N.i.D.S.

NATIONAL INCOME DYNAMICS STUDY

User Guide: Administrative Datasets: Waves 1 & 2

SALDRU
southern africa labour and development research unit
UNIVERSITY OF CAPE TOWN

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

Introduction

This document describes the process of creating Schools Administrative Datasets for the NIDS Wave 1 & 2 respondents. The names of schools attended by respondents are asked in the NIDS questionnaires. These schools are then matched to Department of Education (DoE) registered lists of schools in South Africa. A detailed description of the matching process is given below. It includes a description of the inherent limitations associated with conducting such an exercise. Researchers are urged to fully familiarize themselves with the limitations associated with using this data. Additional variables related to schools (e.g. distance to school) are available in the NIDS Secure Data Center. Researchers interested in utilizing the Secure Administrative Data should make an application to do so by emailing: nids-survey@uct.ac.za.

NIDS Wave 1 and 2 Data Files

Input Files

- Secure NIDS data information on school name and location
- National list of ordinary schools – Q2 2010. Downloaded from the Department of Basic Education (DBE) website accessed in 2010[1]: www.education.gov.za/EMIS/EMISDownloads.

Output Files

- Wave1_indAdmin_Public.dta
- Wave2_indAdmin_Public.dta

Correct Citation of the Data:

Southern Africa Labour and Development Research Unit. 2012. National Income Dynamics Study Administrative Dataset, Waves 1&2 (2008, 2010-2011). [dataset]. Version 1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2012. Cape Town: DataFirst [distributor], 2012.

Matching process

Initial match of Wave 1 variables

Step 1: Prepare NIDS data for matching

- Merged school names and locations from Adult, Proxy and Child file to construct last school, school in 2007 and 2008 variables.
- Location information and other educational variables also included.
- Names standardized with respect to capitals, trimming, substitutions, misspellings etc. Original school names always kept for checks.

[1] This URL was accessible in 2010. To access the updated data (as at 29/11/2012), see: http://www.education.gov.za/EMIS/EMISDownloads/tabid/466/Default.aspx

Step 2: Prepare DoE data for merging

- Obtain DoE National Education Management Information System (NATEMIS) data from DBE website.
- Transferred excel sheet of DBE schools information into Stata.
- Duplicates on NATEMIS assessed and the one with the least information deleted.
- School name and location standardized; original names kept for checks.

Step 3: Matching

- The matching process was iterative, involving both manual and automated matching.

Current school (2008)

1. Match current school on name and location. Those that matched were saved and duplicates in this file checked manually for the most appropriate match.
2. Match on name and province – those that matched uniquely were assumed correct and saved. Those that matched to duplicate schools in the DoE data were checked manually for most appropriate match.
3. Match remaining schools to the closest 20 schools to their household in Wave 1 (i.e. through geographical proximity).
4. Final matching done for last few cases where the name and location of school not matched was similar to a school matched.

School in 2007

- The school in 2007 variable was initially matched from the current school variable then the procedure for matching school in 2008 was repeated. This involved:
- A masterlist of school names and location with their matched school codes was constructed and merged on the name and location of the school in 2007.
- Those respondents who had similar names and locations for their school in 2007 as in 2008 were assigned the code of their school from 2008.
- Steps 1-4 (above) taken in matching current school were repeated.

Last school variable and Wave 2 variables

- Each additional school variable was coded using the same steps described for the school in 2007.
- Each time the 'masterlist of matches of school name and location to code' was expanded. I.e. this list contains all the different spellings of names

and locations that have been assigned a school code.

Last steps and checks

- A data set with multiple records for each respondent was constructed with last school in Wave 1 assigned year==2005 and last school in Wave 2 assigned year==2006.
- This dataset is unique on pid-year.
- Current school name and location was only asked if it differed from the information given for school in the previous year (2007 in Wave 1 and 2009 in Wave 2). These were assigned the match code of the school assigned to the previous year. Similarly, last school in Wave 1 was only asked if the respondent's highest education level differed from that given in Wave 1 or the school name and location information in Wave 1 was not valid.
- Checked that the school name and location of school in previous year was the same.
- Checked that the sample (everyone who answered these questions) was correct.

Variable information

Sample

- Everyone who should have answered the question is assigned a school code (schcd).
    - For the last school variable this requires you to have completed some level of education.
    - For current and previous year's school this requires that you attended school in the respective year.
    - Note that in the child sample, last and current school is combined. This information is captured in the current school variables. Only children who have attended grade R or primary (c2=1or c2=2) are included in the sample.

Variable information

- *Variable naming convention*
    - All last school variables are prefixed with w1_edlstm_ in Wave 1 and w2_edlstm_ in Wave 2. Similarly, school in year 20*X* is prefixed by w1_edXm_ in Wave 1 and w2_edXm_ in Wave 2. Appendix Table A1 lists all the variables available in the public data.
- *Variables*
    - School code (*_schcd) maps uniquely to a DoE natemis number but has no particular meaning itself.
- *Missing codes*

- -17 represents that the respondent's school has not been matched. This does not necessarily mean it cannot be matched, it just has not been up until this point.
- -9 represents that the information provided by the respondent was either don't know, an invalid response or missing.
- -6 represents that the information provided by the respondent was for an educational institution that was not an ordinary school. This includes post schooling institutions, FET/ABET colleges, crèche/daycare facilities and schools for learners with special needs. These were excluded as the DoE masterlist did not include these facilities.

- *Former education department (variables *_exdept) have abbreviated codes that stand for the following:*
  1. Transkei, Bophutatswana , Venda, Ciskei
  2. Gazankulu, Kangwane, KwaNdebele, KwaZulu, Lebowa, QwaQwa
  3. Department of Education and Training
  4. House of Representatives
  5. House of Assembley
  6. House of Delegates
  7. Transvaal, Free State, Cape, Western Cape
  8. New schools
  9. Independent

Important 'health warnings' about the data

- The matching process was iterative, involving both manual and automated matching. As such some schools may be matched to the incorrect school.
- Note also that because we used the 2010 DoE data, some of the schools that were not matched would very likely match to older versions of DoE data as schools may have closed or changed name.
- Everyone is matched to 2010 data. This is not necessarily the year they attended the school.
  - This is particularly true for the last school variable.
  - The variable '*In what year did you successfully complete this grade?* '(h2) can be used to assess how far off these variables are.
  - School characteristics, especially things like pupil/teacher ratios will change over time.
- Sometimes the name and especially the location variables were not very informative and hence matched to multiple schools. In that case a choice had to be made at the discretion of the manual matcher.

Matching Rates

The sample sizes and matching rates for each of the variables matched are presented below:

Table 1: Sample sizes and matching rates

		Don't know	Not an ordinary school	Not matched	Matched	Total	% of valid responses matched
Last school:	Questinonaire						
Wave 1	Adult only	310	142	4,256	8,719	13427	
		2.3%	1.1%	31.7%	64.9%		67%
Wave 2	Adult only	1,116	187	4,636	9,259	15198	
		7.3%	1.2%	30.5%	60.9%		67%
School by year:							
Year	Questinonaire						
2007	Adult only	68	263	397	2,398	3126	
		2.2%	8.4%	12.7%	76.7%		86%
2008	Adult, child or proxy	116	400	865	7,503	8884	
		1.3%	4.5%	9.7%	84.5%		90%
2009	Adult only	716	263	520	2,499	3998	
		17.9%	6.6%	13.0%	62.5%		83%
2010	Adult, child or proxy	52	438	1,610	6,878	8978	
		0.6%	4.9%	17.9%	76.6%		81%

- Adults responded to each of the questions (last school, current school and school in previous year), while Children and Proxies were only asked the name of their school in the year of the survey.
- The high rate of 'don't knows' in 2009 are -2's, i.e. respondents were part of phase 2.
- Once the denominator is restricted to valid responses (those that are not don't know and are ordinary schools), matching rates are above 80% for the school by year variables and 67% for the last school variables.

Table 2: Match rate by questionnaire type

	Adult			Child			Proxy		
	n	%	% of valid responses matched	n	%	% of valid responses matched	n	%	% of valid responses matched
2008:									
Don't know	28	1.0%		48	0.8%		40	15.3%	
Not an ordinary school	274	10.2%		58	1.0%		68	26.1%	
Not matched	312	11.6%		525	8.9%		28	10.7%	
Matched	2085	77.3%	87.0%	5293	89.3%	91.0%	125	47.9%	81.7%
2010:									
Don't know	2	0.1%		23	0.4%		27	10.3%	
Not an ordinary school	328	12.2%		27	0.5%		83	31.8%	
Not matched	446	16.5%		1140	19.2%		24	9.2%	
Matched	2193	81.3%	83.1%	4605	77.7%	80.2%	80	30.7%	76.9%

- Disaggregating the matching rates by questionnaire type (Adult, Child and Proxy) we see high match rates out of valid responses in all types.

- However, the rate of 'don't knows' is far higher (as may be expected) in the Proxy questionnaire than in the Adult and Child questionnaire.
- The rate of responses that are not ordinary schools is high in both the Adult and, especially, the Proxy questionnaires. The Proxy questionnaire asked about any type of educational institution and it is therefore not surprising that the percentage not in an ordinary school is highest for this group. The Adult questionnaire had separate questions for the institution where respondents completed school grades versus diplomas, certificates and degrees.

Appendix

Table A1: List of variables available in the public release administrative data

Wave 1		Wave 2	
Variable name	Variable Label	Variable name	Variable Label
pid	Person identifier	pid	Person identifier
w1_hhid	Wave 1 household identifier	w2_hhid	Wave 2 household identifier
w1_questionnaire	Wave 1 individual questionnaire	w2_questionnaire	Wave 2 individual questionnaire
w1_match	Type of match	w2_match	Type of match
w1_edlstm_schcd	Scrambled school identifier, last school	w2_edlstm_schcd	Scrambled school identifier, last school
w1_edlstm_prov	Province, last school	w2_edlstm_prov	Province, last school
w1_edlstm_quin	School quintile, last school	w2_edlstm_quin	School quintile, last school
w1_edlstm_phase	Education phase, last school	w2_edlstm_phase	Education phase, last school
w1_edlstm_nofee	No fee school, last school	w2_edlstm_nofee	No fee school, last school
w1_edlstm_exdept	Ex department of education, last school	w2_edlstm_exdept	Ex department of education, last school
w1_edlstm_ltrr07	Learner-teacher ratio range in 2007, last school	w2_edlstm_ltrr09	Learner-teacher ratio range in 2009, last school
w1_edlstm_ltrr08	Learner-teacher ratio range in 2008, last school	w2_ed09m_schcd	Scrambled school identifier, school in 2009
w1_ed07m_schcd	Scrambled school identifier, school in 2007	w2_ed09m_prov	Province, school in 2009
w1_ed07m_prov	Province, school in 2007	w2_ed09m_quin	School quintile, school in 2009
w1_ed07m_quin	School quintile, school in 2007	w2_ed09m_phase	Education phase, school in 2009
w1_ed07m_phase	Education phase, school in 2007	w2_ed09m_nofee	No fee school, school in 2009
w1_ed07m_nofee	No fee school, school in 2007	w2_ed09m_exdept	Ex department of education, school in 2009
w1_ed07m_exdept	Ex department of education, school in 2007	w2_ed09m_ltrr09	Learner-teacher ratio range in 2009, school in 2009
w1_ed07m_ltrr07	Learner-teacher ratio range in 2007, school in 2007	w2_ed10m_schcd	Scrambled school identifier, school in 2010
w1_ed07m_ltrr08	Learner-teacher ratio range in 2008, school in 2007	w2_ed10m_prov	Province, school in 2010
w1_ed08m_schcd	Scrambled school identifier, school in 2008	w2_ed10m_quin	School quintile, school in 2010
w1_ed08m_prov	Province, school in 2008	w2_ed10m_phase	Education phase, school in 2010
w1_ed08m_quin	School quintile, school in 2008	w2_ed10m_nofee	No fee school, school in 2010
w1_ed08m_phase	Education phase, school in 2008	w2_ed10m_exdept	Ex department of education, school in 2010
w1_ed08m_nofee	No fee school, school in 2008	w2_ed10m_ltrr09	Learner-teacher ratio range in 2009, school in 2010
w1_ed08m_exdept	Ex department of education, school in 2008		
w1_ed08m_ltrr07	Learner-teacher ratio range in 2007, school in 2008		
w1_ed08m_ltrr08	Learner-teacher ratio range in 2008, school in 2008		