, if False assume idCol is one of ['code', 'id', 'kunta', 'grouper']
- `f`: string, path to where the kunta harmonization module lives

B How to Use the Crosswalks in Stata

If you prefer using Stata instead of Python, there are two options. One is running Python from within Stata, as detailed at `https://www.stata.com/python/`. The more direct approach is to use the provided crosswalks and performing the aggregation "manually." The steps for this are quite simply:

- Download the "long" crosswalks from `https://github.com/JonasMuellerGastell/FinlandKuntaHarmon/tree/master/Data/TransitionsLong`
- Open the crosswalk you want to implement (e.g., 1880 to 1975) in Stata using `import delimited` and rename the `idSource` column into the name of the kunta-codes column in the main data-frame, say `code`
- Implement a many to one merge of the crosswalk in memory with your main data-frame, e.g., if the latter is in the current working directory and is called "analysisData.dta," run

  ```
  merge m:1 code using analysisData.dta
  ```

- For each data column you wish to translate, generate a weighted data-column from the absolute or relative weights column (depending on the type of the data column): e.g., for column "machines", use

  ```
  gen weighted_machines = absValue * machines
  ```

- Perform an `egen` operation to sum across the weighted individual variable values:

  ```
  bysort idTarget: egen machines_translated = sum(weighted_machines)
  ```

- Drop the duplicate rows: `bysort idTarget: gen dup = _n`, followed by `keep if dup == 1` and `drop dup`.

C Additional Details on Data Series Construction, Manual Changes, and Assumptions in Data Processing

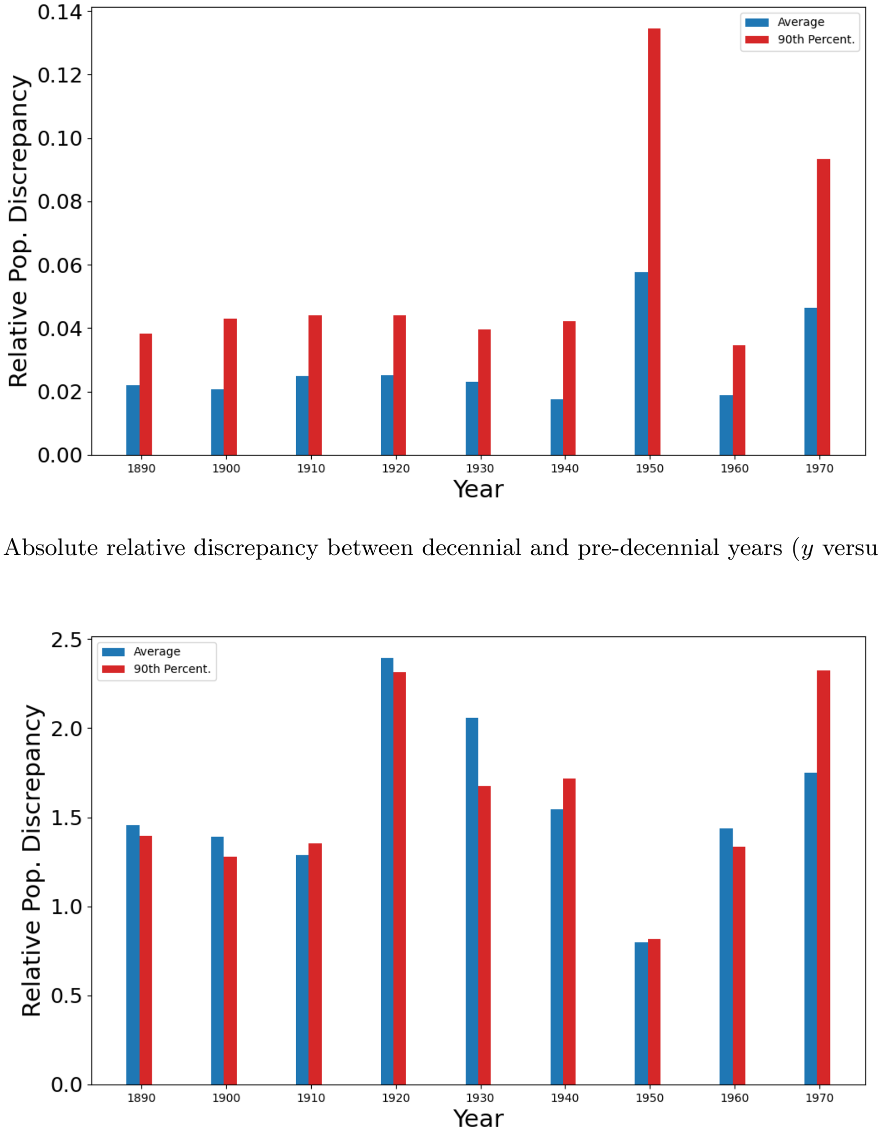
C.1 Aggregated Reporting of Distinct "Kuntas"

One difficulty with the population tables is that some small neighboring municipalities may be recorded jointly, e.g., municipality A and B might both be listed in the index of municipality names but the associated data row in the table is shared between the two municipalities. While this type of "sharing" a row is not a true administrative merge of the municipalities, I have decided to treat it as if it was a merge, given that other data series will likely perform the same aggregation step. Two complications arise. First, if a different data series does *not* perform the same aggregation, a naive translation process would simply drop one of the two underlying entries. To prevent this, the script routines provided check whether all provided municipalities occur in the baseline year's population tables and the script makes a suggestion for what manual merges should be performed in order to bring the baseline data-frame into accordance with the baseline year population table. Second, this type of merge may be especially prone to violate the assumptions necessary for the harmonization procedure to approximate the anachronistic data-frame.

The assumption of uniform or representative population movements is, as noted above, almost certainly only an imperfect approximation. Thus, the more merges are in the translation file, the more inaccuracies will accumulate. The joint reporting of municipalities may be particularly problematic, if the mistakes are more systematic in this category. Consider a rural municipality and a small market town. Assume the two are recorded separately from 1880 to 1920, merged in the table from 1921 to 1951, and separately afterwards. The harmonization procedure in this paper ignores the information available in the prior distinct existence and instead assumes the "split" in 1951 was a uniform/representative split of the population. Thus, we ignore the possibility that the "split' will mirror the original difference between rural and urbanized parts of the municipality. A more assumption-heavy and labor intense harmonization procedure would construct time trends of the two municipalities between the two years and directly model the changes in the variables of interest in the jointly recorded years.

C.2 Additional Details on Population Interpolation

To gain a better gauge of how accurate the interpolation of population counts between decennial census waves is likely to be relative to the actual counts, we can compare the change observed between the interpolated population of the year before the census ($y-1$) and the actual counted census population (in year y). Define the 'average absolute relative discrepancy' as the difference between the maximum and the minimum of the two population figures divided by the average of the two. Figure 6a displays the average across municipalities of the relative absolute discrepancy and the 90% percentile of this statistic in each decennial year. The largest discrepancies occur in 1949-1950 and 1969-1970. The war-time upheaval and resettlement of 10% of the population after ceding Karelia (twice, in 1941 and 1944) will necessarily impact accuracy of the pre-1950 population estimates. This interpretation is supported by Figure 6b which plots the ratio of the measure computed here to the same measure for the years $y-2$ versus $y-1$ (i.e., years ending in 8 and 9). We see that the ratio is usually around 2, but actually smaller than 1 for 1950, indicating that the error stems less from the interpolation procedure and is driven by the overall variance in population across years in this period. The 1970 census is conducted after a heavy round of municipality re-drawing and merging, and the data may hence have suffered relative to other years. As already noted and apparent from the figures, pre-World War 2, average accuracy seems to be very high. However, there are a small number of (small) municipalities that display very large gaps in the 19th and early 20th century. I suspect that record keeping of in- and out-migration out of parishes could have been inaccurate in these particular cases.



(a) Absolute relative discrepancy between decennial and pre-decennial years (y versus $y - 1$).

(b) Ratio of absolute relative discrepancy measures between (y and $y - 1$) and ($y - 1$ and $y - 2$

Figure 6: These figures display information about the accuracy of the population interpolation using relative absolute discrepancy between decennial census years (short y) and the preceding year ($y - 1$). To calculate relative absolute discrepancy take the ratio of the difference between the max and the min of the two years' population figures and the average of this min and max. The result is the relative (to population) absolute (i.e., regardless of which of the two years has larger population) discrepancy and can be interpreted as a percentage. Figure a) shows the average and 90th percentile of the statistic across municipalities, in blue and red bars respectively. Figure b) shows information on the ratio of the discrepancy statistic in the actual census year and a preceding placebo year (i.e., $y - 2$ and $y - 1$). Blue bars show the average of the ratio across municipalities, red bars show the 90th percentile.