Algorithmic Determination of Japanese Ethnic Identity Based on Name

Martin Holmes*

Abstract

The Landscapes of Injustice project[1] is a multi-institutional seven-year research project funded by a Partnership Grant from the Social Sciences and Humanities Research Council of Canada. The project mission is to investigate, document, and analyze the process by which, beginning in 1942, tens of thousands of people of Japanese ethnicity were interned, and their property was seized and disposed of by the Canadian government and its agents.[2] Part of this process involves identifying where Japanese Canadians were living and working prior to their internment, what property they owned, and how the dispossession affected them. This requires that we identify Japanese Canadian individuals across a range of different types of official and unofficial records, often based only on name, on a scale that makes an entirely manual process impractical. The project has therefore developed a semi-automated algorithmic approach to determining whether any name in the records is Japanese or not. This article describes the algorithm in detail, along with its application and limitations.

Introduction

The Landscapes of Injustice project[3] is a multi-institutional seven-year research project funded by a Partnership Grant from the Social Sciences and Humanities Research Council of Canada. The project mission is to investigate, document, and analyze the process by which, beginning in 1942, tens of thousands of people of Japanese ethnicity were interned, and their property was seized and disposed of by the Canadian government

* University of Victoria

[1] Accessed August 31, 2018, http://www.landscapesofinjustice.com/.

[2] See, for example, Sunahara (1981).

[3] Accessed August 31, 2018, http://www.landscapesofinjustice.com/.

and its agents.[4] Part of this process involves identifying where Japanese Canadians were living and working prior to their internment, what property they owned, and how the dispossession affected them. This requires that we identify Japanese Canadian individuals across a range of different types of official and unofficial records, often based only on name, on a scale that makes an entirely manual process impractical.

The determination of ethnicity based on name is a profoundly complex problem; the notion of ethnicity itself is both questionable and highly politicized, and intermarriage in increasingly multicultural populations along with the internationalization of names further complicates the issue. However, the injustices addressed by our project were perpetrated on the basis of ethnicity, and it is vital that we are able to distinguish all the individuals of Japanese origin who appear throughout the land title records, community and city directories, and other primary source documents. For example, Stanger-Ross and Blomley (2017) investigated 292 letters written by Japanese Canadians protesting against the forced sale of their property, most of which inveigh not just against the forced sale but also against the undervaluing of the property itself. To measure the scale of the undervaluation as a whole, it is necessary to examine the history of property sales such that the prices paid for property confiscated from Japanese Canadians can be compared with prices paid for equivalent properties that were owned by individuals of other ethnic backgrounds, as well as the prices paid for properties as they changed hands in the decades leading up to the dispossession. To do this, we need to be able to determine with some certainty who is and who is not a Japanese Canadian property-owner. More broadly, the Landscapes of Injustice project is interested in the composition of immigrant communities such as the “Japantown” around the Powell Street area of Vancouver,[5] where the vast majority of residents were not property owners, simply renters or inhabitants of rooming houses, and the street directories used to build up a picture of the ethnic makeup and diversity of the community frequently provide nothing other than a surname and an initial. Scholars of migration and ethnicity are often working with records such as these, where the absence of specified ethnicity makes quantitative research in particular quite difficult. The approach outlined in this paper may suggest some promising avenues for researchers working on other multiethnic communities

[4] See, for example, Sunahara (1981).

[5] See Jordan Stanger-Ross et al. 2016.

facing similar problems, or working with historical records that would be considerably more informative or valuable if information on ethnicity could be attached to them.[6]

Several factors combine to make the determination of Japanese origin in 1940s Canada a more tractable undertaking than it would be for other times, places, and ethnicities. Japanese immigration to Canada drew from a relatively small range of source locations in a nation that was remarkably lacking in ethnic diversity before World War II; Japanese immigrants seldom intermarried with people of other origins; the population was assiduous in documenting itself in a number of community directories; and the Japanese language itself has helpful distinctive features which are amenable to processing. This article describes an algorithmic process we have developed which uses a variety of evidence from public databases, dictionaries, and graphemic structure to assign a probability that a name is ethnically Japanese, substantially reducing the need for human judgement in the processing of names. This operation depends on a thorough understanding of the migration patterns of the target community, linguistic features of Japanese, the accumulation of results from prior iterations, and human research. We might characterize it as principled algorithmic pragmatism: it is principled in the sense that all the factors which contribute to a decision to assign (or not) Japanese ethnicity to a given name are transparent; it is algorithmic in that the majority of such decisions are made by a computational process; and it is pragmatic in that edge cases and anomalous results are addressed through human research, and the results of that research are fed back into the algorithm.

1. Source Documents

In our research into the dispossession and internment of Japanese Canadians during and after World War II, we are collecting and digitizing a broad range of resources from land title records through city and community directories to trace patterns of residence, occupation and property-ownership among Japanese Canadians before and after the displacement. In working with this diverse material, it is a fundamental requirement that we be able to identify Japanese Canadians and to distinguish them from those of other ethnicities who were their neighbors, their landlords, their tenants, and their business

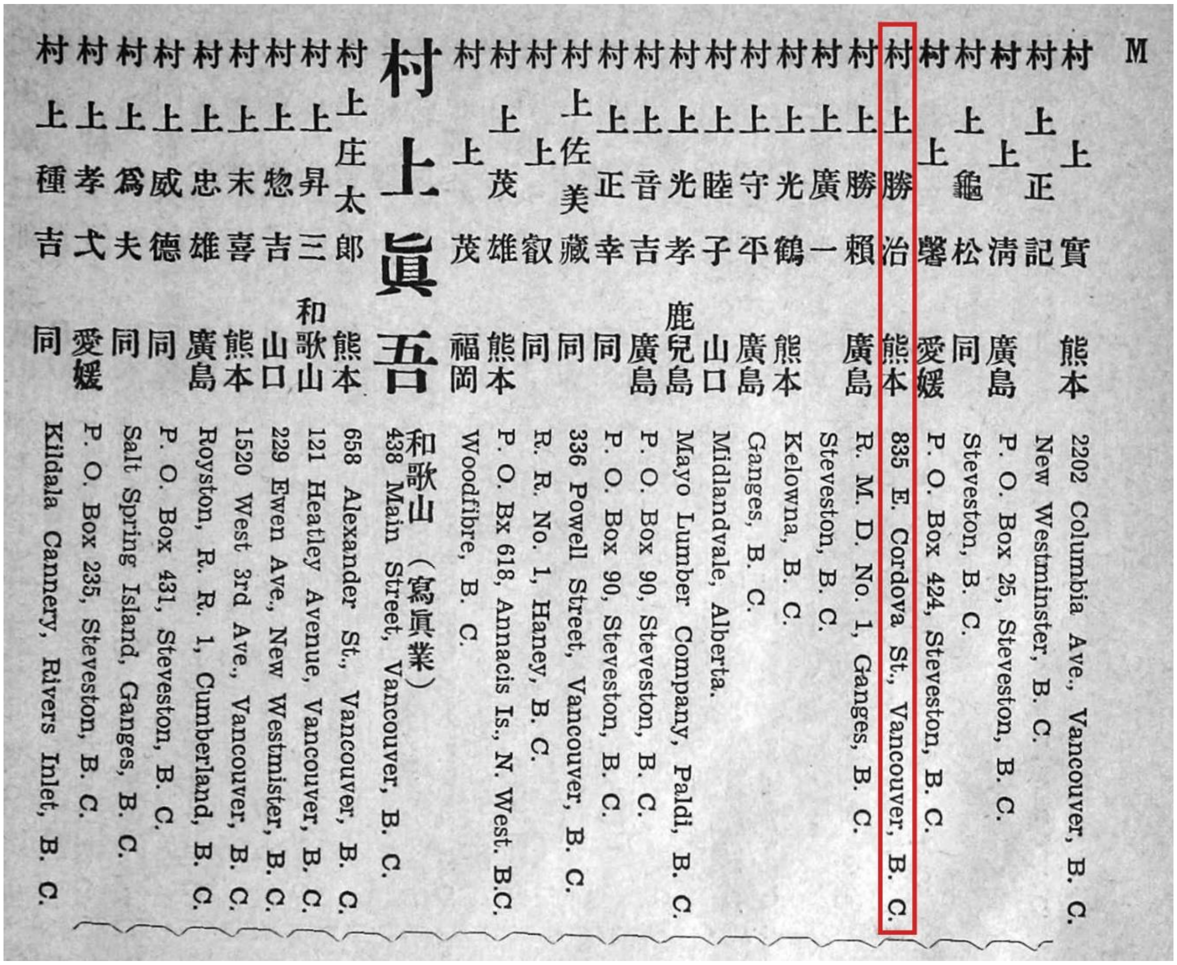
[6] Kobayashi (2017) points out that historical geographers in North America have tended to focus on archival material rather than numbers and statistics, whereas in Japan, the reverse was the case. Statistical work in the Landscapes of Injustice project runs counter to this trend.

colleagues. In the case of some resources this is trivial: a directory of Japanese Canadians such as the *Zai Kanada hōjin jinmeiroku* [All Canada Japanese Expatriate Directory],[7] where all names are in Japanese, and which usually includes information on the Japanese province of origin of the person, can be assumed to contain only names of Japanese origin; see the example in figure 1, where the bordered line shows Katsuji Murakami from Kumamoto Prefecture living at 835 East Cordova Street. (Although names and prefectures are in Japanese, oddly, they are organized according to alphabetical order in the Latin alphabet, rather than Japanese kana order as we might expect.) But in the case of many records, such as the Vancouver Land Titles in which we research property ownership, or the Vancouver street directories from which we establish patterns of residence, names are listed in Latin script, often consist only of a surname and an initial, and are sometimes mistranscribed, mistyped, romanized in a nonstandard fashion, or difficult to read.

In a less academic context, sheer common sense, intuition, and native language expertise would be sufficient to unambiguously identify the vast majority of Japanese names in documents like these. However, our datasets are large and complex, and we lack the resources to address every single name individually; and yet we are concerned to ensure that the assignment of Japanese ethnicity across the dataset is accurate and complete. We have therefore developed a semi-automated approach to assigning Japanese ethnicity which combines existing datasets and custom code to enable automated assignment of ethnicity on the basis of name with a very high degree of confidence to the majority of names in our dataset. The process simultaneously identifies a subset of candidate names that have a lower probability of being Japanese, so that more thorough research can be carried out manually on these names.[8]

[7] 在加奈陀邦人々名錄/ 大陸日報社編纂, the *All Canada Japanese Expatriate Directory* (Tairiku Nippōsha Hensan, 1941).

[8] It is important to stress that our process does not involve named entity recognition; the process of digitizing documents and capturing records in database tables is such that names are already either tagged as names (in XML documents) or captured into name fields in databases (when harvesting other records). The work described below relates to assigning ethnicity to items already known to be names, and usually tokenized into surname and forenames/initials.



M

村上實	熊本	2202 Columbia Ave., Vancouver, B. C.
村上正記		New Westminster, B. C.
村上清	廣島	P. O. Box 25, Steveston, B. C.
村上龜松	同	Steveston, B. C.
村上馨	愛媛	P. O. Box 424, Steveston, B. C.
村上勝治	熊本	835 E. Cordova St., Vancouver, B. C.
村上勝頼	廣島	R. M. D. No. 1, Ganges, B. C.
村上廣一		Steveston, B. C.
村上光鶴	熊本	Kelowna, B. C.
村上守平	廣島	Ganges, B. C.
村上睦子	山口	Midlandvale, Alberta.
村上光孝	鹿兒島	Mayo Lumber Company, Paldi, B. C.
村上音吉	廣島	P. O. Box 90, Steveston, B. C.
村上正幸	同	P. O. Box 90, Steveston, B. C.
村上佐美藏	同	336 Powell Street, Vancouver, B. C.
村上叡	同	R. R. No. 1, Haney, B. C.
村上茂雄	熊本	P. O. Bx 618, Annacis Is., N. West. B.C.
村上茂	福岡	Woodfibre, B. C.
村上眞吾	和歌山（寫眞業）	438 Main Street, Vancouver, B. C.
村上庄太郎	熊本	658 Alexander St., Vancouver, B. C.
村上昇三	和歌山	121 Heatley Avenue, Vancouver, B. C.
村上惣吉	山口	229 Ewen Ave., New Westminster, B. C.
村上末喜	熊本	1520 West 3rd Ave., Vancouver, B. C.
村上忠雄	廣島	Royston, R. R. 1, Cumberland, B. C.
村上威德	同	P. O. Box 431, Steveston, B. C.
村上爲夫	同	Salt Spring Island, Ganges, B. C.
村上孝式	愛媛	P. O. Box 235, Steveston, B. C.
村上種吉	同	Kildala Cannery, Rivers Inlet, B. C.

Figure 1. A page from the Zai Kanada hōjin jinmeiroku [All Canada Japanese Expatriate Directory] showing entries with the surname "Murakami"

2. Previous Work and Approaches

Attempts to determine ethnicity by surname date back to the 1950s, and there is a considerable amount of existing published work on approaches and processes. Much of it is in the field of medicine, where race and ethnicity are important factors in susceptibility to disease, responses to medication, access to healthcare, and other contexts. Fiscella and Fremont (2006) provide a useful overview of previous attempts to use surname analysis, and are healthily skeptical of algorithmic approaches in general—"[a]s a practical matter, no method for obtaining race and ethnicity data can be entirely accurate or bias free" (1483)—and of surname analysis specifically, although they note that curated name lists can be quite effective for some ethnicities in particular: "Lauderdale and Kestenbaum's name list, derived from Social Security records, and validated using the 1990 Census, showed sensitivities ranging from 74 percent for Vietnamese to 29 percent for Filipinos and positive predictive values ranging from 92 percent for Japanese to 76 percent for Chinese."[9] The best predictivity score reported is for Japanese, but even a rate of 92 percent, while presumably very useful in the case of studies involving statistical analysis of large populations, would be unacceptable for our particular project, as we seek, for example, to document the material impacts of every one of the government's forced property sales in a relatively small neighborhood in East End Vancouver.

Shah et al. (2010) carried out a large-scale project to identify South Asian and Chinese residents of Ontario from the general population using name lists which was also relatively effective: "both lists performed extremely well when compared against self-identified ethnicity: positive predictive value was 89.3% for the South Asian list, and 91.9% for the Chinese list. Because surnames shared with other ethnic groups were deliberately excluded from the lists, sensitivity was lower (50.4% and 80.2%, respectively)." Again, such scores are impressive but still fall below our aspiration. As Abrahamse, Morrison, and Bolton (1994, 385) note, "the use of surnames to identify people's origins has inherent conceptual and technical limitations, which confine such use to statistical rather than individual-level applications." Name lists do form part of our process as described below, but are combined with other methods.

3. Factors in Our Favor

There are a number of reasons why the problem for Japanese names, and for this specific population of migrants, is uniquely tractable.[10]

[9] Fiscella and Fremont (2006), 1491. Lauderdale and Kestenbaum's work is discussed in section 4.

[10] The same is not true of the other ethnic categorizations we apply to our data, "Other Asian" (i.e. non-Japanese Asian) and "Other" (the remainder, into which those of European origin fall). For these, we are forced to use a much less sophisticated process; some name lists do exist which can help with the "Other Asian" category, but we cannot apply linguistic constraints as we can with Japanese, and it is harder to create curated name lists for categories which are as broad as this. As a result, the process for these ethnicities is less accurate, involves more human intervention, and takes more time.

3.1 The Population

The source Japanese population itself is relatively mono-ethnic. There was very little voluntary inward migration to Japan prior to World War II (Shin 2010). Although by 1940 there were 1.3 million registered aliens in Japan (Cornelius, Martin, and Hollifield 1994, 376), nearly 99 percent of these were Korean and Chinese conscripted laborers, who can be assumed not to have been able to emigrate from there to Canada during the period in question. It is not unreasonable to conclude, therefore, that virtually all of the individuals coming to Canada from Japan were ethnically Japanese. [1 1]

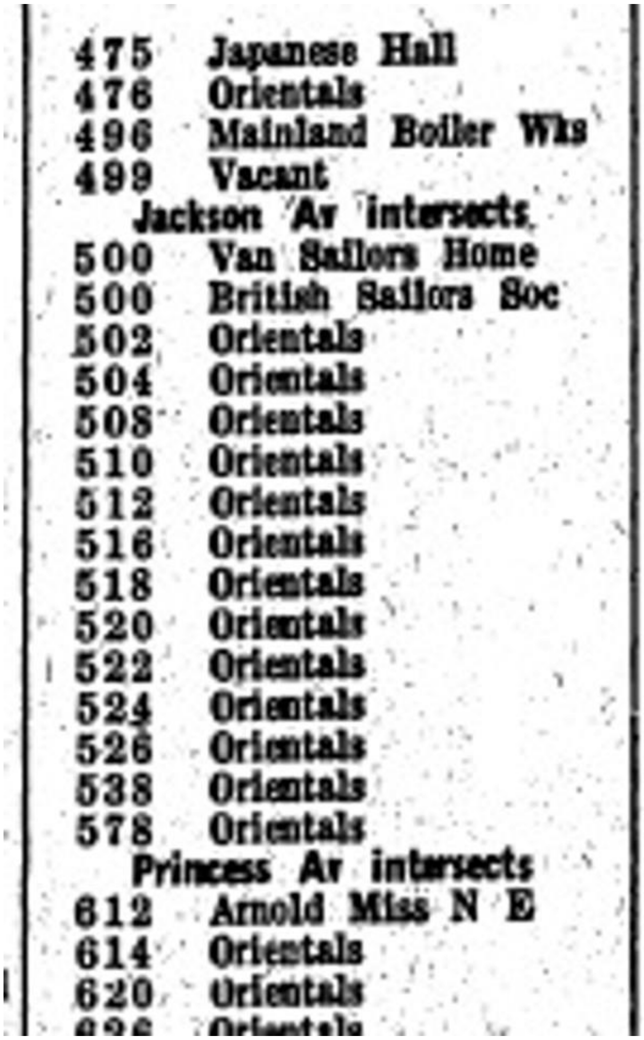
Not only were emigrants from Japan to Canada ethnically homogeneous: they were also drawn from a relatively small number of source locations in Japan. Kobayashi has shown that "Japanese emigrants to Canada have been to a remarkable degree concentrated from two very small geographical areas, the village of Mio in Wakayama Prefecture, and a small cluster of three villages—Hassaka, Oyabu and Kaideima—in Shiga Prefecture" (Kobayashi 1986, 15). Between a third and a half of all emigrants to Canada originated in these villages. Further, Kobayashi's investigation of emigration from the village of Kaideima in Shiga Prefecture to Vancouver has shown that 60–80 percent of all households in the village were involved in emigration (Kobayashi 1984) and that "the village is dominated by eight surnames that represent over seventy percent of the households" (1986, 116). Furthermore, the Hayashi-Lemieux agreement of 1908 severely restricted Japanese immigration to Canada, "effectively limiting immigration after 1907 to the relatives of those who had already become established in Canada" (1986, 11). We can therefore

[1 1] Lie (2001) makes the point that "[a] neglected aspect of the Japanese diaspora, incidentally, is its ethnic heterogeneity, especially the overrepresentation of Burakumin and Okinawans among Meiji-era emigrants, and Korean Japanese and mixed-race people more recently" (23). However, Lie makes finer distinctions with regard to ethnicity than we do. The former social outcasts of the Tokugawa era known as "burakumin" are of course native Japanese characterized by membership of a feudal caste, rather than ethnically distinct, and Lie specifies that he is treating them as "an ethnic group" based not on descent but on "identification and discrimination" (3–4). The Okinawans, although ethnically distinct, make up less than 1% of the Japanese emigration to Canada (see Kobayashi et al., 2018). Okawa (2018) discusses pre–World War II Japan as a "multi-ethnic space."

expect—and in fact we see—high concentrations of a relatively small range of surnames; among the 2,920 Japanese surnames in Latin characters which appear in our current dataset there are only 754 distinct names (3.9 individuals for each surname), whereas in the population as a whole (including Japanese), there are 14,870 distinct surnames from a total of 26,815 (1.8 individuals per surname). Common names include Maikawa/Maekawa (前川): 50 instances; Fukui (福井): 30 instances; Aoki (青木): 27; and Murakami (村上): 22.

This relatively tight-knit population also documented itself in great detail. As Kobayashi notes, there is "a vast wealth of archival data, perhaps not equalled for any other Canadian immigrant society" (Kobayashi 1988, 356). This includes community and national directories such as the directory of Japanese residents in BC published in June 1941 [ビーシー州日本人電話帖兼アドレス便覧].[12] In addition, public city directories of the period are often far more detailed than we would expect modern telephone directories (for instance) to be: a single entry in a Vancouver City directory may include not only the name, address, and telephone number of the householder, but also their occupation, their employer, their spouse's name, and whether they own property they live in. The usefulness of these records is sadly undermined, however, by the fact that Japanese and Chinese residents are sometimes virtually ignored; the single identifier "orientals" appears for many addresses in the Powell Street area in the 1930 directory (see fig. 2). Luckily this is not the case for most years.

[12] ビーシー州日本人電話帖兼アドレス便覧 [BC Shu Nihonjin Denwa Cho Ken Adoresu Binran, BC Japanese Telephone Directory and Address Listing; Vancouver, BC], 1941. This local community phone directory was sponsored by the Union Fish Store at 469 Powell Street, Vancouver. It is held by the Nikkei National Museum (accessed August 31, 2018, http://nikkeimuseum.org/www/item_detail.php?art_id=A3145).



475	Japanese Hall
476	Orientals
496	Mainland Boiler Wks
499	Vacant
Jackson Av intersects	
500	Van Sailors Home
500	British Sailors Soc
502	Orientals
504	Orientals
508	Orientals
510	Orientals
512	Orientals
516	Orientals
518	Orientals
520	Orientals
522	Orientals
524	Orientals
526	Orientals
538	Orientals
578	Orientals
Princess Av intersects	
612	Arnold Miss N E
614	Orientals
620	Orientals
626	Orientals

Figure 2. An extract from the 1930 *Wrigley's British Columbia Directory* showing addresses on Alexander Street in Vancouver

3.2 The Language

3.2.1 Romanization and Phonology

One subset of our materials has no ambiguity at all with regard to ethnicity: the local and national community directories created by Japanese Canadians for their own use. These documents are name/address lists in Japanese as shown in figure 1, and there is no difficulty with assigning Japanese ethnicity to any name appearing in them. In all other records, though, Japanese names are presented in romanized form. A number of different systems of romanization of Japanese exist (Hepburn, Revised Hepburn, Nihon-shiki, Kunrei-shiki, JSL, and more recently Wāpuro), but in the records we are dealing with, no single rigorous approach to romanization can be assumed.[13] Many forms of romanized Japanese names we encounter were transcribed by non-Japanese speakers filling in forms or collecting address information for city directories, so the majority are transcribed phonetically; this means that their effective romanization forms tend to approximate the Hepburn system (although long vowels are not normally marked). There are sporadic exceptions to this; for example the name "Murakami Katsuji" is recorded in the Zai Kanada 1941 directory as 村上勝治 (see fig. 1), but in the Vancouver City Directory for the same year, the forename is romanized as "Katsugi" ("g" in place of "j," which would be pronounced /katsugi/ rather than /katsuji/). Overall, though, variations in romanization such as this result in names which still fit within the range of possible Japanese names; there is a name "Katsugi" in Japanese (加次 or 勝木), although it is not a forename.

Prevalent forms of romanization, then, reflect quite closely the actual phonological structure of the language. Furthermore, the phonology of Standard Japanese is highly distinctive and very constrained. The language has a relatively small phoneme inventory which combines with quite severe phonotactic constraints to result in a very small syllable inventory. This means that for Japanese, it is quite practical (as it would not be for many other languages) to create an orthographic template in the form of a regular expression (a textual search pattern) in which possible Japanese syllables are enumerated, and a word, defined as a sequence of one or more syllables, can be tested against this expression to determine whether it is a plausible Japanese form or not. This is a useful component in our algorithm. We are able to detect with some reliability all name forms which conform to the phonological patterns of Japanese and exclude the vast majority that do not.

3.2.2 Distinctness of Japanese Names as Opposed to Those from Other Linguistic Communities

Many surnames are highly ambiguous with regard to ethnicity. "Park," for example, is a common surname both in Korea and in English-speaking countries; "Chong" may be Korean or Chinese, "Chung" may be Chinese or Vietnamese, "Mann" may be European or Indian, and "Lee" is strikingly broadly distributed, being common in English-speaking

[13] See Reischauer (1940) for a brief account of the dominant systems of romanization prior to 1940.

cultures, Chinese, and Korean. Such examples pose severe difficulties for any algorithmic process attempting to determine ethnicity in the absence of other evidence. Fortunately, there are relatively few ethnically Japanese names which overlap with names from other cultures, and in all the cases we have encountered, it has proved possible to determine from other information whether the name is in fact Japanese. The surname “Jin” (仁, 陣 and many other possible characters) is one example shared with Chinese and Korean, but only one instance of it appears in our data at the time of writing. “Kose” is a possible Japanese surname, but it appears in our Vancouver data coupled with a non-Japanese forename, and only after the Japanese community was moved out of Vancouver; similarly, “Aho,” which appears three times, is joined in two cases with the Scandinavian forename “Helvi” and once with “Edna.” With the exception of a small set of such cases, most conventional Japanese names, especially those with more than two syllables such as “Sugimura” or “Takahashi,” are unambiguous with respect to language.

4. Resources and Tools

4.1 O’Neill’s *Japanese Names*

Our primary resource for individual non-automated research into whether a candidate name could be assumed to be Japanese is P. G. O’Neill’s venerable (and long out of print) reference work *Japanese Names: A Comprehensive Index by Characters and Readings* (O’Neill 1972). This work includes 13,500 surnames and 11,000 personal names as well as historical and geographical names. It was compiled from a list of prior reference works, and supplemented by less formal resources such as the Tokyo telephone directory (viii). As the preface says, it “aims to be comprehensive... but cannot in the nature of things hope to be complete” (vii), and indeed we found many instances of unambiguously Japanese names (such as those included in the community directories in kanji) which were not included in O’Neill. In addition, this resource is not available electronically, which means that it could not be integrated into the automated part of our process. Nevertheless, because of the book’s clear organization and indexing (by character, by reading, and by radical) and its scholarly authority, O’Neill remained an important part of the manual research and verification component of our process as described below.

4.2 ENAMDICT: A Promising But Problematic Resource

The ENAMDICT, produced by the Electronic Dictionary Research and Development Group,[14] is a very large resource containing "Japanese proper names; place-names, surnames, given names, company names, names of artistic and literary works, product names, etc."[15] which originally formed part of the larger EDICT dictionary, but was forked into a separate resource because of its size (it contains nearly 940,000 items at the time of writing). The file is available in XML format. On the face of it, this should be a tremendous asset for the purpose of ethnicity determination based on name. Unfortunately, the dictionary is more complete than we would wish, in that it contains non-Japanese as well as Japanese names (internationally known politicians, celebrities, and other famous people whose names happen to appear in Japanese texts). We generated a subset by discarding all but the surnames and given names, and also excluding all those entries whose syllabic transcription is in Katakana (a syllabary primarily used for foreign words in Japanese), retaining only those in hiragana; this yielded a set of around 322,000 entries, which was still substantial, but which presumably largely excludes non-Japanese names. The very comprehensiveness of this resource meant that it included many rare Japanese names we would not expect to encounter, as well as many whose romanizations collide with non-Japanese names such as the surnames "Dee" (出江), "Hansen" (飯泉), and "Henri" (片理), and the forenames "Anne" (晏音) and "Ben" (勉). In the absence of any information on prevalence or distribution, the value of a "hit" in the ENAMDICT resource was considerably reduced, and we ended up assigning it a very low weight in the overall calculation to reduce the incidence of false positives.

4.3 Jisho.org: An Online Dictionary of Japanese

The online resource http://jisho.org is an ambitious and very usable project which aims to provide a comprehensive guide to Japanese, and it includes material from the ENAMDICT project, as well as many other resources including example sentences and audio files. It is maintained by Kim Ahlström, Miwa Ahlström, and Andrew Plummer, and is aimed mainly at language learners. Its primary value for our purposes lies in convenience

[14] Accessed August 31, 2018, http://www.edrdg.org/. The Electronic Dictionary Research and Development Group, established at Monash University in 2000, is home to the ENAMDICT, but it continues to be maintained by its original creator, Jim Breen.

[15] Accessed August 31, 2018, http://www.edrdg.org/enamdict/enamdict_doc.html.

and speed of access; looking up a name or a name kanji in the Jisho is a trivial task compared with finding it in (say) O'Neill. It also allows searches to be limited to names only. Searches are URL-based (HTTP GET requests), meaning that it would be possible to script lookups as part of an automated system, but we did not do that for two reasons: first, the service does not offer a formal API description, and is therefore presumably not expecting to be afflicted by a sustained barrage of HTTP requests from a remote source; and automated parsing of the results of a query were not really practical, since a single query will typically yield many results, which may include not only Japanese but also non-Japanese names, placenames, and so on (as they would be written in Japanese). For instance, a query for "Ben" produces instances of the full names of individuals (writers and actors), a female given name (without kanji), and many non-Japanese names and place names. Human intelligence is required to derive useful information from such results. Jisho.org therefore became a key resource in manual verification.

4.4 Lauderdale and Kestenbaum's Surname Lists

In 2000, Lauderdale and Kestenbaum published a set of surname lists for six Asian-American ethnicities (China, India, Japan, Korea, the Philippines, and Vietnam), which they created based on US Social Security Administration records of applications for social security cards by individuals born outside the United States prior to 1941. These records, although they do not include ethnicity information directly, do include country of birth, "a viable proxy for ethnicity for Asian Americans" (286). This information was combined with records related to the Medicare program, which include surname and race. Using these datasets, they created a set of 24 surname lists, four for each target ethnicity, one pair of which is conditional on "race information" (Asian ethnicity identification based on external factors), and one pair which is unconditional. In each case, names are judged "predictive" or "strongly predictive." "The progression from predictive to strongly predictive improves accuracy, but at the cost of reduced coverage" (288). The surname lists are (by design) incomplete; names occurring fewer than five times in the dataset were not included, and names which are shared across ethnicities tend to be eliminated: "a surname was included if *at least* 50 percent of persons with that surname were associated with an origin (e.g., Korea) and *less than* 50 percent with any of the other countries. These lists we call 'predictive'. A subset of names from each list was further identified as 'strongly predictive' by using a threshold of 75 percent" (288). This means that all of the most common Japanese surnames in our set are included (although “Maekawa” appears only in this form, and not in the form “Maikawa” which is more common in Vancouver), but of the 685 distinct surnames identified in our dataset as Japanese at the time of writing, only 449 appear in the Lauderdale and Kestenbaum Japanese lists; among the 236 missing are some relatively ordinary surnames such as “Imokawa” (五百川), “Kawashita” (河下), and “Matsubayashi” (松林), alongside some unusual spellings and romanizations (such as “Maikawa”). Some of the names in the list (such as “Chaki” [茶木]) are most probably excluded from Lauderdale and Kestenbaum’s data because they are shared with other ethnicities and therefore have low predictive value, but such cases are relatively few; it seems most likely that the sources of Japanese immigration to the United States (and therefore the concentration of surnames in the Japanese American community) differ substantially from the places of origin which contributed to the Japanese Canadian community. This dataset, then, can provide a useful contribution to the effectiveness of our algorithm, but only in combination with other resources.

5. Approach and Implementation

5.1 Principles

We adopted four core principles to guide our implementation of the algorithm:

Principle 1: Make a firm decision on every potential Japanese name.

For other more problematic ethnicities and ethnicity groupings, we have been forced to create “provisional” categories for names which are ambiguous, but we want to avoid this if possible with Japanese names. Given the nature of the language, the immigrant community, and the available documentation, as discussed above, this is definitely feasible, and so far, we have been able to make a firm decision on every candidate name.

Principle 2: Automate where possible, but allow no false positives.

The objective of the algorithm is to automate the assignment of Japanese ethnicity as far as possible, but we do not tolerate any false positives (non-Japanese names assigned Japanese ethnicity). For this reason, the algorithm is tuned and the results checked such that only names with a high degree of certainty are matched; a subset of names with lower certainty are identified and flagged for manual research.

Principle 3: Fall back to human research where required.

There are multiple sources and resources which can be used to do manual verification of names for which the automated process does not generate the level of certainty we require for a definitive assignment. The process of manual checking is described below.

Principle 4: Feed research results back into the algorithm.

When human research has definitively determined whether a given name is or is not Japanese, the results of that research can be integrated into the algorithm through, for example, curated lists of attested Japanese and non-Japanese names from the community. This process helps to refine the algorithm itself, reducing the number of future problem results requiring research.

5.2 The Process

5.2.1: Manual Checking

The entire process described by this paper begins from the basis of human research on individual names. A so-called “gold” set of Japanese names that we trusted to be correct, drawn from our actual datasets (and therefore reflecting the actual makeup of the Japanese Canadian community in question), allowed us to begin evaluating the various lists and other resources to be incorporated into the automated process. Our initial dataset was the database of land titles applying to properties in the Powell Street (Vancouver) area, which was known as Japantown (日本町) in the years leading up to the dispossession. This database contained 7,469 names of individuals who appeared as owners (buyers or sellers) on title documents.[16]

[16] Inevitably, the same individuals from the Japanese Canadian community are recorded in multiple source documents: community directories, city directories (by person and by address), fishing boat licenses, and land title records include the same people repeatedly. While in the majority of cases it would be practical to determine whether a record in one source refers to the same individual as a record in another source, this is often more difficult than one might think. For that reason, in our algorithmic/curated ethnicity assignment process, we make no effort to identify individuals, and instead focus on assigning ethnicity to names alone. For various specific research questions addressing subsets of the data, such as those focusing on the economic consequences of the internment and property seizures on a specific family or a specific group, we do of course identify individuals, putting additional resources into the necessary research to achieve this.

The first stage was to identify all the candidate Japanese surnames in this dataset. An initial search was done using a regular expression designed to capture most plausible romanized versions of Japanese syllables:

```
^(a|i|u|e|o|ka|ki|ku|ke|ko|kyu|kyo|ga|gi|gu|ge|go|sa|sha|shi|
su|se|so|sho|za|ji|zu|ze|zo|ta|cha|chi|chu|tsu|te|to|da|de|do
|na|ni|nu|ne|no|ha|hi|fu|he|ho|hya|hyu|hyo|ba|bi|bu|be|bo|pa|
pi|pu|pe|po|ra|ri|ru|re|ro|ma|mi|mu|me|mo|wa|ya|yi|yu|ye|yo|n
)+$
```

This regular expression served to identify names consisting of sequences of one or more of these syllables; it consists of a list of all the possible syllables in standard Japanese in the forms in which they are likely to appear in Latin script, and specifies that a name consist of one or more of these syllables and nothing else.[17] This divided the name list into two sets, which were then examined manually. Each name that did NOT match the regular expression was checked to ensure that it was not a Japanese name. This process enabled us to discover a handful of cases in which names had been mistranscribed in our original data, or had been misspelled in the original documents. For example, the name "Suzuki," attested as such from other records, had been written in one document as "Sukzi"; the former is a sequence of Japanese syllables ("su," "zu," "ki") whereas the latter is not ("kzi" is not a possible syllable in Japanese). We did not find any examples of confirmed Japanese names that did not match the regular expression, apart from such errors.

Next, we looked at the names matching the regular expression, and checked each one individually using the following process (which was developed and refined as we worked):

The surname and (if available) the forename(s) were checked in O'Neill's *Japanese Names*. If both appeared, and there was no other reason to suppose the name might not be Japanese, it was assigned to Japanese ethnicity. If the surname appeared in O'Neill, but the forename was only an initial, we accepted the name as Japanese if any other examples of the same surname in confirmed names already existed in our data.

[17] Obviously, there are more "efficient" ways to frame this regular expression, but this long-winded approach is easier to understand and maintain, and there appears to be no time penalty incurred in execution from its verboseness.

In cases where the surname did not appear in O'Neill precisely in the form it took in our data, we broke it down into plausible components. For example, the name "Umakoshi," which does not appear in O'Neill, most likely breaks down into Uma + koshi. We then identified kanji characters in O'Neill with the same pronunciation which appear in other surnames; for instance, in this case, we found 馬島 (Umashima) and 越山 (Koshiyama). We then constructed plausible combinations of name kanji which could be pronounced in the same way as the candidate name, and searched other resources such as online telephone directories and dictionaries (*Jisho.org* provides a useful name-search function) for instances of any of those forms as unambiguous surnames. In this case, we were able to find "Umakoshi" (馬越) appearing many times in the Tokyo telephone directory. We also checked our community directories of Japanese residents, finding an instance of Tasoji Umakoshi residing at 656 East Cordova Street in 1941. Finally, we checked other sources such as immigration records, passenger lists and so on, for additional confirmation. Where forenames were recorded, a similar process was applied. Once a surname or forename was confirmed to be Japanese, it was added to a list of Japanese surnames or forenames known to exist in our dataset. These lists were used as part of the later algorithmic process.

In our dataset of 7,469 names of property owners, we identified 594 as Japanese; among these were 201 distinct surnames. However, an additional 324 distinct surnames matched the regular expression, since many non-Japanese names happen to be constructed of syllables that fit the pattern. Names such as "Beaton," "Page," and even "Robinson" fit the pattern, since "be," "a," "to," "n," "pa," "ge," "ro," "bi," and "so" are all plausible Japanese syllables (while most syllables in Japanese are open-ended, the nasal "n|m|ŋ" alone may constitute a syllable). Each such case where no plausible Japanese name exists, or where other evidence (such as forename) confirms that the name is not Japanese, was added to a list of confirmed non-Japanese names. This includes cases such as "Zen," which is a possible (although not very common) Japanese surname, but is also an Italian surname; in our data, it appears with the forename "Giuseppe," but it never appears in any of the Japanese Canadian records. The resulting curated list of non-Japanese names later became a component in our automated algorithm.

This process, while highly accurate, is very time-consuming. Significant difficulties for manual checking arise out of the complex relationship between the Japanese (kanji) forms of names and their phonological realization (and hence their romanization, as found in our English-language source documents). For example: The name romanized as "Ide" is attested in the *Jisho.org* database as being written in the following ways: 伊出, 伊達, 井手, 井上, 井出, 井田, 居出, 射手, 出田, and 生出; O'Neill (221) adds also 出 and 井代. In attempting to confirm the Japanese identity of the surname "Ide" by examining Japanese language records, then, we must search for twelve different forms. Conversely, a single one of these forms, 井上, is attested by *Jisho.org* with the pronunciations Iue, Ikami, Igami, Inai, Inae, Ine, Inei, Inoue, Inouezaki, Inouta, Inoe, and Ueno, meaning that it might possibly appear in an unexpected location in a Japanese-language directory if positioned under one of its alternative pronunciations. Such difficulties in dealing with our initial dataset provided considerable motivation for developing a faster, more automated process that could accurately classify a large proportion of input names, leaving only a minority of problem cases to be checked in this exhaustive way.

5.2.2 The Algorithm

The process, as laid out the diagram in figure 3, essentially works by assigning a score to any romanized name (consisting of surname and forename) based on a number of tests which use some of the resources described above. The algorithm is tuned continuously based on results from each dataset we process. At the time of writing, this is how the calculation works:

First, the surname is checked against our curated list of names known to be non-Japanese from our study population. If it matches one of these names, the score is zero and the process ends.

Next, the surname and forename are both checked against the regular expression for Japanese syllable sequences. A surname match scores 3 points. If (and only if) the surname matches, and the forename is not an initial, it is also checked, and a match scores an additional 2 points.

The surname is then checked against a specially processed version of Lauderdale and Kestenbaum's lists of Asian surname probabilities. Our version of their database weights the unconditional probability value three times higher than the probability value which is conditional on Asian origin having been established by other means. This test returns a score of between 0 and 8.

If the surname appears in our curated list of attested Japanese surnames from the target population, a score of 3 is added.

If the forename appears in our curated list of attested Japanese forenames, then another 3 is added.

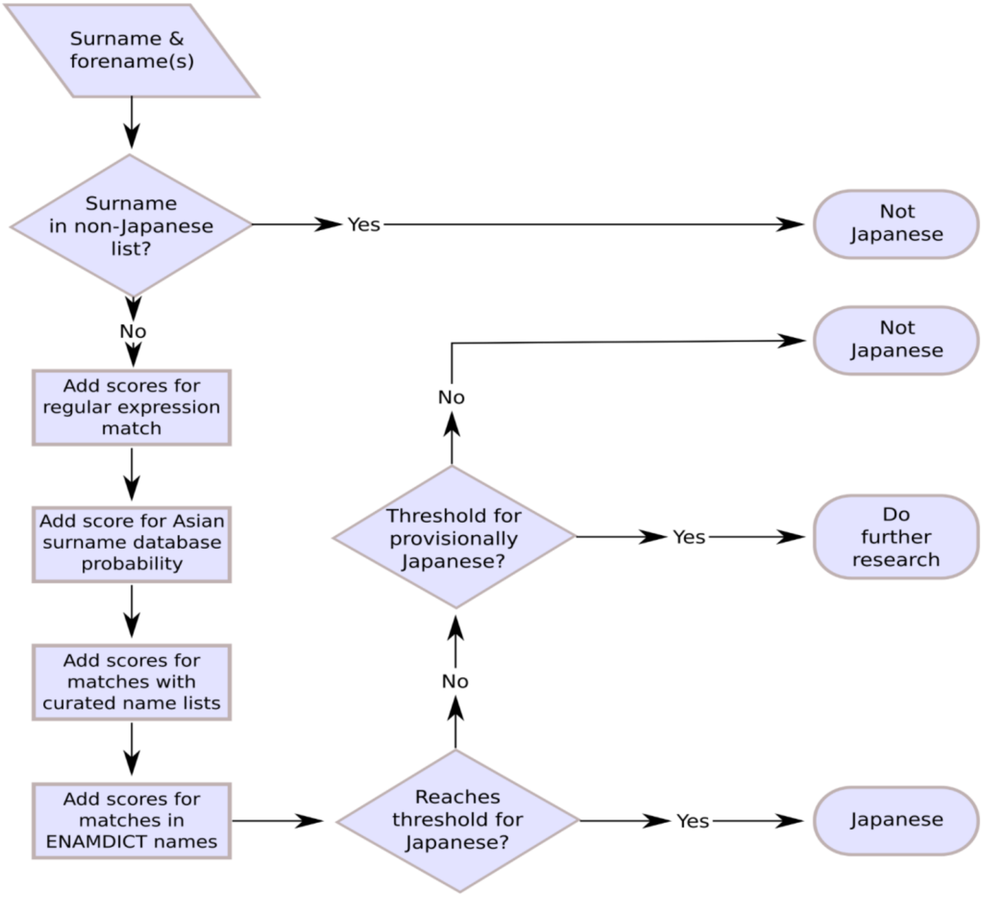

Surname & forename(s)
Surname in non-Japanese list?
Yes
Not Japanese
No
Add scores for regular expression match
Add score for Asian surname database probability
Add scores for matches with curated name lists
Add scores for matches in ENAMDICT names
Reaches threshold for Japanese?
Yes
Japanese
No
Threshold for provisionally Japanese?
Yes
Do further research
No
Not Japanese

Figure 3. The algorithm compiles a score for each name using a range of resources.

Finally, the surname and forename are both checked against a curated version of the ENAMDICT, which (as described above) is processed to contain only native Japanese surnames and forenames. A match for a surname scores one point, and a match for a forename also scores one point. These values are low because, as discussed above, despite the effort to strip out non-Japanese names from the ENAMDICT data, it is so comprehensive that it includes a large number of very obscure Japanese names that overlap with non-Japanese names. If we increase the weight given to this factor, the result is a larger number of false assignments to the Provisional Japanese category; if we eliminate the factor completely, the result is that one or two names that should be assigned to Japanese ethnicity instead end up in the Provisional Japanese category.

The sequence of operations by which the score is accumulated is not significant (figure 3 simply represents the sequence that we currently use). The result of the process is a score of between 0 and 21. Two tests are then applied to determine whether the name falls within a range where we can confidently assert that it is Japanese, or whether it can be deemed “provisionally Japanese,” meaning that further research is required, following the manual process described above. The scores from each factor, along with the cutoff scores, have been tuned steadily based on the results we see from each new dataset; currently a score of six or above generates a confident assertion that a name is Japanese. The cutoff for the provisional Japanese label is rather more complicated, because it involves some additional calculation of the probability that a name is Asian in origin; this depends not only on Lauderdale and Kestenbaum’s data for other Asian ethnicities but also on other curated name lists and resources not described here.

5.2.3 Outcomes

When this process is run against a document set or a database, there are two specific outputs. First, there is a list of all the names tested, sorted into their proposed ethnicity assignments. This includes enough identifying information (such as database record IDs) to enable the researcher to find the original source data easily. The researcher can then scan the lists of confident Japanese and non-Japanese ethnicity assignments to check for any anomalies, and then begin working on the provisional category, which requires detailed investigation. As each such name is investigated and a determination made, curated name lists can be updated and other modifications made to the algorithm such that in the future that name should be assigned to the appropriate set with high confidence. The algorithm is typically re-run periodically as these names are dealt with, to ensure that no side effects of such changes are seen.

The other output from the process differs depending on whether the input dataset is an SQL database (such as our land title record set) or a set of XML documents (as our transcribed street directories are). In the case of a database, the process generates a set of SQL commands which can be run against the database to assign the appropriate ethnicity automatically to the names in the output. This means that once all the decisions have been made and the problem cases dealt with, the database can be updated to assign the appropriate ethnicity to all the names in a single operation. If the input dataset is a collection of XML documents, updated versions of the original documents are created automatically, with the ethnicity assignments encoded as attributes on the tagged names themselves in the documents. In both cases, although the assignment of ethnicity for specific names is automated, it is only enacted at the end of the research process, and individual names that remain problematic can be assigned to a special provisional category pending further investigation if necessary.

The following is an example of how the process works in practice with a specific dataset which at the time of writing is just being processed. This is an SQL database containing land title records for the Maple Ridge area of British Columbia, one of our study areas in which Japanese farmers were active prior to the dispossession. Results of the first run of the algorithm:

Total names processed:	3,199
High-confidence Japanese names found:	119
Low-confidence Japanese names found:	24
Names assigned to other or no ethnicity:	3,056

After an hour or so of research on the low-confidence names (see 5.2.1 “Manual Checking”), the next run of the algorithm produced:

Total names processed:	3,199
High-confidence Japanese names found:	119
Low-confidence Japanese names found:	4
Names assigned to other or no ethnicity:	3,076

The remaining four difficult cases will be subject to further research.

6. Conclusions and Future Work

Although it is clear that determining ethnicity through name analysis is inherently unreliable and problematic, and normally suitable only for large-scale low-precision data analysis applications, our experience has shown that given the particular resources available to us, along with the many unusual features of this particular population, semi-automated assignment of Japanese ethnicity to names from our target population is practical, helpful, and time-saving. This process enables us to deal rapidly with high-confidence cases (both those which are definitely Japanese and those which are definitely not) and focus our precious research time on the much smaller number of cases requiring additional research; then the results of that research can be fed back into the algorithm to improve the outcomes for future cases. Above all, the use of a principled, documented process gives us an audit trail for each name we deal with; when a name is deemed to be Japanese, we can demonstrate exactly why, and cite the resources on which the decision was based. When we pass on data to other researchers on the project, they can be confident that ethnicity assignments are trustworthy.

Such a process may be generalizable to other situations in which historians are dealing with ethnically distinct subpopulations that are internally relatively homogeneous, do not intermarry much with other ethnicities, and in particular, whose names are typically characterized by linguistic features which distinguish them from other groups. Icelandic surnames, for example, are almost invariably patronymics ending in “-son” or “-dóttir”, while there are fewer than three hundred Korean surnames, and only five family names account for over half the Korean population. However, with other datasets, such an approach would be much more difficult; a very diverse range of ethnic groups use names derived from major religions such as Christianity and Islam, making them particularly difficult to distinguish. In such cases, curated name lists may still be helpful.

As part of our work, we also classify non-Japanese names into “Other Asian” and “Other” categories, and that part of the process is far more problematic and time-consuming for a variety of reasons, including the overlap of common surnames across ethnicities, the relative paucity of useful name lists, and above all the fact that we do not have access to the rich array of other documentation such as the community directories listing names in the original source languages. We rely more heavily on the data from Lauderdale and Kestenbaum, which includes lists for names originating in China and India (the two other primary sources of Asian migration in our target populations). We also maintain curated lists of other Asian names confirmed in the target population. But the results include a much higher proportion of assignments to the “Provisional Other Asian” category, requiring more human research to resolve them.

Now that we have a dataset of considerable size with reliable ethnicity assignment, we are considering the possibility that a productive alternative approach may be available through machine learning. This would involve training a classifier on the existing data and applying it to new datasets. Some drawbacks are immediately obvious. Machine classifiers typically work on long texts which they tokenize into words, using the words (or stemmed versions) as predictive features, but our name data consists only of one or two words, so this obviously will not work. However, we could tokenize the input names into 2- and 3-letter segments instead; this might allow the classifier to respond to graphemic subsequences characteristic of the names, in the same way that the regular expression component of our algorithm does. We do not expect such an approach to be as reliable as the current algorithm, but we intend to test it; even if the results are not acceptable, they may be good enough to be included as an additional predictive factor in the existing algorithm.

Our efforts to classify names by ethnicity for our research purposes have benefited from the support of the Partnership Grant program of the Social Science and Humanities Research Council of Canada, which provides generous funding and demands interdisciplinary cooperation. As a result, historians on our project have been able to work with programmers to develop a systematic and rigorous approach to a process that—though fundamental to quantitative analysis (and even to mere descriptive counting of the members of the communities being studied)—often suffers from the ad hoc and very likely imprecise determinations of individual researchers working with their own data sets. As we continue to refine our processes for name classification, these will be made freely available through the digital resources of the Landscapes of Injustice project, enabling historians working alone to benefit from the tools we have developed, and perhaps to initiate wider discussion of and collaboration on systematic approaches to an aspect of research method that many historians may have taken for granted.

Acknowledgements

I would like to thank Dr. Jordan Stanger-Ross for his detailed and very helpful feedback on an earlier version of this article.

References

Abrahamse, Allan F., Peter A. Morrison, and Nancy Minter Bolton. 1994. “Surname Analysis for Estimating Local Concentration of Hispanics and Asians.” *Population Research and Policy Review* 13, no. 4 (December 1994): 383–98. doi:10.1007/BF01084115.

Cornelius, Wayne A., Philip L. Martin, and James Frank Hollifield. 1994. *Controlling Immigration: A Global Perspective.* Stanford, CA: Stanford University Press, published in association with the Center for U.S.-Mexican Studies, University of California, San Diego.

Fiscella, Kevin, and Allen M. Fremont. 2006. “Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity.” *Health Services Research* 41(4p1): 1482–1500. doi:10.1111/j.1475-6773.2006.00551.x.

Kobayashi, Audrey. 1984. “Emigration to Canada and Development of the Residential Landscape in a Japanese Village: The Paradox of the Sojourner.” *Canadian Ethnic Studies* 16(3): 111–31.

Kobayashi, Audrey. 1986. “Social Consequences of Regional Diversity among Japanese Immigrants to Canada: A Preliminary Review.” In *Asian Canadians: Contemporary Issues*, edited by K. Victor Ujimoto and Josephine Naidoo. Ontario: N.p.

Kobayashi, Audrey. 1988. “Regional and Demographic Aspects of Japanese Migration to Canada.” *The Canadian Geographer* 32(4): 356–60.

Kobayashi, Audrey. 2017. “Historical Geography in the Service of Social Justice.” Distinguished Historical Geography Lecture, Annual Meeting of the American Association of Geographers, Boston, April 4.

Kobayashi, Audrey, Reuben Rose-Redwood, Sonja Aagesen, and the Landscapes of Injustice Research Collective. 2018. “Exile: Mapping the Migration Patterns of Japanese Canadians Exiled to Japan in 1946.” *Journal of American Ethnic History 37(4): 73–89*.

Lauderdale, Diane S., and Bert Kestenbaum. 2000. “Asian American Ethnic Identification by Surname.” *Population Research and Policy Review* 19:283–300. doi:10.1023/A:1026582308352.

Lie, John. *Multiethnic Japan.* 2001. Cambridge, MA: Harvard University Press.

Okawa, Eiji, and the Landscapes of Injustice Research Collective. 2018. “*Japaneseness* in Racist Canada during the First Half of the Twentieth Century: Immigrant Imaginaries during the First Half of the Twentieth Century." *Journal of American Ethnic History* 37(4): 10–39.

O'Neill, P. G. 1972. *Japanese Names: A Comprehensive Index by Characters and Readings.* 1st ed. New York: John Weatherhill.

Reischauer, Edwin O. 1940. "Rōmaji or Rōmazi." *Journal of the American Oriental Society* 60(1): 82–89. doi:10.2307/594565.

Shah, Baiju R., Maria Chiu, Shubarna Amin, Meera Ramani, Sharon Sadry, and Jack V. Tu. 2010. "Surname Lists to Identify South Asian and Chinese Ethnicity from Secondary Data in Ontario, Canada: A Validation Study." *BMC Medical Research Methodology* 10(1), article no. 42. doi:10.1186/1471-2288-10-42.

Shin, Hwaji. 2010. "Colonial Legacy of Ethno-Racial Inequality in Japan." *Theory and Society* 39(3–4): 327–42. doi:10.1007/s11186-010-9107-3.

Stanger-Ross, Jordan, and the Landscapes of Injustice Research Collective. 2016. "Suspect Properties: The Vancouver Origins of the Forced Sale of Japanese-Canadian-owned Property, WWII." *Journal of Planning History* 15(4): 271–89. doi:10.1177/1538513215627837.

Stanger-Ross, Jordan, Nicholas Blomley, and the Landscapes of Injustice Research Collective. 2017. "'My land is worth a million dollars': How Japanese Canadians Contested Their Dispossession in the 1940s." *Law and History Review* 35(3): 711–51. doi:10.1017/S073824801700027X.

Sunahara, Ann Gomer. 1981. *The Politics of Racism: The Uprooting of Japanese Canadians during the Second World War*. Toronto: J. Lorimer.

Wrigley's British Columbia Directory 1930. 1930. Volume 40. Vancouver: Wrigley Directories Limited. Available online from the Vancouver Public Library at https://bccd.vpl.ca/index.php/browse/title/1930/Wrigley%27s_British_Columbia_Directory.