Working Papers
www.mmg.mpg.de/workingpapers

Max Planck Institute for the Study of Religious and Ethnic Diversity

Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften

MMG Working Paper 19-05 • ISSN 2192-2357

Daniel Hiebert* (University of British Colombia)

Superdiversity and Cities - Technical Report

Daniel Hiebert (University of British Colombia)
Superdiversity and Cities - Technical Report

MMG Working Paper 19-05

Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften,
Max Planck Institute for the Study of Religious and Ethnic Diversity
Göttingen

ISSN 2192-2357 (MMG Working Papers Print)

Download: **www.mmg.mpg.de/workingpapers**

MPI zur Erforschung multireligiöser und multiethnischer Gesellschaften
MPI for the Study of Religious and Ethnic Diversity, Göttingen
Hermann-Föge-Weg 11, 37073 Göttingen, Germany
Tel.: +49 (551) 4956 - 0
Fax: +49 (551) 4956 - 170

www.mmg.mpg.de

info@mmg.mpg.de

Abstract

This report describes the materials incorporated into the original *Superdiversity* visualization website (**www.superdiv.mmg.mpg.de**), launched late in 2018. This includes the sources of data, the properties of the variables and categories used to produce the website, and the design principles and methods used in the visualizations presented on the website.

Keywords: Superdiversity, migration, Vancouver, Sydney, Auckland, data visualization, linked data

Author

DANIEL HIEBERT is a Professor of Geography at University of British Colombia who specializes in issues of urban social geography and public policy. His research interests focus on immigration policy, the integration of newcomers into the housing and labour markets of Canadian cities, and the consequences of the growing 'superdiversity' of Canadian society. Much of this research is conducted collaboratively, working with partners in government and non-government organizations.
Contact: Daniel.hiebert@ubc.ca

Contents

Background

The *Superdiversity* visualization website (**www.superdiv.mmg.mpg.de**) was formally launched at the International Metropolis Conference convened in Sydney 29 October – 2 November 2018. Its co-creators – Steven Vertovec, Daniel Hiebert, Paul Spoonley and Alan Gamlen – developed the site through in a series of conversations and tasks that were originally sparked by a special Academy on Urban Superdiversity, organized by the Max Planck Institute for the Study of Religious and Ethnic Diversity, that was held in Berlin in 2015. An attempt to convey the diversification of cities in interesting ways was part of the Academy, essentially through a three-dimensional model of data about Vancouver, and it struck us that we should consider powerful and engaging means of conveying the story of diversifying cities. Gradually, the idea of a website emerged, which would gather data on diversity in several cities and present this information accessibly. We considered several key factors in deciding which cities to include in the project: the relevance of cities for comparison (i.e., we needed cities that shared important characteristics but also differed from one another); the quality, availability and comparability of data; and the already existing collaboration among the co-authors. The project began to reach fruition when we met in May, 2018, over several days, to coordinate our objectives for the website and begin the general design process. By the end of this meeting, we had settled on the five visualizations to be included in the site, and the overall narrative logic of the site. The narrative is based on the use of scrollytelling — a combination of 'scrolling' and

* While I wrote this technical report, it summarizes the work of a large team. The idea of a superdiversity website was first raised by Steven Vertovec, who brought Alan Gamlen, Paul Spoonley, and me together in the first instance. We are all co-creators of the website and we are immensely grateful to the Max Planck Institute for the Study of Religious and Ethnic Diversity for providing the funding required for this work. We have also benefitted from the help of a larger group. Meghan Kelly provided crucial ideas for the original website design. We employed highly capable assistants for data management and analysis: Aateka Shashank for Vancouver, and Ernest Healy for Sydney. I give special thanks to Robert Didham of StatsNZ for not only explaining the nature of New Zealand data and answering our many questions, but also for arranging intricate special data tabulations to be supplied freely for this project. I also thank the Data Distribution Team in the Research and Evaluation branch of Immigration, Refugees, and Citizenship Canada, and their colleagues at Statistics Canada, for extracting Canadian data and supplying it to this project freely. Of course, Stamen Design played a major role and made excellent contributions to all aspect of website design (special thanks to Vinay Dixit, Logan Williams, and Benny Lichtner). Lastly, the website is hosted at the Max Planck Institute, and Rami Higazi has been instrumental in uploading and maintaining the site.

'storytelling' that combines interactive digital multimedia to provide a descriptive account while increasing the granularity of information as the user moves through the site. We defined the detailed design of the website over the next few months, with the help of Stamen Design, while we also began acquiring data. The latter process was slower than anticipated and presented us with many difficult choices and technical challenges about variables and categories. We also learned that our expectation that Australia, Canada, and New Zealand collect the same migration data was overly optimistic. We therefore had to accept the fact that some of the elements in our website had to vary according to location and the specificity of data collection, despite our efforts to present three completely comparable stories. We released an early version of the website to the public in November, 2018, but continued to refine it over the next few months. Subsequently, in April 2019, we created a separate website, *Superdiversity in Canadian Cities* (**www.superdiv-canada.mmg.mpg.de**), which includes Vancouver, Toronto and, eventually, will also include Montreal and three other Canadian cities. This technical report addresses both websites.

Note that in this paper, we do not discuss the conceptual ideas behind the website -- such as our definition of superdiversity, or its practical implications. This report is dedicated to the more mundane task of providing users of the website with all the information they need to understand the sources of data and the methods of converting that information into interactive visual tools.

Types and sources of data

We used two main types of data in building the website: administrative and survey or census-based. The former data are collected and, generally, maintained, by immigration ministries, while national statistical agencies are responsible for the latter. We therefore had to create two pathways to acquire the data incorporated into the website.

Administrative data on flows of migration

Nation states maintain administrative data on visas granted to individuals entering their territory. These visas specify the reason for admission and either grant the applicant permanent residence or indicate an expiry date. For the initial *Superdiver-*

sity website, we only acquired data on permanent entries[1] to the three countries, and requested that the data be grouped into large categories. We were able to use the same major categories for Canada and Australia but, unfortunately, not for New Zealand. In all three cases, these data are recorded and maintained by the ministries responsible for immigration (Immigration, Refugees, and Citizenship Canada; Home Affairs (Australia); and Immigration NZ which is part of the Ministry of Business, Innovation and Employment (New Zealand).

For Canada and Australia, the major categories are:

- Economic immigrants: the Principal Applicant and accompanying family members, for any of the admission categories employing economic selection. These include: skilled workers; skilled trades; business immigrants; and individuals nominated by sub-national governments for their potential economic contribution.
- Family class immigrants: individuals nominated by close family members already living permanently in the host country. This may include spouses, children and, in some cases, parents, grandparents and more distant relatives. Note that family class immigrants are sponsored by a mixture of those who initially arrived as economic immigrants or refugees, and also by individuals born in the receiving country (e.g., in the case of a spouse or an adoption).
- Humanitarian: this broad group includes those who arrived as asylum seekers and have been granted the right to permanent settlement, resettled refugees (through government or private sponsorship), and others who may be outside the definition of the Refugee Convention, but who are granted permanent residence based on humanitarian grounds (e.g., the risk of spousal violence if they return to their source country).
- Other: individuals admitted for other reasons. Both Canada and Australia have several categories of admission that do not fit the three main pathways, but the number in these groups is typically small.

Information on immigration to New Zealand is collected using different categories. In that country, permanent residents are categorized according to the first visa they

1 Data for this project only pertain to entries, and not exits, so the website shows gross, not net, permanent migration. While Australia and New Zealand maintain detailed records on emigrants, this is not the case for Canada. Also note that the definition of 'permanent' varies across countries. In Canada, for example, this term is only assigned to visas that are open-ended and that enable holders who fulfil certain obligations to become naturalized citizens, while in New Zealand, anyone holding a visa with a duration of at least one year is considered a permanent resident.

obtained to come to New Zealand. This may have been as a visitor or tourist, to work, or to study (all forms of temporary visas although some "temporary" work and study visas which allow the applicant to remain in New Zealand for 12 months or more, mean that they are classified as Permanent Residents). Some individuals are admitted directly to New Zealand as permanent residents (the 'Residence' category on the website or as part of the "Skilled Migrant Category"). Also, the Other category is much larger in New Zealand than Canada or Australia. In New Zealand, a large proportion of those included in this category are actually New Zealanders returning after a long period outside the country, many of them who are New Zealand citizens who have been living and working in Australia.[2] Another significant number in this category are citizens or residents of Australia who have exercised their right to move to New Zealand as stipulated in the Trans-Tasman Accord between the two countries. Finally, as in the cases of Canada and Australia, the 'Other' category includes those who simply do not fit any of the other forms of admission.

It is also important to note that the bureaucracies of the three countries began recording immigration information, in digital form, at different times. We have been able to acquire data for Canada that indicates both category of entry and source country, stretching back to 1980.[3] New Zealand also maintained digital data for different source countries from that year, but only began to incorporate the category of entry in 2003. Finally, we were only able to obtain detailed origin/category data for Australia dating back to 1991.

The data on source countries, for Vancouver and Sydney, is based on the country of last permanent residence, while it is based on country of citizenship for New Zealand.

Data for these flow visualizations were coded on individual spreadsheets for each year, with origin countries and their respective continents arranged as rows and the categories of entry as columns. Stamen Design used several tools to build the code that converts the data into interactive visualizations, and decided on the colour palette (see the section on website code and design, below). The overall design, with the graph situated above the shifting rectangle of source countries, was defined through a series of conversations between Stamen and the co-authors of the project.

---

2 New Zealand also maintains records on citizens and residents who leave the country (see keanewzealand.com).

3 Earlier Canadian immigration records have been digitized, back to the early 1950s. However, these records do not indicate the category of admission, and only include the basic demographic characteristics of the person and their country of origin.

In order to protect individual privacy, special rounding algorithms were applied to migration data. For Canada, zero values were retained, but values between 1-4 were simply assigned a value of 3. Values greater than 5 were random-rounded to base 3 (i.e., to any number within 3 of the original value). Similar systems were employed for migration data from Australia and New Zealand.

Extra graphics for the *Superdiversity in Canadian cities* website

For this site, we acquired additional data on temporary residents and also at the metropolitan scale. At the national scale, we have been able to obtain information on four types of temporary visas:

- Temporary Foreign Worker Program (TFWP): Canadian employers submit applications to the national government to hire workers from outside the country. These applications are subjected to a Labour Market Impact Assessment to determine whether the migrant would displace a Canadian worker. Individuals entering Canada through this program are tied to a specific employer. The range of occupations associated with the TFWP is very wide and includes, for example, university professors and agricultural workers.
- International Mobility Program (IMP): Typically, individuals hoping to find work in Canada apply to the IMP directly and are not tied to specific employers. The IMP includes a number of reciprocal forms of migration (e.g., agreements that enable young people to embark on a 'working holiday,' signed by Canada and another country), and is also used to recruit individuals who are considered important for Canada's economy or culture.
- Study: Individuals granted visas to reside in Canada for the purpose of education. Typically these visas are only required for educational programs of at least 6 months in duration – individuals enrolled in shorter programs apply for tourist visas that are valid up to 6 months and these are not included in our data.
- Humanitarian: These are temporary visas granted to individuals who are seeking asylum in Canada. Those who receive a positive decision are ultimately offered Permanent Residence (and would therefore be included in the graphic on Permanent Resident admissions), while those whose applications are denied are required to leave Canada.
- Note that there are other categories of temporary visas granted by Canada, but we were not able to gather data on them.

We were, however, able to replicate nearly all of the data on permanent and temporary visas at the metropolitan scale, and therefore include additional flow visualizations for both Vancouver and Toronto. This is possible because individuals applying for visas (or in the case of the Temporary Foreign Worker Program, employers submitting applications) specify their intended destination within Canada. This information is considered reasonably reliable for all of the permanent and temporary categories, except in the case of asylum seekers. These individuals typically enter Canada at airports or border crossings and often do not have a specific or single destinations in mind. Given that IRCC does not have confidence in the accuracy of these data, they were withheld.

Survey data

All of the remaining visualizations on the website are based on data collected in national censuses. Normally, all three countries use the same census cycles, collecting extensive information every five years, in years ending with a 1 or 6. This regular pattern was interrupted in New Zealand, however, by the major earthquake suffered by Christchurch in February, 2011, and the national census was postponed until 2013 (the section of StatisticsNZ which is responsible for the census is based in Christchurch). One result is that the date of the 2016 census was also postponed until March, 2018 – and the new data were unavailable at the time our website was built. Given this series of events, our website incorporates data from 2016 for Vancouver and Sydney, and 2013 for Auckland.

The British colonial governments of Canada, Australia, and New Zealand all held censuses, or required colonial governments to conduct some form of a census, and when these countries became independent, they used these earlier instruments as templates for their own population surveys; this shared history also explains the synchronized timing of the census of the three countries. However, despite their similar origins, early censuses differed between the three countries and their conceptions of which data should be gathered took distinct evolutionary paths. Therefore, although the three contemporary censuses retain important similarities, the questions asked vary, as do the categories given to respondents, as well as the way information is coded. For example, for reasons lost in the mists of time, Canada chose not to include a question on religious affiliation for the new surveys held in years ending in a 6,

when it shifted from decennial to quinquennial censuses, but to do so in those held in years ending in a 1 (Australia and New Zealand maintain the question in all census cycles).[4] Questions on ethnic origin/ancestry differ across the censuses, as does the overall structure of questions about indigenous peoples, in keeping with the changing political significance of indigeneity for the three countries. Finally, each country has introduced questions on issues that have, for whatever reason, attracted national attention (e.g., the composition of 'step-families', household divisions of labour, and same-sex couples, in Canada).

We therefore had to exercise a great deal of choice in selecting information we deemed as comparable as possible across the three censuses. In general, we were interested in the following categories of information:

- Demographic: age, gender.
- Immigration history: admission period, country of origin and, for Canada,[5] category of admission (note that we have attempted to include information on temporary residents but, in all cases, census enumerators have not been able to obtain a high response rate for these individuals, so this information should be treated with caution).
- Mobility.
- Identity: ethnic origin/ancestry, religious affiliation (2011 only), language spoken at home.
- Socioeconomic situation: education, employment, income, housing.

While this seems a rather limited range of information, our project necessitated that we explore all the possible intersections of these data fields and this generated a set of enormous statistical tables. The collected files of raw data were approximately one gigabyte, which was subsequently reduced by over 90 percent in the process of selecting the precise data incorporated in the website.

Different parts of the website employ different census populations. When possible, we have used the total population but, in cases where we incorporate labour market statistics, we have only included the working age population (18-64 years old). Also, note that while we have used the latest census data available, parts of the website also

4 The Canadian quinquennial census was introduced in 1956; prior to that time Canada only conducted decennial censuses. The 1956 census did not include several standard questions, likely for budgetary reasons. The quinquennial census was introduced into New Zealand by the 1877 Census Act. The only exceptions were in 1931 (depression), 1941 (war) and 2011 (earthquake).

5 We also plan to acquire this information for Sydney.

include data from earlier censuses. In all cases, the methodology for earlier data have been the same as those described for contemporary data. That is, we only elected to use earlier data that are strictly comparable across time periods.

Note that for every variable used, there was a 'non-applicable' group. In some cases, we were able to exclude these individuals from tables while, in others, we resorted to the use of empty cells in tables. At times, this was a challenging distinction, since the absence of individuals in certain categories is a meaningful statistic (e.g., no people from a particular ethnic group in a particular geographical area), while in other cases blank values are best removed from the analysis (e.g., people who did not provide an answer on the religion question).

Variables and categories used

Canada

In all cases, for Canada, the population was limited to individuals residing in private households (excluding those in institutional settings, such as people in military bases, hospitals, or under incarceration). For the most part, all information in the Canadian census is based on the self-reported answers provided by respondents, such as their gender, age, and ethnic origin. However, for two of the variables incorporated in the study, Statistics Canada pursues administrative databases and inserts information from those sources. This includes information on the sources and amount of income received (linkage to tax files), and the year and category of immigration (linkage to immigrant landing records). These linkages are based on the names and detailed birthdates of individuals and are very high in accuracy but nevertheless are unable to match every single individual. Statistics are subsequently imputed, based on sophisticated multivariate algorithms, for missing values. That is, if a person indicates that they were born in another country on their census form, an effort is made to locate the landing record of that person. If this fails, a variety of answers supplied by the person is used to impute the category of entry of the person. The precise imputation algorithms are not released to the public.

The Canadian census was conducted in May, 2016, in two parts. All Canadian households were required to complete a 'short form' with limited questions, mainly about demography and language, providing a 100 percent survey of the population (98.4 percent response rate). Another, much longer questionnaire, was sent to a sam-

ple of one quarter of all households (97.8 percent response rate).[6] Most of the information used on the website is derived from this very large sample. In each household, one person was expected to fill in the census form (the Census Reference Person), providing information about each individual in the household as well as information about the household as a whole (e.g., ownership status).

In the remainder of this section, the specific census questions are reproduced, along with the categories we used.

Gender: "What is this person's sex?" (Male, Female).

Age: "What is this person's date of birth and age?" (classified into: 18-25; 25-34; 35-44; 45-54; 55-64; 65+).

Language: "What language does this person speak most often at home?" (classified into: English and/or French; Other).

Immigration history: "Where was this person born?" and "Is this person now, or has this person ever been, a landed immigrant?" Together, these questions are used by Statistics Canada to decide whether to search for the person in Canada's immigration database. If the person did immigrate to Canada, their year of arrival and category of admission are added as fields. (year of arrival was classified into: before 1980; 1980-1990; 1991-2000; 2001-2010; 2011-2016 and admission category was classified into: Economic; Family; Humanitarian/Refugee; and Other). Note the file linkage is also used to identify non-permanent residents (individuals holding a temporary visa) and an awkward category of individuals who were born abroad but are not officially immigrants (i.e., those granted Canadian citizenship at birth regardless of where they were born, such as the children of diplomats).

Ethnicity 1 (ethnic origin): "What were the ethnic or cultural origins of this person's ancestors?" This is one of the most complex variables in the Canadian census. Individuals are entitled to name as many ancestries as they wish, but only the first four are incorporated into the master database (respondents are given 44 characters to write the names of their origin groups). An individual, therefore, may have only one origin or more and, given the complex history of migration to Canada, many individuals provided multiple origins. In the Masterfile of the census, each individual is assigned four fields of data for their ethnicity and, for those with a single origin, three will be blank. This means that any count of ethnic origins will yield a number greater than the population, since individuals with multiple ethnic origins are counted more than once and, in fact, up to four times. This complex variable was only used for one purpose, in the mapping visualization on website: to rank small

6 **https://www12.statcan.gc.ca/census-recensement/2016/ref/response-rates-eng.cfm**

geographical areas according to their degree of ethnic diversity. In this case, for each area, we counted the number of ethnicities indicated by residents, either singly or in combination.

Ethnicity 2 ('visible minority'): In Canada, "The Employment Equity Act defines visible minorities as 'persons, other than Aboriginal peoples, who are non-Caucasian in race or non-white in colour'. The visible minority population consists mainly of the following groups: South Asian, Chinese, Black, Filipino, Latin American, Arab, Southeast Asian, West Asian, Korean and Japanese."[7] The rationale for classifying Canadians in this way is explained in the background statement that precedes the census question establishing visible minority status: "This question collects information in accordance with the Employment Equity Act and its Regulations and Guidelines to support programs that promote equal opportunity for everyone to share in the social, cultural, and economic life of Canada." The question posed to respondents is: "Is this person?" … with a field of answers specified (White, South Asian, Chinese… ending in Other). Additional questions ask respondents if they are Indigenous. Statistics Canada compiles this information into a single variable called 'Visible Minority Status' which classifies individuals into the major visible minority categories mentioned earlier and an omnibus category, Non-visible Minority, that, unfortunately, includes those of European origin as well as Indigenous peoples – groups with quite different socioeconomic circumstances. While acknowledging this important limitation, we employed the Visible Minority Status for the website, as it provides a simple and widely used (in Canada) summary of ethnic origins. Moreover, categories for this variable do not overlap – as we have seen for the case of ethnicity – so the count for the variable matches the total population. (classified according to the categories used by Statistics Canada).

Mobility: "Where did this person live 5 years ago, that is, on May 10, 2011?" (classified as in the same geographical region vs. in a different one).

Birthplace of parents: "Where was each of this person's parents born?" This is followed by questions for each of the person's parents. (classified as both parents born in Canada vs. at least one parent born in another country). This information was used, along with year of arrival for first-generation immigrants, to create a composite variable, called 'Generation status', with these categories: Third-generation Canadians (born in Canada to Canadian-born parents); Second-generation Canadians (born in Canada with at least one immigrant parent); Immigrants pre-1980;

7 http://www23.statcan.gc.ca/imdb/p3Var.pl?Function=DEC&Id=45152

Immigrants 1980-1990; Immigrants 1991-2000; Immigrants 2001-2010; Immigrants 2011-2016; Non-Permanent Residents and Others.

Education: This information was gathered through a series of questions, starting with "Has this person completed a high school (secondary school) diploma or equivalent?" and ending with "Has this person completed a university certificate, diploma or degree?" (with sub-questions identifying the type of degree). (classified into: Did not complete high school; High school diploma; Postsecondary education without a university degree; University degree).

Employment status: "During the week of Sunday, May 1 to Saturday, May 7, 2016, how many hours did this person spend working for pay or in self-employment?" and "During the week of May 1 to May 7, 2016, was this person on temporary lay-off or absent from his/her job or business?" Answers to these questions are used in simplified variables that indicate whether a person is employed (including active self-employment), unemployed, or not in the labour force. (classified as a binary: employed vs. not employed).

Home ownership: "Is this dwelling owned by you or a member of this household (even if it is still being paid for)?" (classified as a binary variable).

Housing affordability: This is based on a composite variable created by Statistics Canada, which incorporates information on the income of all housing members (ascertained via linkage with tax records) and the cost of housing (self-reported). Various ratios are provided but we selected 30 percent as our threshold indicating affordability. That is, we created a binary variable distinguishing between those paying up to 30 percent of their gross income on housing vs. those paying 30 percent or more. (classified as: living in affordable housing vs. not).

Income deciles: Census information on family income is based on a linkage to tax records (i.e., the recorded income of each member of a family is compiled from their tax returns). The distribution of family income is divided into deciles at the national scale, and cutoff points are defined for each decile. Subsequently, the income of every geographical unit in Canada is reported by decile. Poor areas would have a preponderance of their population in lower deciles, in this case, while the majority of people in affluent areas would be in the upper deciles. We used this variable in the mapping visualization (see below).

Low income: There are several measures of low income used in the Canadian census. We chose to use the After-Tax Low Income Measure (AT-LIM), calculated at the household scale. For each unit of household size (i.e., 1-percon, 2-person, etc.), the Median is calculated for the national distribution of after-tax income. All indi-

viduals residing in households with less than half of the Median value (at a specific household size) are classified as experiencing low income. (classified as a binary variable for individuals, with categories for below vs. above the AT-LIM threshold).

Religion: This field of information is only gathered on 10-year cycles in Canada, in the censuses that occur in years ending with 1. In this case, we were forced to use the 2011 census. The relevant question was, "What is this person's religion?" Respondents were instructed to provide either a single religion/denomination or to indicate that they had no religious affiliation. Responses were classified into approximately 100 categories, and we simply used this system for our purposes. Unfortunately, the degree of precision varied in the classification system created by Statistics Canada, with many categories used to distinguish between a wide variety of Christian faiths, but few for other faiths. For example, all Muslims are included in a single category, as are all Buddhists.

New Zealand

The New Zealand census was taken on March 5, 2013 and involved two forms, one for each individual and a separate form for each dwelling. In contrast to the Canadian case, the entire New Zealand census is given to all respondents so coverage is intended to be 100 percent of the population (in practice, the response rate for the 2013 census was 92.9 percent).[8] It is also worth noting that all persons, whether in private households or institutionalized settings, are included. For the 2013 census, Statistics New Zealand did not employ linked administrative data for information on immigration and, therefore, we were unable to obtain cross-tabulations of census data that included category of admission.

Census questions

Gender: "Are you?" (options for male, female).

Age: "When were you born?" (spaces to fill in the day, month, and year) (classified into: 18-25; 25-34; 35-44; 45-54; 55-64; 65+).

Mobility: "Where did you usually live 5 years ago, on 5 March 2008?" (classified as a binary, with in the same geographical area vs. outside the area).

Immigration history: "Which country were you born in?" (example countries listed, along with spaces to fill in one not included in the list), "If you live in New Zealand

8 http://archive.stats.govt.nz/browse_for_stats/population/census_counts/PostEnumerationSurvey_MR13.aspx

but were not born here, answer this question. When did you first arrive to live in New Zealand?" In contrast to Canada, information for the immigration variable in New Zealand is self-reported and does not include the category of admission.

Ethnicity 1: "Which ethnic group do you belong to? Mark the space or spaces which apply to you." Respondents are presented with eight example categories and "Other," with three lines of 12 spaces provided to indicate their ethnicity or ethnicities. Note that the first example category is "New Zealand European," generally used to indicate membership of the majority ethnic group in New Zealand but in reality, this descriptor is used to apply to those born in New Zealand ("Pakeha") as well as those migrants who might also identify as "European". The other response categories are more specific (e.g.,NZ Maori, Samoan, Chinese). For each individual, up to six ethnicities are retained in the census master file. As in the case of Canada, this means that the sum of ethnicities in New Zealand is much greater than the total population, since many respondents indicated multiple ethnicities. We used this complex and detailed classification of ethnicity for our mapping visualization.

Ethnicity 2 (aggregate variable): We also needed a simpler definition of ethnicity for other visualizations (the 'bubble chart' and the dashboard). As in the Australian case, special tables were extracted from the New Zealand census using simplified (i.e., aggregated) ethnic categories, based on our specification. The categories used are: Māori; Pacific Peoples; European; Middle Eastern; Chinese; Southeast Asian; Indian; Asian not further defined (i.e., excludes those from the preceding four Asian groups); Latin American; African; and Other. Note that respondents indicating multiple ethnicities within one of these categories (e.g., British, German, and French) would only be counted once, while those indicating ethnicities across categories (e.g., Māori and British) would be counted once for each category. Also note that, in contrast with the Canadian case, this method enabled us to retain a category for the Indigenous people of Aotearoa/New Zealand (Māori).

Language: "Mark as many spaces as you need to answer this question. In which language(s) could you have a conversation about a lot of everyday things?" Respondents are given a set of options that include English, Māori, Samoan, Sign language, and 38 spaces to write in another, or other, languages. We converted this information into a binary variable. (classified as able to speak English vs. not able).

Religion: "What is your religion?" Respondents are given options for no religion, five major religions, and 38 characters to write in their religious affiliation if it does not match one of the choices. There is also a special box indicating several specific Christian denominations. Up to four religions affiliations are retained in the master

file of the census. We used the standard classification system produced by Statistics New Zealand for our project.

Home ownership: "Do you yourself own, or partly own, the dwelling that you usually live in (with or without a mortgage)?" (classified as owners vs. others).

Education: As in the Canadian case, a composite variable was created out of the answers to two questions: "What is your highest secondary school qualification?" and "Apart from secondary school qualifications, do you have another completed qualification?" (classified into: Did not complete high school; High school diploma; Postsecondary education without a university degree; University degree).

Income: The question posed in the census is quite complex and will not be quoted directly, but asked respondents to indicate their total, pre-tax, personal income from all sources for the year ending March 31, 2013. A set of income bands is provided, along with options for income loss and zero income. The 14 bands begin with $1-5,000, and end with $150,000 plus. (We used all the income bands in our analysis, but excluded individuals with negative or zero income).

Low income: We created a binary variable for individuals above vs. below NZ$30,000 in annual income.

Employment: Respondents were asked several questions about their employment status (e.g., whether they worked in the past week for pay, or without pay; and whether they were self-employed). Information from these questions was used by Statistics New Zealand to create a composite variable indicating whether individuals are not in the labour force, and if they are in the labour force, whether they are employed (including self-employment), or unemployed. (classified in our study as a binary variable, employed [including self-employment] vs. all others).

Australia

The Australian Bureau of Statistics (ABS) conducted its most recent Census of Population and Housing on August 9, 2016, achieving an individual response rate of 94.8 percent.[9] As in the case of New Zealand, the Australian census posed the full set of questions to the entire population. And, as in the case of Canada, the Australian Bureau of Statistics has embarked on an ambitious effort to link the census with administrative databases and other social surveys. Generally speaking, census results

9 https://www.abs.gov.au/websitedbs/d3310114.nsf/home/Independent+Assurance+Panel/$File/CIAP+Report+on+the+quality+of+2016+Census+data.pdf

are based on self-reported answers supplied by respondents. Subsequently, however, the ABS links census information to other sources. Most notably for this project, this includes the *Australian Census and Migrants Integrated Dataset, 2016*, which enables the same abilities to cross-tabulate information on immigration with other socio-economic variables as we have seen with Canada.[10]

Gender: "Is this person male or female?" (classified as binary).

Age: "What is the person's date of birth or age?" (classified into: 18-25; 25-34; 35-44; 45-54; 55-64; 65+).

Mobility: "Where did the person usually live five years ago (at 9 August 2011)? (classified as a binary, same vs. different geographical area).

Immigration history: "In which country was the person born?" and "In what year did the person first arrive in Australia to live here for one year or more?" On parental immigration status, "In which country was the person's father born?" followed by the same question about the person's mother. (classified into a composite variable matching the one for Canada: 3rd+ generation Australian, 2nd generation Australian, immigrated before 1980; 1980-1990; 1991-2000; 2001-2010; and 2011-2016).

Language: "Does the person speak a language other than English at home?" (classified as a binary, English vs. non-English).

Ethnicity 1: "What is the person's ancestry? Provide up to two ancestries only." Respondents are given several categories to select plus 36 characters to write in each of their ancestry groups. Again, this means that the sum of the number of ancestries provided by respondents is greater than the total population. As in the other two cases, this variable was used to measure the number of ancestries per geographical area. We made the assumption that the concept of 'ancestry' in Australia is similar to that of 'ethnicity' in New Zealand (the Canadian question includes both terms).

Ethnicity 2 (aggregated): The same procedure was employed to define this variable as explained in the New Zealand discussion, but with a larger number of categories. In the Australian case, these are: Australian; New Zealander; Aboriginal; Oceanian; Western European; Eastern European; Middle Eastern; Southeast Asian; East Asian;

10 Statistics Canada and the ABS take somewhat different approaches to yield the same result. In Canada, linkage to external datasets occurs while census information is being processed and administrative information is entered into the master census database. For example, for immigrants, category of admission is incorporated as a census variable even though there is no question about this issue in the census form. In Australia, this type of linkage happens later, after all the census information supplied by respondents has been processed. In the Australian case, a new microdata file is created, based on the merged data of two previous databases.

South Asian; Central Asian; American (i.e., all of the Americas); Sub-Saharan African; and Not stated.

Religion: "What is the person's religion?" A list of nine possibilities, plus 'no religion' was provided along with 36 spaces to write another answer. Note that the question implies just one answer but this is not directly specified. (Classified according to the ABS system).

Education: as in the census questionnaires developed in Canada and New Zealand, respondents were asked a series of questions about their educational attainment, starting with whether they hold a high school diploma and continuing on to higher levels of qualification. (Classified into: Did not complete high school; High school diploma; Postsecondary education without a university degree; University degree).

Income: "What is the total of all income the person usually receives?" The respondent is provided with 13 income ranges that each specify weekly and annual income, plus the option to tic a box for either negative or zero income. (Classified using the 13 categories specified by ABS).

Low income: The same threshold was used for Australia as in the New Zealand case, with a binary variable defined as persons with more vs. less than $30,000 annual income. As in the New Zealand case, we excluded those with negative or nil income.

Employment: "Last week, did the person have a job of any kind?" There are also close to a dozen other questions relating to employment and together they enable the ABS to classify a person as not in the labour force, in the labour force and employed, or in the labour force and not employed. (Classified as a binary, employed vs. not employed).

Visualizations of census data

Sankey diagram

These visualizations are normally used to show flows with line widths proportional to volume. They have been adapted for social data, and are particularly useful for visualizing patterns of interaction between two complex variables that are each divided into many categories. Our Sankey explores the relationship between ethnicity/ancestry, and religious affiliation – using the entire population. As noted, the question about religious affiliation is only posed every 10 years in the Canadian census and

was not included in 2016. Therefore, the Sankey for Vancouver shows data from the 2001 and 2011 censuses. In Australia and New Zealand, religious affiliation is a core census question and always included, and therefore we have more recent data for Auckland and Sydney.

The Sankeys for the three cities also differ in terms of the granularity of ethnic/ancestry and religious categories. We were able to obtain more detailed breakdowns of these variables for Vancouver and Sydney than the Auckland case, so the former Sankeys contain more information than the latter.

In each case, data for the two variables were assembled via the 'Table builder' tool provided by the ABS or as special tabulations donated by IRCC in the case of Vancouver, and StatisticsNZ for Auckland. Our team subsequently formatted the data for each city into an Excel spreadsheet with ethnic/ancestry groups as rows and religious affiliations as columns. Different degrees of indentation were used for the row and column headings to indicate hierarchical relationships between categories (e.g., for the Vancouver, 2011 visual, that English is a subset of British Isles, which in turn is a subset of European identity).

Stamen Design created the code to convert the Excel matrices into Sankeys, using a combination of Javascript libraries and their own work (see below). The code for visually highlighting particular groups on the Sankey, and the magnifying tool, is an innovation designed by Stamen.

Bubble chart

We settled on a simple chart to provide an overview of the relationship between ethnic diversity and socio-economic outcomes. Given that one of the outcomes we included is employment, we elected to concentrate this visualization on the working age population (18-64 years old) rather than everyone. We explored several designs and settled on what we have come to call a 'bubble chart'. This was designed entirely in-house by Stamen and built using Python code.

Data for the bubble chart were assembled in the form of cross-tabulations between the Ethnicity 2 variable (Visible Minority groups for Vancouver, and aggregated ethnic/ancestry groups for Sydney and Auckland), and four socio-economic indicators: University degree; Employment; Low income; and Home ownership. Each of these indicators was treated as a binary variable for this purpose. We repeated the data extraction process for one sub-group: those who were immigrants arriving in the most recent period (for Vancouver and Sydney, 2011-16, and 2011-13 for Auckland).

Data for this chart were formatted into Excel spreadsheets and directly imported into the website by Stamen.

Bivariate maps

We created two sets of maps for each of the cities at the finest geographical scale available (with Vancouver divided into 3,452 areas, Sydney into 11,169, and Auckland into 11,767). The first set of 'traditional' maps utilized variables collected for the rest of the project, but this time, gathered at the small-area scale. The variables are: Immigrants; Recent immigrants (2011-16 for Vancouver and Sydney, and 2011-13 for Auckland); High-income population (defined for Vancouver as in the top two deciles of household income, and for Auckland and Sydney as individuals with more than $100,000 per year personal income); and the three largest minority groups for each city.[11] Note that the population universe for these data was the total population. Raw data were exported to an Excel spreadsheet (i.e., with geographical areas as rows and the six variables as columns, and raw values for each cell) and then standardized using a Location Quotient (LQ) statistic. The LQ for each cell of the large table is calculated by dividing the percentage of the variable in the particular geographical area by the percentage of the variable for the total population. For example, if 10 percent of the people in the metropolitan area identify as Chinese in origin, the LQ for a geographical area with 5 percent Chinese would be 0.5, and one with 25 percent Chinese would be 2.5. It is therefore simply a measure of relative concentration. An Excel spreadsheet with LQ values for each variable and each geographical area was developed. The map visualizations of these six variables were created using Mapbox software, adapted for our purposes by Stamen Design, which involved merging the area-based data with polygon shapefiles provided by the statistical agencies of the three countries. A blue colour ramp was chosen to show the degree of concentration of each variable across the geographic areas, with Mapbox instructed to classify the LQ values into 15 categories (quantiles), ranked from the lowest to highest LQs (darkest blue for the highest values). In practice, maps are dominated by low-intensity coloured areas because all zero values (i.e., where a particular group is

11 For Vancouver, this is based on the Visible Minority variable (Ethnicity 2) and includes those of Chinese, South Asian, and Filipino origin. For Sydney this is based on the ancestry variable (Ethnicity 1) and includes those of Indian, Italian, and Chinese ancestry. For Auckland, this is based on the ethnicity variable (Ethnicity 1) and includes those of Samoan, Chinese, and Māori ethnicity.

absent from a geographical area) are in the lowest quantile, and zero values are very common given the extraordinarily detailed geographical areas.

A second set of 'superdiversity' maps has also been created, which are designed to highlight areas of high social complexity. These maps required several steps of calculation. To begin with, raw data were extracted for each geographical area for the total population and for the following variables: Count of the number of ethnicities/ancestries present in the area (Ethnicity 1); Percent of the population that is new to the area over the past 5 years (i.e., 1 minus the percent of the population that had remained in the area); Income deciles; Immigration/generation status (i.e., 3+ generation, 2nd generation, and cohorts of immigrants); Educational attainment (in the four categories explained earlier); and, for Vancouver, Immigration category (Economic, Family, Humanitarian). The ethnicity variable was subjected to a standardization procedure by dividing the number of ethnic groups present in an area by the population of the area, in order to ensure that the differential populations of areas was neutralized (otherwise, areas with larger populations, other things being equal, would have the greatest diversity). The mobility variable was left as is for this stage. The other four variables were all recalculated using a Simpson's Generalized Index of Entropy (SI), which is 1 minus the sum of squares all the fractions of groups in a particular area. For example, if two groups are equally distributed in an area, the SI would be $1-(.5^2+.5^2)=0.5$. This number indicates 1 minus the probability of successfully guessing the group that a person living in the area identifies with. Higher SI values indicate greater diversity. For example, if one group accounted for the entire population of an area the index would be $1-1^2=0$. If there were 10 groups of equal population in an area, the SI value would be $1-(10*.1^2)=.9$ (indicating a 10 percent chance of guessing the group affiliation of a randomly chose individual in the area).

This step yielded a new Excel spreadsheet with a value for each of the six variables, for each of the rows of geographical areas, which indicated either the degree of diversity of the area or, in the case of the mobility measure, the degree of 'churn' in the population of the area. These values represent: Ethnic diversity; Mobility (churn); Income diversity; Immigration/generation diversity; Educational diversity; and in the case of Vancouver, Immigrant category diversity.

In order to enable the use of bivariate choropleth mapping – which illustrates the degree of diversity of two variables at the same time – the Mapbox software required these diversity/mobility values to be ranked as low/medium/high values, since the software is only capable of depicting a total of nine colours in any given map. Accordingly, each columns of values in the Excel spreadsheet (i.e., each varia-

ble) was ranked into three equal-sized sub-groups and each cell was assigned a blank for missing data or a value of 1, 2 or 3.

Finally, the transformed matrix of data was processed using Mapbox and additional coding provided by Stamen into the 'superdiversity' maps shown on the website.

Multivariate maps

Users of the original website are given the option to engage with an experimental tool that enables them to build maps that show as many as seven variables at the same time. This is accomplished by depicting each geographical area in Vancouver, Sydney, or Auckland with an underlying base colour, a height, and superimposing a bar on each area that can vary by height, opacity, colour, and thickness. This tool was designed and created by Alan Gamlen in collaboration with the Immersive Visualization Platform, at Monash University (IVP).

When users click on the graphical hyperlink to use this tool, it opens in a separate website. The IVP tool uses the same data as the bivariate map visualization: diversity and mobility indices. However, the IVP tool also includes a longitudinal dimension, based on similar data for earlier census periods (1996, 2006, and 2016 for Vancouver; 2001, 2006, 2011, and 2016 for Sydney; and 1996 and 2013 for Auckland). For each of the cities, there is a page under the 'help' tab of the control panel, which provides detailed definitions of the variables and categories used for each census year.

Intersectionality dashboard

Users are invited to explore the relationship between diversity and socio-economic outcomes in much greater detail in the final visualization of the website. The nature of the dashboard visualization is similar to that of a multiple regression model or, more accurately, a set of regression models. First, users select values from a set of independent, or control, variables, and then they see results for a number of dependent, or response, variables. The control variables are: Age; Sex; Ethnicity (Visible Minority for Vancouver, and the Ethnicity 2 variables for Sydney and Auckland); Immigration history (the composite variable for generation and immigration cohort); and for the Vancouver case, Immigration category.[12] The response variables are all

12 We plan to add this variable to the Sydney case.

based on statistical likelihoods. That is, once a person with particular characteristics is imagined (e.g., a 35-44 year-old refugee male who arrived in Canada in the 2001-2010 period), six dials provide a visual image of the probability that the individual: has a university degree; is employed; is not experiencing low income (i.e., has at least a moderate income); speaks English (for Vancouver, English or French) at home; has secured affordable housing; and has realized home ownership. In each case, the dial is set with the average for the entire working-age population as the vertical (middle) value, and the probability for each indicator is shown as a deviation from this point. Therefore, for example, if the selected type of person is more likely to own a home than the average working-age person, then the image on the dial would be to the right of the centre point.

The process of assembling all the information required to produce this visualization was arduous. First, data for each combination of control variables had to be cross-tabulated separately for each response variable (the Beyond 20/20 software used by Statistics Canada for creating tables, for example, could not handle the data if all control and response variables were cross-tabulated simultaneously). This required a special order submitted to each of the statistical agencies of the three countries. The resulting tables had to be specially formatted for a series of calculations. Each table was comprised of a set of rows that each represented a unique combination of all of the control variables (for the Vancouver case, this table had 10,800 rows). Each row was associated with five columns that indicated the values on each of the control variables, and then a series of columns for the number of observations and the values for each response variable.

Once this was completed, the raw values for each response column were transformed into z-scores, separately for the total population, females, and males. In practice, this means that when a user chooses the category 'woman' on the drop-down menu of selection variables, all data are automatically scaled such that the middle point of the dials indicates the average expected value for all working-age females on the response variable in question. Therefore, for example, if the user chose to configure the selection variables for a 45-54 year-old male, economic immigrant arriving in the 1990s, one of the dials would show the likelihood that person has a university degree compared with all working-age males in total.

The design for this visualization involved the superdiversity project team and Stamen.

Website design and coding

As noted at several points in the report, designing the website occurred through multiple conversations between the authors and Stamen Design. Generally, Stamen was responsible for converting our ideas into the look and feel of the site, including the colour palate, drop-down menus, and other functionality.

The web application was built by Stamen in Javascript/HTML/CSS using the React.js framework and the d3 data visualization library. Python was used to clean up and format the data for some of the charts. Each visualization component was custom-built using some combination of React.js and d3. The map graphic also used mapbox, together with react. The following javascript libraries were also used in building the site: d3; d3-sankey; d3-selection; intersection-observer; jquery; mapbox-gl; mathjs; nodelist-foreach-polyfill; react; react-cursor-position; react-dom; react-mapbox-gl; react-scripts; and react-scrollable-anchor.

Plans to update website data

We originally planned the website as a single project, done once. However, the effort to produce it, its functionality and the ability to convey a story from complex big data have convinced us that its value will be best realized by updating data as they become available. Accordingly, we are experimenting with adding the capacity to update data on the new, Canadian website. This is a simpler process given the consistency of census categories and data collection more generally, over time (i.e., given that we are just dealing with one immigration ministry and statistical agency).

Stamen Design has already written a plug-in program that enables us to update the flow data for permanent and temporary migration. This involves an exacting process of ensuring that new data are defined and formatted exactly in keeping with the existing data used to create the website. We are exploring whether it is also possible to build plug-in modules for the four census-based visualizations.

If these experiments are successful, and we are able to secure funding for this purpose, we hope to apply the concept of plug-in modules to the original website and to extend the opportunity to do the same for other relevant cities.