(Working Paper 07/02)

Commuting to School in Leeds: How useful is the PLASC?

Kirk Harland and John Stillwell

Version 1.0
May 2007

School of Geography, University of Leeds,
Leeds, LS2 9JT, United Kingdom

Preface

Children's daily travel behaviour is dominated by the journey to school. In some cases, this movement takes only a few minutes and involves no means of transport other than foot; in other instances, the journey can be over substantial distances, be extensive in duration and involve some form of public or private transport. The combination of journeys taking place is likely to have a substantial impact on traffic congestion, particularly since the morning peak coincides with that associated with the journey to work. What datasets exist that allow us to measure and understand this behavour?

The National Pupil Database (NPD) is a relatively new information system that contains data collected through the Pupil Level Annual School Census (PLASC) since 2001 - 2002 academic year that can be used to identify where every state-educated school pupil is living and, as a consequence, what the spatial dynamics of commuting to school are. By examining the PLASC data for a Local Authority (LA) (*Education Leeds*), the paper seeks to explain what variables are available and how interaction data are identified. Moreover, the paper demonstrates a series of rigorous data validation checks and a suite of interpolation methods for estimating data that are either incorrect or unknown. In order to accomplish this, a database structure that allows annual PLASC data files to be linked together and to other school dataset is required. The database system and checking /interpolation methods described here provide a model for *Education Leeds* and other LAs to follow. Whilst the paper stops short of reporting any spatial analysis of the data contained in the Leeds system, it does acknowledge the commuting data that are available for Scotland from the 2001 Census Special Travel Statistics (STS) but advises that the investment in developing the PLASC data is likely to be more worthwhile than lobbying Office for National Statistics (ONS) to extend the STS in 2011 beyond Scotland to the rest of the UK.

Acknowledgements

This paper has been produced as part of an ESRC CASE postgraduate studentship funded by the ESRC and *Education Leeds* and we acknowledge this financial support. We are also grateful to *Education Leeds* for the supply of PLASC and other datasets, and to Heather Eyre in particular, for her continued advice and support.

Contents

List of Figures

List of Tables

Glossary of Terms

AFPD All Fields Postcode Directory

CMPO Centre for Market and Public Organisation

DfES Department for Education and Skills

ERA Education Reform Act

ERD Entity Relationship Diagram

ESRC Economic and Social Research Council

FSM Free School Meal

LA Local Authority

NPD National Pupil Database

OA Output Area

OFSTED Office for Standards in Education

ONS Office for National Statistics

OOP Object Orientated Programming

OS Ordnance Survey

PLASC Pupil Level Annual School Census

PLUG PLASC/NPD User Group

SCAM Small Cell Adjustment Method

SEN Special Education Need

SILC Special Inclusive Learning Centre

SMS Special Migration Statistics

SQL Structured Query Language

STS Special Travel Statistics

SWS Special Workplace Statistics

UPN Unique Pupil Number

Commuting to School in Leeds: How Useful is the PLASC?

1 Introduction

Education provision and performance vary across the world. Whilst the focus in developing countries is on making basic education available to everyone, the subject of education in the most developed nations has become more politically orientated with, for example, debates about accessibility to state schools through admissions procedures and ethnic and social segregation in schools (Gorard & Fitz, 2000; Johnston et al., 2005). Researchers in the UK have now begun to examine issues such as the critical relationships between demographic trends and school places and the links between schools and house prices in school territories.

Geographers have long been interested in identifying and explaining the spatial dimensions of education using indicators such as personal qualifications and examination attainment. Much less attention has been paid by researchers to the way in which the demand for schooling at different age levels is met by the supply of primary and secondary schools and the consequences for pupils commuting between their place of residence and their place of study under particular admissions policy regimes. Estate agents report that education is now one of the most important reasons why people move house. Freedom of school choice, introduced by Labour under the 1988 Education Reform Act (ERA), means that perceived good schools become oversubscribed, creating major problems for the schools concerned and for their LA when it comes to formulating fair admissions policies. One consequence has been the relative increases in the house prices in areas with perceived good schools. More affluent parents have moved closer to ensure access to school places for their children at perceived good schools, thus creating less opportunities for places to be assigned to those pupils from less affluent families. The situation is causing LAs to consider using electronic balloting or lottery-style methods of allocating children to schools and there is considerable political interest and press coverage of this issue at the moment across the UK.

It is also pertinent to recognise the importance that this Government attaches to evidence and particularly to ‘evidence-based policy making’. This means that information which can be derived from the analysis of reliable data about the demand for and supply of education is required. In the United States of America, the Na-

tional Center for Education Statistics (NCES) collects and collates a multitude of surveys and longitudinal data for education at elementary, secondary and postsecondary levels, including fiscal and non-fiscal data at pupil, school and state level. In England and Wales, the Department for Education and Skills (DfES) collates education data into a central database called the National Pupil Database (NPD).

This paper takes a closer look at the NPD and the pupil level data collection Pupil Level Annual School Census (PLASC), with a view to using the latter for analysis of the magnitude, composition and spatial patterns of commuting to school in Leeds since 2001, which is the particular focus of our research. In fact, the PLASC datasets described in Sections 3 and 4 of the paper are those data supplied by the Local Authority (LA), *Education Leeds*, and are supplemented by other sets of data on schools, attainment and preferences. The paper aims to clarify the nature and content of the Leeds PLASC data and to demonstrate how essential it is to check and clean the data prior to its analysis. The structure of the database that has been designed for storing, linking and checking the various datasets provided by *Education Leeds* is outlined in Section 3 whilst the methods used for verification and interpolation are explained in detail in Section 4 of the paper. This section is likely to be of particular interest to *Education Leeds* and to other LAs lacking sufficient resources to commit to data validation processes. The paper does not include any spatial analysis of the PLASC data for Leeds; preliminary analysis is reported in Harland & Stillwell (2007). However, the detailed documentation of PLASC data contained in this paper does provide the opportunity for some comparisons to be made with travel to study data available from the 2001 Census of Population for Scotland and allow conclusions to be drawn about whether Special Travel Statistics (STS) in 2011 might usefully be extended to cover the rest of the Great Britain. We begin, however, with a synopsis of the national context of data collection in this sector.

2 The National Pupil Database and Pupil Level Annual School Census

The NPD is a relatively new dataset created in 2002 and contains individual pupil records for all state educated school children (Ewens, 2005). It is updated on an annual basis with additions in excess of 8 million individual pupil records collected

by each LA in England and Wales and is maintained by the DfES (Jones & Elias, 2006). Access to the NPD has recently been provided through a central gateway funded jointly by the DfES and the Economic and Social Research Council (ESRC) and managed by the Centre for Market and Public Organisation (CMPO) at the University of Bristol where the PLASC/NPD User Group (PLUG) is based (Burgess et al., 2006). The NPD is stored in a relational database structure with several different datasets capable of being linked together using either a UPN or a unique establishment identification number to allow for both temporal and cross-sectional analysis, creating a powerful information resource for policy formulation (Jones & Elias, 2006).

Completion of the Pupil Level Annual School Census (PLASC) is statutory for all state maintained primary, secondary and special schools under section 537A of the Education Act 1996 (Jones & Elias, 2006). The DfES began collection of the data in 2002 and it now forms the cornerstone of the NPD. Individual schools are required to submit a PLASC return to the LA on the third Thursday of January each year. The return consists of entries for every pupil on role with data such as home postcode, ethnicity, Special Education Need (SEN) status and Free School Meal (FSM) eligibility, plus information relating to the school and its staff. In actual fact, the data collection of pupil information is no longer referred to as PLASC because a tri-annual data collection procedure called the School Census with a modular structure was introduced in 2006 for secondary schools and will be introduced in 2007 for primary schools (DfES, 2006b). One of the three data collections will still be carried out in January, with two further collections on the third Thursday in May and the third Thursday in September augmenting the January collection (DfES, 2006a). The tri-annual data collections coincide with the three school terms and will enable more effective tracking of pupil migrations; specifically moves between homes and moves between schools, throughout the year.

On receipt of the PLASC data in January each year from 2002 to 2006, the Leeds LA collated all the entries from individual schools and carried out data validation. Details relating to the collection of PLASC data in Leeds were obtained during meetings with *Education Leeds* data management staff. Commonly occurring problems detected in the validation process include duplicate pupils in the dataset, especially when a pupil has changed school mid-year and appears to be attending two institutions at the same time, mistyped entries where a data entry

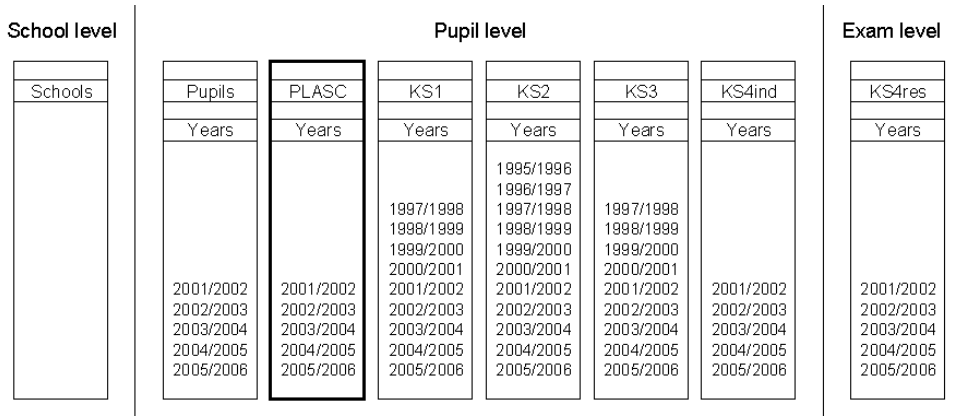


Figure 1: National Pupil Database tables overview, December 2006
Source: adapted from Jones & Elias (2006, p.7)

does not match an expected list and data omissions where no entry is made. These issues are referred back to the schools concerned for rectification. This was, and still remains for the tri-annual collection, an iterative process until the data are as robust as possible. On completion of the data checks by the LA, the PLASC data are submitted to the DfES in March for further cross-LA border validation, initiating another round of iterative data corrections to rectify any issues pertaining to multiple LAs. The most frequently occurring problems at this stage are pupils that have moved schools, leaving a school in one LA to attend a school in a different LA area and are shown to be attending both institutions in different LA areas at the same time. After compilation into a national PLASC dataset by the DfES, the collected data are integrated into the NPD tables where they can then be linked to the other data tables at school, pupil and exam level (Figure 1) through the use of the UPN and the unique establishment identification number.

The varying number of years of collected data tables can be seen in Figure 1 which also shows how the tables in the NPD are represented at three different levels. At the *school level*, only one table exists and this holds data relating to individual schools such as school capacity, number of computers and number of staff by qualification. At the *pupil level*, six tables exist, each holding data for pupils at the different stages of their school careers. The 'Pupils' table is a master table holding information on every pupil appearing in the NPD. This differs from the 'PLASC' table because not every pupil in the NPD attends a state school and is therefore not subject to the statutory pupil census.

Pupils attending independent schools sit some, if not all, of the Key Stage examinations and therefore must have a recorded UPN to enable linking between tables. Pupils may also move out of the state education system into private education at some point in their schooling and it is important that the NPD be structured in a way that allows these moves to be tracked. Therefore, the 'PLASC' table contains only pupils from the state education system at the point in time when the data were collected, and the 'Pupils' table contains a record of all pupils in the NPD, both in the state sector and in the private sector. Each of the Key Stage (KS) tables records the results of Key Stage tests at the individual pupil level. The *exam level* has only one table which contains data relating to exam results at a school level rather than at the individual pupil level. KS tables have been collected since 1995/96 whilst PLASC tables were first assembled in 2001/02.

Figure 2 is an alternative representation of the datasets held in the NPD but in this case the individual years of data are represented by coloured blocks. Figure 2 shows that 'Key Stage 2' ('KS2' in Figure 1) is the dataset with the most years of collected data and the 'Foundation Stage Profile' is the dataset with the least. Additionally, Figure 2 also shows how each dataset links to others within the database building up the temporal dimension. Note that to follow a cohort of pupils through the database, it is not a simple task of linking one collection year to the next as the KS tables are collected with different time spans between them. For example, following the path highlighted by the dashed lines shows that the link between Key Stage 4 results tables collected in '04/05' and Key Stage 3 results for this particular cohort of pupils skips the collection year '03/04' and links to collection year '02/03'. This is because the Key Stage 3 tests are taken in year 9, two years before the Key Stage 4 tests taken in year 11. Following the path back further to the link between the year 7 progress tests and the Key Stage 2 tests, it can be seen that the link is just one year as the tests are taken in successive years.

This example illustrates the complex nature of the linkages between the different datasets with differing time spans between tests and datasets having differing numbers of years of recorded results (some starting in 1995/96 and some as late as 2002/03). This has lead to data in certain years being unable to be linked to result tables in other datasets within the database, limiting the temporal aspect of any analysis performed on this early cohort of pupils. The unlinkable datasets are limited to the three data tables highlighted in red, and only one of these data tables

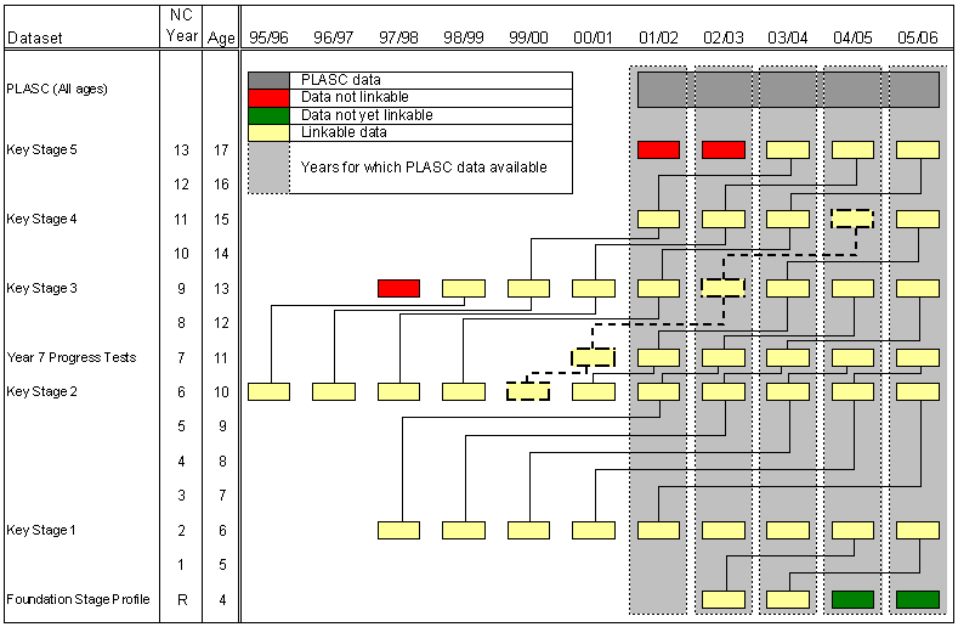


Figure 2: National Pupil Database datasets and linkage, December 2006
Source: adapted from Jones & Elias (2006, p.6)

is not linkable to the PLASC dataset. The datasets highlighted in green are unable to be linked in 2006 but will become linkable in subsequent years as new data are added for 2006/07, 2007/08 and so forth.

PLASC and the master 'Pupil' table form one part of the NPD holding the pupil characteristics and the UPN. The second half of the NPD is constructed using the attainment tables Key Stage 1 through to Key Stage 5 (represented by the 'KS1', 'KS2', 'KS3' and 'KS4ind' tables in Figure 1). "In particular, the attainment data includes

- Levels attained in reading, writing and mathematics at Key Stage 1;
- Levels attained and the marks achieved in English, Mathematics and Science at Key Stage 2 and Key Stage 3;
- GCSE or equivalent results;
- AS and Advanced level results." (Jones & Elias, 2006, p.5).

The PLASC data collection and the master 'Pupil' table combined with the attainment tables form the NPD. With the data being compiled from as far back

as 1995/96 for some of the datasets, the ever-increasing longitudinal search and analysis possibilities play an important role in educational model formulation as noted by Jones & Elias (2006, p.6). "Statistically, such data allow for more precise formulation of models to consider the effect on educational outcomes of, say, pupil attributes or circumstances, policy interventions, and so on." Furthermore, Ewens (2005, p.4) comments that "the National Pupil Dataset is amongst the most important national innovations in data collection in the recent past. Its potential is considerable and the scope for development is also considerable." However, he also notes that the NPD is still a relatively new and unknown dataset and holds many pitfalls for inexperienced or 'unwary' users.

3 Data Supplied by *Education Leeds*

3.1 PLASC data

Whilst Section 2 has introduced the national database in which the PLASC are embedded, the PLASC data utilised in this study have been supplied by *Education Leeds* for six years, 2000/01 through to 2005/06. The data have been supplied in a Microsoft Access database with data for each year in a separate table that contains the fields as shown in Table 1. It is evident that not all fields are present across all years of the PLASC data and not all field types are consistent across the years when they are present. However, it is possible to convert the inconsistent data types using either explicit or implicit 'cast' functions without loss of data to ensure data compatibility in the temporal dimension. Cast functions are native functions in a programming language and are designed to convert data from one datatype to another datatype with minimal data loss.

The building of longitudinal properties of the pupil data collection provides for powerful analysis of the data, especially when considering policy formulation and its impacts. However, the structure of the data must be stable throughout the duration of collection, otherwise it is possible to introduce truncation of data items leading to possible loss of data and inconsistencies across years. The PLASC data for Leeds are currently stored in a Microsoft SQL Server database at *Education Leeds* as one part of a much wider data resource for the Leeds Metropolitan District, in addition to their integration in the NPD. The data supplied for this study have been extracted from this SQL Server database into Microsoft Access for-

2001		2002		2003		2004		2005		2006	
Name	Type	Name	Type	Name	Type	Name	Type	Name	Type	Name	Type
Dummy Pupil ID	Number (LONG)	Dummy Pupil ID	Number (LONG)	Dummy Pupil ID	Number (LONG)	Dummy Pupil ID	Number (LONG)	Dummy Pupil ID	Number (LONG)	DUMMY PUPI	Number (Double)
DFEE	Number (DOUBLE)	Estab	Number (LONG)	Estab	Number (INTEGER)	Estab	Number (INTEGER)	Estab	Number (INTEGER)	ESTAB	Number (DOUBLE)
YEAR	Text (5)	Actual-NC-YearGrp	Text (2)	NCYear-Actual	Text (255)	NCYear-Actual	Text (2)	NCYear-Actual	Text (2)	NCYEAR-ACTU	Text (2)
DATE-_OF_BI	Date/Time	Dateof-Birth	Date/Time	DOB	Date/Time	DOB	Date/Time	DOB	Date/Time	DOB	Date/Time
GENDER	Text (10)	Gender	Text (1)	Gender	Text (255)	Gender	Text (1)	Gender	Text (1)	GENDER	Text (1)
ETHNI-CITY	Text (10)	Ethnic-Group	Number (BYTE)	Ethnicity	Text (255)	Ethnicity	Text (4)	Ethnicity	Text (4)	ETHNI-CITY	Text (4)
MOTHER-_TON	Text (10)	Mother-Tongue	Text (3)	First-Language	Text (255)	First-Language	Text (3)	First-Language	Text (3)	FIRST-LANGU	Text (3)
MEAL	Text (10)	Free-School-Meal	Text (1)	FSM-Eligibility	Text (255)	FSM-Eligibility	Text (1)	FSM-Eligibility	Text (1)	FSM-ELIGIBI	Text (5)
DATE-_OF_AD	Date/Time	Dateof-Joining	Date/Time	Entry-Date	Date/Time	Entry-Date	Date/Time	Entry-Date	Date/Time	ENTRY-DATE	Date/Time
POST-CODE	Text (10)	Home-Postcode	Text (8)	Postcode	Text (255)	Postcode	Text (8)	Postcode	Text (8)	POST-CODE	Text (8)
XCOORD	Number (DOUBLE)	X	Number (DOUBLE)	X	Number (DOUBLE)	x	Number (DOUBLE)	X	Number (DOUBLE)	X	Text (50)
YCOORD	Number (DOUBLE)	Y	Number (DOUBLE)	Y	Number (DOUBLE)	y	Number (DOUBLE)	Y	Number (DOUBLE)	Y	Text (50)
		SEN-Stage	Text (1)	SEN-Status	Text (255)	SEN-Status	Text (1)	SEN-status	Text (1)	SEN-STATUS	Text (1)
		Registra-tionType	Text (1)	Enrol-Status	Text (255)	Enrol-Status	Text (1)	Enrol-status	Text (1)	ENROL-STATU	Text (1)
				InCare	Text (255)	InCare	Text (1)	InCare	Text (1)	INCARE	Text (5)
						Primary-SENType	Text (4)	Primary-SENType	Text (4)	PRIMA-RYSEN	Text (4)
						Secondary-SENType	Text (4)	Secondary-SENType	Text (4)	SECON-DARYS	Text (4)

Table 1: The structure of the pupil level PLASC tables supplied by *Education Leeds*, 2001-2006

mat. Despite the supplied PLASC data tables originating from the same source, the methods of extraction must have been inconsistent to produce the difference in data types between the fields shown in Table 1.

The first field shown in Table 1 for all years is 'Dummy Pupil ID' (truncated to 'DUMMY PUPI' in year 2006). This field has a datatype of 'Number (LONG)' for all years except 2006 where it has a datatype of 'Number (DOUBLE)'. Each datatype requires a specific amount of computer memory space be set aside to accommodate the data that datatype variables or fields will contain. A common misuse of datatypes in databases is to set fields to be a datatype requiring more computer memory space than is actually necessary. This process artificially increases the database size and slows down any queries performed on the database (Self & Dunckley, 2003a). Thus, it is important to standardise all tables that hold similar data using the appropriate datatypes to represent the data during the database design phase.

The most common difference across the PLASC tables is the length associated with text field datatypes. In 2003, all the text datatype fields are set to 255, the maximum length for a text field in Microsoft Access and also the import wizard default. The inconsistency that would cause the most disruption to analysis is the change of datatypes between the coordinate fields from 'Number (DOUBLE)' through years 2001 to 2005 to 'Text (50)' in year 2006. Any attempt to geocode the home locations for the pupils in year 2006 without first altering the datatype of the coordinate fields would fail, due to them not being a numeric datatype.

Table 1 summarizes the structure of the PLASC data provided by *Education Leeds*. The list below expands the description of each field, whilst the content of the tables will be considered in more depth in Section 4. The variables available in all years are as follows:

Dummy Pupil ID/DUMMY PUPI henceforth referred to as the UPN and, as with the original UPN, this tracks a child through their school career, allowing temporal analysis. The original UPN has been replaced with a dummy UPN in order to ensure that pupil records are untraceable and thus meet the strict data disclosure policies of *Education Leeds*.
Generic field name 'UPN'.

DFEE/Estab/ESTAB is the unique identification number associated with the institution the child attends.

Generic field name 'Estab'.

YEAR/ActualNCYearGrp/NCYearActual/NCYEARACTU refers to the actual year group of schooling the child is currently attending.
Generic field name 'NC_Year'.

DATE_OF_BI/DateofBirth/DOB is the pupil's date of birth.
Generic field name 'DOB'.

GENDER/Gender is the pupil's gender.
Generic field name 'Gender'.

ETHNICITY/Ethnicity/EthnicGroup is the coded representation of the pupil's ethnic orientation. The code structure changed in 2003, moving from a 14 category structure to a 20 category structure that is comparable with the ethnicity code structure used in the 2001 Census.
Generic field name 'Ethnicity'.

MOTHER_TON/MotherTongue/FirstLanguage/FIRSTLANGU is the pupil's main language and is coded into six categories.
Generic field name 'First_Language'.

MEAL/FreeSchoolMeal/FSMEligibility/FSMELIGIBI is a flag that relates to the pupil's eligibility to claim a Free School Meal (FSM), although this does not identify that the pupil actually takes up the FSM.
Generic field name 'FSM'.

DATE_OF_AD/DateOfJoining/EntryDate/ENTRYDATE is the date the pupil began attending the institution referred to in the DFEE/Estab field.
Generic field name 'Entry_Date'.

POSTCODE/HomePostcode/Postcode is the pupil's home postcode.
Generic field name 'Postcode'.

XCOORD/X/x is the geocoded x coordinate of the pupils home. The geocoding of both X and Y coordinates has been undertaken by *Education Leeds* and uses Address Point to extract the easting and northing grid coordinates for each pupil's home address.
Generic field name 'X'.

YCOORD/Y/y is the geocoded y coordinate of the pupil's home.
Generic field name 'Y'.

The variables available for the years 2002 through to 2006 are as follows:

SENStage/SENStatus/SENSTATUS is a flag that shows if the pupil has a Special Education Need (SEN).
Generic field name 'SEN'.

RegistrationType/EnrolStatus/ENROLSTATU is a flag showing whether the pupil is currently enroled at the institution.
Generic field name 'Enrolment_Status'.

The variable available for years 2003 through to 2006 is as follows:

InCare/INCARE is a flag to indicate if the pupil is currently in care.
Generic field name 'In_Care'.

The variables available for years 2004, 2005 and 2006 is as follows:

PrimarySENType/PRIMARYSEN is a coded indicator of the pupil's primary Special Education Need (SEN).
Generic field name 'Primary_SEN'.

SecondarySENType/SECONDARYS is a coded indicator of the pupil's secondary Special Education Need (SEN) if the pupil has more than one SEN.
Generic field name 'Secondary_SEN'.

3.2 Supplementary data

In addition to the pupil level data supplied by *Education Leeds*, several spreadsheets in Microsoft Excel format have also been received. These are shown in Table 2 together with a summary description of the contents. The additional data contained in the spreadsheets are not intended to provide a comprehensive time series equivalent of PLASC but will be used in later analysis to help quantify the attractiveness of schools. The resulting attractiveness values will be utilised in the modelling techniques applied to the PLASC data. The first three files contain spreadsheets with information relating to the LA unique identifier, the establishment unique identifier, a teacher category, number of full-time male, part-time male,

File Name	File Contents
2004 teacher PLASC.xls	Summary of teaching and non-teaching staff at each school
2005 teacher PLASC.xls	
2006 teacher PLASC.xls	
2004 school PLASC.xls	Summary of school attributes
2005 school PLASC.xls	
2006 school PLASC.xls	
net capacity.xls	Maximum overall school capacity and an actual attendance value from an unstated year
School_Admissions_Limits.xls	Maximum capacity figures for the intake year for each school
overall inspection grade.xls	Current OFSTED inspection grades
2004 all prefs.xls	Raw parental school choice data and corresponding school allocation

Table 2: Excel spreadsheets containing data supplied by *Education Leeds*

full-time female and part-time female staff members in each individual category. The next two files contain information about schools. The 2004 school spreadsheet contains the LA unique identifier, followed by the establishment unique identifier, establishment name, teaching phase, type of intake, type of governance, lowest national curriculum year and highest National Curriculum year. The 2005 and 2006 school spreadsheet contains the same fields as 2004 with the addition of a count of computers at each institution.

The 'net capacity.xls' file contains the unique establishment identifier, establishment name, the net capacity of the school and the number of pupils on role. From the latter two columns, the number of surplus places and the percentage of surplus places in each school have been derived. 'School_Admissions_Limits.xls' contains both the unique identifier of the establishment and its name along with the maximum number of pupils that can be admitted to the intake year group; 'overall inspection grade.xls' again contains both the unique establishment identifier and the establishment name with information on the Office for Standards in Education (OFSTED) framework that the school was last inspected on and the grade that was achieved.

Finally, '2004 all prefs.xls' contains information on the child, the preferences of schools they would like to attend and also the allocation process. Information contained in the sheet relating to the child is the category, or type of school, the

child will attend into in the coming academic year (r = reception/infant school, j = junior school and h = high school), a unique identifier for the child made up of a combination of the category identification letter and an incremented number, gender, current institution attended and its name, whether the child has a SEN and location coordinates geocoded from the child's home postcode. The preference information is held in the form of a preference number (1 to 5), the unique identifier of the corresponding chosen establishment and a yes/no flag to indicate whether the child has a sibling at the corresponding establishment. Information held relating to the allocation of institution is the unique establishment identifier and name of the allocated school, the stage of allocation and finally the reason for allocation of institution in relation to the LA's admissions policy.

3.3 Database design

The use of databases to organise vast quantities of data is widespread in today's business and public service sectors. Self & Dunckley (2003b, p.7) define a database as "a collection of data stored in a computer system." Different database models exist for structuring how the data are stored. The most common models being: hierarchical, network, relational, object-relational and object. Currently, the most widely used model is the relational model first presented by Codd (1970) that was superior to those in use at the time of invention. It was only after the birth of the relational data model that the hierarchical and network models were produced to fit the current systems and enable comparisons to be drawn. The design process associated with relational databases promotes data independence from other data and also separation from information systems that use the data through the application of Structured Query Language (SQL). The database design process also encourages the description of the data using its own natural structure rather than imposing an arbitrary structure onto the data (Codd, 1970).

The development of object and object-relational databases has been troubled by a lack of general uptake. Object databases are designed to be used with Object Orientated Programming (OOP) languages such as Java, Visual Basic.Net or C++ and use the same design model as the actual programming language. The object relational database model uses the same relational structure but with an object orientated mapping schema to make integration with OOP more effective. Both these database models have their advantages and disadvantages. Object databases

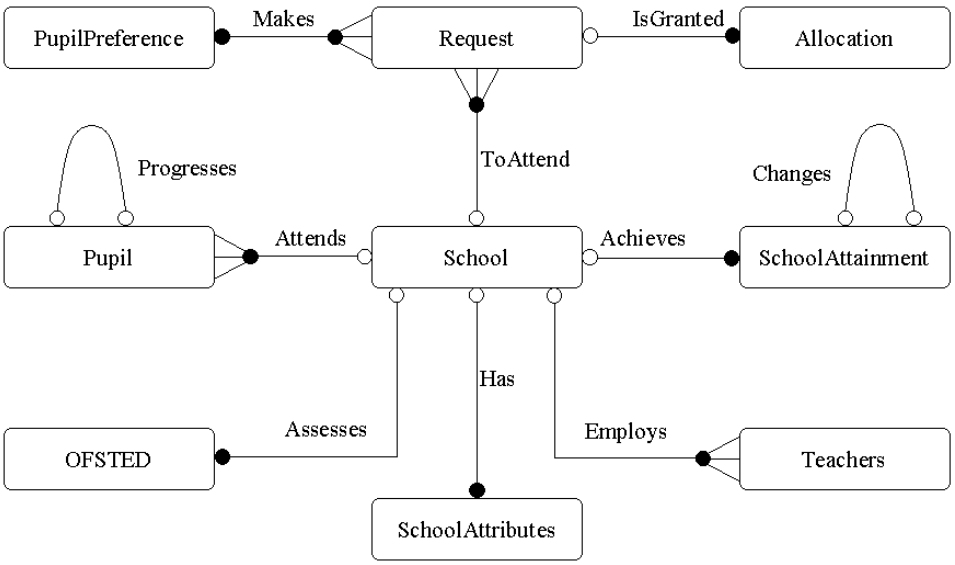
hold the record for the largest database and the world's fastest data ingestion rate. However, the encapsulation paradigm that is the fundamental building block of both OOP and object databases works against the general acceptance of the object database model over the relational model. Encapsulation demands that the data be hidden away inside an object and only be available through predefined programmed interfaces, whereas the relational model allows access to the data, for viewing and editing, as long as the user has sufficient security privileges. The prospect of using a high-level programming language such as Java or C++ to access data held within a database is a daunting prospect for many database users, whereas the use of a simple query builder or easy to learn descriptive language such as SQL is much more attractive; thus the relational model continues to thrive.

Because of the wider acceptance of the relational data model and because both the NPD and the locally held pupil database at *Education Leeds* are relational databases, this study will use the relational database model as the main database structure. The relational database design process consists of two main stages. The first stage is to produce an Entity Relationship Diagram (ERD) using the rules associated with entity relationship modelling. This can be broken down into three steps as presented by Dawson (2001):

1. Identify the entities and attributes.
2. Determine the relationships.
3. Combine the entities and relationships into one diagram.

The second stage is to go through a process called data normalisation which removes duplicate data and can be used to check that the original ERD has the correct structure. The process of normalisation breaks larger tables down into smaller ones that relate to each other. However, it is possible to over-normalise a database which then becomes too defragmented with large numbers of defined relationships which, in turn, increases processing times for queries. The ideal balance is to reduce data duplication to a minimum and eliminate unused space in fields whilst maintaining the maximum size for tables. For an in depth discussion, refer to Dawson (2001). The final ERD of the database design established to house the data described in Section 3 is shown in Figure 3.

To understand this diagram it is important to have a grasp of the general notation for ERDs. Each rounded box is an entity. An entity is not a table in a database



Entity types

Allocation (Pref_UPN, Estab, Stage, Reason)

Request (Pref_UPN, Preference, Estab, Sibling)

PupilPreference (Pref_UPN, Pupil_Category, Gender, Current_Estab, SEN, X, Y)

Pupil (UPN, Estab, NC_Year, DOB, Gender, Ethnicity, First_Language, FSM, Entry_Date, PostCode, X, Y, SEN, Primary_SEN, Secondary_SEN, Enrolment_Status, In_Care, Ethnicity_Stand, Postcode_Stand)

School (Estab, Estab_Name)

SchoolAttributes (Estab, Phase, Intake, Governance, Lowest_NC_Year, Highest_NC_Year, Computers, Net_Capacity, Intake_Year_Capacity)

SchoolAttainment (Estab, KeyStage, English, Math, Science, AveragePointScore)

OFSTED (Estab, Framework, Grade)

Teachers (Estab, Teacher_Category, Male_FT, Male_PT, Female_FT, Female_PT)

Figure 3: Entity Relationship Diagram of the Leeds study area database

but is a conceptual representation of a set of related data attributes that can be grouped together to form a logical block representative of a real world object (Self & Dunckley, 2003b). Indeed, an entity can end up being the diagrammatic design representation of a single table or it can represent multiple tables that have the same structure and contain similar attributes. The entity structure is displayed in the list at the bottom of the ERD under the heading entity types. Each entity has an entry in the entity type list and a list of attributes contained by the entity type are shown within the parentheses that follow. The underlined attributes at the beginning of the list indicate an attribute or a combination of attributes that will be unique for every row within a table realisation of the entity. The unique row identifier for a table is referred to as the primary key. Other types of table keys exist, such as alternate and foreign keys, however a full explanation of these features fall outside of the remit for explanation here but a full description can be found in Self & Dunckley (2003b) or Dawson (2001).

The lines between the entities represent relationships between the entity types. Each relationship has a name assigned to it which is descriptive of the way the entity types relate to each other. For example, in Figure 3, the 'Pupil' entity has a relationship of 'Attends' with the 'School' entity describing that pupils attend a school. The circles at the end of the relationship lines show relationship participation. If the circle is black then participation in the relationship is compulsory. Every row in a table represented by an entity with compulsory participation must participate in the relationship. If the circle is white then the participation is optional, indicating that rows in represented tables may or may not take part in this relationship.

The end type of a relationship line dictates whether that particular side of the relationship has a one or a many involvement with the relationship. If it ends in a participation circle then that side of the relationship is a 1: indicating that only one row of any represented table will be involved in the relationship at any one time. If the end of the relationship ends with the participation circle and then three lines opening out to the entity (sometimes referred to as a crow's foot) then this indicated that the relationship at this end is an n: or many relationship and many rows of a represented table may take part in the relationship at the same time. Using the example of the 'Attends' relationship a 'Pupil' may only attend one 'School' thus the 'School' end of the relationship is a :1. However, a 'School'

has many attending 'Pupil' entries thus the 'Pupil' end of the relationship is an n:. The 'Attends' relationship is described as an n:1 or many to one relationship. Two further relationship types can be defined, these are 1:1 (one to one) or n:m (many to many). An example of a one to one relationship is the 'progresses' relationship in Figure 3. Because this relationship begins and ends on the 'Pupil' entity, it indicates that only one row in a table representation of the entity type 'Pupil' can refer to one row in either the same table or another table of the 'Pupil' entity type. The many to many relationship type has not been required for the design of this database, therefore it will not be discussed in any detail. For further information on ERD notation and general database design, refer to Dawson (2001) or Self & Dunckley (2003a,b,c,d).

The implementation of the above database design reduces the amount of duplicated data to a minimum while maintaining flexibility in the database structure. Table 3 displays the tables created within the database, the entity structure that each table is based upon and the source of the data. Examining Figure 3 and Table 3 together shows the data structure for each of the tables in an entity type and how tables from different entity types relate to each other. The 'School' entity type has only one table based upon it which is also called 'School', although this table draws data from three spreadsheets supplied by *Education Leeds* and is also supplemented by data extracted from EduBase, the DfES online database of education institutions. The 'School' table is the master school table and contains an entry for every school that has been open during the period 2001 through to 2006.

The 'School' table can be linked to the three tables representing the 'SchoolAttribute' entity type through the relationship 'Has'. This relationship utilises the 'Estab' fields and forms a link between the master 'Schools' table and the changable 'SchoolAttributes' that were associated with a particular school in a given year. Because not all schools are open in all years, the 'Has' relationship participation is optional at the 'School' side but compulsory on the 'SchoolAttribute' side as every school must appear in the master table.

The 'Pupil' entity type has six tables based on it, each one holding the data collected in one year and taken from the PLASC data supplied in Microsoft Access database format by *Education Leeds*. Each 'Pupil' table contains a field with the identifier of the school attended called 'Estab'. This field can be joined with the corresponding field in the 'School' table also called 'Estab' to fulfill the re-

Entity Name	Table Name	Data Origin
School	School	2004 school PLASC.xls
		2005 school PLASC.xls
		2006 school PLASC.xls
		with additional information for closed schools taken from the DfES EduBase database
SchoolAttributes	School_2004	2004 school PLASC.xls
		net capacity.xls
		School_Admissions_Limits.xls
	School_2005	2005 school PLASC.xls
		net capacity.xls
		School_Admissions_Limits.xls
	School_2006	2006 school PLASC.xls
		net capacity.xls
		School_Admissions_Limits.xls
Pupil	PLASC_2001	PLASC databases
	PLASC_2002	
	PLASC_2003	
	PLASC_2004	
	PLASC_2005	
	PLASC_2006	
SchoolAttainment	Results_2001	DfES Attainment tables
	Results_2002	
	Results_2003	
	Results_2004	
	Results_2005	
	Results_2006	
PupilPreference	PupilPreference	2004 all prefs.xls
Request	PreferenceRequest	
Allocation	PreferenceAllocation	
OFSTED	OFSTED	overall inspection grade.xls
Teachers	Teachers_2004	2004 teacher PLASC.xls
	Teachers_2005	2005 teacher PLASC.xls
	Teachers_2006	2006 teacher PLASC.xls

Table 3: Tabular realisation of entities and origin of data content

lationship ‘Attends’. The ‘UPN’ field in the ‘Pupil’ tables can be joined to the ‘UPN’ fields in other ‘Pupil’ tables to form a temporal join across several years of data fulfilling the ‘Progresses’ relationship. The optional participation of the ‘Progresses’ relationship at both ends is required to cater for pupils starting school thus

not being present in previous years' data or leaving school so not being present in subsequent years' data.

The 'SchoolAttainment' entity is realised into six separate tables holding results of Key Stage 2, 3 and 4 examinations taken from the DfES attainment tables for the years 2001 through to 2006. The 'Estab' field in the 'School' table can be linked to the 'Estab' field in any of the 'SchoolAttainment' entity type tables although as schools are closed over the period 2001 to 2006 they do not appear in the 'SchoolAttainment' tables thus the optional participation of the relationship 'Achieves' on the 'School' side. The 'Estab' field in the 'SchoolAttainment' tables can be used to join the tables across years to assess 'Changes' in the performance of schools. However, the participation of both ends of this relationship have to be optional to allow new schools to be opened and existing schools to be closed.

The '2004 all prefs.xls' is split into three different logical entity types through the application of the normalisation process. The three different entity types are realised into three tables called 'PupilPreference', 'PreferenceRequest' and 'PreferenceAllocation'. The 'PupilPreference' table contains all the data relating to the pupil including a 'Pref_UPN' field which unfortunately does not allow a join to the 'UPN' field in the 'Pupil' entity type tables. However, the 'Pref_UPN' field does allow a join to the 'PreferenceRequest' table realising the 'Makes' relationship. The 'PreferenceRequest' table holds information about the requests made to attend institutions by a pupil. Each pupil will make one or more requests and so the 'Makes' relationship is a one to many relationship with both sides having a compulsory participation. Using the 'Estab' field in the 'PreferenceRequest' table the 'ToAttend' relationship can be made to the 'School' table. Not every school has to be part of a request, indeed no pupils may request to attend any given school therefore the 'ToAttend' relationship has an optional participation on the 'School' side. The 'IsGranted' relationship allows a join to the 'PreferenceAllocation' table although not every request is granted, as pupils may make up to five ranked requests but they can only be allocated to one school, making the 'PreferenceRequest' side of the 'IsGranted' relationship optional.

The 'OFSTED' table relates on a one to one basis with the 'School' table using the 'Estab' fields and holds information on the school's last inspection grade. Not every school has a current inspection grade as inspections are only carried out every three years. Some new schools have not yet been inspected, making the participa-

tion in the 'Assesses' relationship optional on the 'School' side. The 'Employs' relationship joins the 'School' table with the 'Teacher' entity type tables using the 'Estab' fields. Optional participation on the 'School' side of the 'Employs' relationship is due to the closing of schools. Thus they do not employ any teachers in subsequent years.

Having esablished the database design and structure for the research, the following section will deal with data content. It will begin by examining some general summary statistics for the data before presenting an analysis of errors and omissions and finally considering the temporal aspect of the data.

4 Content of the PLASC Data

4.1 PLASC school and teacher data

Three types of data are recorded in the PLASC returns: pupil data, school data and teacher data. The school data includes the unique establishment ID, the phase of schooling that is catered for by the institution, intake type, governance, lowest and highest National Curriculum years taught and the number of computers in the school. The school tables contained in the study area database created for this project are supplemented with data on each school's net capacity and intake year capacity. The phase of schooling provided by an institution falls into one of three categories: primary school (PS), secondary school (SS) and special school normally referred to as Special Inclusive Learning Centre (SILC). The intake type of schools is more complex and a brief overview of the development of the education system is required to understand the diversity contained in this field.

The development of the education system from 1900 through to the introduction of the 1988 ERA had been fraught with a lack of specific guidance at a national level. The Hadow report in 1926 called for compulsory schooling up to the age of 15 for children and suggested a comprehensive school system (Armytage, 1970). The Spen report in 1938 also called for compusory schooling for children but suggested a school leaving age of 16 and favoured a selective tripartite school system containing grammar, secondary and technical schools (Lawrence, 1992). The onset of the Second World War delayed education reforms until the historic 1944 Education Act that is widely accepted as the Act that "laid the foundation for the modern education system" (Statham et al., 1991, p.42). Although this may be the case,

a lack of cohesive legislative guidance at a national level lead to many different education systems being implemented in parallel at a local level for many years, the remnants of which are reflected in the diversity of school types still available in England and Wales.

Table 4 shows the valid intake type codes for schools and a brief description of each code. The code 'COMP' refers to 'comprehensive' schools, the favoured secondary school system suggested in the Hadow report of 1926. Comprehensives do not select their pupil intake on academic aptitude (Armytage, 1970). Up to 1976, the comprehensive school system, the tripartite school system and several systems that fell somewhere between the two all existed in different LAs across England and Wales. The Labour party's 1976 Education Act brought in legislation to officially end both the tripartite school system and selection by ability in state schools (Gordon et al., 1991). However, resistance to the abolition of the tripartite system of schooling in some LAs and changing political views has lead to state run 'grammar', 'technical' and 'secondary modern' schools remaining. These are reflected by the codes 'SEL1', 'SEL2' and 'SEL3', although none of these school types exist in the Leeds study area. The code 'SEL4' refers to schools that are selective on religious grounds. The code 'SPEC' refers to SILCs, where the school's intake consists of children with SEN.

Code	Decription
COMP	Comprehensive
SEL1	Selective (Grammar)
SEL2	Selective (Secondary Modern)
SEL3	Selective (Technical)
SEL4	Selective (Religion)
SPEC	Special

Table 4: School intake types

Definitions of the governance status of the different types of schools found in England are supplied by Teachernet (2007) and are summarized in Table 5. Schools with a 'community' governance type are controlled by the LA. The buildings and land are LA owned, the staff are employed by the LA and the LA decides and implements the school admissions policy. 'Voluntary aided' schools are normally church schools, with the land and buildings being owned by a charitable foundation. The governing body contributes to the costs of running the school, employs

the staff and has responsibility for the schools admissions policy.

'Voluntary controlled' schools are similar to voluntary aided schools although the LA employs the staff and has responsibility for the admissions policy of the school. 'Foundation' schools were formerly known as grant maintained schools. School land and buildings are owned by the governing body or by a charitable foundation in this case. The governing body employs the school staff and has responsibility for the school's admissions policy. 'City academies' and 'city technology colleges' are funded through schemes directly from the Government with academies being partially funded through private investment. 'Non-maintained' schools are not-for-profit charitable schools and 'independent' schools are supported by fee-paying students and are not maintained by state funding.

Code	Decription
CO	Community
VA	Voluntary Aided
VC	Voluntary Controlled
FO	Foundation
IN	Independent
NM	Non-Maintained
CT	City Technology College
CA	Academies

Table 5: School governance codes

The lowest and highest National Curriculum years are the lowest and highest year groups that the establishment provides education for. The computers field is simply a count of the computers available for use at the school. Table 6 summarizes the number of schools within the study area over the six years of PLASC data. The number of schools in the Leeds study area show a consistent decrease in numbers, especially in the primary phase (PS) where 19 schools have been closed. Secondary schools (SS) and SILC (SP) both have the lowest number of school closures with three, however, when put in context of the total number of schools in each phase, the SILCs have declined in number by a third.

The teacher data counts the number of staff at a school by qualification category. The first field contains the unique establishment identifier of the school. The second field contains the category of teacher that is being counted. Table 7 shows the valid teacher category codes with a brief description of each category. The code

Phase	2001	2002	2003	2004	2005	2006
PS	244	244	240	241	230	225
SS	43	43	43	43	41	40
SP	9	10	10	10	6	6

Table 6: Number of schools by phase and year

'QT' indicates a qualified teacher and should include the head or acting head. The code 'NQ' relates to staff employed at the school that are not qualified teachers and are not on a course to become a qualified teacher. The code 'LQ' refers to staff that are not yet qualified teachers but are on a training course leading to qualified teacher status, although, students at the school on teaching practice should not be counted. The code 'ET' is for any teacher employed to meet the needs of ethnic minority pupils and 'LT' is for teachers that teach English as an additional language. The remaining four fields in this dataset contain the counts of the full-time female staff, full-time male staff, part-time female staff and part-time male staff in the relevant staff qualification category.

Code	Description
QT	Qualified teacher
NQ	Not qualified teacher
LQ	Teacher in training
ET	Teacher of ethnic minorities
LT	Teacher of English as an additional language

Table 7: Teacher category codes

4.2 PLASC variables

The PLASC data for individual pupils are contained in six tables within the database, each one representing an annual data collection as discussed in Section 3.3 and illustrated in Table 1. The 'UPN' field is a unique numeric identifier that is retained by the pupil throughout his/her school career and enables temporal analysis of the data in the pupil tables. The 'Estab' field contains a numeric identifier which is the unique identification number associated with the institution the child is currently attending.

Data on the year group the pupil are currently attending is held in the National

Curriculum year group field, 'NC_Year'. The valid entries for this field are shown in Table 8, however these year groups include additional school years to those defined as compulsory schooling. Under Section 8 of the Education Act 1996, compulsory school age is defined 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 year next following,
>
> (b) if he attains that age on that day, or
>
> (c) (unless paragraph (a) applies) if that day is the school leaving date next following his attaining that age." (HMSO, 1996).

Therefore the National Curriculum year groups for children of compulsory school age are reception and years 1 to 11.

School Type	Key Stage	Year Group	Pupil Age on 31 August
Nursery		N	3
Primary (Infants)	Foundation	R	4
		1	5
	1	2	6
Primary (Juniors)		3	7
		4	8
		5	9
	2	6	10
Secondary		7	11
		8	12
	3	9	13
		10	14
	4	11	15
		12	16
	5	13	17

Table 8: National Curriculum year groups and corresponding pupil ages

The field 'DOB' holds the date of birth for each pupil and the 'Gender' field holds a single character either an 'M' for male or 'F' for female pupils. The 'Ethnicity' field contains a coded representation of the ethnic origin of each pupil, however the categories changed both in number and structure in 2003. The field's content, category and the associated problems caused by the change are considered

in Section 4.3. 'First_Language' contains the coded representation of the pupil's primary spoken language as shown in Table 9. There are six codes: two relating to whether English is known or believed to be the pupil's first language, two relating to the pupil's first language being one other than or believed to be other than English. The final two codes indicate that the information on first language has been refused by either the pupil or their parents or that the information has not yet been obtained.

Code	Description
ENG	English
ENB	Believed to be English
OTH	Other
OTB	Believed to be other
REF	Refused
NOT	Information not obtained

Table 9: First language codes

FSM eligibility is stored in the 'FSM' field and is recorded as either true when the pupil is eligible to receive FSMs or false when the child is not eligible to receive FSMs. FSM entitlement in the Leeds study area is subject to the following conditions "Children whose parents receive Income Support, Income-based Job Seekers Allowance, or support under Part VI of the Immigration and Asylum Act 1999 are entitled to free school meals. Children who receive these benefits in their own right are also entitled to free school meals"(Leeds, 2006).

The 'Entry_Date' field contains the date that the pupil began attending the institution indicated in the 'Estab' field. The home postcode of the pupil is stored in the 'Postcode' field and must be a valid alpha-numeric entry. Two fields, 'X' and 'Y', contain the geocoded coordinates of the pupil's home address. The geocoding process has been completed by *Education Leeds* using the Ordnance Survey dataset Address-Point. However, *Education Leeds* data management staff confirmed that not all pupil addresses could be matched using Address-Point and those that could not be matched were geocoded using the centroid of the pupil's home postcode.

4.3 PLASC variable changes

Each of the PLASC pupil tables has the same structure although not all the data attributes were collected in all years. In 2003, the ethnic origin codes entered into the 'Ethnicity' field were changed to make them compatible with the codes used in the 2001 Census (Godfrey, 2004). Table 10 shows the 16 ethnic group categories used in the Census of Population 2001 and the 14 original broad ethnic codes used in PLASC. The 20 new broad ethnic codes used in PLASC since 2003 can also be seen in Table 10. *Education Leeds* use extended codes to further distinguish between the Pakistani category, these consist of:

- APKN - Pakistani
- AMPK - Mirpuri Pakistani
- AKPA - Kashmiri Pakistani
- AOPK - Other Pakistani

Godfrey (2004) has examined the implications of the change in ethnic codes using the national data from PLASC 2002 and 2003. He shows how some pupils changed between broad ethnic classifications between these years. One explanation is that these errors may be due to large amounts of mis-entered data in the 2002 collection due to the numeric coding that was employed in that year. However, Godfrey dismisses this explanation as unlikely for explaining all the category changes which involved approximately one record in every 50 for some categories. A second explanation is that the selection of ethnic category is seen as a political statement and that the choice may have been made based on a person's perceived nationality rather than their ethnic background. However, Godfrey (2004) suggests that a third explanation may be that the introduction of the new categories clarified the data requirements leading to more accurate classification by the respondents. Whatever the reason for the discrepancies, it is clear that the change in ethnic categories leads to some disruption in the data.

In 2002, variables were added to the data collection for Special Education Need (SEN) status, the 'SEN' field and for enrolment status, the 'Enrolment_Status' field. The SEN status variable consists of four single character codes (Table 11). 'N' indicates that the child has no SEN. 'A' indicates that a child has a SEN and the school should provide additional support (individual support sessions or small

Census 2001	PLASC 01-02		PLASC 03-06	
Group	Code	Description	Code	Description
Bangladeshi	50	Bangladeshi	ABAN	Bangladeshi
Indian	30	Indian	AIND	Indian
Any Other Asian Background	90	Other	AOTH	Any Other Asian Background
Pakistani	40	Pakistani	APKN	Pakistani
African	21	Black-African	BAFR	Black-African
Caribbean	20	Black-Caribbean	BCRB	Black-Caribbean
Any Other Black Background	22	Black Other	BOTH	Any Other Black Background
Chinese	60	Chinese	CHNE	Chinese
Any Other Mixed Background	90	Other	MOTH	Any Other Mixed Background
Mixed: White and Asian	12	White Other	MWAS	White and Asian
Mixed: White and Black African	12	White Other	MWBA	White and Black African
Mixed: White and Black Caribbean	12	White Other	MWBC	White and Black Caribbean
	99	Data not sought	NOBT	Information Not Yet Obtained
Any Other Background	90	Other	OOTH	Any Other Ethnic Group
	98	Preferred not to say	REFU	Refused
British	10	White UK	WBRI	White-British
Irish	11	White European	WIRI	White-Irish
Any Other White background	11	White European	WIRT	Traveller of Irish Heritage
Any other White Background	12	White Other	WOTH	Any Other White Background
Any Other White Background	12	White Other	WROM	Gypsy/Roma

Table 10: Ethnicity codes used in Population Census 2001 and PLASC 2001/02 and 2003/06

group sessions) to the child from within the school's main budget. 'P' requires that the child be given support from an appropriate external service as well as from within the school.

Code	Description
N	None
A	School action
P	School action plus
S	Statemented
Q	School action plus and statutory assessment*

*Category only used in 2003 and 2004 data collections.

Table 11: SEN status codes

The school, external support provider and parents are all included in the strategy creation and implementation for the child's continued support. 'S' means that a child has been assessed and found to need additional support that cannot be reasonably met by a school's normal resources. Detailed definitions for the stages of SEN status can be found at http://eduaction.com/. In 2003 and 2004, an additional code of 'Q' was used relating to pupils that require aid from an external service and are also undergoing a statutory assessment. From 2005 onwards, this code was not used and the pupils were reallocated into either category 'P' or 'S' as required. In 2002, the categories for data collection were numeric, ranging from 0 to 5 with 0 representing a child with no SEN and 5 representing a child with a statement of SEN. Data collection in 2002 returned mostly the numeric codes in 2002 but also some of the new character codes to be adopted as standard from 2003. The inconsistencies in the collection in 2002 and the category change that occured in 2003, coupled with the exclusion of the 'Q' category after 2004 demands that a standardisation method be adopted to allow comparison of SEN data between years. Table 12 shows the reclassification codes to be used and how they translate to the codes used for PLASC data collection in each year. The reclassification adopted here is based on the reclassification strategy used by DfES statisticians in 2004 (DfES, 2004).

The enrolment status variable introduced in 2002 contained four codes as shown in Table 13. An enrolment status of 'C' indicates that the pupil is registered for normal attendance at the indicated school in the 'Estab' field. A status of 'G' indicates that the pupil is not registered at the school but is attending some classes. The PLASC guidance notes (published each year on the Teachernet website at http://www.teachernet.gov.uk/) state that pupils that are not registered at a school should not be recorded in the PLASC return (Teachernet, 2006). However, soft-

Generic SEN Category	Included SEN Categories
No special provisions - code N	2002 numeric code - 0 and missing 2003 onwards and 2002 non-numeric - N and missing
SEN without statement - code A	2002 numeric codes - 1, 2, 3 and 4 2003 onwards and 2002 non-numeric - A, P and Q
SEN with statement - code S	2002 numeric code - 5 2003 onwards and 2002 non-numeric - S

Table 12: Generic SEN status codes to be used for yearly comparison

ware limitations may make it impossible to exclude a pupil who is attending a school as a guest pupil. Therefore the 'G' entry in the enrolment status field allows these 'guest' pupils to be excluded from any subsequent analysis of the PLASC data. Enrolment status 'M' indicates that the pupil is registered at the school in the 'Estab' field and takes the majority of his/her classes at that school but is also registered and takes some classes at another institution. The enrolment status 'S' indicates that a pupil attends the institution in the 'Estab' field for some classes but is registered at another institution for the majority of his/her classes.

Code	Description
C	Current (single registration at this school)
G	Guest (pupil not registered at this school but attending some lessons or sessions)
M	Current Main (dual registration)
S	Current Subsidiary (dual registration)

Table 13: Enrolment status codes

In 2003, a true/false field was added to the PLASC to indicate whether a child was in care on the date of the data collection, the 'In_Care' field. Two further fields were added to the PLASC in 2004 to augment the SEN status field. These fields are called 'Primary_SEN' and 'Secondary_SEN' and record the primary and secondary SEN type that must be selected from the coded list shown in Table 14. A SEN type should only be entered in the 'Primary_SEN' or 'Secondary_SEN' fields if a 'SEN' status of either 'schools action plus', (code 'P'), or 'statemented', (code 'S'), has been recorded. For pupils with no 'SEN', (code N), or a 'SEN' type of 'school

action', (code 'A'), the 'Primary_SEN' and 'Secondary_SEN' fields should be left blank. Pupils with a statement of SEN, indicated by a SEN status of 'S', a SEN type of 'OTH' should not be recorded (Teachernet, 2004).

Code	Description
SPLD	Specific Learning Difficulty
MLD	Moderate Learning Difficulty
SLD	Severe Learning Difficulty
PMLD	Profound and Multiple Learning Difficulty
BESD	Behaviour, Emotional and Social Difficulty
SLCN	Speech, language and Communication Needs
ASD	Autistic Spectrum Disorder
VI	Visual Impairment
HI	Hearing Impairment
MSI	Multi-Sensory Impairment
PD	Physical Disability
OTH	Other

Table 14: SEN type codes

The SEN types displayed in Table 14 can be grouped into four broad categories as portrayed by Teachernet (2004). The first category is 'Cognition and Learning Needs' and includes the SEN types 'SPLD', 'MLD', 'SLD' and 'PMLD'. 'Behaviour, Emotional and Social Development Needs' is the second broad category and only includes the SEN type of 'BESD'. The third category is 'Communication and Interaction Needs' and includes 'SLCN' and 'ASD' SEN types. The fourth broad category is 'Sensory and/or Physical Needs' and includes the SEN types 'VI', 'HI', 'MSI' and 'PD'. Each of these broad categories and the individual SEN types have guidelines for if and how a pupil should be recorded in the PLASC data collection as having a SEN.

4.4 PLASC errors and omissions: identification and rectification

When attempting to apply the primary key status to the UPN field in the PLASC pupil tables, two problems were discovered. The first problem was that 4,012 pupil records out of the total 116,506 in the 'PLASC_2001' table had no UPN assigned to them. Although 163 schools had pupils attending them with a missing UPN, 2,638 of the pupils with missing UPNs attended one of only 15 schools, of which

five had 100% of their pupils with missing UPNs. This large amount of missing data, especially in a crucial field such as 'UPN', casts doubt as to whether the data for this year is usable for temporal analysis.

The second problem was that duplicate values existed in the UPN field in the 'PLASC_2006' table. A further examination of the 'PLASC_2006' table revealed that 15 pupils were registered at two different schools. The 'Enrolment_Status' field showed that the 15 pupils with duplicate registrations had a registration code 'C' with one school indicating that they were registered as a current single registration at that school and additionally had a registration status of 'G', a guest registration at a second school. The pupils having duplicate registrations all had SEN and were shown as attending a SILC as the main education establishment and a mainstream school as a guest. Recording pupils that attend an institution as a guest is not recomended in the PLASC collection guidance notes (Teachernet, 2006). Software limitations make it impossible to exclude such pupils from the data collection but guest registered pupils should be excluded from any data analysis. Therefore the duplicate entries showing guest registration have been removed from the database.

Table 15 shows the total record counts for each of the PLASC pupil tables. It shows that the number of pupils attending schools in the Leeds study area is steadily declining year on year from 116,506 in 2001 to 110,332 in 2006, a reduction of 6,174 pupils or 5.3%; an average decline of approximately 1% per year. However, the raw PLASC tables include children attending pre-school classes and those studying for further education qualifications, in addition to those pupils of compulsory school age with which this study is concerned. Thus, the field 'Compulsory School Age' in Table 15 shows the counts of records of pupils of compulsory school age, the National Curriculum years of reception and year groups 1 through to 11. The count of compusory school age pupils also reflects a steadily decreasing number from 102,629 in 2001 decreasing to 97,802 in 2006, a fall of 4,827 pupils or 4.7%. A reduction in the number of pupils of this magnitude is a cause for concern when it is evident that the largest number of pupils on role at any one school in the 2006 PLASC data is 1,964. This study is only concerned with state educated pupils of compulsory school age as defined by the Education Act 1996 (HMSO, 1996) and therefore any pupils not attending the National Curriculum reception year or year groups 1 through to 11 will be discounted from the

database.

Table	Total Record Count	Compulsory School Age
PLASC_2001	116,506	102,629
PLASC_2002	115,134	102,262
PLASC_2003	114,090	101,336
PLASC_2004	112,690	100,269
PLASC_2005	111,518	99,014
PLASC_2006	110,332	97,802

Table 15: Record counts for the PLASC pupil tables

4.4.1 Data omissions

The 'Primary_SEN' and 'Secondary_SEN' fields in the PLASC pupil tables are left null for pupils if they do not have a SEN. For all other fields, data should be recorded and any field with a missing entry can be regarded as an omission error in the data. Table 16 shows the omission errors in the PLASC pupil tables after the exclusion of pupils not of compulsory school age and indicates the worst year for data quality is 2001. The data in 2001 has 2,484 UPN omissions, reduced from 4,012 by the exclusion of pupils not of compulsory school age, 1,336 ethnicity omissions, 2,888 first language omissions and 13,294 FSM omissions. This is the only year of data that suffers with omission errors in any of these fields. Although the 2001 data has many problems in other fields, it has the second least number of geographical coordinate omissions. Coordinate omissions are highest for year 2004 at 12,205 and lowest in 2002 at 330. Although coordinate omissions are substantial in 2003, 2006 and, to a lesser extent, in 2005, postcode omissions are relatively low across most years except 2001 which has 549. 2006 is the best year with zero postcode omission errors.

With the longitudinal data that are available here, it is possible to use interpolation routines to minimise the number of omissions. Some pupil characteristics represented in the PLASC pupil tables are subject to change over time. Burgess et al. (2006) indicate that FSM, SEN status, location and school attended are all time variant characteristics whereas gender, ethnicity, within-year-age, first language and date of birth should all remain constant. However, it is the location attributes that are most critical when considering questions relating to the jour-

Field	2001	2002	2003	2004	2005	2006
UPN	2,484	0	0	0	0	0
Estab	0	0	0	0	0	0
NC_Year	0	0	0	0	0	0
DOB	0	0	0	0	0	0
Gender	0	0	0	0	0	0
Ethnicity	1,336	0	0	0	0	0
First_Language	2,888	0	0	0	0	0
Entry_Date	0	0	0	0	0	0
Postcode	549	98	89	22	9	0
X	781	330	6,974	12,205	1,192	6,100
Y	781	330	6,974	12,205	1,192	6,100
FSM	13,294	0	0	0	0	0
Enrolment_Status	*	0	0	0	0	0
SEN	*	0	0	0	0	0
In_Care	*	*	0	0	0	0

*Indicates data not collected for this attribute in this year

Table 16: Ommisions from the PLASC pupil tables for pupils of compulsory school age

ney to learn and the daily commuter behaviour of school age children. It is these attributes that are the most compromised by omission errors.

Table 16 indicates that the postcode data are subject to substantially less omissions than the coordinate data. It is sensible to assume that if a pupil's postcode data remains constant across several years then it is most likely that the child has not moved home. It is then reasonable to interpolate any missing coordinate data in one year with data from another year so long as the postcode data remains constant. Furthermore, if the postcode is missing for a pupil record in one year but remains constant in the years directly before and after the year in which the data are missing, it is reasonable to assume that the child has not moved home. This enables the data for the missing postcode to be interpolated from the postcode data contained in either the year directly before or after the instance of missing data.

Although interpolation techniques can be used on the location data because the two assumptions discussed above are reasonable, interpolation is not suitable for use on the FSM data in 2001. As stated by Burgess et al. (2006), FSM eligibility is changeable based on the circumstances of the child or of the child's parents or legal guardian. In this case, it is not reasonable to assume that if a child's FSM

eligibility is not known for a year, it is the same as either the previous or following years' status. Circumstances for that child may have changed, altering the status of the child's FSM eligibility. Thus, little can be done to improve on the missing data in the 'FSM' field for 2001.

The problem relating to the missing UPN data in the 'PLASC_2001' table may be improved through the use of interpolation given two assumptions. The first assumption is that if a pupil record in the 2002 data does not appear in the 2001 data and the 2002 pupil record shares a common date of birth with a record in the 2001 data with a missing UPN, then these records represent the same pupil. The second assumption is that if more than one pupil record appears in the 2002 data that does not appear in the 2001 data but they all share a common date of birth but only one of the 2002 records shares a common postcode with the 2001 record with a missing UPN, then these records represent the same pupil. These assumptions are reasonable as a pupil's date of birth should remain constant and location attributes are unlikely to change. It is possible that one pupil may exit the study area after the 2001 data collection with the same date of birth as a student who enters the study area before the 2002 data collection and in this case an incorrect assignment would occur. It is also possible, but even more unlikely, that a pupil may move out of a postcode and another pupil may move into that postcode with the same date of birth, again provoking an incorrect assignment. However, using the code for the establishment attended and the National Curriculum year attribute in the data across both years, it is possible to limit the chances of incorrect assignment.

Table 17 shows the omission errors for the location attributes and the UPN of the PLASC pupil tables after the data interpolation routines have been completed and checked. It also shows the original omission error counts and the difference between the two (raw omission counts minus corrected omission counts). The interpolation of the UPNs for the 2001 table has reduced the amount of missing data substantially from 2,484 down to 559 although the result is not perfect. A total of 135 of the pupil records with missing UPN numbers in 2001 attended Farnley Park High School, 60 atteded Manston St James Church of England School, 51 attended Summerfield Primary School, 44 attended Allerton Grange High School, 40 attended Austhorpe Primary School and 30 attended Bruntcliffe High School. The remaining 199 pupil records with missing UPNs were distributed in much smaller numbers among a further 62 schools. The high number of missing UPNs

at secondary schools is to be expected due to pupils in year 11 at the end of 2001 leaving school and thus not being recorded in the PLASC returns for subsequent years, leaving no chance to interpolate these data. Pupils at school leaving age at the end of 2001 account for over a quarter of the remaining missing UPNs. There is no way to bring full resolution to this problem as the data collection in this year was optional and is not included in the NPD. However, the reduction in missing UPNs for 2001 will allow for that data to be used in more extensive analysis later in this study. The collection in the academic year 2000/2001 was a preparatory exercise conducted by *Education Leeds* to test the data collection system for the statutory collection the following year.

Interpolation of the location attributes has been more successful in some cases than others. The postcode data showed very little improvement for 2005 with the number of omissions falling by just two from nine to seven. Other years faired better with the greatest improvement being in 2001, where 434 out of 549 were identified. The coordinate data showed more improvement with the missing values in 2004 decreasing from 12,205 omissions to just 136. These improvements in data content are very important with the location attributes of the pupil records being critical for the accuracy of the study of commuting behaviour.

Field	2001	2002	2003	2004	2005	2006
UPN raw data	2,484	0	0	0	0	0
UPN after interpolation	559	0	0	0	0	0
Difference	1,925	0	0	0	0	0
Postcode raw data	549	98	89	22	9	0
Postcode after interpolation	115	78	66	15	7	0
Difference	434	20	23	7	2	0
X raw data	781	330	6,974	12,205	1,192	6,100
X after interpolation	481	190	202	136	173	191
Difference	300	140	6,772	12,069	1,019	5,909
Y raw data	781	330	6,974	12,205	1,192	6,100
Y after interpolation	481	190	202	136	173	191
Difference	300	140	6,772	12,069	1,019	5,909

Table 17: UPN and location attribute omissions from the PLASC pupil tables before and after data interpolation

Although the ethnicity field is one that is expected to remain constant over

time as the ethnic origin of an individual cannot change, it is not suitable to apply the cross-year interpolation methodology to this field. The results of the analysis by Godfrey (2004) on the effects of the changing ethnicity codes in 2003 have been discussed in the Section 4.3. Godfrey explains that respondents gave different ethnic origin data in the PLASC in collection year 2002 to the collection year 2003 when the ethnic code changes were implemented. Although Godfrey offers several plausible reasons for the changes in response, there is no way of being certain which year of data is the more accurate. Using interpolation routines, data from either side of the code change implimentation would produce different results and therefore the validity of such interpolation is questionable.

Despite interpolation not being suitable in this situation, a lookup table between the ethnic categories exists (Supplied by *Education Leeds* Data Management Department) which allows a conversion of the pre-PLASC 2003 ethnicity responses into the corresponding classification in the new ethnic code system. A conversion of the codes will enable some cross-year analysis to determine the extent of disruption brought to the study area data by the changes in ethnic codes. This subsection has examined missing data in the PLASC pupil tables and ways to interpolate missing attributes through the use of the datasets' temporal dimension. Even though the application of interpolation techniques produces a more complete dataset, errors other than omission errors may still persist. Isolating alternative sources of error is the focus of the next subsection.

4.4.2 Data errors

To analyse for possible errors within the PLASC pupil tables, three general approaches can be used. The first approach is to apply general 'reality' checks to the data looking for mis-entered attributes that fall outside of an expected range. A list of reality checks for this dataset is as follows.

1. Calculate the age of the pupils in each year of the PLASC data using the 'DOB' field and examine the maximum and minimum ages of the children to see if they fall within the expected 1 to 16 range.

2. Ensure the contents of the fields 'Ethnicity', 'First_Language', 'Gender', 'SEN', 'Primary_SEN', 'Secondary_SEN' and 'Enrolment_Status' only contain the correct category data items for each respective year.

3. Ensure that the 'Estab' field only contains valid identification numbers for schools in the study area.

4. Check each pupil is shown to be attending an 'Estab' that provide education for the phase of schooling appropriate to the 'NC_Year' of that pupil.

5. Ensure that each pupil record 'Postcode' field contains a valid UK postcode, using the appropriate All Fields Postcode Directory (AFPD). The AFPD is available to download from UKBorders and updated versions are released quaterly with version specifications and user guidlines. The name and structure of the postcode directory changed in 2006, current releases are referred to as the National Statistics Postcode Directory.

Table 18 shows the number of pupils that are not of compulsory school age when using the 'DOB' field to calculate pupil age on the date of the PLASC collection for that year. Examining these records further showed that in the year 2003, 2004 and 2005, all pupils fell within the age range of either 4 to 17, with very few pupils being aged 4 or 17. In 2006, two pupils had reached the age of 18 but the lowest age remained 4. Further still, in the years between 2003 and 2006, all the pupil records with either low or high ages attended school in either the first or the last National Curriculum year group.

	2001	2002	2003	2004	2005	2006
Pupils	100	158	22	20	88	27
Invalid Postcodes	866	285	240	155	203	274

Table 18: Pupils not of compulsory school age and with invalid postcodes

These anomalies in pupil age could be due to either children being advanced for their age or needing extra time to prepare for their final key stage 4 examinations and so being held back for further schooling. In 2002, two pupils of ages 2 and 3 were recorded in the data. They were shown to attend a SILC in 'NC_Year' 1 and had a SEN status of 'S', indicating that they had a statement of special educational need. It is most likely that the 'NC_Year' of these pupils had been recorded incorrectly and so they have been excluded from this year of data. All other pupils were in the age range of 4 to 17. In 2001, the data contained 12 records for pupils under the age of 4. These records could not be tracked into subsequent years either because of a missing UPN or because the allocated UPN did not appear in

subsequent years' data, thus these records have been excluded from further analysis. Nine records in 2001 had an age over the normal expected age for compulsory schooling. One of these records showed an age of 41 although this proved to be a typographical error and could be corrected using data from the 2002 PLASC. The remaining eight records had 'NC_Year' entries of 2. These are believed to be mistyped 'NC_Year' entries that are meant to represent pupils in 'NC_Year' 12, which would correspond with the 'DOB' field. These records have been excluded from the database.

Carrying out checks of the content of the fields 'Ethnicity', 'First_Language', 'Gender', 'SEN', 'Primary_SEN', 'Secondary_SEN' and 'Enrolment_Status' in each dataset revealed that only the expected category values were found in each of the fields. In addition, all years of data were checked to ensure no other pupil records existed with an 'Enrolment_Status' of 'G' indicating that the student attended the school as a guest and should be excluded from analysis. The 'Estab' field only contained unique identifiers for schools in the study area LA that were open in the year of the relevant PLASC collection. Comparing the 'NC_Year' field with the phase of school, either infant/primary school, secondary school or special school, indicated that one pupil in year 2002, four in 2004 and two in 2005 were shown to attend schools that did not provide education at the National Curiculum year indicated. All of these records could be attributed to mis-entered values and easily corrected.

Using the appropriate year AFPD, the postcode entries of each pupil were checked to ensure that they represented a valid 'UK' postcode in all years. The results can be seen in Table 18 and show that 2001 is the worst year for invalid postcode entries with 866. The years from 2002 through to 2006 range between 155 to 285 invalid postcode entries. On closer examination, many of the pupils with invalid postcodes also had blank location coordinates and so these records have been excluded from the database through all years.

4.4.3 Coordinates checks

The second approach to error analysis is to examine the location attributes of each pupil to ascertain if the geocoding of the pupil addresses has been carried out with suitable accuracy and precision. Geocoding is the process of applying a spatial reference to data that has no spatial reference. For example, spatial coordinates

can be applied to a pupil record if the postcode is known. By using the AFPD, the pupil record can be joined to the postcode information in the AFPD which contains spatial coordinates for the centroids of all UK postcodes. The postcode coordinates can then be used for spatial analysis or mapping of the pupil record. There is a more extensive discussion of the geocoding process in Gatrell (2002).

Having established the validity of the postcodes in the previous series of checks, it is possible to establish if the accuracy and precision of the geocoding of pupil record location coordinates is satisfactory. The commute to school normally involves short distances, it is therefore important that the accuracy and precision of the geocoded pupil coordinates be established to promote confidence in the journey distances that will be the focus of subsequent analysis. In meetings with *Education Leeds* Data Management staff, it was expressed that the geocoding of pupil addresses was undertaken using Ordnance Survey (OS) Address-Point database, theoretically geocoding pupils to within 0.1 metre of their home address (OS, 2007). An example of this can be seen in Figure 4 which shows pupil coordinates from the PLASC 2003 data plotted on a backdrop of OS MasterMap data for part of Leeds. It can clearly be seen that the geocoded pupil points fall inside dwelling structures in the MasterMap data, initially suggesting that the precision of geocoding is high. Although whether the pupil points fall inside the correct dwelling structures cannot be verified.

It is possible to check the level of accuracy and precision in the geocoded data by calculating the distance between the pupil coordinates, geocoded using Address-Point, and the postcode centroid coordinates contained in the AFPD. The distance measurement can be calculated using Pythagoras' Theorem since all the coordinates of the pupil records and the postcode centroids are provided in 1 metre British National Grid format. The distance between the geocoded location of a pupil's dwelling and the known postcode centroid can be examined to see if it exceeds a threshold value. For each pupil record where the distance from the pupil's geocoded coordinates to the postcode centroid of the pupil's postcode exceeds the threshold value, the validity of the geocoding cannot be assured.

The definition of an arbitrary threshold value to indicate whether a pupil record has been geocoded within an acceptable distance of the actual postcode centroid is not a simple matter. A single arbitrary distance would not be suitable in all cirumstances. Whilst postcodes contain a small number of delivery address points, their

Source: Map created with OS MasterMap data

Figure 4: Pupil coordinates geocoded using OS Address-Point

boundaries are not defined explicitly. In urban areas, the delivery address points are more dense, and usually smaller, than in suburban or rural areas, leading to smaller postcode areas in urban centres than in suburban areas and smaller postcode areas in suburban areas than in rural areas. Thus, an arbitrary distance that indicates a pupil record falls reasonably close to a rural postcode centroid would be too large for use with urban postcodes and vice versa. It is therefore necesary to calculate a relative threshold value for evaluating the proximity of pupil record to postcode centroid coordinates.

Figure 5 shows two examples of a method used for calculating the relative threshold verification value, one for a pupil record in an urban area and one in a rural area. The black postcode centroid in each case is that which is associated with the pupil record undergoing verification. The black star represents the location of the coordinates associated with one pupil record. To calculate the relative threshold value, the closest four postcode centroids to the postcode centroid associated with the pupil record are found and the distance to each of these is calculated, shown in the diagram by the four lines labeled A, B, C and D. The average is calculated and shown in the examples in Figure 5 as red circles. If the associated pupil location

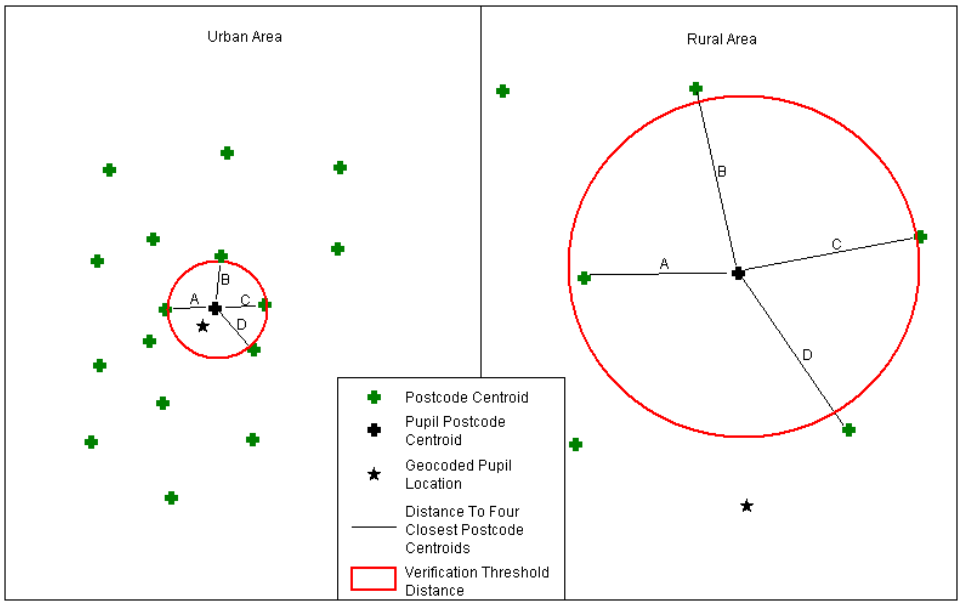
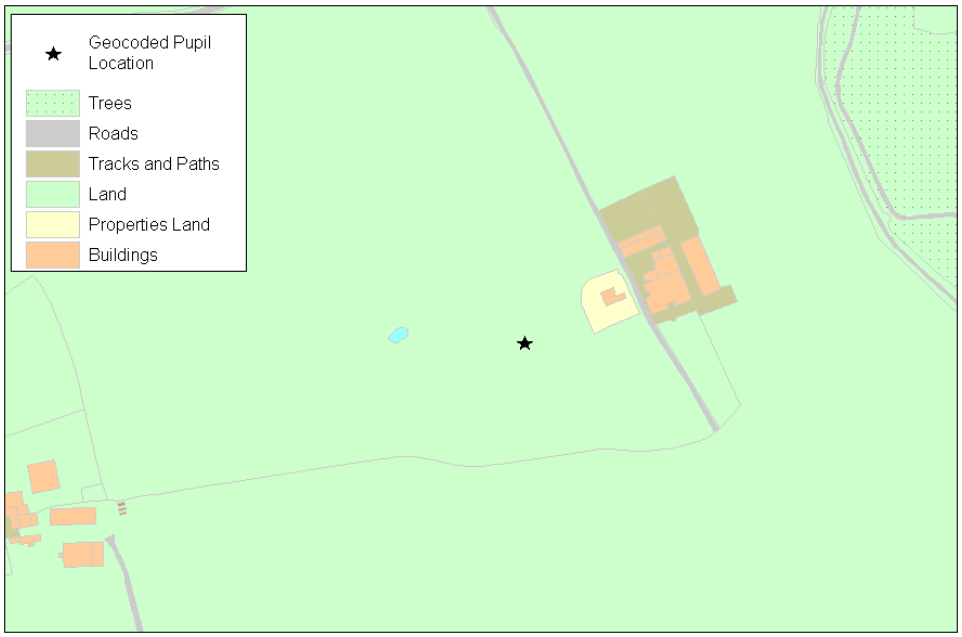


Figure 5: Pupil location verification method

falls within the threshold calculated, inside the red circle, as is the case with the urban example, then the coordinates are within an expected distance and can be considered as having been verified with good geocoding precision and accuracy. However, if the coordinate location of the pupil record is outside the circle, as in the case of the rural example, then a problem with the pupil coordinates may exist. It is important to note the extent that the threshold value changes between the urban and rural examples with respect to the density of postcodes, highlighting the need for a relative threshold value rather than a single arbitrary value to be used in all cases.

Figure 6 shows that pupil coordinates verified by the method exemplified in Figure 5 do not always fall at expected locations when mapped against the OS MasterMap data. A lack of accuracy and precision in geocoding databases is the central focus of the work by Ward et al. (2005). From their work carried out in Iowa, they conclude that geocoding is less accurate in rural areas than in urban areas and whether the work is carried out by commercial firms or in-house using location databases has little influence. In cases such as the one in part of Leeds illustrated in Figure 6, it is not possible to correct the location of the pupil. Simply placing the pupil in the nearest household is not plausible solution he/she may not reside in that household, or there maybe multiple dwellings in which residence is



Source: Map created with OS MasterMap data

Figure 6: Pupil location geocoded by *Education Leeds* and verified by proximity to postcode

possible and therefore the process of error correction could introduce more errors. For this reason, pupil records that are within a reasonable distance of their postcode centroid but do not fall inside a dwelling will not be corrected.

Although a decrease in geocoding accuracy for urban areas is present in the PLASC data supplied by *Education Leeds*, another issue that can be corrected to some degree is also present. The top part of Table 19 shows the total number of valid pupil records and the number of pupil records that achieved the status of verified location when checked against the relative threshold value for the pupil postcode. On initial inspection, the years 2003 and 2004 seem to have a high number of verified pupil coordinates with respect to the total number of valid pupil records. However, on further examination 88,762 pupil records in 2003 and 82,309 in 2004 share the same coordinate locations as the postcode centroid for the pupil records, 88.5% and 82.9% respectively. Clearly this indicates that the level of geocoding precision across all years, and particularly in 2003 and 2004, is not as high as would be expected from using the OS Address-Point database.

The short nature of each pupil's commute to school requires that the coordinate

locations used for both pupil home and school locations should be as accurate as possible across all years of available data. Therefore, using the longitudinal perspective of the data and applying the assumption that if a pupil's postcode does not change between years they have not moved home, it is possible to interpolate the coordinate locations of pupils in years with less precise geocoded data with data from years with more precise geocoding. The results of the interpolation process can be seen in the bottom part of Table 19. A clear improvement in both the verified pupil coordinates and the number of pupils geocoded with the postcode centroid is achieved. The number of pupils geocoded with the postcode centroid falls from 88,762 to 12,070 in 2003 and from 82,309 to 11,808 in 2004 whilst, at the same time, the number of verified coordinate locations increases.

Year	2001	2002	2003	2004	2005	2006
Total valid records	101,320	101,140	100,283	99,325	98,114	96,984
Verified coordinates	99,562	100,025	100,011	98,882	97,585	96,158
Pupil = postcode	5,725	9,608	88,762	82,309	13,036	8,236
After coordinate interpolation						
Verified coordinates	100,431	100,348	100,122	99,175	97,917	96,399
Pupil = postcode	4,842	6,789	12,070	11,808	7,183	6,960
Locations not verified	889	792	161	150	197	585

Table 19: Pupil coordinate precision

Although the coordinate interpolation process increases both the geocoding precision and the number of pupil records with verified pupil locations, there are still a small number of pupil records with valid postcodes but without verified location coordinates. These records could be a source of significant error when calculating pupil commuting distances. Indeed, some of the pupil records with unverified coordinates show a distance to their associated postcode centroid of over 40 kilometres. The confidence in the geocoding of these pupil records is low and therefore they will be geocoded to their appropriate postcode centroid to minimize the possibility of location errors.

4.4.4 Summary of data cleaning

Table 20 shows a summary of the total number of pupil records in the PLASC pupil data tables at different stages of the data cleaning process. The first row shows the

total number of pupil records contained in the raw data supplied by *Education Leeds* in each of the six years. The second row shows the number of pupil records in each year representing pupils of compulsory school age. The third row shows the total numbers of pupils in each year after errors and omissions in the data have been removed. The fourth and fifth rows show the total number of exclusions due to errors or omissions and the percentage of exclusions expressed as a percentage of compulsory school age children, respectively. The percentage of exclusions due to errors or omissions in the data shows a consistent drop over the six years of data indicating that the collection and data validation procedures used in the study area are consistently improving over time.

	2001	2002	2003	2004	2005	2006
Raw data	116,506	115,134	114,090	112,690	111,518	110,332
School age	102,629	102,262	101,336	100,269	99,014	97,802
Cleaned	101,320	101,140	100,283	99,325	98,114	96,984
Excluded	1,309	1,122	1,053	944	900	818
Excluded %	1.28	1.10	1.04	0.94	0.91	0.84

Table 20: Summary of valid pupil records in the PLASC pupil tables

4.4.5 Checks over time

The third approach to error analysis is to examine each pupil record from a temporal perspective. Using the temporal dimension of the dataset and the ability to track each pupil across years with the UPN, the data can be examined to reveal inconsistencies in fields that should remain constant across years (Burgess et al., 2006). The aim of temporal analysis is to find typographic errors that fall within the expected range for a field or more serious errors involving the UPN. Although the 'Ethnicity' field should remain consistent across all years, the changes in the ethnicity categories between the years 2002 and 2003 (Godfrey, 2004) make this field unsuitable for use in temporal error detection.

Temporal error detection examines each pupil record from year to year to see if the progression of the pupil record through the dataset is as expected. The fields that will be used for this analysis are as follows:

- 'UPN' should remain constant across all years tracking the pupil throughout the school career.

- 'Gender' should remain constant across all years.
- 'DOB'should remain constant across all years.
- 'NC_Year' is expected to increment year on year. Exceptions to this may be encountered if a pupil is held back a year for further tuition. However, changes in the 'NC_Year' field of more than 1 year are unexpected and should be investigated further.

The fields 'Estab' and 'Postcode' can be used to verify whether a pupil record with temporal inconsistencies in the above listed fields represents the same pupil record with typographic errors or if a possible error exists with the UPN.

Table 21 shows the individual temporal inconsistencies in the PLASC pupil tables. There are a significant number of temporal inconsistencies in both the 'Gender' and 'DOB' fields. The 'NC_Year' field displays inconsistencies where pupils either regress backwards or progress more than one year forward. A definite spike in the number of temporal inconsistencies in the 'NC_Year' field occurs between 2002 and 2003 with 7,312 unexpected changes. However, on closer inspection the majority of these records do not progress between 2001 and 2002 and then make a two year jump between 2002 and 2003. For example a pupil in 'NC_Year' 3 in 2001 remains in 'NC_Year' 3 in 2002 but then in 2003 the pupil appears to jump to 'NC_Year' 5 showing unexpected progression. In these cases, it is clearly due either to a typographic error or a record that has not been updated in 2002 and the information can easily be interpolated between years to ensure consistency.

	Count of records		
Time period	DOB	Gender	NC_Year
2001-02	605	307	954
2002-03	430	114	7,312
2003-04	377	124	81
2004-05	384	109	257
2005-06	313	92	209

Table 21: Individual field temporal inconsistencies in the PLASC pupil tables

To distinguish between typographic errors and records with possible UPN problems, the records with temporal inconsistencies in more than one field need to be isolated. Table 22 shows a summary of records with multiple inconsistencies

across years. In the years between 2001/02, 2002/03 and 2003/04, there is one pupil record in each period with inconsistencies in all three of the tested fields. Examining these three records further reveals that in addition to the date of birth, gender and National Curriculum year of the pupils changing, both the home location of the pupils and the school which they attend also change. It is probable that these three records each represent two different pupils and could indicate an error with the UPN, therefore these three records will be excluded from all years data in the database.

	Number of inconsistencies	
Time period	All 3 fields	Any 2 fields
2001-02	1	119
2002-03	1	79
2003-04	1	5
2004-05	0	14
2005-06	0	12

Table 22: Count of pupil records by multiple temporal inconsistencies in the three 'key' fields

The remaining temporal inconsistencies can be processed to find any values that are obvious typographic errors. For example, a pupil record that progresses through the years of PLASC data with a 'Gender' entry of 'M' for all years but one where the fields has an entry of 'F' and both location and institution data remain constant, can be assumed to be a typographic error and the data interpolated from another year. For the 'Gender' and 'DOB' fields, this is a simple matter of a direct comparison and adjustment, so long as enough temporal data are available to make a valid interpolation.

For the 'NC_Year' field, the situation is a little more complicated. The natural progression of a pupil through the year groups is to increment one year group annually. Therefore, when examining records for typographic errors, it is necessary to consider the year group that the pupil should attend in the current year, the year group attended in the previous year and the year group that is attended the following year. For example, a typographic error in the 'NC_Year' field would be indicated by the year group progression 1-2-2-4-5-6. The pupil seems to spend two years in year group 2 and then jump unexpectedly to year group 4. This is

unlikely, a more realistic explanation is that the year group for the pupil in the third entry has been mis-entered as a 2 and should be a 3 which would then make the year group progression a natural 1-2-3-4-5-6. However, if a pattern of 1-2-2-3-4-5 was encountered, this could be brought about by a pupil having to resit a year of schooling and may not be an error in the data. In cases like these, the values cannot be corrected. To enable the reception year group to be included in this analysis the year group category has been changed from 'R' to '0'.

Table 23 shows the total number of temporal inconsistencies remaining in the PLASC pupil tables after data interpolation. The difference column shows how many records have been interpolated for each field in each year. The most significant change is the 'NC_Year' field in the period 2002/03 where 7,206 records have been interpolated reducing the number of inconsistencies to be similar to other periods. The least amount of change has occured in the 'DOB' field. This is due to the 'DOB' entries for pupils changing multiple times across the years 2001 through 2006, making it impossible to evaluate which entries are correct and which are likely to be errors, thus very little correction can be made. However, with the comparison between the age of the pupil and the National Curriculum year attended, carried out earlier in this section, it can be assumed that the variation in the 'DOB' field does not distort the age of the pupils to the extreme. Therefore, the variations can be tolerated although the reliability of the data in this field is questionable and any analysis carried out using this field should be interpreted carefully.

	DOB		Gender		NC_Year	
Time period	Count	Diff	Count	Diff	Count	Diff
2001-02	605	0	58	-249	62	-892
2002-03	429	-1	68	-46	106	-7,206
2003-04	376	-1	66	-58	51	-30
2004-05	384	0	42	-67	47	-210
2005-06	313	0	64	-28	124	-85

Table 23: Temporal inconsistencies in the PLASC pupil tables after temporal interpolation

Unfortunately the remaining inconsistencies in the data cannot be resolved further. However, because the locational accuracy and precision of these records has been successfully validated, and given the importance of retaining as many valid location observations in the database as possible, these inconsistencies can be ac-

cepted. They are relatively small in number and will have an inconsequential effect on analysis results disaggregated by gender or National Curriculum year group. However, they will have a positive effect on the more critical analysis concerning distance calculations imperative to understanding the commuting behaviour of compulsory school age pupils.

4.4.6 Bias detection

The final stage of error detection is to ensure that the exclusion of pupil records has not been biased towards one particular school in any one year. Table 24 shows the final distribution of pupil record exclusions in all years. The 'Excluded records' column shows the final number of excluded records. The 'Highest % rate of school roll' column shows the highest number of excluded pupil records by establishment expressed as a percentage of the total number of compulsory school age pupils on role at each establishment (excluded pupil records and valid pupil records). The 'Highest % rate of all exclusions' shows the highest number of excluded pupil records by establishment expressed as a percentage of the 'Excluded records' column.

Year	Excluded records	Highest % rate of school role	Highest % rate of all exclusions	Number of schools
2001	1,310	10.64	4.20	251
2002	1,124	10.87	4.80	239
2003	1,055	11.10	5.02	225
2004	946	6.11	4.44	212
2005	900	5.79	4.00	199
2006	818	3.17	3.18	182

Table 24: Summary statistics for the excluded pupil records

Although the excluded records are slightly higher in Table 24 than in Table 20 because of the exclusion of records with temporal inconsistencies, it is clear that the number of pupil records excluded in each year falls consistently from 1,310 in 2001 through to 818 in 2006. It is also clear that the percentage calculations also decrease as does the number of schools involved, from 251 in 2001 down to 182 in 2006. The percentage calculations show that the excluded pupil records do not originate at one school. The percentage calculations based on school roll are im-

portant because they display that one single school will not be impacted too greatly by the number of pupil records excluded from the database, although in 2001 and 2002, the same SILC does have 11% of its pupils excluded from the database. However, this establishment was by far the highest in both years. In 2003, 11% of a primary schools roll is excluded. This is closely followed by another primary school with 9.5% of its role excluded. These two schools have by far the highest number of excluded records in 2003. In 2004, 2005 and 2006 there are no schools that show a dramatic exclusion rate.

The percentage calculations based on the total number of excluded records are of interest to identify whether a particularly high number of the total exclusions originate from one school. This is different to those based on school role because if, for example, a school has 1,500 pupils and 150 are excluded from the database, this is only 10% of its role but in 2006 this would represent 18.3% of the total excluded records. Another smaller school with 100 pupils on roll could have 15 pupil records excluded, 15% of its roll, however this would only represent 1.8% of the total number of excluded records in 2006. The percentage ratios are very dependent on the size of school and therefore both must be considered to ensure bias has not been introduced into the data through the exclusion of invalid pupil records. In this case, the percentage of all excluded records does not rise above 5.2% in 2003 with the lowest being 3.2% in 2006. Although the excluded records will effect subsequent analysis of the data, the excluded records are dispersed relatively evenly throughout the database, thus limiting the effects.

4.5 PLASC data summary

Table 25 shows the final count of valid pupil records for each year. In the second section of the table, the pupil record counts are disaggregated by National Curriculum year to show how many pupils attend each year group. The progression of a cohort can be tracked year on year by looking at the subsequent year's data and progressing the year groups up one year. For example, year group R in 2001 has 7,845 pupils, this year group then progresses to year group 1 in 2002 where 8,272 pupils are recorded in the data. As an example of cohort progression, the reception year group 'R' in 2001 is highlighted by the grey cells throughout their school career. Following each year group through the data reveals that four cohorts leave the dataset with more pupils than were recorded when they entered the dataset. The

cohorts with increasing pupil numbers have relatively low increases and all enter and leave the dataset in the primary phase of education, before year 7. All cohorts of pupils that leave the dataset in the secondary phase of education decrease in pupil numbers.

Category	2001	2002	2003	2004	2005	2006
All Valid Pupil Records						
	101,319	101,138	100,281	99,323	98,114	96,984
National Curriculum Year						
R	7,845	7,771	7,874	7,826	7,411	7,496
1	8,220	8,272	8,001	7,879	7,906	7,420
2	8,374	8,320	8,232	7,989	7,874	7,890
3	8,493	8,368	8,158	8,167	7,957	7,842
4	8,929	8,416	8,316	8,118	8,173	7,969
5	9,040	8,944	8,388	8,259	8,084	8,167
6	8,747	8,878	8,872	8,350	8,266	8,103
7	8,589	8,526	8,709	8,631	8,221	8,141
8	8,629	8,583	8,515	8,694	8,610	8,212
9	8,360	8,554	8,551	8,467	8,718	8,599
10	8,317	8,318	8,535	8,502	8,495	8,747
11	7,776	8,188	8,131	8,441	8,399	8,398
Gender						
F	49,638	49,538	48,976	48,589	48,052	47,421
M	51,681	51,600	51,305	50,734	50,062	49,563
First Language						
ENG	90,706	92,869	91,587	90,214	88,518	85,081
ENB	0	17	16	106	34	931
OTB	0	11	73	169	115	1,236
OTH	10,613	8,241	8,604	8,830	9,410	9,437
REF	0	0	0	1	0	9
NOT	0	0	1	3	37	290
Free School Meal						
Eligible	20,577	20,490	19,880	20,204	20,089	19,174
Not Eligible	80,742	80,648	80,401	79,119	78,025	77,810
Special Education Need Status						
N	*	79,015	83,234	83,290	82,674	80,146
A	*	18,647	13,658	13,080	12,908	14,574
S	*	3,476	3,389	2,953	2,532	2,267

* Data not collected in this year.

Table 25: Summary of the PLASC pupil records

The general downward trend in cohort size as each cohort progresses through the data, as seen in Table 25, cannot be explained as a by-product of pupil record exclusions carried out in the error detection work. The number of exclusions decreases between years 2001 and 2006 as shown in Table 24, which would theoretically leave more pupils in each year group in latter years contradicting the downward trend found here rather than supporting it. The overall decrease in pupil numbers shown in the total number of pupils recorded in the PLASC data is not solely due to a declining number of pupils starting school, although a general decline in pupil numbers in the reception year is clear between 2001 with 7,845 and 2006 with 7,496. The subsequent decrease in numbers as the pupil cohorts progress through their school careers suggests that pupils are leaving the state school system in Leeds, possibly to attend private secondary schools.

Disaggregation of the data by gender in Table 25 shows a relatively even distribution although the number of male pupils is consistently higher than the number of female pupils in all years. The disaggregation by first language shows an interesting trend. Because of the lack of entries in the 'ENB' (first language believed to be English) and 'OTB' (first language believed to be other than English) in 2001 and the marked increase in 2006 these fields are added to the corresponding 'ENG' and 'OTH' fields for consideration. The number of pupils speaking English as a first language declines from 90,706 in 2001 to 86,012 in 2006 after a small increase in years 2002 and 2003, this is in keeping with the general decline in pupil numbers. However, pupils that speak or are believed to speak a language other than English are 10,613 in 2001 and increase to 10,673 in 2006. In the intermediary years, there is a sharp fall in 2002 to 8,252 but this steadily increases to 9,525 in 2005. The slow increase in numbers of pupils speaking or believed to speak a language other than English is contra to the general decline in pupil numbers and would initially indicate an increase in the foreign population of the study area.

The disaggregation by FSM eligibility has a slight decline in pupil numbers in 2002 and again in 2006, whilst the years 2003, 2004 and 2005 remain relatively constant. The number of pupils not eligible for FSM show a more consistent decline with one small increase in 2003. Disaggregation by SEN status shows a decrease of approximately 5,000 in the number of pupils with a SEN without statement, category 'A', between 2001 and 2002. At the same time, the number of pupils with no SEN, category 'N', increases by approximately 4,200. This blip in

the data may be explained by the change in category coding around this time which may have lead to the reclassification of some pupils.

5 Commuting Data from the 2001 Census Special Travel Statistics (STS)

In this section we provide a short synopsis of the flow data available from the 2001 Census by way of context, and examine in more detail the data on commuting to place of study as an alternative set of data to PLASC.

The Census of Population, collected every decade and collated by the Office for National Statistics (ONS) for England and Wales, General Register Office for Scotland (GROS) and the Northern Ireland Statistics and Research Agency (NISRA), has three datasets explicitly containing data on flows between an area of origin and an area of destination. One of these datasets is the Special Migration Statistics (SMS) which contains information on the population moving usual residence in England, Wales, Scotland and Northern Ireland in the year prior to the Census. The Special Workplace Statistics (SWS) in England, Wales and Northern Ireland and the STS in Scotland, are concerned with commuting patterns and the journeys that are undertaken on a regular daily basis to work in the case of the UK and also to study in the case of Scotland.

Each of the flow datasets obtained from the 2001 Census has tables generated at the different geographical levels shown in Table 26. The geographical units become smaller as the data level increases. The level 1 data are created at district geography, a course geographical scale suited to national studies and summary statistics. The level 2 data are available at the ward geography which some might argue is still a relatively coarse geographical scale. For example, the Leeds district with a population of 715,402 in 2001 (extracted using Casweb Table KS001 'Usual resident population' using the cell 0001 'All people') contains only 33 wards. The level 3 data are available for the finest geographical area in the 2001 Census, the Output Area (OA). OAs are constructed from postcode units containing approximately 125 houses and the Leeds district contains 2,439 OAs. Stillwell et al. (2005) summarize the interaction data available at each level, indicating that as the geographical scale becomes more detailed, the level of data disaggregation becomes less refined. For example, the 2001 SWS dataset at level 1 contains seven tables

with disaggregation detail such as 'living arrangements by employment status by sex' whereas at level 3, only one table is available relating to 'method of travel to work'. Thus, at the finer geographical scales, only more aggregated data variables are available.

Country	Level 1	Level 2	Level 3
England	London Boroughs (33) Metropolitan Districts (36), Unitary Authorities (46), other Local Authorities (239)	CAS wards (7,969)	Output Areas (165,665)
Wales	Unitary Authorities (22)	CAS wards (881)	Output Areas (9,769)
Scotland	Council Areas (32)	ST wards (1,176)	Output Areas (42,604)
Northern Ireland	Parlimentary Constituancies (18)	CAS wards (582)	Output Areas (5,022)
Total	Districts (426)	Interaction wards (10,608)	Output Areas (223,060)

Source: Stillwell & Duke-Williams (2007, p.2)

Table 26: Geographical unit levels for 2001 Census SMS/SWS/STS data

Census data have traditionally excluded information on school-aged childrens' daily commutes to school whilst at the same time have held detailed information on adults' journeys to work in the SWS dataset. However, the inclusion of the STS dataset for Scotland instead of the SWS dataset in the 2001 Census has broken this tradition. In addition to asking a question about place of work, the 2001 Census in Scotland asked about place of study. Whilst the STS dataset is similar in the 16+ ages to that of the SWS dataset collected across England, Wales and Northern Ireland for those of working age, the STS age categories include those for children under the age of 16. The age category breakdown for children at the geographical level 1 is 0, 1-2, 3-4, 5-9, 10-11, 12-14 and 15. At level 2, this is aggregated to just three categories 0-4, 5-11 and 12-15 and at level 3, the age choice is simply divided into all people 'age 16-74 in employment' and 'other persons'. Table 27 shows the tables available in the STS with the counts available at each level in each table. Level 1 tables have more counts available than those at level 2 and 3 and it can also be seen that there are seven tables available at level 1, six available at level 2 but only one available at level 3.

One important advantage of the STS dataset for Scotland compared with the SMS for the rest of the UK is that it has not been subject to the Small Cell Adjustment Method (SCAM), a disclosure prevention method implemented by the ONS

Table	Attributes	Level 1 count	Level 2 count*	Level 3 count
STS 1	Age by sex	183	114	-
STS 2	Family status by sex	258	135	-
STS 3	Method of travel to place of work or study (by sex at level 1)	186	62	50
STS 4	NS-SEC (by sex at level 1)	174	58	-
STS 5	Industry (by sex at level 1)	201	52	-
STS 6	Ethnic group by sex	120	-	-
STS 7	Employment status by sex	51	51	-

*Also refers to attribute counts for postal sectors

Table 27: Tables, attributes and attribute counts in the STS 2001

across the SMS and SWS 2001 Census datasets for England, Wales and Northern Ireland (Stillwell & Duke-Williams, 2007). The SCAM, although not published by the ONS, is believed to replace values of 1 or 2 with either 0 or 3. If a value of 1 is to be replaced, then it has twice the probability of becoming a 0 rather than a 3; if the value to be replaced is a 2, then it has twice the probability of becoming a 3 rather than a 0. In this way, it is assumed that the overall flow adjustment will even out, though unfortunately the adjusted zero values are not distinguished from the recorded zero values. It is not only the fact that the data have been manipulated to prevent disclosure that causes a problem, but also the uncertainty that this introduces into any analysis incorporating the data. Interaction data by their nature contain high numbers of small values, especially those for flow counts at fine geographical scales. Thus, SCAM means that much data analysis carried out using the Census SMS and SWS datasets at fine geographical scales is subject to a high level of uncertainty.

The Scottish STS data have not been subject to SCAM and the inclusion of school-age children means that this new dataset is well suited for analysis of the daily commute to school in Scotland. Indeed, an examination of the STS dataset by Fleming (2006) confirms that school-age pupils in Scotland are less likely to travel from a rural to an urban area to study than adults are to attend their place of work or study for example. However, detailed analysis of the data on school-age commuting patterns revealed several drawbacks to the STS (Harland et al., 2006). The study indicated that the most suitable geographical level for using the STS is level 3, whilst showing the difficulty in isolating schools as individual destinations,

given that the data are for commuting flows between OAs. It has to be assumed that pupils move from the centroid of an origin OA to a school which is located at the centroid of the destination OA. Whilst 85% of schools in Scotland do not share an OA with another school, the remaining 15% are present in an OA with more than one school. This makes determining the exact destination point for school pupils travelling to that OA difficult and it is not possible to distinguish the flows to each school from one another. A further difficulty is the coarse age breakdown of the data at this fine geographical level, making it impossible to distinguish primary school age children from secondary school-age children in a precise way.

In general terms, the journey to school is relatively short, posing an additional problem when calculating the distance travelled to school, particularly when the precise school location is not known. Under these circumstances, one approach is for each pupil in an OA to be assigned to the centroid of the OA at the origin of the journey and distance is calculated to the weighted centroid of the destination OA of the school. The median distance travelled to school in Dundee is calculated as 0.7 km for primary school children and 1.2 km for secondary school children. When dealing with distances of this short nature, the need for locational accuracy is essential and using the weighted centroid of an OA will inevitably introduce significant inaccuracies into the calculations. Furthermore, many of the journeys to school take place within an OA. In this case, an estimated distance to school can be produced using the calculation of half the radius of a circle that represents the exact area of the OA. This again introduces inaccuracies into distance calculations. In addition, the only data variable available for analysis at this particular geographical scale is 'mode of transport to place of work or study'. This is restrictive considering variables for which variations in distances travelled are most interesting (e.g. ethnicity, social status and gender) are only available at level 1 and level 2, as shown in Table 27. Interestingly 'mode of transport' is not a variable collected in the PLASC between 2001 and 2006, although the variable was introduced in the 2007 tri-annual data collection. However, the variable is only collected in the January data collection and not in May or September. Because of variations in weather conditions and daylight hours throughout the year a single collection of the 'mode of transport' variable in the middle of winter when daylight hours are at a minimum, will probably not reflect pupil transport choices at other times in the year when the weather is more clement and daylight hours are much longer.

Although the STS dataset is a valuable source of information for commuting to school in Scotland, the limitations of data variables at the appropriate geographical scale, in conjunction with the inaccuracies introduced into distance calculations by using OA centroids as a proxy for pupil and school locations, presents limitations when using the dataset for studying the interaction of pupils with schools on a daily basis. Furthermore, the Census STS can only be used for analysis of commuting in 2001 in Scotland.

6 Conclusions

The daily commute to school involves large numbers of people and deserves more attention by researchers than it has received hitherto. In Leeds alone, there are well over 100,000 children travelling daily to state primary and secondary schools, many accompanied by parents or guardians. The impact of these flows on traffic volumes is considerable. Local adult commuters in Leeds find the journey to work during school holidays significantly easier.

At a national level, the NPD has been established to collect together details of school pupils. The PLASC is one dataset in the NPD in which each student has a unique identifier, enabling pupil records to be linked across tables within a time slice or through time, thus allowing each student to be tracked through the longitudinal series. The fact that the location coordinates of home and school can be geocoded for each pupil in the PLASC data enables cross-sectional patterns of commuting to be examined for any combination of the attributes assigned to the individual. This means that PLASC data can be used to study school territories for boys and girls, for different ethnic groups, for different year or age groups and for different types of school, for example. Moreover, the tracking of pupils from one PLASC to another creates immense potential for longitudinal analysis of individual cicumstances. Clearly, PLASC data provides the opportunity for spatial analysis other than that of the commute to school and work on variations in attainment is a particular focus. Additionally, there is a huge potential for using the data to examine patterns of pupil mobility between schools and tracking child home migration rates becomes possible.

The PLASC datasets that have been supplied by *Education Leeds* have allowed us to investigate the reliability and accuracy of particular attributes of the data

through time. Various methods of checking and interpolation have been reported in the paper to emphasise what is required to limit errors and inaccuracies where possible. When checking is undertaken, the postcodes and geographical coordinates are amongst those attributes that are most prone to error or inaccuracy. In response, and using the longitudinal information available, methods have been derived to improve the accuracy of the geocoded points that represent the origins from which the pupils journey to school. The application of look-up tables and data interpolation techniques has greatly increased the quality and reliability of the final processed dataset with minimal exclusion of unreliable data (approximately 1% across all six years). To facilitate the processing and analysis of the temporal datasets, a relational database has been specifically designed to accomodate the features of the PLASC and the related data required for this study. This database structure provides a valuable model for organising the datasets that LAs are collecting to facilitate checking and interpolation. The next step would be to construct a user-friendly interface that would allow users to build queries, undertake checks, perform interpolations and extract data required for analysis or for reporting. The suggestion here is for the development of an Education Planning Support System (EPSS) that might be expanded in due course, to contain analysis, modelling and projection modules (see Geertman & Stillwell (2003) for further details on planning support systems).

Finally, the paper has examined the PLASC in contrast to interaction datasets available from the 2001 Census of Population, namely the STS dataset for Scotland. Although the latter dataset contains useful flow counts by mode of travel down to the level of output areas, distances between these small areas have to be computed between area centroids with assumptions made about intra-area distances. In contrast, PLASC data for England and Wales allow more precise specification of the origins and destinations of pupil commuters. Equally important is the fact that STS are only available for census dates whilst the PLASC data are available on an annual basis from 2001 and will be available on a triannual basis from 2007.

Thus, whilst it is tempting to consider lobbying ONS to develop a set of STS in 2011 for England and Wales along the lines of the STS data for Scotland in 2001, our advice from this intial exploration of the PLASC data for Leeds is to advise the DfES to allow the PLASC data to be developed as a national resource

for data on commuting to school along with data on other dimensions of education. The interaction data alone have enormous potential for informing education analysts and policy makers. However, if this route is to be followed, more effort and resources are required at the LA and school levels to ensure that the data are collected, collated, checked and processed so as to minimise the error and, in so doing, maximise the utility of these rich datasets.

References

Armytage, W. H. G. (1970). *Four Hundred Years of English Education*, 2nd Edition. Cambridge University Press, Cambridge.

Burgess, S., Briggs, A., McConnell, B., Slater, H. (2006). 'School Choice in England: Background Facts'. *Working Paper 06/159*, Center for Market and Public Organisation, University of Bristol.

Codd, E, F. (1970). 'Relational Model of Data for Large Shared Data Banks'. *Communications of the ACM* **13**(6):377 – 387.

Dawson, R. (2001). *Relational Databases Design and Use*, 3rd Edition. Group D Publications Ltd., Loughborough, United Kingdom. ISBN:1874152098.

DfES (2004). 'National Curriculum assessment and GCSE/GNVQ attainment by pupil characteristics, in England, 2002 (final) and 2003 (provisional)'. *National Statistics First Release 04/2004*, Department for Education and Skills.

DfES (2006a). '2006 School Census: Maintained Secondary Schools Technical Specification'. URL: http://www.teachernet.gov.uk/_doc/9096/School Census 2006 Specification School and Pupil v1.91.doc

DfES (2006b). 'National Pupil Database: The Future'. URL: http://www.bris.ac.uk/Depts/CMPO/PLUG/userguide/catherine.ppt#258,2,National Pupil Database: The Future

Ewens, D. (2005). 'The National and London Pupil Datasets: An introductory breifing for researchers and research users'. *DMAG Breifing 2005/8*, Data Management and Analysis Group, Greater London Authority, City Hall, The Queen's Walk, London, SE1 2AA.

Fleming, A. D. (2006). 'Scotland's Census 2001 Statistics on Travel to Work or Study'. *Occasional Paper 12*, General Register Office for Scotland, Statistics Customer Services, Dissemination and 2001 Census Analysis Branch, General Register Office for Scotland, Ladywell House, Ladywell Road, Edinburgh, EH12 7TF.

Gatrell, A. C. (2002). *Geographies of Health An Introduction*. Blackwell Publishers Ltd, Oxford.

Geertman, S., Stillwell, J. (eds.) (2003). *Planning Support Systems in Practice*. Springer, Heidelberg.

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

Gorard, S., Fitz, J. (2000). 'Investigating the determinants of segregation between schools'. *Research Papers in Education* **15**(2):115 – 132.

Gordon, P., Aldrich, R., Dean, D. (1991). *Education and Policy in England in the Twentieth Century*. The Woburn Press, London.

Harland, K., Duke-Williams, O., Stillwell, J. C. H. (2006). 'Commuting to School: an investigation of 2001 Census STS and PLASC data'. Presentation at GIS-RUK'06 University of Nottingham.

Harland, K., Stillwell, J. C. H. (2007). 'Using PLASC data to identify patterns of commuting to school, residential migration and movement between schools in Leeds'. *Working Paper 07/03*, School of Geography, University of Leeds.

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

Johnston, R., Wilson, D., Burgess, S. (2005). 'England's multiethnic educational system? a classification of secondary schools'. *Environment and Planning A* **37**:45 – 62.

Jones, P., Elias, P. (2006). 'Administrative data as a research resource: a selected audit'. *Economic & Social Research Council Regional Review Board Report 43/06*, Warwick Institute for Employment Research.

Lawrence, I. (1992). *Power and Politics at the Department of Education and Science*. Cassell, London.

Leeds, E. (2006). 'Frequently Asked Questions'. URL: http://www.educationleeds.co.uk/aboutEL/faqanswer.aspx?section=8&FAQid=26

OS (2007). 'Address-Point Technical Information'. URL: http://www.ordnancesurvey.co.uk/oswebsite/products/addresspoint/techinfo.html

Self, S., Dunckley, L. (2003a). 'Relational Database Systems: Development of database systems'. *Computing for Commerce and Industry, course text M876 Block 4*, Computing Department, The Open University, Walton Hall, Milton Keynes, MK7 6AA. ISBN:0749293578.

Self, S., Dunckley, L. (2003b). 'Relational Database Systems: Information systems'. *Computing for Commerce and Industry, course text M876 Block 1*, Computing Department, The Open University, Walton Hall, Milton Keynes, MK7 6AA. ISBN:074929342X.

Self, S., Dunckley, L. (2003c). 'Relational Database Systems: Relational theory'. *Computing for Commerce and Industry, course text M876 Block 2*, Computing Department, The Open University, Walton Hall, Milton Keynes, MK7 6AA. ISBN:0749293470.

Self, S., Dunckley, L. (2003d). 'Relational Database Systems: Using SQL'. *Computing for Commerce and Industry, course text M876 Block 3*, Computing Department, The Open University, Walton Hall, Milton Keynes, MK7 6AA. ISBN:0749293527.

Statham, J., Mackinnon, D., Cathcart, H., Hales, M. (1991). *The Education Fact File*. The Open University, London.

Stillwell, J. C. H., Duke-Williams, O. (2007). 'Understanding the 2001 UK census migration and commuting data: the effect of small cell ajustment and problems of comparison with 1991'. *Journal of the Royal Statistical Society A* **170**:1–21.

Stillwell, J. C. H., Duke-Williams, O., Feng, Z., Boyle, P. (2005). 'Delivering Census Interaction Data to the User: Data Provision and Software Development'. *Working Paper 05/01*, School of Geography, University of Leeds and School of Geogrphy & Geosciences, University of St Andrews.

Teachernet (2004). 'Data collection by type of SEN'. URL: http://www.teachernet.gov.uk/wholeschool/sen/datatypes/

Teachernet (2006). 'Primary Notes of Guidance 2006'. URL: http://www.teachernet.gov.uk/docbank/index.cfm?id=8813

Teachernet (2007). 'Teaching in England: Types of schools in England'. URL: http://www.teachernet.gov.uk/teachinginengland/detail.cfm?id=497

Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M., Wolter, C., Airola, M., Mix, W., Colt, J. S., Hartge, P. (2005). 'Positional Accuracy of Two Methods of Geocoding'. *Epidemiology* **16**:542 – 547.