A contextual database at the district level using French census and other administrative data

The example of the “Trajectoires et Origines 2” contextual database

Giulia Ferrari et Rosa Weber

DOCUMENTS DE TRAVAIL 300

A contextual database at the district level using French census and other administrative datas

The example of the “Trajectoires et Origines 2” contextual database

Giulia Ferrari and Rosa Weber

October 2024

ined
INSTITUT NATIONAL
D'ÉTUDES DÉMOGRAPHIQUES

https://doi.org/10.48756/ined-dt-300.1024

A contextual database at the district level using French census and other administrative data
The example of the “Trajectoires et Origines 2” contextual database

Giulia Ferrari and Rosa Weber
Institut National d'Études Démographiques (INED)

Abstract

This paper presents a comprehensive contextual database designed to enhance the understanding of demographic phenomena by integrating socioeconomic measures at multiple geographic levels in France. The database was developed to support the "Trajectories and Origins 2" (TeO2) survey and contains approximately 400 indicators derived from the 2018 French population census, harmonized historical census data, and other administrative sources. These indicators provide valuable context on immigration, socio-economic structures, and living conditions, allowing for a fine-grained analysis of demographic behaviors. By linking geographic context to survey data, this database offers a robust tool for examining the impact of the local environment on individual life trajectories. The database covers various geographic levels including IRIS, municipalities, and employment zones, and is planned to be made available to researchers through a secure data hub. Future work will focus on developing a visualization tool for easy data access and updating the database with new data releases.

Keywords: Contextual Database, Socioeconomic Indicators, Spatial Demography, Small Area Indicators, Census Data Analysis

1. Introduction

Information about the context in which people live[1] is increasingly important in understanding demographic phenomena (Caporali et al. 2016). In addition, socioeconomic measures characterizing the place of residence of survey respondents at a fine geographic level (i.e.: sub-regional) could add a macro dimension to individual profiles, broadening the interpretation of demographic behavior. Individual demographic phenomena can indeed be strongly influenced by the surrounding context.

Ethnic residential segregation is persistent in Western societies, including in France (McAvay and Safi 2018; Verdugo and Toma 2017): immigrant origin populations have higher chances than natives to live in neighborhoods with a higher share of immigrants and in poorer, more socially deprived, neighborhoods. Extensive work has also showed stable levels of ethnic school segregation (causally intertwined with ethnic residential segregation).

There is a relatively long-lasting tradition of demographic studies that examine the factors influencing social processes at multiple levels (especially in US and Nordic countries). Even though national and sometimes regional macro indicators are commonly used in the demographic literature (eg: Fertility: Campisi et al. 2020, Nisén et al. 2020), studies integrating a finer contextual perspective are rarer (for some exceptions, see Alcohol-related mortality and suicide: Winkleby et al. 2006, Martikainen et al. 2004, Agerbo et al. 2006. Migration: Beauchemin & Bouquier 2005. Fertility: Campisi et al. 2022; Chetty and Hendren 2018, Chetty, Hendren and Katz 2016).

The scarcity of small area-contextual indicators in the literature is partly due to difficulties in obtaining and accessing neighborhood and municipality level data. However, there is a need for contextual data encompassing different geographic levels. When considering how the context impacts fertility behaviour for instance, family, health and education indicators may be considered at a broader level than the availability of childcare facilities: policies for the former are often decided at the regional level, whereas policies regarding childcare facilities are often set at the municipality level.

In light of this backdrop, the aim of this paper is to present a database of contextual indicators characterizing all French mainland districts and municipalities derived from the French population census and other French administrative sources using the example of the contextual database built for the *Trajectories and Origines 2* (TeO2) survey (INED, INSEE 2019-2020[2], Beauchemin et al. 2023).

The current version of the contextual database contains about 400 indicators at different geographic levels[3] based on the 2018 French population census, harmonized previous census data and other administrative sources. The indicators created were directed by the survey's Scientific Board and are linked to the TeO2 survey's main topic (i.e. immigrants' and their descendants' individual trajectories and living conditions) and questions regarding the availability of administrative data at a subnational

[1] The place where they work or go to school may also be of interest.

[2] https://teo.site.ined.fr/en/

[3] These are aggregated units for statistical information commonly used by INSEE and that were retained by the scientific board to provide as much information as possible, without being redundant. The smallest geographic unit is the neighborhood, called *IRIS* (N=48 409, mean size 1,400 inhabitants), followed by municipality (*commune*, N=34 849); *bassin de vie* (N=1 641), that groups several municipalities where the inhabitants have access to the most common facilities and services; *arrondissement* (N=320), a district subdivision of the *département*; *zone d'emploi* (N=287), the geographic area within which most of the active population resides and works, and in which companies can find most of the labor force needed to fill the jobs offered; and *département* (N=96, mean size 700 000 inhabitants), a regional subdivision.

level[4]. The geographic levels included are: the IRIS, municipality, *bassins de vie* indicators, *arrondissement*, *zones d'emploi*, *départements*.

2. Data Sources

The main source of information used in constructing the indicators in the contextual database is the French population census. Census data at the individual level are needed to compute macro indicators at any given geographic level. However, depending on the country, access to individual-level census data can be complex and costly. In France, individual census data is available for a fee to scientists who have submitted a research proposal to, and received the authorization of, the Statistical Secrecy Committee, in accordance with the legal procedure and through CASD (*Centre d'Accès Sécurisé aux Données*, a secure data hub).[5]

In France, the census, started in 1801, was performed every 5 years up to 1936. From 1946 to 1999, the periods between one census and the next varied from 6 to 9 years (1946, 1954, 1962, 1968, 1975, 1982, 1990, 1999). Since January 2004, the traditional exhaustive count has been replaced by annual census surveys. Municipalities with less than 10,000 inhabitants are still covered by an exhaustive census as in the past but once every 5 years instead of every 8 or 9 years. The municipalities of 10,000 inhabitants and over are now covered by an annual survey but on a sample of 8% of the population, dispersed over the area. After 5 years, the entire territory of these municipalities has been covered and the results of the census are calculated on the basis of a sample of 40% of the population thus constituted (INSEE 2024). Census data are made available for research purposes by the French National Institute of Statistics (INSEE). Saphir has been used in constructing contextual data that cover a longer time frame. Saphir comprises a sample of the population from the complementary sample of the census. The French National Institute of Statistics has harmonized about forty variables over the time period 1962-2019 which pertain to the individual, household, and housing unit-level (INSEE 2023).

Current contextual characteristics are computed using the 2018 census. Indicators characterizing districts in the past are computed using Saphir. Saphir dataset consists of harmonized files covering universal censuses of 1968, 1975, 1982, 1990, 1999 as well as the yearly rotating samples from 2004 onwards.

The contextual database also contains measures derived from the 2019 "Permanent Database of Facilities" (*Base Permanent des Equipements,* INSEE 2024). It is an open-access database at the district level built by INSEE that gathers a wide range of equipment and services, commercial or not, accessible to the public throughout France on January 1st of each year. It covers 188 different types of services and facilities, divided into seven major categories: personal services, businesses, education, health and social services, transport-travel, sports, leisure and culture, and tourism. Among the set of facilities available, 148 have been selected for this contextual database.

[4] Upon the country of interest and the topic of the research, other contextual measures can be either derived from the census individual and housing data or gathered from freely available public data.

[5] A terminal allows remote access to a secure infrastructure where confidential data is stored. Provided a scientist has obtained access to the data, he or she connects to the server with a chip card and his or her fingerprint. He or she can see and use the data, but not directly export it. Once the computations are made, the scientist has to submit an export request to CASD, where experts verify the compliance with the confidentiality rules (i.e. output has to avoid any possible personal identification). If the confidentiality rules are respected, the scientist receives his or her output by email.

An indicator of "Localized Potential Accessibility to Medical Facilities" (*Indicateurs d'Accessibilité Potentielle Localisée*, DREES 2020;) was also included. This was developed in 2013 by DREES (Office for Research, Studies, Public Policy Evaluation and Statistics) and IRDES (The Institute for Research and Information in Health Economics) to measure the spatial adequacy between supply and demand for primary health care at the municipal level. The indicator is based on health insurance data (SNIIR-AM) and population data from INSEE. In this contextual database, 11 indicators of access to different health professionals were considered. The geographic scale available is that of the municipality and, being a synthetic measure, it was not combined into larger geographic units.
In addition, the database includes the "French Deprivation Index", an indicator of social disadvantage considered as a combination of material and social disadvantages, computed in 2018 by INSERM (National Institute of Health and Medical Research 2019). The geographic scale available is that of the municipality and, being a synthetic measure, it was not aggregated to form larger geographic units.
Finally, we included some measures computed by INSEE on Living Standards, Income Distribution and Poverty based on both the 2019th reported and disposable income, available at all geographic levels (INSEE 2019).

3. Method to compute contextual indicators

In this section, we describe the method used to compute the indicators derived from the population census. The indicators from other sources were freely accessible and already computed.
Concerning recent data based on the annual census surveys, in France there are four types of files available: individual and housing data derived from the main census, and individual and household data derived from the complementary census[6]. In the construction of the contextual database, we used all four files.
Files containing individual-level data from the French population census are included in a huge database (of 30Gb of size) that is divided into smaller files. Still, even these smaller files are too big to be processed all at once. Therefore, we have split them up, creating a file for each *département*.[7]
The Saphir dataset is derived from the complementary census. In this database, as in the regular census, the initial files are too voluminous to be processed at once and they had to be split up by creating a file for each *département* before the calculations.
For each *département*'s file, we have calculated the weighted population size (i.e. denominators of the percentages and rates to be calculated) for each geographic unit of interest (see note 3), in order to retain one line for each unit. Weight variables are provided in the census files. This means completing the indicator creation programs for all individuals counted in the geographic level under

[6] The main census concerns all the questionnaires collected. It is therefore exhaustive for municipalities with less than 10,000 inhabitants and covers about 40% of the dwellings in municipalities with 10,000 inhabitants or more. It produces a "detailed file" containing all the dwellings and individuals surveyed. The second phase of the statistical analysis, called "complementary", is intended to produce variables that are more complex to construct. These are the variables that describe the family structure of households: precise determination of the reference person in the household, identification, where appropriate, of families within the household and the composition of these families. The data are derived from automatic procedures, completed by human intervention for the most complex or special cases. This processing is long and costly, which is why it only concerns a sample of the questionnaires collected.)

[7] We excluded the *départements* outside metropolitan France (overseas *départements*). Moreover, to avoid the possibility of calculating indicators that are potentially biased because of denominators that are too small (i.e. population), districts with a population of 10 inhabitants or less were removed from the main database.

consideration. Variables are constructed by summing up the weighted number of individuals concerned and relating it to the relative denominator.

In practice, the computation of indicators of the contextual database is straightforward: these are percentages and rates with respect to the weighted population size of each considered geographic level. For instance, the percentage of immigrants in the chosen IRIS (*Ilots Regroupés pour l'Information Statistique*) is obtained by dividing the sum of the weighted number of individuals born abroad and residing in that IRIS, by the weighted number of people residing in that IRIS:

$$i=1\ldots\mathcal{N}\ inhabitants\ of\ the\ geographical\ unit$$
$$k = 1 \ldots \mathcal{J}\ geographical\ units$$
$$ipondi = survey\ weight$$
$$\text{immi\_dicho} = 1 \text{ if immigrant, } 0 \text{ otherwise}$$

POP_IND: Number of inhabitants of the geographic unit (i.e. population size)

$$\text{pop\_ind}_k = \sum_{i=1}^{n_k} \mathbb{I}$$

POP_IND_POND: Weighted number of inhabitants of the geographic unit (i.e. weighted population size)

$$\text{pop\_ind\_pond}_k = \sum_{i=1}^{n_k} ipondi_i$$

PCT_IMMI: Percentage of immigrants

$$\text{pct\_immi}_k = \frac{\sum_{i=1}^{n_k} immi\_dicho_i * ipondi_i}{\sum_{i=1}^{n_k} ipondi_i} \times 100$$

Aggregated index:

$$\text{pct\_immi}_L = \frac{\sum_{k=1}^{n_L} pct\_immi_k * pop\_ind\_pond_k}{\sum_{k=1}^{n_L} pop\_ind\_pond_k}$$

Once the indicators are computed, we iterate the indicator-creation programs over all the individuals counted in the considered geographic level. Subsequently, synthetic files are generated by replacing individual data with the aggregates derived from each geographic unit. These synthetic files are then stacked and six separate files are created for all the geographic units of metropolitan France. Due to the substantial size of the original files, these operations, while straightforward, necessitate a significant amount of time to be executed.

The final contextual database comprises 48,409 lines for IRIS-level indicators, 34,849 lines for municipality-level indicators, 1,641 lines for *bassins de vie* indicators, 320 lines for *arrondissement* indicators, 287 lines for *zones d'emploi* indicators, and 96 lines for *départements* indicators.

4. Dataset contents

The current version of the contextual database includes indicators derived from the French population census, INSEE's "Permanent Database of Facilities" *(Base Permanent des Equipements*), DREES's and IRDES's "Localized Potential Accessibility to Medical Facilities" (*Indicateurs d'Accessibilité Potentielle Localisée*) , INSERM's "French Deprivation Index", and INSEE's measures on Living Standards, Income Distribution and Poverty.

We have only included a subset of all computable indicators derived from the census. Specifically, our focus has been on indicators relevant to migration-related questions. These indicators are particularly valuable for analyzing the life trajectories and conditions of immigrants and their descendants in comparison to native populations. However, they may also prove useful in other socio-demographic research domains.

The contextual database contains 359 indicators at the district, municipality and broader geographic levels. The list of the retained census-based measures is as follows (189 indicators):

Age structure

- *Percent below 18, above 60 and 65*

Structure by nationality

- *Percentage of foreigners, of naturalized immigrants, of immigrants*
- *Percentages of immigrants by (groups of) country of birth over the total population and over the total immigrant population*
- *Country groups: Maghreb, Sub-Saharan Africa, Southern Europe, Other EU countries, Southeast Asia*
- *Single countries*[8]*: Algeria, Morocco, Portugal, Tunisia, Italy, Spain, Turkey, Germany, United Kingdom, Belgium, Senegal, Romania, Madagascar, Ivory Coast, Vietnam, China, Cameroon*

Socioeconomic structure

- Educational attainment
- *Percentage of students attending a school outside their municipality, of people with a primary level of education, of people with at least a secondary school degree*
- Activity rate
- *Activity rate of people aged 15 and over, of females aged 15 and over, and of people aged 15 to 24*
- Employment rate
- *Employment rate of people aged 15 to 65, of females aged 15 to 65, and of people aged 15 to 24*
- Unemployment rate
- *Unemployment rate of people aged 15 and over, of females aged 15 and over, of people aged 15 to 24, and unemployment rate of immigrants*
- Status of employed persons
- *Percentage of self-, part-time, temporary, in atypical jobs, blue-collar, executive employed, of long-term job seekers, and of labor force that work in their municipality of residence*

Housing

- Type of dwelling
- *Percentage of people living in a detached house, in social housing, homeowners*
- Density of the dwelling
- Number of persons in the dwelling
- Household type
- *Percentage of single parents, of dual-earner couples*

[8] The choice of the countries was based on the resident population ranked by country of birth, selecting the 17 countries with resident population above 100,000 people. The countries differ slightly for the two contextual databases due to data availability issues.

5. Linkage to survey data: the example of the Teo2 database

The main purpose of the contextual database described above is to characterize the geographic areas where survey respondents currently reside, work, study, etc., as well as where they have lived in the past. Originally constructed to complement the "Trajectoires et Origines 2" (TeO2) survey (INED, INSEE 2019-2020, Beauchemin et al. 2023), this section outlines the linkage procedure using this survey as an example. The TeO2 survey includes geographic codes for respondents' current and previous accommodations, their first independent accommodation, and their accommodation at age 15.

The deterministic linkage with the contextual database is, in principle, straightforward because it was conducted using the geographic code (IRIS, municipality, etc.) as the linkage key. However, we encountered some linkage issues when the geographic codes changed over time due to merges or divisions of certain geographic areas. These challenges were addressed with the assistance of conversion tables provided by the French national statistical institute.

The database covering a longer time span was merged not only on the level of the geographic unit but also the year (in which individuals address was collected to span their residential trajectories over the life course). Considering that yearly census information is only available from 2004 forward, these matches were based on the closest census year, prioritizing more recent information. There remains an approximative element but considering that neighborhoods often do not change within short time spans, we consider that the information is highly accurate.

6. Further steps: Dissemination and update

The indicators described above are very useful to characterize a geographic level of interest in terms of its population demographic and socio-economic structure. However, given the large amount of indicators and the detailed geographic level at which they have been computed, it might be hard or even impossible to visualize them in an integrated manner. We are therefore currently designing a Shiny[9] app that will allow researchers to select the indicator and the geographic level, and to visualize its gradient on the map of France. This dashboard will have an internet address and will be accessible to everyone.

The entire contextual database is not yet accessible to researchers who work with other datasources than TeO2. However, it will become accessible in the near future through the CASD (Centre d'Accès Sécurisé aux Données, a secure data hub).

Additionally, we aim to update the database with each new annual census release and other administrative data releases.

The database described above has the potential to be constructed from other censuses and administrative sources and can be linked to all surveys that have collected geographic codes (such as at the municipality level) of respondents' places of residence. Therefore, the codes used to build the dataset are provided as an appendix to this paper to assist those interested in creating similar databases with other data sources.

[9] Shiny is an R package to build interactive web apps straight from R

References

Agerbo, E., Sterne, J. A. C., & Gunnell, D. J. (2007). Combining individual and ecological data to determine compositional and contextual socio-economic risk factors for suicide. Social Science.

Beauchemin, C., & Schoumaker, B. (2005). Migration to cities in Burkina Faso: Does the level of development in sending areas matter? World Development, 33(7), 1129–1152. https://doi.org/10.1016/j.worlddev.2005.04.007

Beauchemin, C., Ichou, M., Simon, P., & Dutreuilh, C. (2023). Trajectories and Origins 2 (2019–2020): A Survey on Population Diversity in France. Population, 78(1), 11-28.

Billari, F. (2004). Becoming an Adult in Europe: A Macro(/Micro)-Demographic Perspective. Demographic Research, Special 3, 15–44. https://doi.org/10.4054/DemRes.2004.S3.2

Black, R., & Skeldon, R. (2009). Strengthening Data and Research Tools on Migration and Development. International Migration, 47(5), 3–22. https://doi.org/10.1111/j.1468-2435.2009.00575.x

Campisi, N., Kulu, H., Mikolai, J., Klüsener, S., & Myrskylä, M. (2020). Spatial variation in fertility across Europe: Patterns and determinants. Population, Space and Place, 26(4). https://doi.org/10.1002/psp.2308

Campisi, N., Kulu, H., Mikolai, J., Klüsener, S., & Myrskylä, M. (2022). A Spatial Perspective on the Unexpected Nordic Fertility Decline: The Relevance of Economic and Social Contexts. Applied Spatial Analysis and Policy. https://doi.org/10.1007/s12061-022-09467-x

Caporali, A., Kluesener, S., Neyer, G., Krapf, S., Grigorieva, O., & Kostova, D. (2016). The Contextual Database of the Generations and Gender Programme: Concept, content, and research examples. Demographic Research, 35, 229-252.

Dang, A., Goldstein, S., & McNally, J. (n.d.). InternalMigration and Development in Vietnam;

Denney, J. T., Wadsworth, T., Rogers, R. G., & Pampel, F. C. (2015). Suicide in the City: Do Characteristics of Place Really Influence Risk?*: Suicide in the City. Social Science Quarterly, 96(2), 313–329. https://doi.org/10.1111/ssqu.12165

Doignon, Y., Ambrosetti, E., & Miccoli, S. (2021). The spatial diffusion of fertility decline in Egypt (1950–2006). Genus, 77(1), 23. https://doi.org/10.1186/s41118-021-00131-9

DREES (2020), APL https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/lindicateur-daccessibilite-potentielle-localisee-apl

Franklin, R., & Plane, D. A. (2004). A Shift-Share Method for the Analysis of Regional Fertility Change: An Application to the Decline in Childbearing in Italy, 1952-1991. Geographical Analysis, 36(1), 1–20. https://doi.org/10.1111/j.1538-4632.2004.tb01120.x

Goldscheider, F. K. (n.d.). Realizing the Potential of FFS1 with Contextual Data.

Grigoriev, P., Jasilionis, D., Stumbrys, D., Stankūnienė, V., & Shkolnikov, V. M. (2017). Individual- and area-level characteristics associated with alcohol-related mortality among adult Lithuanian males: A multilevel analysis based on census-linked data. PLOS ONE, 12(7), e0181622. https://doi.org/10.1371/journal.pone.0181622

INSEE (2019), Household income and poverty, https://www.insee.fr/en/statistiques/6455956?geo=METRO-1

INSEE (2023), Saphir https://www.insee.fr/fr/information/1303688

INSEE (2024), Base Permanent des Equipements https://www.insee.fr/en/metadonnees/source/serie/s1161

INSEE (2024), Census https://www.insee.fr/en/metadonnees/source/serie/s1321

Inserm (2019), FDep https://public.opendatasoft.com/explore/dataset/indice-de-defavorisation-sociale-fdep-par-iris/information/?flg=fr&q=croix&location=13

IRDES (Barlet, M., Coldefy, M., Collin, C., Lucas-Gabrielli, V.) (2020). L'accessibilité potentielle localisée (APL) : une nouvelle mesure de l'accessibilité aux médecins généralistes libéraux. https://drees.solidarites-sante.gouv.fr/publications/etudes-et-resultats/laccessibilite-potentielle-localisee-apl-une-nouvelle-mesure-de-0

Jasilionis, D., Grigoriev, P., Stumbrys, D., & Stankūnienė, V. (2020). Individual and contextual determinants of male suicide in the post-communist region: The case of Lithuania. Population, Space and Place, 26(8). https://doi.org/10.1002/psp.2372

Kulu, H., & Washbrook, E. (2014). Residential context, migration and fertility in a modern urban society. Advances in Life Course Research, 21, 168–182. https://doi.org/10.1016/j.alcr.2014.01.001

Martikainen, P., Ki, N. M., & Blomgren, J. (n.d.). The Effects of Area and Individual Social Characteristics on Suicide Risk: A Multilevel Study of Relative Contribution and Effect Modification.

Matysiak, A., Sobotka, T., & Vignoli, D. (2021). The Great Recession and Fertility in Europe: A Sub-national Analysis. European Journal of Population, 37(1), 29–64. https://doi.org/10.1007/s10680-020-09556-y

Neyer, G. R. (2003). Gender and generations dimensions in welfare-state policies (WP-2003-022; 0 ed., p. WP-2003-022). Max Planck Institute for Demographic Research. https://doi.org/10.4054/MPIDR-WP-2003-022

Nisén, J., Klüsener, S., Dahlberg, J., Dommermuth, L., Jasilioniene, A., Kreyenfeld, M., Lappegård, T., Li, P., Martikainen, P., Neels, K., Riederer, B., te Riele, S., Szabó, L., Trimarchi, A., Viciana, F., Wilson, B., & Myrskylä, M. (2021). Educational Differences in Cohort Fertility Across Sub-national Regions in Europe. European Journal of Population, 37(1), 263–295. https://doi.org/10.1007/s10680-020-09562-0

O'Reilly, D., Rosato, M., Connolly, S., & Cardwell, C. (2008). Area factors and suicide: 5-year follow-up of the Northern Ireland population. British Journal of Psychiatry, 192(2), 106–111. https://doi.org/10.1192/bjp.bp.107.040360

Schoeni, R. F., Bianchi, S. M., Hotz, V. J., Seltzer, J. A., & Wiemers, E. E. (2015). Intergenerational transfers and extended family roster: A new substudy of the Panel Study of Income Dynamics. Longitudinal and Life Course Studies, 6(3). https://doi.org/10.14301/llcs.v6i3.332

Spielauer, M. (2004). The contextual database of the Generations and Gender Program: Overview, conceptual framework and the link to the Generations and Gender Survey (WP-2004-014; 0 ed., p. WP-2004-014). Max Planck Institute for Demographic Research. https://doi.org/10.4054/MPIDR-WP-2004-014

Vikat, A., Spéder, Z., Beets, G., Billari, F. C., Bühler, C., Désesquelles, A., Fokkema, T., Hoem, J. M., MacDonald, A., Neyer, G., Pailhé, A., Pinnelli, A., & Solaz, A. (n.d.). Generations and Gender Survey: Concept and Design.

Winkleby, M., Cubbin, C., & Ahn, D. (2006). Effect of Cross-Level Interaction Between Individual and Neighborhood Socioeconomic Status on Adult Mortality Rates. American Journal of Public Health, 96(12), 2145–2153. https://doi.org/10.2105/AJPH.2004.060970

Appendix

We provide a brief example of a Stata script to calculate the percentage of immigrants living in a specific municipality. This same method is applicable to compute all other indicators.
Firstly, we start with the crude variable on immigrant census respondents:

· IMMI : Indicator of immigrant

Code	Label
1	Immigrant
2	Non Immigrant

We then create a binary variable for immigrants

```
gen immi_dicho = (immi == 1)
```

Then, we adjust for the census weight (ipondi):

```
g immi_dicho_w=immi_dicho*ipondi
```

Next, we aggregate the weighted count of immigrants for each municipality:

```
bysort com_resid: egen com_immi_dicho_w=total(immi_dicho_w)
```

Finally, we compute the percentage of immigrants for each municipality by dividing the weighted number of immigrants in that municipality by the weighted number of inhabitants and multiplying by 100:

```
bysort com_resid: gen com_pct_immi=com_immi_dicho_w/com_tot_w*100
```

This script demonstrates the process for calculating the percentage of immigrants in a given municipality, which can be applied to derive other indicators as well.