The Cathie Marsh Centre for Census and Survey Research

Preparation of Pupil Level Annual School Census data for the analysis of Internal Migration

CCSR Working Paper 2009-03
Naomi Marquis, Stephen Jivraj
Naomi.Marquis@postgrad.manchester.ac.uk,
Stephen.Jivraj@postgrad.manchester.ac.uk

The study of internal migration is central to our understandings of demographic change, but data used to measure migration within the UK are limited. While patient records are useful for studying migration for the population as a whole and by age and sex, they lack other characteristics important in understanding migration. Census data contain socio-economic and ethnic detail, but are limited by decennial collection. A relatively new data source, the Pupil Level Annual School Census (PLASC) may provide an alternative measure of internal migration which has the potential to fill some of the gaps left by current methods. This paper begins the process of exploring the usefulness of the PLASC for this purpose by considering the internal quality of the data.

www.ccsr.ac.uk

Preparation of Pupil Level Annual School Census data for the analysis of internal migration

Naomi Marquis and Stephen Jivraj

Cathie Marsh Centre for Census and Survey Research, School of Social Sciences, University of Manchester, Manchester M13 9PL, UK

naomi.marquis@postgrad.manchester.ac.uk,

stephen.jivraj@postgrad.manchester.ac.uk

Preface

Internal migration is a central component of demographic change, but data currently used to measure migration in the UK are limited. Patient records are used for studying trends by age and sex, but lack important economic and ethnic detail. While previous research into the role of economic status and ethnicity on migration has relied heavily on transition data from the National Census, this too is limited both by its decennial nature, and by the recoding of migrants characteristics only after a move has taken place.

It is clear then that there is a gap in current methods, which may be filled in part by a relatively new data source, the Pupil Level Annual School Census, or PLASC. The PLASC contains individual records for all state-school registered children in England, and is updated yearly with details of each pupil's home location. The linking of

annual datasets provides a longitudinal record which can be used to explore residential movement over time, while ethnicity and household income indicators add potential.

While limited in its usefulness by only capturing the movement of households containing school-aged children it is argued that as no data source provides an ideal picture, where data exists it should be utilised alongside other data sources to provide a better picture of migration patterns.

This paper provides a first step in exploring the potential of the PLASC as a measure of migration through rigorous consideration of its internal quality. Moreover, the paper details interpolation techniques that can be used to estimate data that are omitted or incorrect.

Keywords

Pupil Level Annual School Census, PLASC, Internal Migration, Ethnicity, Income, England, Interpolation, Data Quality.

Acknowledgements

The authors would like to thank Mark Brown, Nissa Finney and Nikos Tzavidis for their useful comments on earlier drafts of this paper, also to Susan Lomax for her technical support.

We are grateful to the Department for Children, Schools and Families for the supply of PLASC datasets (request numbers DR081027.02 and DR090209.02).

Naomi Marquis's research is funded by the ESRC and Oldham Rochdale Housing Market Renewal Pathfinder as part of a CASE postgraduate studentship. Stephen Jivraj's research is also funded by an ESRC postgraduate studentship. We acknowledge this financial support.

Contents

1. **Introduction** 1

2. **PLASC data supplied by DCSF** 2

3. **Pupil attributes** 3

4. **Population Coverage** 8

5. **Errors and Omissions in Pupil Matching Reference** 9

6. **Dealing with Errors and Omissions in Pupil Attributes** 12

7. **Errors and Omissions in Residential Postcode and Census Output Area** 20

8. **Assessment of potential movers** 22

9. **Conclusions** 27

1. Introduction

1.1. This paper details the content and structure of the Pupil Level Annual School Census (PLASC), and steps taken in preparing data as received from the Department for Children Schools and Families (DCSF, formerly Department for Education and Skills: DfES) for the analysis of internal migration. The PLASC is a relatively new data source, providing a continuous time series of migration data. It forms part of the National Pupil Database (NPD), records details of all state-school registered children in England, and is updated annually. While limited in its usefulness by only capturing the movement of families with school-aged children, who are less likely than other groups to migrate, its potential comes from the fact that it contains socio-demographic indicators, as well as postcode level geographical detail. The first section of this paper deals with the content and coverage of PLASC files, with subsequent sections exploring the quality of pupil attribute data, and interpolation techniques used to improve quality upon the linking of annual records.

1.2. Measuring internal migration is central to understanding demographic trends in the UK. However, it is more difficult to measure than other components of population change. Within the UK there is no compulsory register-based system for measuring migration; therefore migration analyses are often conducted using administrative data, decennial census data, or other large-scale surveys. Traditionally, migration estimates in the UK have been produced using National Health Service (NHS) patient records (ONS, 2005).

NHS records are used in studying trends for the population as a whole and by age and sex. However, propensity to move, distance of move and destination, which are influenced by a combination of demographic, economic and social attributes, are not recorded. In particular, studies have consistently indicated that socioeconomic status and ethnicity are of central importance to migration trends, and current policy interest is in whether and how different social groups migrate.

1.3. Previously, researchers studying the relationship between these factors and migration have relied heavily on analysis of the decennial census, which has included information on residential address one year prior to Census enumeration since 1961. By the nature of the Census, data are limited to the period immediately prior to each decennial data collection, which may be atypical periods for movement. One example of this is where migration patterns observed in the 1991 Census differed from much of the rest of that decade due to a period of economic recession. Census data are further limited by the recording of socio-economic attributes only *after* a move has taken place, while pre-migration attributes may be important in understanding movement.

2. PLASC data supplied by DCSF

2.1. The PLASC data source is derived from an administrative form completed electronically by each state school in England. The individual level census covers all pupils who were in the state education system at the time of data collection, for whom the census is statutory. School aged children who do

not attend state establishments are not included in individual level tables. Data are collected locally by Local Education Authorities (LEAs) and compiled by the DCSF. Between 2002 and 2005 the pupil level census was conducted annually on the third Thursday of January. In 2006 tri-annual data collection was introduced for secondary schools, and in 2007 for primary schools (DfES, 2006).

2.2. On receipt, each LEA collates the data from its schools and carries out a validation process. Data are then submitted to the DCSF who further validate content before compiling each school's submission into a national dataset. Individual level PLASC data are integrated into the NPD, alongside attainment data at pupil, school and exam level, where they can be linked to other data through the use of unique pupil and establishment numbers. Jones and Elias (2006) provide an overview of the contents of each of the tables which make up the NPD. PLASC data are available from for the academic year 2001-2002 onwards.

2.3. The data utilised in this study are for the spring census in each year between 2002 and 2007 covering the academic years 2001-2 through to 2006-7, and were supplied by the DCSF as individual files for each census.

3. Pupil attributes

3.1. PLASC fields which are useful to consider when examining residential movement are those relating to individual pupil attributes, such as age, gender, ethnicity, first language, and free school meals status, and also the

geographic information attached to each pupil record. Available locational information includes detail on the pupil's school, and the postcode and associated output area (OA) and Super Output Area (SOA) of their home address. Descriptions of the PLASC fields of interest in examining pupil residential movement are given below. Where the field is referred to by an alternative name in different years, it is that given in bold which will be used henceforth.

Table 1 Pupil Attributes

Label	**Description**
age	Pupil's age (in years) at the start of the school year.
entry	Date the pupil started attending their current establishment.
estab	Unique identification number referring to the educational establishment attended by the pupil.
eth	Detailed coding of the pupil's ethnicity. This field is not available in the 2002 PLASC table.
ethg	Condensed ethnicity field. While the field is present in all years, the coding system changed in 2003 to a structure comparable with that used in the 2001 Census.
flang (lgrp)	Pupil's first language, coded into six categories based on whether English is believed or known to be the pupil's first language.
fsm	Marker of whether a child is in receipt of free school meals.
gend	Pupil's gender.
la	Identification number of the Local Education Authority in which the school is located.
laest	Local Education Authority number and establishment number combined.
mob	Pupil's month of birth.
month	Month part of the pupil's age at the start of the school year.
post	Pupil's home postcode.
oa	Census Output Area associated with the pupil's current residential address.
pmr	Anonymised version of the Unique Pupil Number (UPN). As with the original field, the PMR is unique stays with the pupil throughout their time at school.
yob	Pupil's year of birth.

3.2. The Pupil Matching Reference (PMR) field contained in the data is an anonymised derivative of a pupil identifier, the Unique Pupil Number (UPN), which is retained by each pupil throughout their school career. Pupil records for each census can be matched by this identifier to create a longitudinal record for each pupil, allowing for powerful temporal analysis of pupil attributes, and potentially a unique tool for examining residential movement at a fine geographic level that includes socio-demographic indicators.

3.3. In order to usefully link attributes, they must be stable over time. That is, information about pupils must be recorded in the same way each year so that comparison across years can be made. Table 2 details where PLASC fields are absent between 2002 and 2007. Gaps in the table represent years where a particular field is not found. For example, the 'lgrp' (language group) field is found only in 2007, though the 'flang' (first language) field in earlier years contains the same information, in the same format, thus detail for this attribute can be compared across years.

Table 2 Absence of PLASC Fields

Code	**Attribute**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
pmr_	Pupil Matching Reference						
ac_	Academic Year						
age_	Age at Start of Academic Year						
board_	Boarder						
enrol_	Enrolment Status	X					
entry_	Entry Date						
estab_	Establishment Number						
eth_	Ethnicity	X					
ethg_	Ethnic Group						
ethsc_	Ethnicity Source						

flang_	First Language						X
fsm_	Free School Meals Eligible						
gend_	Gender						
la_	Local Authority						
laest_	Local Authority / Establishment Number						
lgrp_	Language Group	X	X	X	X	X	
mob_	Month of Birth						
month_	Month Part of Age at Start of Academic Year						
oa_	Census Output Area of Residence						
post_	Postcode of Residence						
soa_	Super Output Area						
yob_	Year of Birth						

3.4. Full birth date information, although recorded in original PLASC tables, is of a sensitive nature, thus is not released for this work. However, month and year of birth are available in the 'mob' and 'yob' fields respectively, while the 'month' field provides further birth date related information.

3.5. Ethnicity information is contained in two fields. The detailed 'eth' (ethnicity) field contains 98 ethnic categories, though the number of categories available at the point of entry differs by LEA. The 'eth' field is absent from 2002 data, but categories are consistent across later years. The field 'ethg' (ethnic group) represents a condensed form of 'eth'. While this attribute is found in all years, the codes used to record ethnic group changed in 2003 to make them more comparable with 2001 Census codes (Godfrey, 2004). The different structure of 2002 and post-2002 categories means that they are not directly comparable. The 'flang' (first language) field contains a coded representation of a pupil's first spoken language. There are six codes relating to whether the pupil's first language is known or believed to be English, or

other than English, or whether the information is not known by the school. A more detailed language variable is available, but due to its sensitive nature is not released for this work.

3.6. Claimants of free school meals are flagged in the 'fsm' field (recorded as true where a pupil is known to be eligible and false where they are not known to be eligible).[1] It is important to consider that this field is a marker of whether a pupil is known to be eligible for free school meals (FSM) based on the above criteria. That is, the school has been made aware of eligibility through a valid claim for FSM being made. It is likely that some pupils will be eligible for FSM based on the above criteria but will not use their entitlement. The FSM marker on pupil records has been widely used in educational research as a proxy of low household income, and is similarly used in this paper. Nonetheless, some authors have questioned the usefulness of the FSM eligibility flag as an indicator of household socio-economic status finding it to be 'coarse and unreliable' in identifying all deprived pupils (Kounali et al, 2007; see also Hobbs and Vignoles, 2007),

3.7. Through the PLASC, the DCSF collect information on full home address, though it is only the postcode part of the address which is available here for analysis. As well as full postcode ('post') the Census Output Area ('oa') and

[1] As at April 2006 (the threshold for the final PLASC year included in this study, entitlement to FSM was subject to the pupil's parent or legal guardian being in receipt of one or more of the following benefits: Income Support (IS), Income-Based Jobseeker's Allowance (IBJSA), Support under Part VI of the Immigration and Asylum Act 1999 , Families in receipt of Child Tax Credit will also qualify provided that (a) they are **not** entitled to Working Tax Credit, and (b) their annual income, as assessed by Her Majesty's Revenue and Customs does not exceed £14,155 as at 6 April 2007 (this is subject to annual review), Guarantee element of State Pension Credit. Children who receive IS or IBJSA in their own right are also entitled to receive free school meals (Teachernet, 2006).

Super Output Area ('soa') associated with each postcode are also attached. The availability of pupil home postcode allows for aggregation to any higher level geography.

4. Population Coverage

4.1. One advantage of PLASC data for this project is that it provides a census of all children of compulsory school age who are in the state education system. Although the number of pupils in state education varies by LEA, the England average in 2002/3 was around 92% of all eligible pupils (ONS, 2004). The PLASC is therefore updated with almost 8 million records each year. It is important to consider, however, that patterns of residential movement may be different for pupils not in state education, particularly due to the likely improved household economic status of children attending fee paying establishments, thus assumptions cannot be made about this group. In areas where the number of school aged children not in state education is high, PLASC data may not be suitable for examining residential movement.

4.2. Pupils not of compulsory school age are also included in PLASC tables. These include pre-school aged children in state 'early years' education, and those over the age of 16 in state further education. However, coverage of such pupils is not comprehensive, and thus analysis of the residential movement of children using PLASC data can only usefully include those pupils of compulsory school age.[2]

[2] The period of compulsory schooling in England is defined under Section 8 of the Education Act 1996 as follows: "A person begins to be of compulsory school age when he attains the age of five [and] a person ceases to be of compulsory school age at the end of the day which is the school leaving date for any calendar year (a) if he attains the age of 16 after that day but before the beginning of the school

4.3. It is important to note that while comprehensive in its coverage of school aged children, PLASC data cannot be used to generalise about the movement of the wider population. Indeed, households containing school-aged children are less likely than others to migrate than other groups (Bailey and Livingston, 2005; Meen et al, 2005).

5. Errors and Omissions in Pupil Matching Reference

5.1. When attempting to link files by unique pupil number (PMR), two problems were discovered. Firstly, a number of records had no PMR assigned to them. Secondly, duplicate values existed in the PMR field in some years. Without a PMR, it was not possible to link records, resulting in the exclusion of such records.

5.2. Each pupil should have a unique PMR - an anonymised derivative of the Unique Pupil Number which follows the pupil through their time in state education, regardless of the point of entry or whether there have been gaps in their time in the state school system. It is reported in documentation received with the data that those pupils without a PMR were nursery aged children, who were included in PLASC collections, but who were not allocated a UPN until the time of their starting compulsory education, except in earlier data collections (2002 to 2004). On examining the ages of pupils with missing PMR, it was found that whilst the majority were indeed of nursery age, some were not. All, however, were recorded as primary-school aged children

---

year next following, or (b) if he attains that age in that day, or (C) (unless paragraph (a) applies) if that day is the school leaving date next following his attaining that age." (HMSO, 1996).

(below age 12). Given the large number of pupil records with no PMR, it would have been impossible to ascertain who each record belonged to in order to link it to the longitudinal file, thus all records with no PMR were excluded. As analyses of residential movement using PLASC data will exclude pupils outside the ages of compulsory schooling (for whom coverage is not full), most records removed here would otherwise have been removed at a later stage, therefore their removal at this stage does not pose a threat to the quality of the data. The removal of records belonging to pupils of compulsory school age is potentially more problematic, though such records amounted to less than 0.01% of the total, thus it is not considered that their removal will impact on future analyses.

5.3. A small number of records within some years were found to have duplicated PMRs. That is, the same PMR was attached to multiple records. In order to link records by PMR to create a longitudinal dataset, it is necessary for the PMR field to uniquely identify each record. In total, 943 records were found not to have a unique PMR. Of these, 907 duplications were found in 2003, making up less than 0.01% of all records in this year. Where a PMR was found to be duplicated, steps were taken in order to determine which of the records was most likely to be the 'correct' record for the pupil, assuming that the records did not represent different pupils. It is possible that where a pupil had changed schools, or had a gap in state schooling (i.e. left the system and returned at a later date), a duplicate record may have been created – if, for example, the pupil record is incorrectly retained on a school's system when no longer present.

5.4. Had one of each pair of records contained more missing data than the other, it would have been sensible to retain that which was most complete. However, for records examined here, no record could be eliminated in this way. For this data, three main criteria were used in determining the record to retain. Firstly, where the dates of entry on duplicated records differed, the record with the latest date was retained. The assumption here is that an old record has been left on the system, and that the most recent record is more likely to be correct. This first stage resulted in the selection of one record of every pair for 846 records. Of the records remaining, six pairs were exact duplicate records. Whichever of these were chosen would have made no difference to the final dataset, thus an arbitrary decision was taken to retain the second record. At this stage, 88 pairs of records with non-unique PMR entries remained, all of which contained matching entry dates but differences in other fields. In some records, entries in all fields except those relating to location ('la', 'laest', 'post') were identical in the year of duplication, but the 'laest' for the previous or following year matched only one of the records. In these cases, the record with non-matching 'laest' was selected, on the assumption that duplication had occurred after a residential or school move, thus the retention of the record with different location information maximises the capture of movement. In total, 74 records were selected using this criterion. In the remaining records, the above criterion could not be applied as entries in fields other than location differed. Decisions about the treatment of these records were made on a case-by-case basis, considering the recorded location information in previous and following years. Table 3 summarises the

reasoning for duplicate record removal, while Table 4 shows the number of records with missing or duplicated PMR each year, and the percentage of records excluded due to errors in the PMR field.

Table 3 Sequential reasoning for duplicate removal

Reason for Removal (see key)	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**Total across years**
1	-	846	1	-	2	-	849
2	-	1	-	5	-	-	6
3	-	47	-	27	-	-	74
4	-	13	-	1	-	-	14
Number duplicated each year	-	907	1	33	2	-	943
Total records each year	7738453	7740153	7733278	7685317	7669121	7622662	

Key to table 2

Reason Number	**Explanation**
1	Where 'entry' field on the two records differed, that with the most recent date was retained.
2	Where records were exact duplicates, the second record in was retained.
3	Where all fields except location attributes ('la', 'estab', 'laest', 'post') were the same in both records, but the 'laest' entry for the previous and/or following year matched one record, that with non-matching 'laest' was retained.
4	Most records selected using above criteria. Remaining cases selected on a case-by-case basis, based on consideration of the recorded location information in previous and following years.

Table 4 Records with missing or duplicated PMR resulting in exclusion

Records	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
Initial Records	7738453	7740153	7733278	7685317	7669121	7622662
Missing PMR (Excluded)	0	0	0	520501	317599	319239
Duplicated PMR (excluded)	0	907	1	33	2	0
Total Number Excluded	0	907	1	520534	317601	319239
Percent Excluded	0	0.02	<0.01	6.77	4.14	4.19
Retained Records	7738453	7739246	7733277	7164783	7351520	7303423

6. Dealing with Errors and Omissions in Pupil Attributes

6.1. For all PLASC fields discussed here, valid entries are expected for all pupils in all years in which they are present in the state school system. Three types

of errors are considered here. The first is omitted values. Any field not containing data where the pupil is in attendance is considered an omission. Such fields differ to those where 'Information not Obtained', or 'Parent/Pupil Refused' is entered. Although of little use in future analysis, such entries are valid. The second type of data error is where invalid values are entered in yearly PLASC tables. These are entries which fall outside of the expected range for a particular variable. For most, where the expected range is small, these errors are easily spotted. Gender, for example, should be recorded as 'M' for male, or 'F' for female. A single occurrence of the letter 'T' within the dataset is an obvious error. The third type of error considered here is temporal inconsistency. Inconsistent records are those where a field which is expected to remain constant over time does not. The 'gend', 'mob', 'month' and 'yob' fields are expected to remain constant, while a year on year increment is expected in 'age', as this is always the age as recorded at the start of each school year. The 'fsm', 'post' and 'estab' attributes are expected to change over time. The temporal dimension of the linked dataset makes it possible to track pupil attributes across years, examining the occurrence of inconsistencies, and detecting whether the progression of a record is as expected. Table 5 (below) details the occurrence of each error by field.

6.2. Examination of the data reveals that the number of records affected by errors, omissions and unexpected temporal inconsistencies is very low. Some degree of input and reporting error is to be expected, but this is low for all fields and for all years, considering the size of the dataset. This low occurrence of error is suggestive of good internal quality. However, the number of omissions is

much greater for some fields than others. Omissions are very low for all fields except ‘eth’, ‘ethg’ ‘flang’ and’board’, where they appear much higher, but even errors in each of these fields account for less than 1% of records in any year. Omissions are greater in later years, suggesting that data quality has worsened slightly over time, rather than improving, though omissions in ethnicity fields account for the vast majority of this increase, suggesting an isolated problem with these fields in later years.

6.3. As with omissions, inconsistencies in pupil attributes over time are more prominent for ethnicity fields (with around 8% of records containing inconsistent ethnicity detail). As discussed earlier, it is expected that recorded ethnicity may change over time so this phenomenon is not necessarily reflective of poor data quality, and in fact, further investigation reveals that most ethnicity changes are from unknown to an allocated ethnicity, or vice versa.

6.4. The linking of files enables the usage of interpolation routines to reduce the impact of missing and incorrect values in attributes that are expected not to change. Here, an interpolation routine was applied to all fields expected to remain consistent over time, or to follow a particular progression. The procedure used is similar to that developed by Harland and Stillwell (2007), in their work using PLASC data to forecast pupil roll numbers and commuting patterns for the Leeds LEA. For example, where a pupil’s ‘yob’ is missing in one year, but a constant value is available in other years, it is sensible to assume that this is the correct ‘yob’ for the child. When working

with such a large dataset, it is not necessary to retain a 'yob' value for each year a pupil is present. Instead, it is useful to create a single variable that represents a pupil's most likely *true* 'yob'. In cases where a value has been omitted, or an invalid or incorrect entry made in one or more years, this can effectively be eliminated where there is another value entered *that occurs more often* by selecting this 'majority value' to become the pupil's ascribed 'yob'. Where there is no majority value, but values (other than missing) are present, it is sensible to select the most recently recorded value to become the final 'yob' value. While this is less robust than imputing a majority value, it is assumed that the latest entry will be a correction of a previously incorrectly-recorded value.

6.5. The creation of a single value for attributes that should remain constant aims to improve the quality of these attributes. However, where an invalid value is the only value for a pupil, it is not automatically eliminated using this procedure, thus is recoded later as missing. Similarly, if an attribute is omitted for every year that the pupil is present, their final attribute value will be recorded as missing. Applying interpolation techniques, where possible, is found to reduce the number of missing and invalid pupil attributes, and where input errors have resulted in incorrect entries in some (but not all) years, the impact of these errors is lessened.

6.6. Such interpolation techniques cannot be used on attributes that may be subject to change. Burgess et al. (2006) note that 'fsm' may alter based on changes in the circumstances of a pupil or their parent or guardian, thus it is not

reasonable to assume that a value for the previous or following year should be the same as a year in which the value is missing or invalid. Geographical attributes ('estab', 'post' and 'oa') are subject to change based on changes in school or home address, and the 'post' field is also subject to Royal Mail postcode changes, which occur without a residential move.

6.7. The benefit of using this interpolation procedure is most evident for the ethnicity fields with errors in these fields accounting for over 99% of all errors. On allocating each record with 'eth' and 'ethg' values based on the majority value given across years, the number of records with missing data is substantially reduced. However, as 'eth' is not available in 2002, and the 2002 'ethg' codes are not compatible with later years, pupils present only in 2002 are left with missing ethnicity information in the final dataset. The impact being that around 680,000 (6%) of records contain no ethnicity detail. The prevalence of errors and omissions in ethnicity fields is of concern when examining the potential of the PLASC for measuring residential movement, especially as it is this dimension which is missing from other data sources. However, 94% of nearly 11 million records do contain ethnicity detail, thus the PLASC undoubtedly provides a useful source of this information. Table 5 details the occurrence of error by field for each year, while Table 6 contains the number of records cleaned for each field.

Table 5 Occurrence of Data Errors and Omissions by Field

Attribute	Inalid						Omitted						Inconsistent	
	2002	2003	2004	2005	2006	2007	2002	2003	2004	2005	2006	2007	Total	No majority
Age†	290,628	286,331	296,025	36,272	3,342	4,203	2	1	1	1	25	4	14,883	3,377
Ethnicity*		-	-	-	-	-		307,022	224,541	171,855	169,235	162,876	877,627	846,617
Ethnic group*	12	692	447	-	-	162	207,076	307,009	224,522	171,880	169,235	162,885	548,061	768,028
Source of eth info.	6	-	-	-	-	1	6,886,834	2,712	12,039	1,924	36,003	101,868		
First language*	6	-	-	-	1	352	7,929	18,921	15,671	16,174	76,247	75,926	506,930	197,964
FSM eligibility	-	-	-	-	-	-	12,938	6,219	5,952	6,477	31,718	47,106		
Gender	1	-	-	-	-	-	5	-	-	-	13	-	41,016	-
Month of birth	-	-	-	-	-	-	2	1	1	1	25	-	48,501	11,786
Month part of age	-	-	-	-	-	-	2	1	1	1	25	4	48,501	11,786
Year or birth±	200,896	197,719	206,919	8,473	6,333	7,357	1	1	1	1	25	-	15,840	4,126
Total	491,549	484,742	503,391	44,745	9,676	12,075	7,114,789	641,887	482,729	368,314	482,551	550,669	2,101,359	1,843,684

*For ethnicity, ethnic group, and first language, omitted entries include those recorded as 'refused'.
Ethnic group: Some invalid codes already grouped together as ‘pre-2002 code’, but recorded here as invalid.
†For the purpose of this work, age range considered valid is 0-18.
±As above, range considered valid is the range of birth years for pupils aged 0-18 in each year.

Table 6 Number of records before and after interpolation

		Age	Ethnicity	Ethnic Group	Source of Ethnicity Information	First Language	Free School Meals Eligibility	Gender	Month of Birth	Month Part of Age at Start of Academic Year	Year of Birth
2002	Total records	7,738,453		7,738,453	7,738,453	7,738,453	7,738,453	7,738,453	7,738,453	7,738,453	7,738,453
	Records with field entry	7,738,451	0	7,531,377	851,619	7,730,524	7,725,515	7,738,448	7,738,452	7,738,451	7,738,452
	Records with error*	310,537	1,585,364	1,472,767	6,886,840	489,298	12,938	27,110	47,636	47,622	224,053
	Records with field entry after interpolation	7,732,024	6,943,835	6,944,972		7,725,898		7,738,447	7,738,451	7,738,435	7,734,359
	Records with missing entry after interpolation	6,429	794,618	793,481		12,555	12,938	6	2	18	4,094
	Percentage of errors fixed	97.9%	49.9%	46.1%	100.0%	97.4%	0.0%	100.0%	100.0%	100.0%	98.2%
	Percentage of records cleaned	99.9%	n/a	89.7%	0.0%	99.8%	0.0%	100.0%	100.0%	100.0%	99.9%
2003	Total records	7,739,246	7,739,246	7,739,246	7,739,246	7,739,246	7,739,246	7,739,246	7,739,246	7,739,246	7,739,246
	Records with field entry	7,739,245	7,432,224	7,432,237	7,736,534	7,720,325	7,733,027	7,739,246	7,739,245	7,739,245	7,739,245
	Records with error*	309,010	1,338,943	986,209	2,712	539,385	6,219	32,318	52,129	51,261	220,251
	Records with field entry after interpolation	7,731,622	7,589,370	7,590,808		7,731,450		7,739,246	7,739,245	7,738,375	7,733,745
	Records with missing entry after interpolation	7,624	149,876	148,438		7,796		0	1	871	5,501
	Percentage of errors fixed	97.5%	88.8%	84.9%	100.0%	98.6%	100.0%	100.0%	100.0%	98.3%	97.5%
	Percentage of records cleaned	99.9%	98.1%	98.1%	0.0%	99.9%	0.0%	100.0%	100.0%	100.0%	99.9%
2004	Total records	7,733,277	7,733,277	7,733,277	7,733,277	7,733,277	7,733,277	7,733,277	7,733,277	7,733,277	7,733,277
	Records with field entry	7,733,276	7,508,736	7,508,755	7,721,238	7,717,606	7,727,325	7,733,277	7,733,276	7,733,276	7,733,276
	Records with error*	319,106	1,308,990	924,149	12,039	564,077	5,952	37,107	53,958	50,968	229,320
	Records with field entry after interpolation	7,725,952	7,617,042	7,618,860		7,729,143		7,733,276	7,733,276	7,730,284	7,727,687
	Records with missing entry after interpolation	7,325	116,235	114,417		4,134		1	1	2,993	5,590
	Percentage of errors fixed	97.7%	91.1%	87.6%	100.0%	99.3%	100.0%	100.0%	100.0%	94.1%	97.6%
	Percentage of records cleaned	99.9%	98.5%	98.5%	0.0%	99.9%	0.0%	100.0%	100.0%	100.0%	99.9%
2005	Total records	7,164,783	7,164,783	7,164,783	7,164,783	7,164,783	7,164,783	7,164,783	7,164,783	7,164,783	7,164,783
	Records with field entry	7,164,782	6,992,928	6,992,903	7,162,859	7,148,609	7,158,306	7,164,783	7,164,782	7,164,782	7,164,782
	Records with error*	52,410	1,224,727	839,728	1,924	555,152	6,477	36,780	52,210	49,242	25,329
	Records with field entry after interpolation	7,163,864	7,073,392	7,075,588		7,161,143		7,164,782	7,164,783	7,161,813	7,163,815
	Records with missing entry after interpolation	919	91,391	89,195		3,640		1	0	2,970	968

		Age	Ethnicity	Ethnic Group	Source of Ethnicity Information	First Language	Free School Meals Eligibility	Gender	Month of Birth	Month Part of Age at Start of Academic Year	Year of Birth
	Percentage of errors fixed	98.2%	92.5%	89.4%	100.0%	99.3%	100.0%	100.0%	100.0%	94.0%	96.2%
	Percentage of records cleaned	100.0%	98.7%	98.8%	0.0%	99.9%	0.0%	100.0%	100.0%	100.0%	100.0%
2006	Total records	7,351,520	7,351,520	7,351,520	7,351,520	7,351,520	7,351,520	7,351,520	7,351,520	7,351,520	7,351,520
	Records with field entry	7,351,495	7,182,285	7,182,285	7,315,517	7,275,273	7,319,802	7,351,507	7,351,495	7,351,495	7,351,495
	Records with error*	18,379	1,212,031	835,082	36,003	668,050	31,718	37,841	51,086	48,100	22,298
	Records with field entry after interpolation	7,351,117	7,265,121	7,267,377		7,339,194		7,351,513	7,351,504	7,348,516	7,350,118
	Records with missing entry after interpolation	403	86,399	84,143		12,326		7	16	3,004	1,402
	Percentage of errors fixed	97.8%	92.9%	89.9%	100.0%	98.2%	100.0%	100.0%	100.0%	93.8%	93.7%
	Percentage of records cleaned	100.0%	98.8%	98.9%	0.0%	99.8%	0.0%	100.0%	100.0%	100.0%	100.0%
2007	Total records	7,303,423	7,303,423	7,303,423	7,303,423	7,303,423	7,303,423	7,303,423	7,303,423	7,303,423	7,303,423
	Records with field entry	7,303,419	7,140,547	7,140,538	7,249,884	7,201,555	7,256,317	7,303,423	7,303,423	7,303,419	7,303,423
	Records with error*	20,053	1,127,312	776,962	53,540	676,942	47,106	35,807	46,146	43,377	22,067
	Records with field entry after interpolation	7,300,493	7,215,908	7,216,941		7,257,686		7,303,423	7,303,423	7,300,643	7,302,206
	Records with missing entry after interpolation	2,930	87,515	86,482		45,737		0	0	2,780	1,217
	Percentage of errors fixed	85.4%	92.2%	88.9%	100.0%	93.2%	100.0%	100.0%	100.0%	93.6%	94.5%
	Percentage of records cleaned	100.0%	98.8%	98.8%	0.0%	99.4%	0.0%	100.0%	100.0%	100.0%	100.0%

7. Errors and Omissions in Residential Postcode and Census Output Area

7.1. The usefulness of PLASC data to examine residential movement relies heavily on the quality of the geographical information available in the data, both in terms of locating pupils and tracking any moves they may make. Within the PLASC datasets, the 'post' field refers to the pupil's reported residential postcode, while the 'oa' field, which is a Census Output Area marker, is added later by DCSF.

7.2. Using the appropriate All Fields Postcode Directory (AFPD, National Statistics Postcode Directory from 2006 onwards) for each year, pupils' recorded residential postcodes were checked to ensure that they represented valid UK postcodes in each year. That is, that the postcode was in use at the time it was recorded, and had not been terminated. Terminated postcodes are those which are no longer used by Royal Mail delivery. The most common reason for terminating postcodes is due to postcode reorganisation, or the demolition or redevelopment of buildings. From Table 7 it can be seen that the number of missing postcodes in each year is consistently low at around 1% of records, with the number of postcode entries found to be invalid similarly low, at between 1 and 2% each year. Invalid postcodes may occur as a result of typographic errors at the point of entry, errors in reported postcode, or continued use of a postcode following its termination.

Table 7 Quality of Postcode Field

	Valid Postcode (%)	Missing Postcode (%)	Invalid Postcode (%)	Total Records
2002	96.9	1.2	1.9	7745581
2003	98.3	0.2	1.5	7739246
2004	98.5	0.2	1.3	7733277
2005	99.1	0.1	0.8	7143366
2006	98.8	0	1.2	7316112
2007	98.3	0.7	1	7303423

7.3. As pupils' home postcodes are liable to change over time, it is not possible to use earlier interpolation procedures to remedy errors and omissions. Postcodes found to be invalid are therefore recoded as 'missing', excluding them from further analysis.

7.4. From 2003, when Census Output Areas were developed, the AFPD (used for postcode validation) also contains the Output Area associated with each valid postcode. These were checked against the Census Output Area ('OA') fields attached by the DCSF to each pupil record. PLASC recorded OAs were found not always to be the same as those recorded in the AFPD as being associated with particular postcodes. Furthermore, a small number of postcodes used in the PLASC were found not to have a corresponding OA. Further investigation of these postcodes revealed that all were 'large user' postcodes, usually assigned to workplaces rather than to residential addresses. Although such postcodes may be incorrect, they are retained in the data as the postcodes themselves are valid.

7.5. In order to improve the quality of OA level detail contained in the PLASC, the OAs originally recorded in the data were replaced with those from the AFPD. For 2002, where a postcode-OA look-up was not available, postcodes

were matched to the AFPD 2003 OA's. Where a postcode had been terminated prior to 2003, and thus did not have a related OA, the field was left blank in 2002.

8. Assessment of potential movers

8.1. The usefulness of PLASC data in measuring migration is based on the assumption that a change in postcode from one time period to the next is reflective of a change in home address. Having removed invalid postcodes, as above, the creation of a 'movetype' for each time period (2002-2003, 2003-2004, 2004-2005, 2005-2006, 2006-2007) allows assessment of the type of move that is assumed to have taken place within that period. For the purpose of this paper, move-types are categorised as no move, move within ward, move between wards but within district, and move within region but between districts, and between regions, but can be user-defined as required. Pupils who were absent from the system in one or both years are not allocated a move type.

8.2. In order to make move-type analysis more robust, a procedure adapted slightly from that developed by Burgess and Key (2008) was applied to all pupils with a change of postcode between two years. This procedure is intended to distinguish between changes of postcode that result from genuine residential moves, and those which occur as a result of postcode redistricting, or typographic errors. Postcode redistricting refers to process whereby Royal Mail change the postcodes within an area to accommodate new

developments, while typographic errors may occur at the point of data input, or if incorrect information is provided to the school.

8.3. The procedure to clean postcode changes for redistricting and typographic errors involved a number of stages. It is argued that where a postcode changes to one which is very similar, this is likely to be a result of Royal Mail making changes to the postcode, or of a typographic error occurring in one year (with the change probably reflecting a correction). The first stage of postcode-change cleaning, then, involved the removal of such postcode changes from the 'movers' group. Postcode changes are deemed not to be moves where: where there is a change in the length of a postcode by one character, with all other characters remaining unchanged ('a' in table 8); where the first or last two characters are coded in reverse compared with the previous postcode ('b' in Table 8); where either of the first or last two characters *only* change ('c' in Table 8); the first and last two characters remain unchanged ('d' in Table 8).

8.4. In their work to separate postcode changes that reflect residential moves from those that do not, Burgess and Key (2008) argue that apparent moves over very short distances are often unlikely to reflect actual moves. Following Burgess and Key's lead, moves between distances of less than 100 metres are here discounted. For this purpose, euclidean, or 'straight-line' distances are calculated using the Pythagorean theorem, thus it is distances of less than 100 metres between former and latter postcode centroid which are excluded.

8.5. The penultimate stage of postcode-change cleaning was to discount as residential moves occurrences where more than eight pupils are found to move between the same two postcodes. It is considered that such 'group' postcode changes are unlikely to reflect true *en masse* residential moves, but are more likely reflect Royal Mail postcode changes. The 'more than eight' cut-off is proposed by Burgess and Key (2008) as a sensible maximum family size. Thus, where all pupils move from one postcode to another, and 'all' is eight or fewer, it is likely that a residential move involving siblings has occurred.

8.6. The final stage is an extension of that described above, and again considers the likelihood of more than eight pupils moving. The distinction here is that where more than eight pupils move 'out' of a postcode, that postcode is not re-used. As with the previous stage, such 'moves' are likely not to reflect migration, but postcode redistricting.

Table 8 Moves excluded at each stage of postcode-change cleaning

	2002-03		**2003-04**		**2004-05**		**2005-06**		**2006-07**	
	Number	**%**	**Number**	**%**	**Number**	**%**	**Number**	**%**	**Number**	**%**
Total movers before cleaning	691672		659923		648090		579747		625554	
Character change (a)	159	0.02	160	0.02	173	0.03	181	0.03	135	0.02
Character change (b)	2671	0.39	2365	0.36	2344	0.36	1988	0.34	2433	0.39
Character change (c)	8	0.00	4	0.00	6	0.00	2	0.00	-	
Character change (d)	3355	0.49	2966	0.45	2866	0.44	2659	0.46	3153	0.50
Short-distance postcode change	41651	6.02	39476	5.98	37487	5.78	29445	5.08	25896	4.14
Group move between postcodes	458	0.07	819	0.12	1131	0.17	1115	0.19	372	0.06
Group move, postcode no longer used	1026	0.15	939	0.14	613	0.09	699	0.12	492	0.08
Total 'movers' excluded	49328	7.13	46729	7.08	44620	6.88	36089	6.22	32481	5.19

Note: Postcode-change exclusion is sequential, thus a change which is excluded at the first stage is not considered at subsequent stages.

8.7. Table 8 details the percentage reduction in moves for each part of postcode change cleaning. As expected, applying this procedure to PLASC data

reduces the number of recorded moves. The effect is greatest for moves within-ward (Figure 1), less for moves between wards but within the same district (Figure 2), and there is very little effect for moves within the same region but between districts (Figure 3), and those between regions. This is as expected, given that where former and latter postcodes are very similar, they are likely to be in the same ward or district. Interestingly, rates of movement following the removal of non-move postcode changes for pupils aged 5 to 15 in the PLASC show greater consistency with 2001 National Census rates for people aged 5 to 15. This is particularly true for within-ward moves, but also apparent for between-district moves. However, for between ward but within-district moves, where the impact of postcode change cleaning is minimal, PLASC recorded migration rates are found still to be higher than moves between 2000 and 2001, as recorded in the Census. This, though, is not in itself cause to question the validity of PLASC migration rates, as the Census is known to underestimate migration (Rees et al, 2002). For further comparison of PLASC migration rates and those recorded elsewhere, see Jivraj and Marquis (2009).

Figure 1 Within-ward migration rates for 5-15 year olds

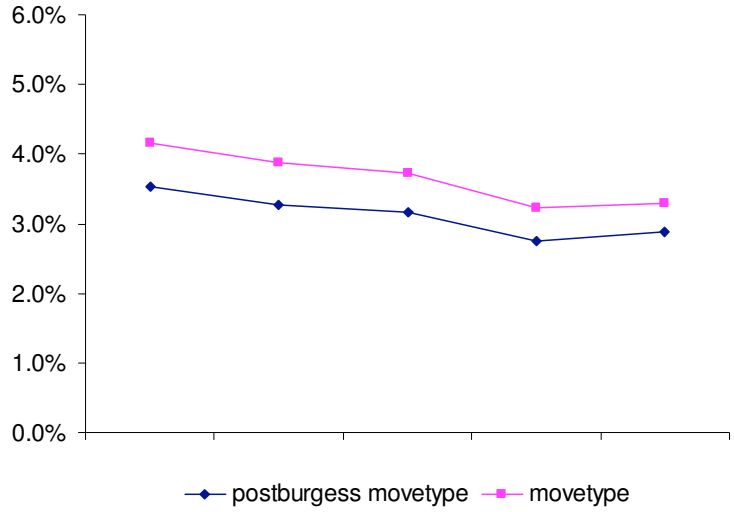


Figure 2 Between-ward but within-district migration rates for 5-15 year olds

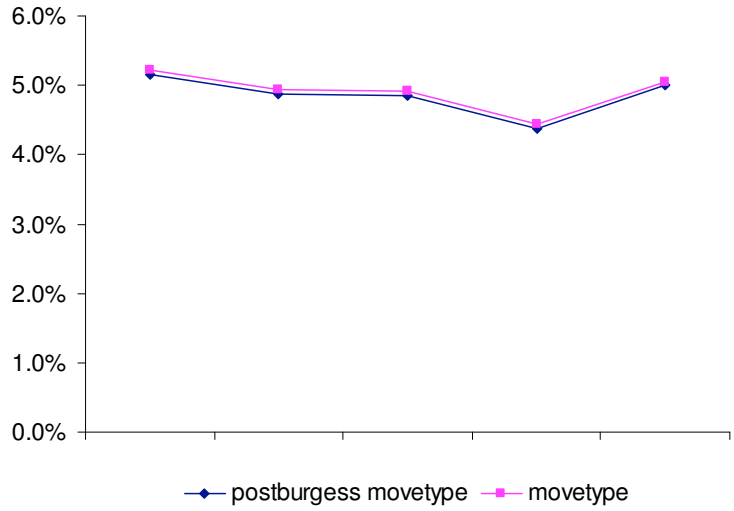


Figure 3 Between-district but within-region migration rates for 5-15 year olds

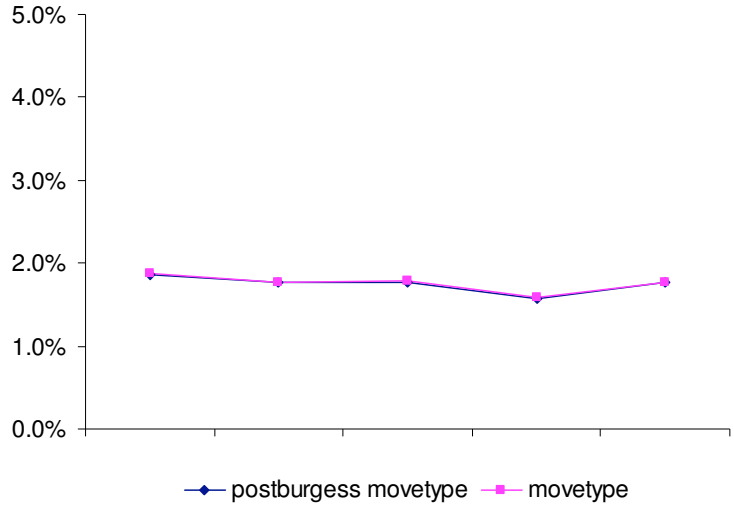
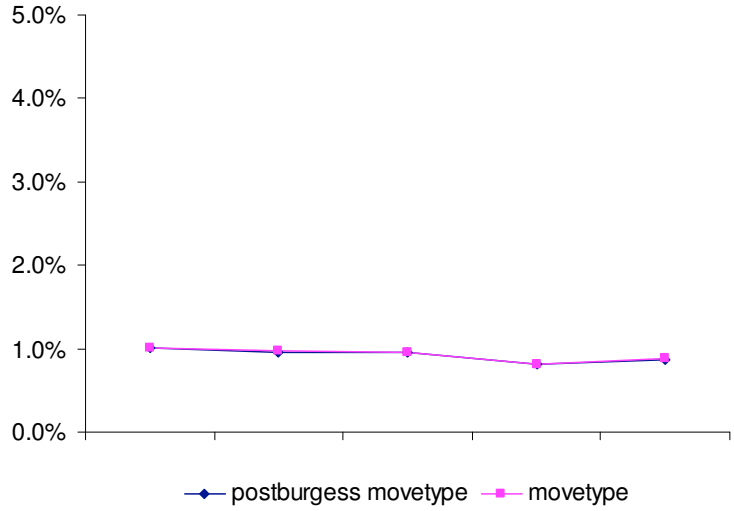


Figure 4 Between-region migration rates for 5-15 year olds



9. Conclusions

9.1. This report has detailed the content, coverage and quality of Pupil Level Annual School Census (PLASC) data supplied by the DCSF for the analysis of internal migration. The potential for PLASC to be used to measure migration comes from the inclusion of home postcode for each pupil, in a dataset which is updated yearly. Interpolation routines, as discussed, were used to improve the quality of data where records were found to contain

errors or to be inconsistent overtime. The creation of a set of 'final' attributes for each pupil, based on their entries over time, improves the quality of inconsistent records, and those with missing or invalid entries, though for most fields, improvement was slight due to the low occurrence of error.

9.2. Errors with postcode information supplied in the PLASC are low. For pupils who changed postcode between one year and the next, cleaning procedures have allowed the exclusion of postcode changes that are likely not to result from a residential move, thus improving the quality of movement as measured through the linkage of annual PLASC tables.

9.3. Overall, it is fields containing ethnicity information that are found to suffer most from error, though even the absence of data for 2002 results only in 6% of all records being without this extended or condensed ethnicity information after interpolation. Similarly, while temporal inconsistencies are highest for ethnicity attributes, such records only account for 8% of almost 11 million records. The absence of ethnicity information in other datasets, and the relatively low occurrence of error within the current data mean that the PLASC provides a potentially rich new data source for examining residential movement, though it would benefit from further investigation of errors in ethnicity fields. In particular, it would be useful to ascertain whether the quality of ethnicity fields may be improved in future PLASC releases.

References

BAILEY, N. & LIVINGSTON, M. (2005) Determinants of individual migration: an analysis of SARs data. *Working Paper 3.* Scottish Centre for Research on Social Justice

BURGESS, S., BRIGGS, A., MCCONNELL, B. & SLATER, H. (2006) School Choice in England: Background Facts. *Working Paper 06/159.* Centre for Market and Public Organisation, University of Bristol.

BURGESS, S. & KEY, T. (2008) School quality, school access and the formation of neighbourhoods. *http://www.bris.ac.uk/Depts/CMPO/PLUG/events/181108/burgess.ppt.*

DFES (2006) National Pupil Database: The Future. *http://www.bris.ac.uk/Depts/CMPO/PLUG/userguide/catherine.ppt.*

GODFREY, R. (2004) Statistical Topic Note: Changes in Ethnicity Codes in the Pupil Level Annual School Census 2002-2003. *http://www.dfes.gov.uk/rsgateway/DB/STA/t000455/index.shtml.*

HMSO (1996) Education Act 1996, Act of Parliament. Her Majesty's Stationery Office and Queens Printer of Acts of Parliament, London.

HOBBS, G. & VIGNOLES, A. (2007) Is free school meal status a valid proxy for socio-economic status (in schools research)? . *CEEDP, 84.* London, UK., Centre for the Economics of Education, London School of Economics and Political Science.

JIVRAJ, S. & MARQUIS, N. (2009) he Pupil Level Annual School Census: A new approach to measuring internal migration of school pupils in England. *http://www.ccsr.ac.uk/erm/2009-04-02/documents/PLASC_ERMpresentation_020408.pdf.*

JONES, P. & ELIAS, P. (2006) Administrative data as a research resource: a selected audit. *ESRC Regional Review Board Report 43/06.* Warwick Institute for Employment Research.

KOUNALI, D., GOLDSTEIN, H., ROBINSON, A. & LAUDER, H. (2007) The Probity of Free School Meals as a Proxy Measure for Disadvantage.

MEEN, G., GIBB, K., GOODY, J., MCGRATH, T. & MACKINNON, J. (2005) *Economic segregation in England: Causes, consequences and policy,* Bristol, Policy Press.

ONS (2004) Pupils and teachers: by type of school, 2002/03'. *Regional Trends 38.*

ONS (2005) Making a population estimate in England and Wales. *National Statistics Methodological Series No. 34.*

SIMPSON, L. & AKINWALE, B. (2007) Quantifying stability and change in ethnic group. *Journal of Official Statistics,* 23**,** 1-25.

TEACHERNET (2006) Changes to free school meals eligibility criteria from 6 April 2006. *http://www.teachernet.gov.uk/_doc/9530/FREE%20SCHOOL%20MEAL%20ELIGIBILITY%20CRITERIA%20FROM%206%20APRIL%202006.doc.*