Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada's Symposium 2005
Methodological Challenges for Future Information Needs

USING THE POSTAL CODE AS MERGE KEY FOR INDEPENDENT DATA FILES: MATCHING DATA FROM THE CANADIAN CENSUS AND AN ADMINISTRATIVE FILE OF SCHOOL TEST SCORES IN QUEBEC

Soundiata Diene Mansa, Jean-Guy Blais[1]

ABSTRACT

This article proposes a procedure for hybridizing independent data files with existing common identifiers. The procedure is a response to the challenge often presented by the limitations of empiricism, which prevent researchers from making sound inferences. The technical and methodological challenge in question is to match a file of school test scores with a set of contextual variables from the Census. This article describes the key steps in the procedure. The procedure's value becomes clearer when one considers that producing complex indexes, such as the low-income index and the socio-economic environment index, has traditionally meant requesting geocoding from Statistics Canada. The proposed procedure is presented as an alternative.

KEY WORDS: Postal code, file matching, data extraction, hybrid database

INTRODUCTION

This article outlines a procedure for hybridizing independent data files with prior common identifiers. Researchers such as Gauthier and Turgeon (1997)[2] linked the idea of merging data files to the "greening" of research and predicted a bright future for it because it would allow the construction of hybrid databases, which in turn would make it possible to analyze new or poorly researched issues without having to invest in new data collection operations. The procedure described here was the subject of a paper presented at Statistics Canada's 22nd International Methodology Symposium in October 2005. The procedure's appeal lies in the fact that it responds to the challenge often presented by the limitations of empiricism, which prevent researchers from making sound inferences primarily because the data typically contained in today's administrative files provide such poor coverage.

In concrete terms, we obtained one such file from the *Ministère de l'éducation, des loisirs et des sports du Québec*, which contained the scores of Quebec students on the standard end-of-secondary test of written French for 2001. For each student who took the test, the file contains variables such as the test score assigned by the ministry, the test score assigned by the school, the student's gender, the student's age, the postal code of the student's residence, and the school system (public, private). However, with nothing more than the data available in the test scores file, it is impossible to make inferences that have much validity, since we have very little information about the students' characteristics, especially their living conditions and their educational histories. Consequently, the technical and methodological challenge we set for ourselves was to take the data published by Statistics Canada, which are available through the network of CRÉPUQ member libraries and match the test scores file with a set of contextual variables from the 2001 Census.

[1] - Soundiata, Diene Mansa, Labriprof, Université de Montréal, Canada, H3C 3T4, soundiata@rogers.com
- Jean-Guy Blais, Labriprof, Université de Montréal, Canada, H3C 3T4, jean-guy.blais@umontreal.ca

[2] Gauthier and Turgeon (1997), Recherche sociale, de la problématique à la collecte de données. Les données secondaires. Presses de l'Université du Québec.

A few preliminary technical considerations:

First, the corpus of data we are using was used in a previous study,[3] which we wanted to extend, and we are taking some of the same methodological precautions as the author of that study did to enhance the quality of the inferences that can be made from the corpus. We should also clarify and justify how we chose to delimit the data in the file. We started with a file containing more than 50,000 observations, the scores of all the students across Quebec who took the standard test. Because of some technical constraints associated with the exploratory nature of the proposed match procedure, we decided to confine our analysis to students attending schools in the Montréal urban community. An examination of the characteristics of the corpus of data suggested that the postal code's differentiation index was very low in semi-urban and rural areas. The index seems reasonably high only in major population centres, which was an important criterion for the postal code match strategy on which our procedure is based. A summary table showing the structure of the data in the test scores file after trimming and clean-up is presented below.

Summary table showing the structure of the data

Year	N	Boys	Girls	Public	Private
2001	9133	45%	55%	69,5%	30,5%

The second requirement we had to meet was due to editorial constraints, in particular the small number of pages set aside for our article. In the paper presented at the Symposium, we designed and documented the hybridization procedure as a guided tour, walking through the steps and tools involved in the various stages of the hybridization process. The space limitation imposed on us here will affect the "instructional" quality of the description of the procedure. As a result, readers unfamiliar with the advanced functions of the applications used to extract and match the data (in particular, Beyond 20/20, SPSS and Excel) may have difficulty following and replicating the procedure presented here. The diagram in Figure 1 provides an overview of the procedure and hence a clearer understanding of how the various steps in the process of generating the resulting hybridized database (HDB) fit together. In the remainder of the article, we will describe the five steps in the process, specifying in each case the purpose, tools and output.

1. PRIMARY MATCH

The first step in the procedure is to produce the PC/DA reference list. The idea is to match the postal codes (PCs) in the test score files with their dissemination areas (DAs). There are two possible tools for doing so: Canada Post's Postal Code Conversion File, and the search engine of the Canadian Census Postal Code Analyzer (CCPCA), a user-friendly tool produced by the University of Toronto, available at the following address: http://datacentre.chass.utoronto.ca/census/. We chose the second alternative, the CCPCA. What we obtained, at the end of a well-documented process[4] that we cannot reproduce here, was an output file with no duplicates, since the CCPCA has the very convenient feature of recording each postal code only once.

[3] Blais, Jean-Guy (2003), Étude des différences entre les écoles secondaires du Québec quant aux résultats de leurs élèves à certaines épreuves du ministère de l'Éducation de la fin du secondaire. Research report *CRIFPE-LABRIPROF, Faculté des sciences de l'éducation, Université de Montréal, October 2003.*

4 See 2001 Canadian Census Postal Code Analyser (Data Codebook), a fairly straightforward user manual.

Figure 1: Steps in the data extraction and hybridization process

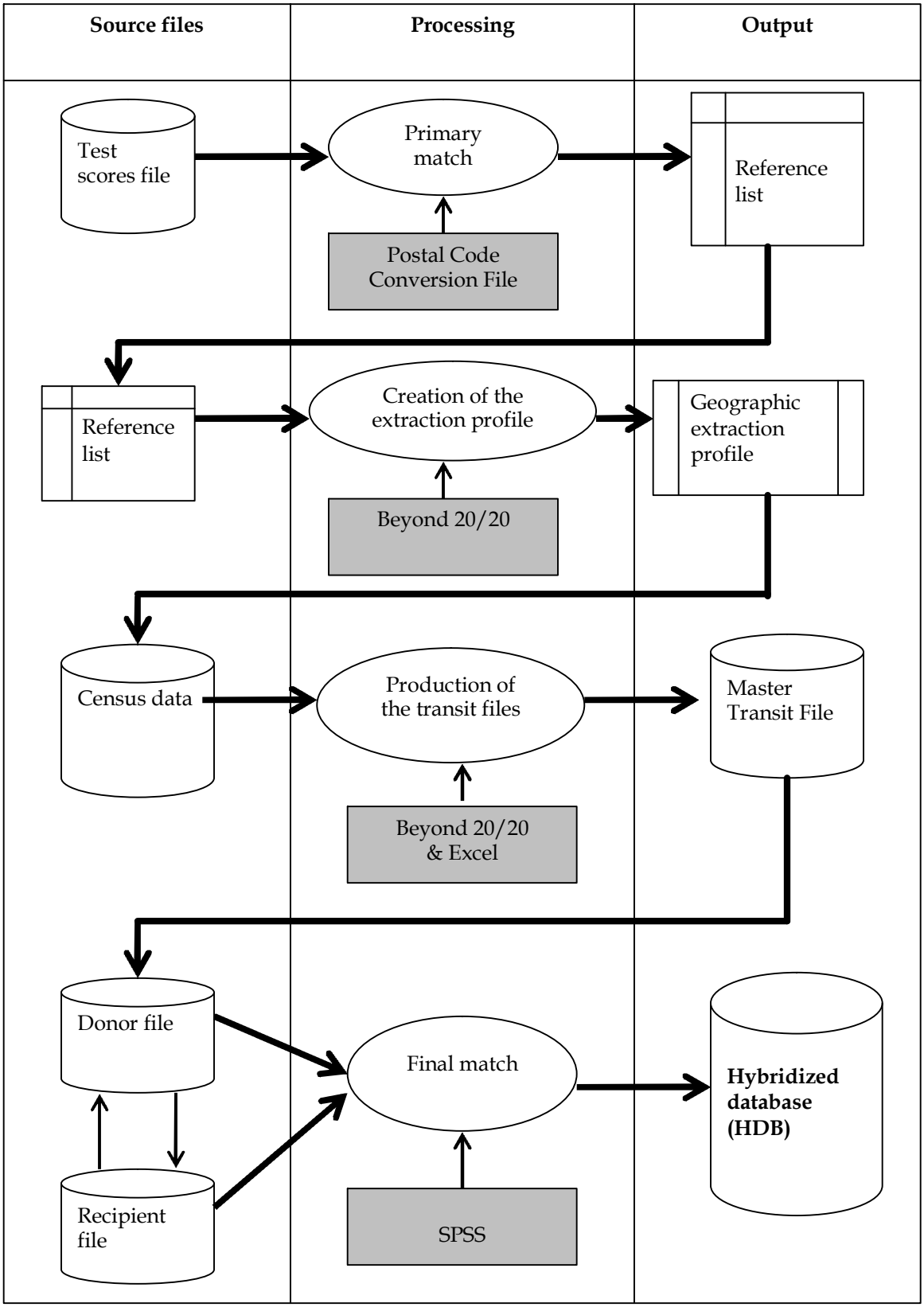
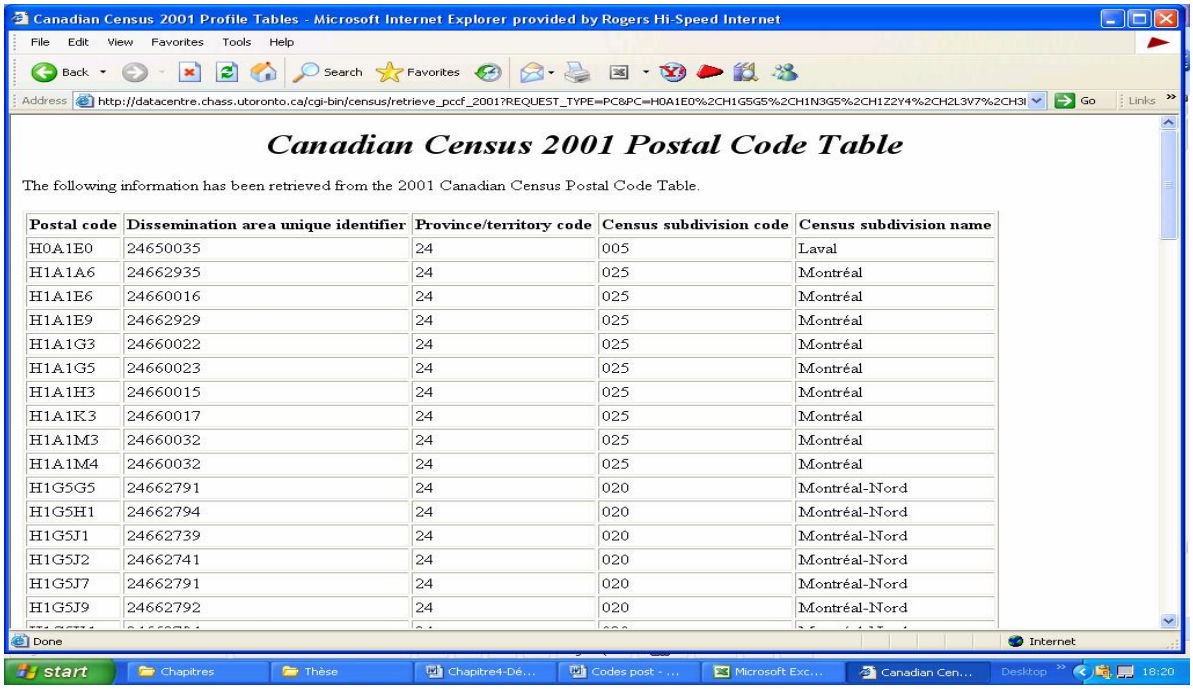


Figure 2: Format of the PC/DA reference list



In the PC/DA table shown in Figure 2, the postal codes from the test scores file (column 1) are associated with various indicators from the census universe, including the dissemination area code (column 2), the province code (column 3), the census subdivision code (column 4) and the census subdivision name (column 5).

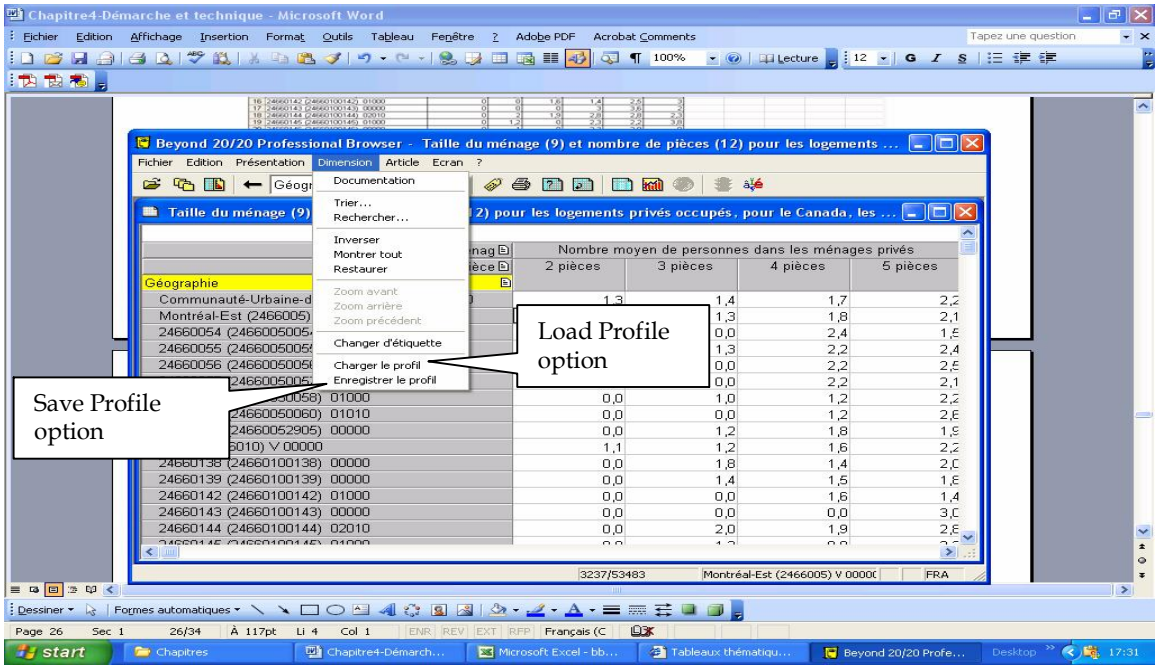
2. CREATING THE GEOGRAPHIC EXTRACTION PROFILE

The second step involves an operation that is not only tedious and delicate but also important and useful for the rest of the process: creation of the geographic extraction profile (GEP). This operation consists in identifying the dissemination area codes in the PC/DA list, selecting them and transferring them to the Beyond 20/20 environment, where they will be stored in memory. It is a demanding undertaking in that it entails navigating the immense universe of census codes and identifying and "masking" all the identifiers that do not match the ones in the reference list. At the end of this second step, we will have a profile of all the dissemination areas (DAs) in a format that is similarly free of duplicates. However, the GEP's most valuable characteristic is its reproducibility, which means that we can load it with just one mouse click each time we want to match it with variables from the various census universes.

This paper does not cover the GEP development process (using Beyond 20/20), which we described and documented in our presentation at the October 2005 Symposium.[5] Nevertheless, Figure 3 presents one screen shot from that process, which shows the main options for loading and storing the GEP.

[5] Cf. presentation by Diene Mansa and Blais (2005) at the 22nd International Methodology Symposium, Statistics Canada, October 2005, entitled "National databases, student outcomes and school performance: Hybridization of census data and student outcomes files – a springboard to a proxy index of school performance".

Figure 3: Beyond 20/20 screen shot showing load and store options for the GEP



3. GENERATING THE TRANSIT FILES

Once the GEP has been created, we are ready to produce the donor file, which will later be merged with the original test scores file. Prior to that, however, we have to go through the desired census universes, select variables and load the GEP. Each time the GEP is loaded, we obtain a file that matches the GEP with the selected variables aggregated at the DA level for all DAs in the profile. This process can be repeated over and over as we explore the census universes, which can result in the accumulation of many specific files (transit files). Using Excel, we can combine them into a master transit file (MTF).

4. GENERATING THE DONOR FILE

The fourth and penultimate step involves combining the various transit files into a single file (MTF). This leads to the production of the donor file, which involves merging the MTF with the PC/DA reference list. The merge operation uses the dissemination area (DA) code as a temporary match key, which is necessary because the postal code is not included in the GEP and because, in the GEP (and the MTF it generates), the DA variable contains observations with no duplicates and can therefore be used as the key for matching with the reference list. We carried out the merge operation to generate the donor file in the SPSS environment, using the DA variable as the temporary match key, which had the expected effect of positioning the postal code variable as a future match key. Along with other technical requirements, this operation necessitated sorting the two files to be merged (MTF and PC/DA reference list) by observation.

5. FINAL MATCH

The last step in the process, final hybridization, involves merging the original test scores file (including duplicates) with the donor file, in which the PC variable is available for use as a match key, since it is already included in the file and the observations contain no duplicates. The PC variable also brings with it all of the environment variables derived from the Census. When the two files have been hybridized, all the observations in the original test scores file will benefit from the "transplant" operation. As in step 4, we used SPSS to carry out this final merger.

CONCLUSION

Using the procedure we have just described, we were able to generate a hybridized database from two completely independent data sources. The procedure is experimental and can certainly be improved. The article describes the key steps in the procedure for generating a hybridized database from two distinct data sources. The procedure is beneficial in that it allows more universal access to and production of adequate statistical data that are better suited to social science research. Its value becomes clearer when one considers that producing complex indexes, such as the low-income index and the socio-economic environment index (which are used in Quebec for such purposes as accounting for the performance of secondary school students), has traditionally meant requesting geocoding from Statistics Canada. Such requests are very expensive and therefore unavailable to most unfunded researchers. The proposed procedure provides an alternative.

REFERENCES

Blais, Jean-Guy (2003), ''Étude des différences entre les écoles secondaires du Québec quant aux resultants de leurs élèves à certaines épreuves du ministère de l'Éducation de la fin du secondaire''. Research report, *CRIFPE-LABRIPROF, Faculté des sciences de l'éducation, Université de Montréal, Octobre 2003*

Diene Mansa and Blais (2005) ''National databases, student outcomes and school performance: Hybridization of census data and student outcomes files – a springboard to a proxy index of school performance'', presentation at the 22nd International Methodology Symposium, Statistics Canada, Ottawa, Canada.

Gauthier and Turgeon (1997), Recherche sociale, de la problématique à la collecte de données. Les données secondaires. Presses de l'Université du Québec, pp 401-430.