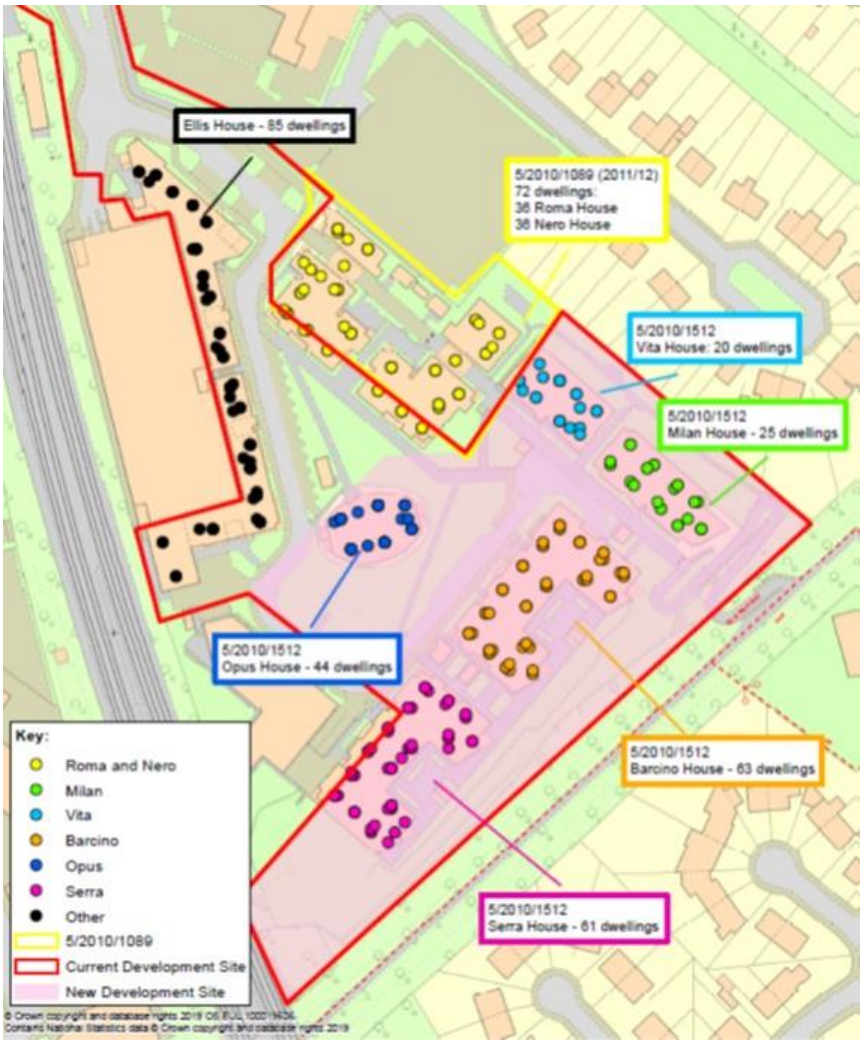
Hertfordshire County Council

Pupil Yield Survey



Methodology for a census of resident mainstream pupil yields from new build housing developments within the boundary of Hertfordshire.

Version 1, 2021

Hertfordshire

Contents

LIST OF TABLES

LIST OF FIGURES

COMMON ABBREVIATIONS

AMR	Annual Monitoring Report
AP	Alternative Provision (Education)
AR	Affordable Rented (dwelling tenure)
ASC	Annual Schools Census
BLPU	Basic Land and Property Unit
CI	Confidence Interval
CIDS	Community Intelligence & Data Science team
CIL	Community Infrastructure levy
COU	Change Of Use
DfE	Department for Education
DAA	Data Access Agreement
DPA	Delivery Point Address
DSA	Data Sharing Agreement
EPA	Education Planning Area
ESC	Education Support Centre
FE	Form of Entry
FoI	Freedom of Information
GDPR	General Data Protection Regulations (2018)
GIS	Geographical Information System
HCC	Hertfordshire County Council
HELAA	Housing & Economic Land Area Availability
ICO	Information Commissioners Office
IDP	Infrastructure Development Plans
LA	Local Authority
LAD	Local Authority District
LEA	Local Education Authority
LLPG	Local Land and Property Gazetteer
LP	Local Plan
LPA	Local Planning Authority
LTA	Long Term Average
MoE	Margin of Error
NCYG	National Curriculum Year Group
NHBC	National House Building Council
NLPG	National Land and Property Gazetteer
NPD	National Pupil Database
NPPF	National Planning Policy Framework
NSA	National Statistics Authority
NSPL	National Statistics Postcode Lookup (ONS)
OM	Open Market (dwelling tenure)
ONS	Office for National Statistics
OS	Ordnance Survey
PAF	Postcode Address File
PAO	Primary Addressable Object
PD	Postcode Directory (ONS)
PE	Point Estimate
PLASC	Pupil Level Annual Schools Census
PRU	Pupil Referral Unit (Education)
PPG	Planning Practice Guidance

PWC	Population Weighted Centroid (of postcode or Output Area)
RM	Royal Mail
RUC	Rural Urban Classification (ONS)
SAO	Secondary Addressable Object
SD	Standard Deviation
SDC	Statistical Disclosure Control
SE	Standard Error
SEN	Special Educational Need
SEND	Special Education Need & Disability
SHLAA	Strategic Housing Land Availability Assessment
SIMS	School Information Management System
SLASC	School Level Annual Schools Census
SMART	Spatial Planning Monitoring, Analysis and Reporting
SNPP	Sub-National Population Projections (ONS)
SR	Social Rented (dwelling tenure)
TCPA	Town and Country Planning Act
PRU	Pupil Referral Unit
UPN	Unique Pupil Number
UPRN	Unique Property Reference Number

1.0 Overview

Hertfordshire County Council (HCC) is responsible for ensuring the provision of a range of services to its resident population and seeks contributions from developments which would have an additional impact on service provision. The process through which contributions, financial or in kind, are generally sought is by the establishment of planning obligations which are intended to make acceptable developments which would otherwise be unacceptable in planning terms.

Historically estimates of early years, primary and secondary pupil yields arising from new housing developments have a varied approach between authorities and developers and data to support these negotiations has often been limited. A more consistent, robust and defensible basis is achieved through informed up-to-date evidence of actual mainstream pupil yields from development, at both the plan preparation and application stages. In 2019 the Department for Education (DfE) published preliminary guidance which specifically addressed the issue of developer contributions towards education requirements, including information about the necessary supporting evidence.

It can be provisionally indicated that the council Pupil Yield Study will include 1,076 developments containing 51,479 dwellings constructed within the boundary of the authority across 19 annual cohorts in the period 2002 to 2020. This assessment will far exceed the number of developments, and dwellings, reviewed in other authorities. The methodology applied by Hertfordshire County Council is presented herein although it is recognised that it is a “live process” and subject to continuous refinement and progression.

In summary and as a result of the emerging Pupil Yield Survey evidence, HCC has reviewed the strategic approach to plan-making to adopt a three tier approach to plan making and has adjusted the Hertfordshire Demographic Model to ensure outputs are supplemented with observed survey data to support the decision-making process. Information about plan making can be found in the Local Plan Engagement Document, while the adjustments to the Hertfordshire Model are detailed in the Guide to the Demographic Model.

2.0 Introduction

Hertfordshire County Council (HCC) is responsible for ensuring the provision of a range of services to its resident population and seeks contributions from developments which would have an additional impact on service provision. Local authorities differ according to the amount of space they have for developments, if they have shrinking or growing populations, whether they are an area of housing growth, the types of development being implemented and where the new builds are taking place.

The process through which contributions, financial or in kind, are generally sought is by the establishment of planning obligations which are intended to make acceptable developments which would otherwise be unacceptable in planning

terms. The HCC approach to seeking obligations is set out in the Guide to Developer Infrastructure Contributions 2021.

Information relating to population yield is necessary for assessing complex infrastructure requirements for the future which in turn can impact on an area's demography. It is broadly acknowledged that estimating the number of children expected to live in a new housing development can be difficult to estimate due to the wide range of factors that affect the outcome (Rockwell et al. 2005).

However, information relating to pupil yield, across all school age stages, from new build housing is necessary for assessing school's capacity, the potential development of new schools and, can result in less land being available for residential developments. There is a significant body of evidence that a high-quality built environment of schools and other settings can also have a direct and positive impact on the quality of learning (Reading Borough Council 2004).

Generally new housing developments place additional pressures on school places through inward migration into an authority and by the redistribution of the existing population into areas where existing schools are at capacity or do not exist within a reasonable distance (Cumbria County Council 2011). Schools need to be located as centrally as possible to serve their catchments and generally also provide a focus for the provision of other community facilities (Cambridgeshire County Council 2009).

Hertfordshire County Council (HCC) is the authority with statutory responsibility for the provision of education services including the provision of sufficient school places for nursery, primary, secondary and sixth form age pupils. Provision must also be made available for children with special needs and childcare spaces in the early year's sector.

In January 2019 HCC undertook a research project into mainstream pupil yields arising from new Hertfordshire residential housing developments. This work has been in development since 2019 and it is intended that the work will continue to be supplemented with up to date information as it becomes available.

The following sections explain the approach taken and describe the methodology in detail. Further enquiries or questions regarding this report should be directed to the following email address: growth@hertfordshire.gov.uk

2.1 Purpose of the Pupil Yield Survey

Local Planning Authorities (LPAs) and developers plan around the strategic overview in their processes such as allocation of land parcels and financial resourcing and increases to modelled mainstream pupil yields raises their risks as well as those of HCC. The high-level approach to plan-making is intended to minimise the chances of underestimating the impact of new development and so should reduce the potential risk of children being without a school place. Equally, having regard to planning legislation, it is important to avoid over-estimating the child yield so as not to seek planning obligations which exceed the impact of a

development. Pupil yield rates from new developments can change over time, dependent in part upon transitional demographic and household characteristics.

In 2019 the Department for Education (DfE) published guidance[1] which specifically addressed the issue of developer contributions towards education requirements, including information about the necessary supporting evidence. The guidance includes several high-level principles, including that pupil yield factors should be based on up-to-date evidence from recent housing developments.

2.2 Emerging Pupil Yield Methodology from the Department of Education

In April and November 2019, the Department for Education (DfE) issued non-statutory guidance on the determination of pupil yield factors from recently completed housing developments. Whilst the DfE have yet to produce a detailed recommended methodology HCC has been in close communication with central government senior project managers, and analysts, associated with the project. HCC has been able to ascertain that, at high level, the methodologies are very similar between the two independent studies.

The DfE continues to indicate that locally held evidence to support any pupil yield method should be used where available. It is incumbent on the county council to assess emerging proposals against evolving methodology and locally held data now. It is acknowledged that this data will develop and become even more robust over time. Prior to presenting the HCC methodology applied it is important to both understand when education contributions are sought by the authority and what factors affect the likely level of contribution. The latter effects the design of any study into Pupil Yield from new housing developments.

3.0 When are education contributions generally sought?

Contributions for education places are commonly sought where schools are already oversubscribed or have been projected to become so. Literature suggests that consideration should be given to projections of pupil growth based on local demography to ensure enough capacity for the existing resident population. This holistic approach, to consider the total proposed developments within an area, is used by some authorities to prevent developers from avoiding contributions through dealing with a site via more than one planning application.

Overall, factors that influence when a contribution is sought include:

- Development size (total number of dwellings).
- Development location.
- School capacity in the area (for primary and secondary aged pupils), allowing for known and projected growth.
- Early years (childcare and pre-school) capacity in the area.
- Development composition, published literature indicates that sometimes the following are excluded from providing an education contribution:

[1] Securing Developer Contributions for Education, DfE, April 2019.

- Bed Sits, Studio and one-bedroom properties.
- Sheltered accommodation.
- Hostels.
- Student accommodation.
- Specialist elderly housing such as rest homes and nursing homes.
- Redevelopment or housing development schemes which do not increase the number of family houses.

Where it is identified that there are insufficient school places then a developer is often expected to provide:

- The full capital cost of providing new education buildings or extending / refurbishing / remodelling existing buildings (including ancillary facilities such as toilets, storage, hall space, additional parking and, staff facilities.
- The full cost of related fittings, furniture and equipment.
- The provision of, or full cost of acquiring, land and/or rights over land required.

Several authorities seek to provide additional places within existing schools as this maintains stability in the school system, provides places in a timely fashion and, achieves the best value for money. However, where the predicted pupil yield from a development is sufficiently large that it exceeds capacity, or where it is not feasible to expand a school, then a new school may be required to address the shortfall.

3.1 Variables affecting pupil yield calculations

General factors for increased pressures on school capacity can occur from a rise in population birth rate, greater inward migration to an authority, parental choice of one school above another and new housing developments (Lancashire County Council 2011). A number of local authority's state that the level of contribution sought from developers depends on the type of housing that is intended to be developed. For example, a development of large family dwellings would be expected to generate a higher number of primary and secondary age children than a development of one- and two-bedroom houses. However, it is not uncommon for there to be uncertainty around the size of large new developments and the mix of housing generated within them (EMIE & NfER 2006). The type of accommodation, tenure and size are broadly acknowledged to influence the child yield as will the locality of a development (Hollis 2005 and the Greater London Authority 2005).

Tenure often relates to two broad groups: social housing and market housing. Social housing is provided by a landlord based on housing need and rents are no higher than target rents set by the government for housing association and local authority rents. Market housing relates to owner-occupied and private rented housing which does not meet the affordability and access criteria for social housing or intermediate housing. EMIE and NfER (2006) reported that it is an accepted convention that pupil yield from new housing varies with the size of properties and many authorities use formulae based on the number of bedrooms. However, in some authorities there are local circumstances in which smaller properties are more densely occupied and yield higher numbers of pupils than might be expected, such as: rented (especially short-term lettings); developments including social housing; flats and; developments in areas of rising house prices.

Overall from a review of published literature common base data required for estimating mainstream pupil yield should, where possible, include the following variables:

- Dwelling Type – identifying flats and houses separately.
- The number of bedrooms in each dwelling.
- Tenure – distinction between social rented and private ownership.
- Number of children by age in each dwelling – for pre-school, primary and secondary school ages (including Post-16 where appropriate).

Some authorities apply discounts for affordable housing. However, to simply discount contributions from affordable dwellings, to reflect their lower market value, does not change the number of pupils likely to arise from the development requiring education. Such discounts would only serve to increase risk both to the provision of sufficient school places for children and to the public purse.

The open market dwellings within a new development are just as open to local private ownership/private rented residents. The presumption is that families in local open market dwellings in the locality are less likely to move into dwellings within a new build development. A discount presumes that families moving locally to new build affordable housing have in the most part reached parity such that they create little further demands on education, this is unlikely to be the case. Housing demands with many councils are significant and for many years HCC has had the lowest proportion of vacant dwellings of all Shire authorities. The significant demand for housing is indicative that backfill of properties with family units containing children will occur. Demand for local school places will not be negated through families moving into new build developments.

Affordable rented (AR) and social rented (SR) dwellings are well known to have higher mainstream yields than that of equivalent bed size open market dwellings although the proportional representation to total dwelling stock is substantially lower. The higher single dwelling yields for AR/SR tenure types results in calculated single dwelling financial contribution costs which are substantially larger than that of open market dwellings.

4.0 Administrative census versus survey and sample

There are several methods which could be used to determine mainstream pupil yields from new build developments such as postal, telephonic, electronic (online form submittals) or “on the ground” door to door. However, the precision, accuracy and confidence of outputs from many of these is dependent upon the sample/survey framework, sample/survey size, response rate, available resources (both human and financial) and type of survey. Survey type relates to random sample selection, weighted, clustered, stratified and so forth with each having their own pros and cons.

The central theme is that a survey, or sample, provides results from which inferences can be made to the population as a whole and therefore must be representative of this population with avoidance of bias at all levels. For example, an electronic survey would exclude those persons without access to the internet, door to door knocking in daytime excludes those persons out at work, voluntary responses only include those persons prepared to spend the time to submit a response and so forth. Surveys tend

to be specific to areas where new developments have occurred, and their robustness is directly proportional to the sampling methodology and response rates. Yields determined from samples or surveys tend to be more specific than demographic ratio methods and take into account factors such as accommodation type (house or flat), size (number of bedrooms) and tenure (affordable and market housing) which are accepted to influence overall child yield from a development. Due to the expense and resource intensity of conducting surveys generally only those developments with a number of dwellings larger than five or ten are often included as this limits the number of sites that have to be visited.

Whilst surveying, or sampling, is a means of providing information with respect to a “whole population” without the need to examine that population in its entirety (termed a census) data determined from a sample permits reliable inferences to be made about the population as a whole only when the Confidence Interval and Confidence Level are known. The Confidence Interval (CI), also referred to as the Margin of Error (MoE), is the plus-or-minus figure usually reported alongside survey results.

For example, if a confidence interval of 3 is applied to a survey and 58% of respondents picks a particular answer it can be "sure" that if the question had been asked of the entire relevant population then between 55% and 61% (58% - 3% and 58% + 3%) would have responded similarly. The wider the confidence interval the more certainty there is that the whole population answer would be within the specified range, however offset against this is that this widening impacts upon the possible range of the answer itself. For example, a confidence interval of 8 when applied to the 58% answer would give a range of between 50% and 66% of the entire population responding similarly.

The Confidence Level (CL) informs as to how “sure” one can be of the survey result, 58% with a range from 55% to 61% in the above example (confidence interval of 3), when applying the answer to the relevant population. The confidence level is expressed as a percentage and represents how often the true percentage of the population would pick an answer which lies within the confidence interval. The confidence level and confidence interval are expressed together such that, for example, one can state a 95% certainty that the true percentage of the population is between 55% and 61%. When a confidence level of 95% is applied then this indicates that one can be 95% certain, at the 99% confidence level one can be 99% certain. The confidence level statistic is commonly referred to as the Type 1 Error risk.

The most commonly used confidence level applied within research is the 95% confidence level. A 95% level of confidence means that 5% of the samples or surveys will be off the wall with numbers that do not make much sense. Therefore, if for example 100 surveys are conducted using the same question, then five of them will produce results that are abnormal. Normally researchers do not worry about this 5% because they do not repeat the same question over and over so the odds are that they will obtain results among the 95%.

There are three factors which determine the size of a Confidence Interval at a given Confidence Level, these are of relevance for general understanding:

- Sample size: The larger the sample size then the higher the certainty that the survey answers truly reflect that of the population itself. As such, for a given confidence level, the larger the sample size then the smaller the confidence interval. However, it should be noted that this relationship is not linear such that doubling the sample size does not halve the confidence interval.
- Percentage of responses: The accuracy depends on the percentage of a sample that selects a particular answer, for example, if 96% of the sample responded "Yes" to a particular question and 4% responded "No," then the chances of error are remote irrespective of the sample size itself. However, if these responses were 52% and 48% then the chances of possible error would be much greater, it is therefore easier to be sure of extreme answers than those that are 50:50. A 50:50 situation, whether in consideration of responses to a question “Yes/No” or the percentage of the population believed to have a particular attribute/characteristic under investigation, represents the likelihood of the largest possible errors. At the 50% level the sample sizes required, for a defined confidence level and confidence interval are much larger than where the attribute, characteristic or percentage of responses is 80%. The “worst case” 50% value is often applied where it is necessary to determine a general level of accuracy for a sample/survey already taken.
- Population size: The mathematics of probability prove that the size of the population under study is generally irrelevant unless the size of the sample exceeds a few percentage points of the total population being examined. The survey system therefore ignores the population size when it is "large" or unknown. Population size is only a factor normally considered when using a relatively small and defined group.

The application of confidence interval calculations assumes a genuine random sample of the relevant population. If a sample is not truly random then the intervals are not reliable, non-random samples usually result from some flaw or limitation in the sampling procedure. Authorities which apply a sample to determine pupil yield from new build developments tend to focus on known completed developments which introduces bias into the assessment. Whilst it is less resource intensive to focus on a specific large development there may be specific characteristics associated with the development, such as typology, which produce yields that are not representative of the “whole” population of new build which occurred over a defined period. Such results would be “indicative” rather than statistically robust measures of the actual mainstream yield arising from the whole population of new build dwellings over time. Application of such estimates should be applied with caution to proposals which do not meet the observed attributes of the surveyed development. Results from randomised new build dwelling surveys should always be published with the CI/CL. For example, a randomised survey determines a yield of 35 primary mainstream pupils per 100 dwellings, the CI is $\pm$ 5% and CL 95% (industry standard). The true population yield from new build dwellings can therefore be determined to lie between 30 and 40 per 100 dwellings. In a development of 1,000 dwellings the yield would therefore be between 300 and 400 mainstream primary pupils which is a substantial range.

HCC has applied an administrative census which removes the error element associated with many of the aspects of surveys and samples, it is a study of all dwellings which satisfy the population inclusion criteria. Where the entirety of a

defined population is surveyed then this is not a sample and no Margin of Error is obtained, the result is specific to the whole population as all individuals within the population have been surveyed for a response. HCC considers the population under consideration to be defined as the number of completed dwellings of specified residential classifications arising from developments solely within the boundary of Hertfordshire County Council. All dwellings included in the population were lawfully erected through the Town and Country Planning system as evidenced by planning permission consent being granted by the relevant Local Planning Authority. Each dwelling included within the population was determined to have a development construction start and completion date. Whilst larger phased developments may not be completed in entirety during the inclusion year, the dwellings included within the population in each study year from such developments were identified as occurring either in entirety, or in a phase, which was associated with the commence of producing residential completions in the period. This inclusion criteria permitted the collation of phased developments starting in the same year, but completing in different years, to the same annual cohort.

The application of an administrative census aligns with the work and methodology currently being applied by the Department for Education; it is therefore a homogenous approach. As such the HCC Pupil Yield Study is a census of the whole population of new build dwellings although this is on condition that the necessary data is required and available on a statutory basis. In this context it is a legal requirement for the information to be collected/provided and therefore the whole population is subject to these conditions such that no bias can be introduced. Consideration of databases held internally within HCC determined that statutory planning/dwellings information could be sourced via SMART Herts whilst mainstream pupils could be determined from the Schools Census return.

4.1 Principle data sources

Three principle sources were identified for the data required in the Pupil Yield Study.

4.1.1 SMART Herts

HCC utilises a monitoring system termed the SMART (Spatial planning, Monitoring, Analysis and Reporting) system which records amongst various factors planning permission applications and dwelling completions. The system is jointly used by HCC, and all the districts, and is a web-based data repository for legally required planning and building related information entered by the districts, building control and, annually provided National House Building Council (NHBC) updates, which enables centralised reporting. SMART Herts therefore provides a centralised repository of data relating to both residential and commercial planning applications and completions within the authority area.

SMART Herts picks up all dwelling gains and losses through the Town and Country Planning system. A new dwelling cannot be constructed outside of the system aside from within Permitted Development rights. However, information on the latter is also collated under Prior Approval applications within the same regime and added to the database. Conversions, such as from an office to a block of flats, are also included within the system. Any enforcement appeals would also be included as HCC applies

a system which checks the Planning Inspectorate website. Any dwelling construction not picked up would therefore result from either human error (generally unlikely as both HCC and the Districts validate the data) or be illegal development. The authority collates completions and permissions data in conjunction with, and primarily on behalf of, the Districts as an evidence base for their Local Plans and statutory returns to Government. The data set provided is therefore considered to represent the whole population of completed developments. Within the authority SMART Herts access is generally via the Environment & Infrastructure Directorate, Planning Infrastructure & Economy, Strategic Land Use team.

4.1.2 THE SCHOOLS CENSUS.

The 1996 Education Act (section 537A) provided a statutory requirement for each school in England and Wales to return a pupil census to the then named Department for Education and Skills (DfES). This was originally known as the Form 7 return and mainly dealt with total pupil numbers although, by 2002 schools were asked for the first time to supply detailed information about each pupil including names and address postcode (January each year). Termed the Pupil Level Annual Schools Census (PLASC) this was replaced in 2007 with the Schools Census which is now the Department for Education's (DfE) largest and most complex data collection exercise. Data is provided to the Department for Education for all pupils on a school's admission register on a termly basis. Data is provided to the Department for Education for all pupils on a school's admission register in accordance with:

- Regulation 5 of the Education (Pupil Registration) (England) Regulations 2006
- The Education Act 1996 - section 434 (1), (3), (4) & (6) and section 458 (4) & (5)
- The Education (Pupil Registration) (England) Regulations 2006
- The Education (Pupil Registration) (England) (Amendment) Regulations 2010
- The Education (Pupil Registration) (England) (Amendment) Regulations 2011
- The Education (Pupil Registration) (England) (Amendment) Regulations 2013

The School Census is a statutory data collection for all maintained nursery, primary, secondary, middle-deemed primary, middle-deemed secondary, local authority maintained special and non-maintained special schools, academies including free schools, studio schools, university technical colleges and city technology colleges in England. Pupil Referral Unit/Alternative Provision (PRU/AP) establishments are legally defined as schools and are also included (comprising pupil referral units, 'AP' academies and 'AP' free schools). Collected data is core to the National Pupil Database (NPD) and accuracy is therefore highly important with zero errors expected by the DfE. HCC as the education authority collates the School Census data on behalf of its schools for submittal to the DfE.

Within Hertfordshire alone the School Census provides over 190,000 individual pupil records of school age children, this excludes the approximately 8,000 pre-school children aged three-and-four years reported in the Private, Voluntary and Independent (PVI) sectors from the Early Years Census return. All records are subject to extensive data validation during the submittal process and local authorities, on behalf of the DfE, actively pursue amendments where validation errors occur and as such finalised data sets are as accurate as possible. Within the authority the finalised schools census data sets are held within the Resources

Directorate, Information and Technology, Intelligence Services, Data Collection Team.

4.1.3 Births data and GP registrations

The Population (Statistics) Act 1938 gave the Registrar General power to collate any information obtained by registrars in the process of birth and death registration which is needed for statistical purposes (some amendments made by the Population [Statistics] Act 1960). The information includes confidential items regarding a birth or death which do not appear in the public register and may be used 'only for the preparation and supply of statistical information'. The Statistics and Registration Service Act 2007 (SRSA) came into force on 1st April 2008. Section 39 of the SRSA governs the confidentiality of personal information held by the United Kingdom Statistics Authority and its executive office (the Office for National Statistics).

All information held by ONS and which relates directly or indirectly to a person (whether living or dead) is protected by section 39 of the 2007 Act. Disclosure of identifying information is an offence, unless an exemption to that offence applies. Section 42 of the SRSA created a new legal gateway between the Registrar General and ONS, enabling the Registrar General to provide ONS with any information entered in any births and deaths register, as well as any other information received by the Registrar General in relation to any birth or death. This includes all categories of information collected as part of the birth and death registration process.

Section 42(4) of the SRSA (as amended by the Health and Social Care Act 2012) includes provision for the ONS to supply information on individual births and deaths for the purpose of assisting the Secretary of State or the Welsh Ministers, or any one of a list of health-related organisations to enable them to produce statistics or carry out statistical analysis. This means that disclosive personal information of the specified type can be passed by ONS to the NHS or other health bodies, including local authorities when acting in their health role only, provided the information is used only for the purpose of producing and analysing statistics.

However, onward disclosures by those bodies of this information to non-listed bodies are not authorised by the SRSA. A full risk assessment must be carried out before making the decision to release identifying data. Within HCC it was identified that Public Health colleagues have a Data Sharing Agreement and Data Access Agreement with NHSDigital/ONS for provision of individual record deaths data from 2006 to present and individual record births data from 2008 to present. It is the sole route by which access could be granted to identifying births information and was considered important for inclusion in the wider study to determine birth prevalence by dwelling type, bed size and tenure.

The authority produces a School Place Planning Forecast, part of the data which underpins the DfE required forecast is GP registrations data for children aged 0 to 7 years by anonymised counts to postcode area. The Pupil Yield Study will cross match postcode sector counts of children aged <5 years to development co-terminus postcodes to produce an annual county wide sample-based assessment of yields in the early years from new build developments. The use of postcode small area

geographies permits the determination of early years yields by new build dwelling type although, to date, much of this work has been suspended with prioritisation of the mainstream yields study. Further work is required to determine whether bed size and tenure distinctions can be determined. These assessments will be essential in the longer term for the accurate location of localised early years services and childcare provision.

4.1.4 GEO & SOCIO-DEMOGRAPHIC PROFILING.

Profiling is based on socio-demographic segmentation tools used by both commercial and non-commercial organisations to better understand their customers/clients. Some tools are Public Sector created specifically for use by authorities to classify their citizens into one of several Groups and detailed Types, and each has its own likely characteristics such as demographics, location, lifestyles, motivations and behaviours. Generally, such analysis is based on household level, not individuals, and can utilise more than 450 data variables sourced from a combination of proprietary, public and, trusted third party sources. Such information is not actual household data; rather it is modelled analysis of expected household characteristics.

Although classifications discriminate between households, it does not mean that the authority has data on individuals residing in households but rather indicates expected characteristics from similar households around the UK. As new properties are built, or converted and inhabited, they are automatically placed into a group which reflects this type (occupants of brand-new homes who are often younger singles or couples with children). As the data footprint of the family increases and improves over time their segment classification will change to better suit their specific lifestyle. This can provide information on the likely characteristics of residents whom occupy new build developments. There are two types of profiling that the authority can apply: socio-demographic and geo-demographic.

There are two key resources applied in profiling; Household and Postcode level data. This data is normally contained in a spreadsheet with a record for every household in Hertfordshire (approximately 500,000) detailing the full address, AddressBase Unique Property Reference Number (UPRN), Ordnance Survey Grid References, and the corresponding Group & Type classification assigned to that household.

The most common way in which geo-socio-demographic data is used in the authority is by taking local data which contains record level information by home address and matching to Group and/or Type in order to determine their characteristics. Due to contractual obligations much of the data at household, or postcode level, cannot be shared outside of HCC although aggregates such as characteristics of people in identified completed new build developments can be released at county level.

4.2 Annual study periods and development size thresholds

In the PYS trial the time period was defined as the annual financial years 1st April 2012 to 31st March 2013 and, 1st April 2013 to the 31st March 2014. Since successful completion of the trial study annual financial period cohort extracts from 1st April to 31st March for each year 2002 through to 2020 have been implemented from SMART

Herts[2]. This permits the longitudinal examination of mainstream pupil yields from unique annual development cohorts across a 19-year period. The PYS annual new build development completions 2002_03 to current financial period, and inclusion of large developments within the 1990's, once complete will far exceed the number of developments reviewed, either by the DfE (within a single local authority area) or any other local authority.

In the PYS trial only developments >=30 dwellings in size were initially included in the study. This occurred due to observed difficulties with successfully geolocating poor quality School Census address records to small area development polygons. However, refinement of the method, to that presented herein, has enabled the inclusion of developments >=10 to <30 dwellings in size within each annual cohort. Developments <10 dwellings in size are excluded based on being deemed "windfall housing". Such dwelling completions are not planned by districts, but they generally help the achievement of district housing trajectories. Windfall housing is commonly disregarded in population projections due to its uncertain nature over the longer term. The inclusion of only those developments >=10 dwellings aligns the Hertfordshire PYS with threshold sizes based on emerging DfE guidance.

Whilst the principle data sources and time periods for study were established consideration was first given to Information Governance and recording the flow of data streams within the county council Information Asset Register prior to further work commencing.

5.0 Information governance and the General Data Protection Regulations (GDPR May 2018)

HCC processes personal information to enable the authority to provide a range of government services to local people and businesses and as such is registered as a Data Controller with the Information Commissioners Office (ICO) under Registration Number Z6406154[3]. A substantial amount of information is provided within this section to ensure that HCC analysts are fully informed as to requirements and how they relate to the PYS. An indicative range of government services which the authority provides, the types of information relevant to these services, sensitive classes of information and, examples of the types of persons that HCC processes data about is given in Appendix 1.

It is also displayed within the privacy statement of the authority's website that HCC analyses existing service data to ensure that the authority can provide the services needed in the future. On occasion this data is compared or combined with population data from other sources, official data from the Office for National Statistics, NHS Digital or, commercial sources. The information is not used to identify individuals, but rather non-identifying aggregates are used to forecast future demand such as for school places, social care and, health trends. There is also public interest in authority finances being appropriately reimbursed by private developers for services that HCC will be required to provide for the future both in support of such developments, and in

[2] Prior to 2011 planning permission applications were recorded in a different system called "DEMONS", data was transferred to SMART Herts following implementation of the latter replacement system in 2011.

[3] https://ico.org.uk/ESDWebPages/Entry/Z6406154

ensuring that the LA can meet its statutory duty to provide sufficient education/child care places.

HCC indicates that dwelling completions datasets for the period 1st January 2012 to 31st December 2013 sourced from SMART Herts are not considered as personal data beyond contact details for land agents, developers and descriptors regarding a development type, bed size and proposed tenure. This information is already within the public domain. However, the project requires the geolocating of anonymised schools census and births information to development polygons to determine aggregate cohort counts. The application of potentially identifying personal information therefore requires appropriate consideration of the General Data Protection Regulations (2018).

5.1 The Information Asset Register (IAR)

In order to be compliant under new Data Protection legislation, HCC needs to maintain an Information Asset Register (IAR) holding "key elements", these are:

- Data items – what personal information are held - such as name, address, email etc and other sensitive data such as health data, criminal records.
- Format of the stored data – for example is this hardcopy, electronically on a purpose–built system, or standard office software such as an Excel spreadsheet.
- How the data is received and transferred – for example: is it collected electronically, paper through the post etc.
- Location – such as, is the data stored on the Shared Drive, internal, or external system or paper.
- Accountability – the person accountable for the service which uses this personal data.
- Access – who can access this data and what restrictions are in place

In order to map the data that HCC holds it is necessary to both understand, and be able to describe, the information or data flow from one location or system to another. For each function, or activity, that involves personal data HCC collects details to identify what happens to it and by which team, department or even third party. The Information Asset Register is a catalogue of the personal data/information HCC holds and processes, where it is stored, how the data/information moves and who the authority shares it with. A form is completed for each function or activity undertaken with the data and it is important that if the data is transferred to another team or department within HCC that this is clearly identified.

The IAR is therefore a table of information relating, initially, to personal and sensitive data collected or held within HCC. It contains the “Key Elements” and the legal basis for collecting and processing the data, together with other useful information. The IAR is updated and permits the consideration of the following aspects:

- Retention periods – ensure that data is being held for the correct time rather than 'forever'.
- Duplication – does the data needs to be held on more than one format
- Legal Basis – there are 6 lawful bases for processing and the relevant one should be identified before processing starts, e.g. to meet a statutory duty of

HCC; to fulfil a contract with the person (data subject); consent of the data subject etc.

- Data Protection Impact Assessment (DPIA) – Needed when there is risk of harm, significant loss of privacy to the individual e.g. social care or health, cloud-based systems changed or implemented from May 2018 etc.
- Privacy Notice (PN) – This should be in place wherever HCC is collecting data direct from individuals. It should state clearly what we will do with the data collected, how long it is kept for and whether it is shared, it should also identify the lawful basis for processing personal data.

In compiling the Information Asset Register it is necessary to:

- Walk through the information lifecycle to identify unforeseen or unintended uses of data. This also helps to minimise what data is collected and how long it is held.
- Make sure the people who will be using the information are consulted on the practical implications.
- Consider the potential future uses of the information collected, even if it is not immediately necessary.

The basis of the information entered into the IAR comes from a Personal Data Information (PDI) form and should not be confused with the requirements of a Data Protection Impact Assessment (DPIA).

5.2 Data Protection Impact Assessment (DPIA) & Personal Data Information (PDI) Form

When HCC implements a new service or technical solution which changes the way the authority collects, stores or uses personal data it is necessary to check whether a Data Protection Impact Assessment is needed. The DPIA is a legal requirement where the following activities are undertaken:

- Processing personal data for a new service.
- Where a data sharing agreement is commenced or amended.
- If any significant change is made to the technology used within an existing service including upgrades or cloud storage.
- When undertaking profiling for service planning or other purposes.

One of the key principles of General Data Protection Regulation (GDPR) is Privacy by Design, that is planning and designing systems and processes to ensure personal data privacy. This includes implementing role-based access, appropriate security and only collecting the data that HCC needs. The quick guide tool indicated that for the Pupil Yield project, which incorporates schools census and births data, a DPIA was required. A DPIA and PDI were completed prior to commencement of the project and recorded the data that would be processed and the benefits/potential risks to both the individuals whose data was affected and to HCC as an organisation.

The completed DPIA further identified the timespan that the collected data was required for, staff access to the information and, technical security and processes required to ensure the data safety. The DPIA and PDI forms were assessed against the Information Commissioners Office (ICO) guidance by the HCC Data Protection

Team, this ensured that the appropriate measures were in place to mitigate identified risks. Further information regarding the DPIA and PDI can be requested from:

- DPIA data.protection@hertfordshire.gov.uk
- PDI information.governance@hertfordshire.gov.uk.

5.3 The Privacy Notice (PN).

The purpose of the Pupil Yield project was to undertake an administrative assessment of child yield per 100 dwellings (primary, secondary and births) arising from new build developments within the boundary of Hertfordshire. No data or information was collected directly from individuals for purposes of the survey. Data utilised in the examination was embedded within the statutory framework for which the authority is required to collect, and expected to project, future service demands. As such it was not necessary to produce a Privacy Notice.

5.4 Births information – Data Sharing Agreement (DSA) and Data Access Agreement (DAA)

NHS Digital is a corporate body established pursuant to section 252 of the Health and Social Care Act 2012 and is the national information and technology partner to the health and social care system[4]. NHS Digital collect and store some information from everyone's health and care records so that it can be used to run the health service, manage epidemics, plan, and research health conditions, diseases and treatments. They process and publish data and information from across the health and social care system in England. Civil Registration data via NHS Digital is replacing Office for National Statistics data supplies which removes the need for ONS Terms & Conditions and named users. This takes place under the legal basis of Section 42(4) of the Statistics and Registration Service Act (2007) as amended by section 287 of the Health and Social Care Act 2012 and Regulation 3 of the Health Service (Control of Patient Information) Regulations 2002.

Since April 2013 the Health and Social Care Act has provided local authorities with the power to perform public health functions. To deliver public health, local authorities need to use available health data sources to get relevant health and social care information. In order to access this information local authority's, require a Data Sharing Agreement (DSA) and a Data Access Agreement (DAA) with NHS Digital, these documents establish the framework within which data can be accessed and analysed amongst other statutory requirements. The births data for each defined local authority is securely distributed to the LA each quarter by NHS Digital together with an annual refresh of the births data containing any required updates.

The Director of Public Health is the Information Asset Owner for the births and deaths data and is responsible on behalf of the Local Authority to NHS Digital for ensuring that the data supplied is only used in fulfilment of the approved public health purposes as set out in the DSA. HCC has both a DSA and a DAA in place with NHS Digital (reference DARS-NIC-35699-L3K3Q-v2.4) and use of provided data is specifically covered within Section 5 (the Purpose). Within the DSA the authority as Data Recipient is recognised as the Sole Data Controller. NHS Digital retains

[4] http://www.isb.nhs.uk/library/standard/128

copyright of the Data, application of births information as applied within this project is therefore acknowledged as: © Copyright 2020, re-used with the permission of NHS Digital (All rights reserved). The authority has a responsibility to ensure that any publication derived from the Data by any party complies with Anonymisation Standard for Publishing Health and Social Care Data guidance and Anonymisation: managing data protection risk code of practice. HCC has undertaken an organisational risk assessment exercise to ensure compliance with these guidelines and a Data Protection Impact Assessment (DPA Registration Number: Z6406154).

An overview of the project and the requirement for access to individual birth address information for the identified and finalised development polygons was submitted to NHS Digital via public health intelligence. At the request of NHS Digital amendments were made to both the Data Sharing Agreement and the Data Access Agreement. These documents which both grant access to the individual births information, and establish the framework within which the information can be used, are held by HCC.

6.0 Pupil Yield Study Overview

Figure 1 displays the overall processes associated with the principle data sets: SMART Herts, School Census and Births/GP Registrations. The initial step was the identification of developments which should be included within each annual cohort 2002_2003 through to 2019_2020.

Once developments satisfying the population inclusion criteria were identified SMART Herts data files relating to each development in each annual cohort were aggregated. Specific development polygons extracted from SMART Herts were used by the HCC GIS team to obtain *AddressBase_Premium* dwelling addresses by specific residential dwelling characteristics. Dwelling counts by type were compared to SMART Herts data sets to ensure totals matched in relation to total number of dwellings and counts by type specific to each permission. Master address files were created for each development and in aggregate for each annual study cohort.

The postcodes arising from the master address files were used to extract specific individual anonymised school census records from the January School Census return 2007 to 2020. For early cohorts between 2002 and 2006 January School Census records were extracted based on co-terminus postcode data. School Census records were address cleansed and Unique Property Reference Number (UPRN) identified. Linking the two data sets based on UPRN established mainstream sector counts by specific dwelling UPRN in new build dwellings over time. Longitudinal mainstream counts in aggregate for each development were determined and the arithmetic mean taken each year to determine the variance of average development yield over time within each study year. This was repeated for dwelling type in addition to dwelling units overall. Development typology was determined, and the analysis repeated to calculate mainstream sector yields for each development characteristic Tier.

The latter parts of the project: ACORN Household, FOI (HMRC & ONS) and SMART Herts Individual Dwelling relates to further work that needs undertaking once all annual cohorts are completed. ACORN is specific to geodemographic and socio-economic profiling of new build development populations. SMART Herts individual

dwelling involves the inclusion of bed size and tenure data for each dwelling completion recorded in the system 2020_2021 onwards. The Freedom of Information Act process relates to the obtaining of UPRN specific bed size and tenure data from HMRC/ONS for those dwellings to which HCC has been unable to assign this information. Each of these elements is discussed in greater depth in the proceeding sections.

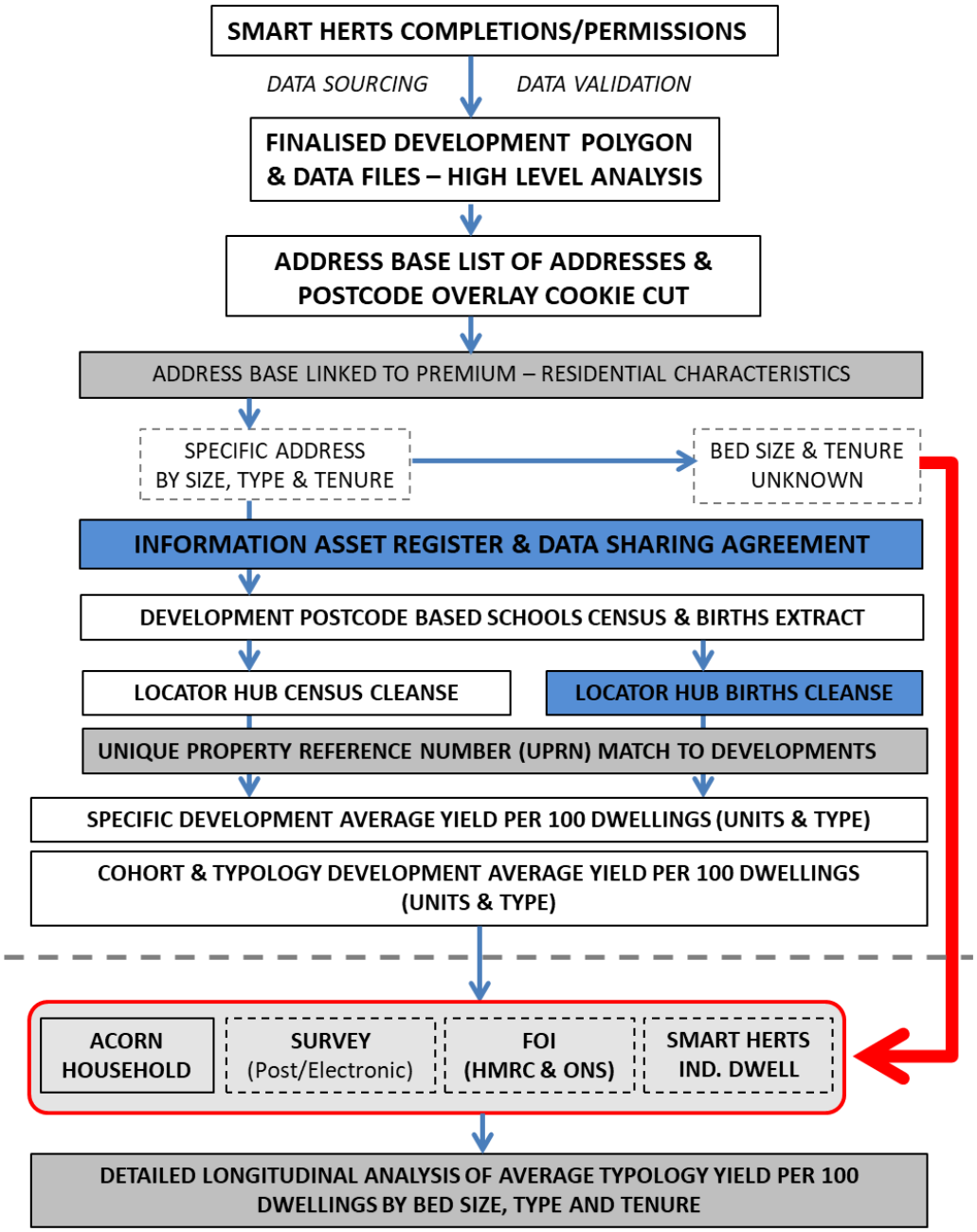


Figure 1. The high-level overall components to the Pupil Yield Study.

.

7.0 SMART Herts data sets processing

Within the authority SMART Herts access is via the Environment & Infrastructure Directorate, Planning Infrastructure & Economy, Strategic Land Use team whom ran

the relevant extract reports from the system. There were three principal data files extracted for each district for each financial year 2002 through to 2020:

(1) Overall Permissions.
(2) Residential Completions.
(3) Size_Type Completions.

This generated 30 data files for all ten districts within each financial year with a total extract of 570 files for the whole study period 2002 through to 2020. Additional extract routines were run to obtain specific development polygons per annum for GIS analysis.

7.1 Overall permissions data files

The initial step undertaken was to establish the population of developments in the boundaries of Hertfordshire, for each annual cohort, which should be included within that specific year's assessment. The overall permissions data files relate to planning permissions indicated as current within the annual financial period under investigation. Developments were included in a specific annual cohort on the basis that residential completions began to be produced in the inclusion year (>0), but not prior to the inclusion year, and where the total gross permitted dwellings was >=10 in count (this excludes developments which had dwellings under construction [UC code] but no completions in the financial period).

This enabled the determination of the annual development cohort >=10 dwellings in size, the total gross proposed gains associated with these developments and, the count of residential completions in the first year of construction. No indication of C2/C4 (older persons, sheltered accommodation, Houses of Multiple Occupation etc) developments is available within these data files and as such all developments meeting inclusion criteria are considered at this point. Defined fields included within the determined annual cohort data table for each financial year were:

- LAD - Local Authority District Name
- LAD CD - Local Authority District Code (Office for National Statistics)
- PPREF – Unique Planning Permission Reference
- Unique Site Ref – The PPREF prefixed with the LAD CD
- Address – the site address as recorded in the system capped at 250 characters
- PDL – Previously Developed Land flag (Yes/No)
- Permission Granted – Date planning permission was granted for the PPREF
- Permission Lapses – Date planning permission for the PPREF lapses
- Permission Started – Date that construction started
- Permission Completed – Date that all works associated with the PPREF were completed.
- Total Proposed Gain – Total number of dwellings proposed within the PPREF
- Total Proposed Loss – Total number of dwellings proposed to be lost (demolition of existing dwelling stock on site)
- Total Proposed Net Gain – Calculated as Total Proposed Gain minus Total Proposed Loss
- Total Proposed Gain >=10 Dwellings – Calculated flag of Yes/No based on Total Proposed Gain

- Completed to Date Gross Completions – the number of dwellings completed to date within the PPREF
- Completed to Date Net Completions – the number of net dwellings completed to date within the PPREF
- Gross Completions in PYS Financial Year – the number of gross completions in the PYS annual extract year
- Net Completions in PYS Financial Year – the number of net completions in the PYS annual extract year
- Gross Outstanding Commitments in PYS Financial Year – the gross number of dwellings which remain outstanding after the PYS financial year. Calculated as Total Proposed Gain minus Completed to Date Gross Completions
- Outstanding Commitments in PYS Financial Year U/C – the gross number of dwellings which are Under Construction but not completed in that year
- Outstanding Commitment in PYS Financial Year N/S - the gross number of dwellings which remain to be constructed following the current financial year
- Net Outstanding Commitments – the net number of dwellings which remain to be constructed following the financial year
- Application Type – Planning Application Type for example, Full or Reserved Matters
- Completions Started Prior to Current Year – calculated flag (Yes/No) as to whether gross residential completions had been produced prior to the current financial year (all rows to be “No” for inclusion in current annual cohort)
- In PYS Annual Cohort – calculated flag (Yes/No) as to whether the unique PPREF is included within that year’s annual cohort. If “Completions Started Prior to Current Year” = No and “Completed to Date Gross Completions” is >0 then flag = Yes
- Outstanding Commitments Check – Yes/No flag to indicate whether the system reported completions in year plus the outstanding commitments equals total gross proposed gain. Where Flag = No data returned to spatial planning for the PPREF for resolution.
- Inclusion Year – The financial year to which the development is allocated e.g. 2011/2012, 2015/2017, 2004/2005 etc

Developments >=30 dwellings and >=10 and <30 dwellings were processed as separate annual cohorts. The final annual overall permissions files formed the basic table to which residential completions, GIS determined addresses and, size_type files were matched using the unique permission reference for each site (PPREF code).

The Overall Permissions files were also used to determine the total number of residential completions associated with all residential developments in the financial year under investigation. All developments were included in a specific annual cohort on the basis that residential completions began to be produced in the inclusion year (>0), but not prior to the inclusion year, irrespective of development total gross proposed gain. Effectively this included “Windfall” housing developments of <10 dwellings in size such that the percentage of dwellings included in those developments >=10 dwellings relative to the total dwellings relating to the specific financial period could be calculated.

Although the PYS is a census of all developments >=10 dwellings in each financial year, and not a representative sample from which inferences within known statistical boundaries are made to the population as a whole, there is merit in determining the percentage of all dwellings built in the population of intertest relative to the total for that year. For example, if a particular financial year had 1,000 dwellings included in the PYS from developments >=10 dwellings in size then, despite being a census including all dwellings which meet the population of interest criteria, criticism could be levelled at the cohort size. If the total dwelling count constructed in that year was 1,200 dwellings, i.e. 200 dwellings occurred from windfall, then it can be observed that the census included cohort was 83.3% (1,000 / 1,200 = 0.833) of all dwellings constructed in the annual period. Despite being diminutive in size the census included 1,000 dwellings cohort would be in excess of 80% of all dwellings constructed in the period which would be greater than the proportions included within randomised samples based on the included population size. This approach increases the evidenced robustness of the overall Pupil Yield Study.

7.2 Residential completions data files

Individual District residential completions data files were aggregated to create singular annual financial year tables, the process undertaken is provided in Appendix 2. As with the overall permissions files fields were of consistent name and format between extract periods and descriptive text data limited to <250 characters in length as required for import to ArcGIS. The consistency of format assisted in the replication of ArcGIS projects for annual survey periods with the replacement of underlying data files and polygons whilst, automated processes could remain consistent for efficiency. The principle data fields were:

- Report Year – the year to which the extract file related
- LAD – Local Authority District Name
- LAD CD - Local Authority District Code (Office for National Statistics)
- PPREF – Unique Planning Permission Reference
- Unique Site Ref – The PPREF prefixed with the LAD CD
- Address – the site address as recorded in the system capped at 250 characters
- Description – a description of the site as recorded in the system capped at 250 characters
- PDL – Previously Developed Land flag (Yes/No)
- Gross Comp in Period – the number of gross dwelling completions in the report year
- Loss in Year – the number of dwelling losses in the report year
- Net Comp in Period – the number of net dwelling completions in the report year
- Wislistperm – A unique identifier which relates the residential completions data to a polygon identifier.

These are “developments” which have involved the conversion or loss of residential dwellings with no replacements occurring. The unique identifier field PPREF was applied in looking up annual completions data to the developments included within the annual cohorts resulting from the overall permissions files. Values added to the

overall permissions files reflect the field names listed above. Additional fields relating to the overall permissions and residential completions data sets were added for:

- Total Proposed Gain Check – a formula to check the number of gross dwellings permitted in the overall permissions files versus the sum of the residential completions associated across one or more years for the development.
- Total Gross Completions – the sum of the gross residential completions observed for each specific development across one or more years specific to the PPREF.
- Match – Yes/No error function to determine whether the total proposed gain check equals the total gross residential completions data.

Where the checks returned divergent values then a mismatch between the overall permissions and residential completions data files occurred. Permissions where this occurred were investigated further, the most common reason for differences were partially superseded permissions or PPREFs which were sub-permissions of a larger development that came forward in parts. In such instances “Estate” files were requested from spatial planning which listed permissions by PPREF associated with complex sites.

Subsequent to the resolution of non-matches for all developments within an annual cohort the permission reference was cross referenced to a C2/C4 report from SMART Herts. Developments of C2 (older persons/residential care homes) and C4 (Houses of Multiple Occupation) do not represent the majority of dwelling stock within the authority from which mainstream school age pupils could arise and are excluded from the Pupil Yield Study. It was noted that C2/C4 development specific data files have not been used until more recently within the authority (2012/13). An additional check was therefore implemented such that the description field for each development was reviewed. Six additional fields were included within the data file for each annual cohort, these were:

- C2 Development from SMART Herts Report – Yes/No flag as to whether a development was identified as C2 from the specific SMART Herts report.
- C4 Development from SMART Herts Report – Yes/No flag as to whether a development was identified as C4 from the specific SMART Herts report.
- C2 Development from Description – Yes/No flag as to whether a development was identified as C2 from the site-specific description field.
- C4 Development from Description – Yes/No flag as to whether a development was identified as C4 from the site-specific description field.
- Include/Exclude in cohort – Include/Exclude Flag wherein developments indicated as C2/C4 were excluded.
- Reason Exclude – Notes field detailing the reason a site has been excluded for future reference.

Other than C2/C4 flagged exclusions the only other accepted reasons for not included a development within an annual cohort was: All addresses in the development Polygon do not have an AddressBase Premium dwelling classification Type which matches the criteria applied for inclusion within the cohort [Section 7]; Development Polygon does not reside wholly within the authority boundary. It is important to note that Residential Completions files also contain completions data for those developments which are specific residential losses only such that there will be

some non-matches between the unique identifier PPREF of the completions data to the Size_Type permissions files. The two data streams, although related, include disparate information and should be treated separately.

7.3 Permissions size_type data files

Individual District permissions size_type data files were aggregated to create singular annual financial year tables, the process undertaken is provided in Appendix 3. As with the overall permissions and residential completions files fields were of consistent name and format between extract periods and descriptive text data limited to <250 characters in length as required for import to ArcGIS. The consistency of format assisted in the replication of ArcGIS projects for annual survey periods with the replacement of underlying data files and polygons whilst, automated processes could remain consistent for efficiency.

Developments built out over more than 1 year and/or where there are multiple providers, dwelling types and/or tenure have multiple rows of data. Figure 2 displays an example size_type processed extract for a singular PPREF wherein there are multiple providers, dwelling types and tenures associated with the development and, the development took more than 1 year to complete.

Figure 2. A development of 68 dwellings wherein there results multiple rows of size_type data due to different providers, dwelling types and tenures.

PP Ref	ResLine Provider	ResLine Tenure Type	Dwelling Type	Number of completions	1 Bed Units	2 Bed Units	3 Bed Units	4+ Bed Units
07/14/0076/F	Housing Association	Social Rented	Flat, Apartment or Maisonette	12	0	12	0	0
07/14/0076/F	Housing Association	Social Rented	House	9	0	0	9	0
07/14/0076/F	Private	Market	House	47	0	0	20	27

The overall permissions and residential completions data sets collated to this point are singular rows of data per unique permission. The initial step undertaken in processing the size-type files was therefore to collate the multiple row data into a singular row for cross referencing to the annual cohort master files. For example, multiple providers were combined to “Private & Housing Association”, multiple tenures to “Open Market & Social Rented”, multiple dwelling types to “Flat, Apartment or Maisonette & House or Bungalow” and so forth. The principle data fields resulting from this process were:

- LAD – Local Authority District Name
- LAD CD - Local Authority District Code (Office for National Statistics)
- PPREF – Unique Planning Permission Reference
- Unique Site Ref – The PPREF prefixed with the LAD CD
- ResLine Provider – The provider type for the development, or part of development to which the size_type data row relates. For example, Private, Housing Association, Local Authority, Unknown
- Dwelling Type – House, Bungalow, Flat/Apartment/Maisonette
- ResLine Tenure Type – Tenure of dwellings within the row of dwelling type data
- Overall Units – Number of dwelling units associated with the tenure, dwelling type and provider row of the dwellings for all/part of the relevant permission

- 1 Bed Units – Count of 1-bed dwelling units
- 2 Bed Units – Count of 2-bed dwelling units
- 3 Bed Units – Count of 3-bed dwelling units
- 4+ Bed Units – Count of 4+ bed dwelling units
- Overall Houses – Number of the overall number of houses completed
- 1 Bed Houses – Count of 1-bed Houses
- 2 Bed Houses – Count of 2-bed Houses
- 3 Bed Houses – Count of 3-bed Houses
- 4+ Bed Houses – Count of 4+ Houses
- Overall Flats – Number of the overall number of Flats completed
- 1 Bed Flats – Count of 1-bed Flats
- 2 Bed Flats – Count of 2-bed Flats
- 3 Bed Flats – Count of 3-bed Flats
- 4+ Bed Flats – Count of 4+ Flats
- CHECK – formula which checks that the number of houses and flats equates to the total number of units which in turn equates to the total number of gross dwellings permitted (overall permissions) and residential completions.

The data items were referenced to the annual master files through linkage of the unique identifier PPREF.

Where the check field indicated a mismatch between the overall permissions and residential completions data associated with a unique permission reference versus the size_type data files for overall units (both in aggregate and per specific year of completions) then further work was undertaken to resolve. This was conducted by both referring the permission to spatial planning, GIS analysis of the polygon and, further research into the overall permissions and residential completions data files for further permissions possibly associated with the overall site.

7.4 Known limitation of the permissions and completions data

The Type (House or Flat), Tenure (Affordable/Open Market) and Bed Size data associated with each specific development as presented within the data files is correct as at the time which planning permission was granted. If there are local amendments to the agreed development mix between the Planning Authority and a developer subsequent to the granting of permission, or as a development progresses, then this will not necessarily be reflected in the permissions data file. Changes to the affordable dwelling element of a development would only be determined by comparing a developments permissions file to information held by district housing authorities regarding stock location.

There is currently no Information and Data Sharing Agreement in place with Districts to access this level of information to pick up any such amendments, however experience within HCC spatial planning indicates that this is not common. Dwelling completions can sometimes be associated to the wrong financial year for various reasons such as lag in paperwork, human error etc, but within the authority, and the Districts, these are thought to be infrequent and tend to be odd dwellings rather than large developments.

The SMART Herts data enables the determination of the location and magnitude, number of dwellings constructed, of each of the permitted and subsequently completed developments. The data also enables a determination of bed size mix, type and tenure of constructed dwellings associated with each development overall.

However, the individual addresses of each of the dwellings within a development is not available as a data extract nor is there relational data for each individual dwelling with regards to bed size, type and tenure. Specific address information is required to be sourced externally via AddressBase and AddressBase Premium products which also lack the detailed relational datasets. Aggregation of the developments to county level therefore provides an overall indication of the number of units completed by bed size, type and tenure over time. Information relating to the overall tenure and bed size is still required in order to compare the overall type, tenure and bed size mix of identified developments to observed mainstream yields from the school census.

The size_type data sets and individual address residential characteristics code from AddressBase Premium could also be applied to determine overall number of houses and flats, separately, by bed size and tenure for comparative analysis. Both the bed size and tenure of individual dwellings and, the addition of multiple size-type row data to the overall permission are work streams which will be conducted once the master annual cohorts are finalised 2002 through to 2020. This is of relevance to the >=30 dwelling cohorts whilst many of the small development cohorts are of singular type, tenure and bed size for which this information can be obtained immediately. However, overall the aggregate counts of dwelling type from the size-type permissions enables cross comparison to residential dwelling classification types determined from the GIS analysis of development polygons.

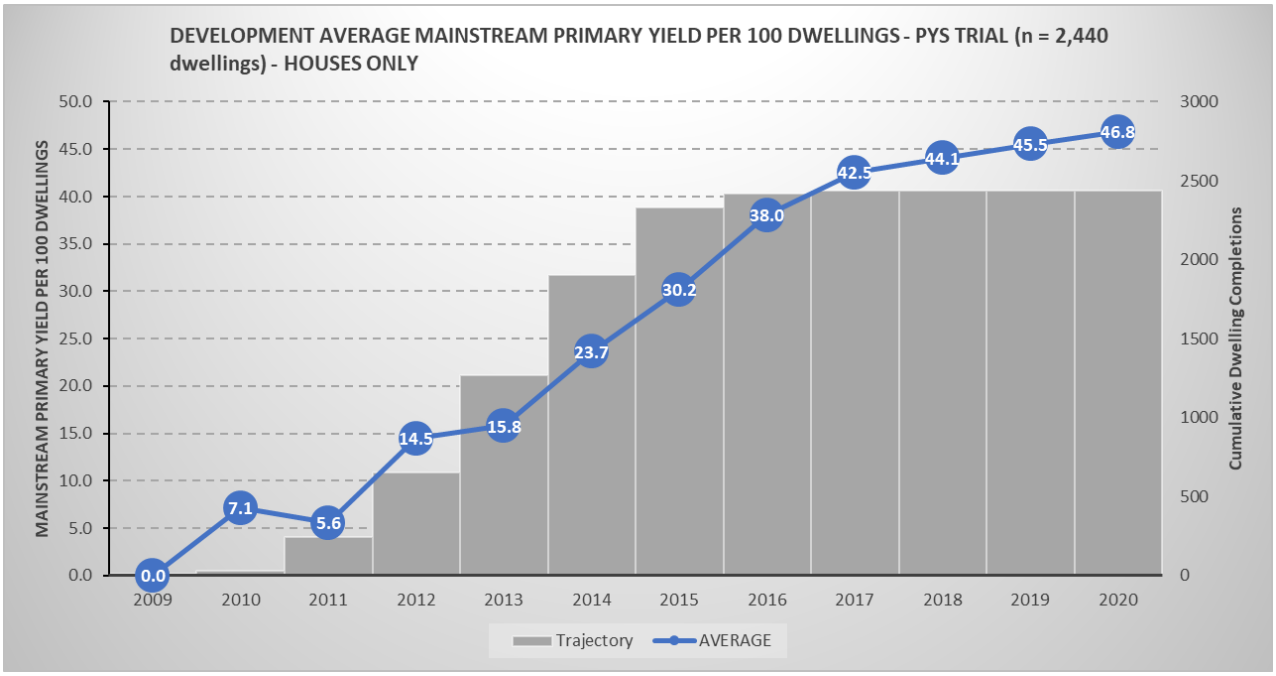
7.5 Trajectory of development completions

Unified counts of aggregate dwelling completions between the overall permissions, residential completions and size_type files permitted the creation of development trajectories specific to each permission. The residential completions files annual gross completions count per annum permits this at a Units Only level of detail whilst the size_type data allows for specific dwelling type (House or Bungalow and, Flat Apartment or Maisonette). Trajectories were constructed from the SMART Herts data sets for each permission within the annual >=30 and >10 to <30 dwellings development cohorts.

The trajectory is important for calculating specific annual yield per 100 dwelling rates from the observed number of cumulative completions in a particular year versus the observed mainstream pupil counts from the school census data sets in that year. This is the case both for specific developments and the determination of development average yield for all permissions within an annual cohort. If only the total number of dwelling completions were applied, then calculated annual mainstream yields per 100 dwellings prior to development completion would be underrepresented. This occurs as the denominator (number of units, houses or flats) would be substantially larger than the actual number of completions which would have occurred at the mid-point of development construction. An example of the determined trajectory and associated mainstream primary yield per 100 dwellings displayed in Figure 3 below. The data presented results from the initial Pupil Yield

Study trial conducted by the authority and relates to 2,440 Houses Only, the principle is the same for Units Only and, Flats Only.

Figure 3. An example of the trajectory determined from SMART Herts data sets for 2,440 Houses Only, and the calculated development average mainstream primary yield per 100 houses, arising from the initial HCC Pupil Yield Study trial.



8.0 GIS analysis of SMART Herts development polygons

The annual cohort master files for each financial year based on the collation of SMART Herts overall permissions, residential completions and size_type files formed the principle PPREF list for GIS analysis. Financial, as opposed to calendar, periods were applied in the extract processes due to the origination of the development completion date. For some developments the completion date is provided by the National House Building Council, or LAD building control, upon issuance of a completion certificate or, via administrative desktop survey. However, in some cases the information is absent, and the completion date is determined via a field survey. If the survey determines a development to be completed, and a date is absent, the end of the financial year date, 31st March, is entered to the completion date field. Consequently, for a proportion of the completed developments it will be known that they were completed in the twelve-month period since the last survey, but this is not accurately transferable to calendar period.

Development polygons were extracted from SMART Herts for all developments on the system 2002 through to 2020. Whilst the polygons give the location of developments all underlying data associated with the shape files other than unique identifier (PPREF) was ignored. Only data from the provided master files was used. This was particularly relevant as the polygons include residential completions which are completions of “losses and gains” whereas the master files are gains only. The polygons associated with the developments were subject to an extensive GIS process. The Hertfordshire County Council corporate GIS solution, ArcGIS, was

used to process and analyse spatial data in support of this task. The 'base' address data used in the initial trial was Ordnance Survey AddressBase Premium Epoch 64 (released 17th January 2019)[5] although updated versions were applied as they became available. GIS processing and analysis of the polygons involved several outcomes:

- Ensure all developments were within the boundary of Hertfordshire County Council.
- Establish an address master cohort for each development, and total count of dwellings, based on residential dwelling characteristics.
- Establish a coterminus and buffer postcode list for each development.
- Produce a map of each development in each annual cohort.

The annual cohort development site polygon data, recorded in the SMART Herts system, was exported as a series of datasets (in ESRI shapefile format) for each Local Authority District in Hertfordshire. These were appended to a pro-forma layer with a spatial extent set for Hertfordshire. Once compiled these polygons were used to spatially select from AddressBase Premium the Basic Land and Property Unit (BLPU) points which they contained. Through matching data from the AddressBase Premium Class Records table to the selected BLPU (Basic Land and Property Unit) using the Unique Property Reference Number (UPRN), and then from the Class Code to the AddressBase products classification scheme table, it was possible to assign a typology (Class Description) to each address contained within each development polygon. Several types were deemed to meet the project criteria in conjunction with customised ONS census defined output tables for unshared dwellings (Table 1).

Table 1. The AddressBase Class Code classification scheme for included residential dwellings.

CLASS CODE	***CLASS DESCRIPTION***
CR06	Public House / Bar / Nightclub
R	Residential
RB	Ancillary Building
RD	Dwelling
RD01	Caravan
RD02	Detached
RD03	Semi-Detached
RD04	Terraced
RD06	Self-Contained Flat (Includes Maisonette / Apartment)
X	Dual Use

Dual use records generally include a residential element, where this is the case, as determined by considering the individual address details and undertaking further research. For example, some dwellings were clearly businesses/residential mixed use such as a Farmhouse, Pub or, a business operating from a home. Dual use BLPU were uncommon and therefore included in the initial selection, relevant

[5] Support and technical documentation for AddressBase Premium can be accessed here: https://www.ordnancesurvey.co.uk/business-and-government/help-andsupport/products/addressbase-premium.html

records were rejected where it became apparent that they did not include residential quarters. A similar approach was taken with CR06 and RB classified records.

Following this a count of BLPU by development polygon was made from which it was possible to identify any development polygon which appeared to contain no residential records. These were few in number and each was investigated to determine what BLPU records were present and reassign an appropriate classification in any cases where it was clear from the details of the planning permission that a non-residential BLPU should be reclassified as a result of the development and associated Change Of Use (COU). Each AddressBase Premium record was attributed with the unique Planning Permission Reference (PPREF) of the development polygon within which it was contained and then filtered to include only those permissions in the annual cohort under consideration. At this point checks were undertaken to ensure that:

- The refined count of development polygons equalled that of the provided master files (based on unique PPREF count), such that there were no polygons absent.
- That there were no overlapping development polygons (all BLPU assigned to one permission [PPREF] only).
- That there was not a lack of BLPU meeting classification criteria within identified developments.

It was observed that some BLPU had not been re-classified since the permitted works were carried out and needed to be updated, some BLPU points ‘missed’ relevant development polygons and needed to be moved with notes indicating as such appended to the unit records. The subsequent step was to assign addresses to the records selected.

AddressBase is essentially the Postcode Address File (PAF) produced by Royal Mail, it is a “flat file” and simple to work with. Whilst there is no inclusive dwelling type information there are classification codes associated with each address record. The available classification codes are:

- C – Commercial (Attracts non-domestic rates and/or use is of a business nature).
- L – Land
- M – Military (Military Defence Site)
- O – Other (Ordnance Survey only)
- P – Parent shell
- R – Residential
- U – Unclassified
- X – Dual use
- Z – Object of Interest

Note that not all residential characteristics will necessarily occur within an area. Individual address records that were contained within the finalised development polygons were selected from AddressBase based on a residential classification code of “R – Residential”, these specific addresses were then linked to AddressBase Premium. Data from the AddressBase Premium Class Records table were matched to the selected BLPU (Basic Land and Property Unit) cohort defined above using the

Unique Property Reference Number (UPRN), and then from the Class Code to the AddressBase products classification scheme table.

AddressBase Premium utilises multiple GIS files thereby requiring a relational database for application and is generally complicated to work with. Through the application of multiple polygons this product permits the determination of the characteristics of each address to provide information both on dwelling type and to further specify the R – Residential classification code from AddressBase. The “point-in-polygon” finalised development addresses previously identified within AddressBase were cross referenced to AddressBase Premium and only those address with the following residential sub-classification extracted:

- Residential
- Dwelling
- Detached
- Semi-Detached
- Terraced
- Self-Contained Flat (Includes Maisonette/Apartment)
- Ancillary Building

The following residential sub-classification codes were specifically excluded:

- Houseboat
- Sheltered Accommodation
- Privately owned holiday caravan/chalet
- Ancillary building
- Caravan
- Car park space
- Allocated parking
- Garage
- Lock-Up Garage/Garage Court
- House in Multiple Occupation
- HMO parent
- HMO bedsit/other non-self-contained accommodation
- HMO not further divided
- Residential institution
- Care/Nursing home
- Communal residence
- Non-Commercial lodgings
- Religious community
- Residential Education

This produced the final list of addresses associated with the validated and included development polygons, by unique site identification number. Prior to exporting finalised address cohorts, development specific buffer and coterminus postcode files were generated.

8.1 Development buffer and coterminous postcode files

Development postcode buffers contain the postcodes associated with a specific permission and those which occur in a 200m range of the boundary. In creating permission specific buffers, postcode layers were superimposed over the

development boundaries, where a postcode polygon intersected/overlay a boundary, or was within 200m, then the postcode was extracted. The relevant PPREF was assigned to each postcode for all developments in each annual cohort. Postcode buffer files were principally applied in the PYS trial study to extract school census records. Individual pupil records were cleansed and geolocated to development permissions using GIS in order to create mainstream pupil counts. These counts were compared to counts from a more specific direct address-in-polygon only school census extracts. It was observed that the more specific address-in-polygon method was as accurate as that of the buffer method and was subsequently applied to the main study.

To determine whether postcodes were wholly coterminous within development sites a combination of Codepoint Polygons and AddressBase Premium records were applied. Using both datasets a postcode was deemed coterminous with a development site if it contained residential addresses (determined using BLPU classes within AddressBase Premium) which fell within the site boundary but no residential dwellings beyond the site. Under this methodology a 'theoretical' postcode (as defined using Codepoint Polygons) which overlaps a development may go beyond the development significantly, but if there are no other residential dwellings aside from the ones within the site then it can be inferred that all addresses with such postcodes are attributable solely to that development site. There are four possible scenarios when determining coterminous postcodes using this methodology:

Scenario 1 – the development site falls completely within a single postcode polygon, and all residential BLPU's within that postcode fall within the development site. Therefore, all dwellings with such postcodes, it can be inferred, would fall within the development site (Figure 4). These development sites are accepted for analysis within the coterminus postcode cohort.

Scenario 2 – all postcodes which overlap the development site fall relatively neatly within it, and therefore all residential BLPU's with such postcodes can be attributed solely to the site. In some instances, postcode polygons may go beyond the site, however if no residential BLPU's are found within such areas then all dwellings within the postcode can still be attributed to the site (Figure 5). These development sites are accepted for analysis within the coterminus postcode cohort.

Scenario 3 – no postcode sits neatly within the development site, however there are no postcodes overlapping the site which contain residential BLPU's which fall both within and beyond the site. As such although 'theoretical' postcode polygons extend beyond the development boundary, the postcodes of all residential BLPU's within the site only belong to dwellings within the site, and therefore such postcodes can be attributed solely to the site (Figure 6). These development sites are accepted for analysis within the coterminus postcode cohort.

Scenario 4 - some postcodes which overlap the development site contain residential BLPU's which fall both within and beyond the development site. As such it is not possible to determine that dwellings within such postcodes solely fall within the development site and therefore these development sites are not accepted for coterminus postcode analysis (Figure 7).

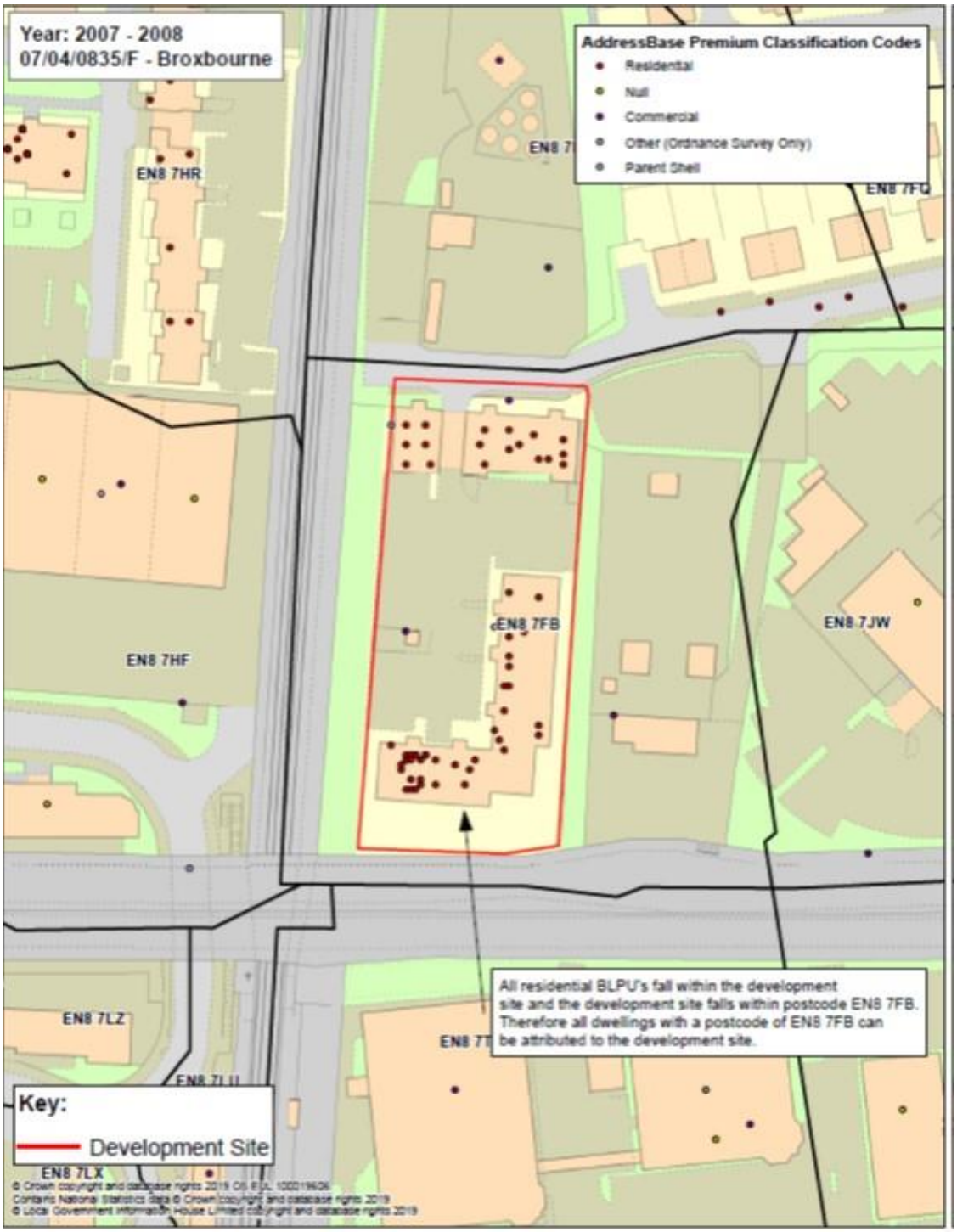


Figure 4. Scenario 1 for determining development coterminus postcodes.

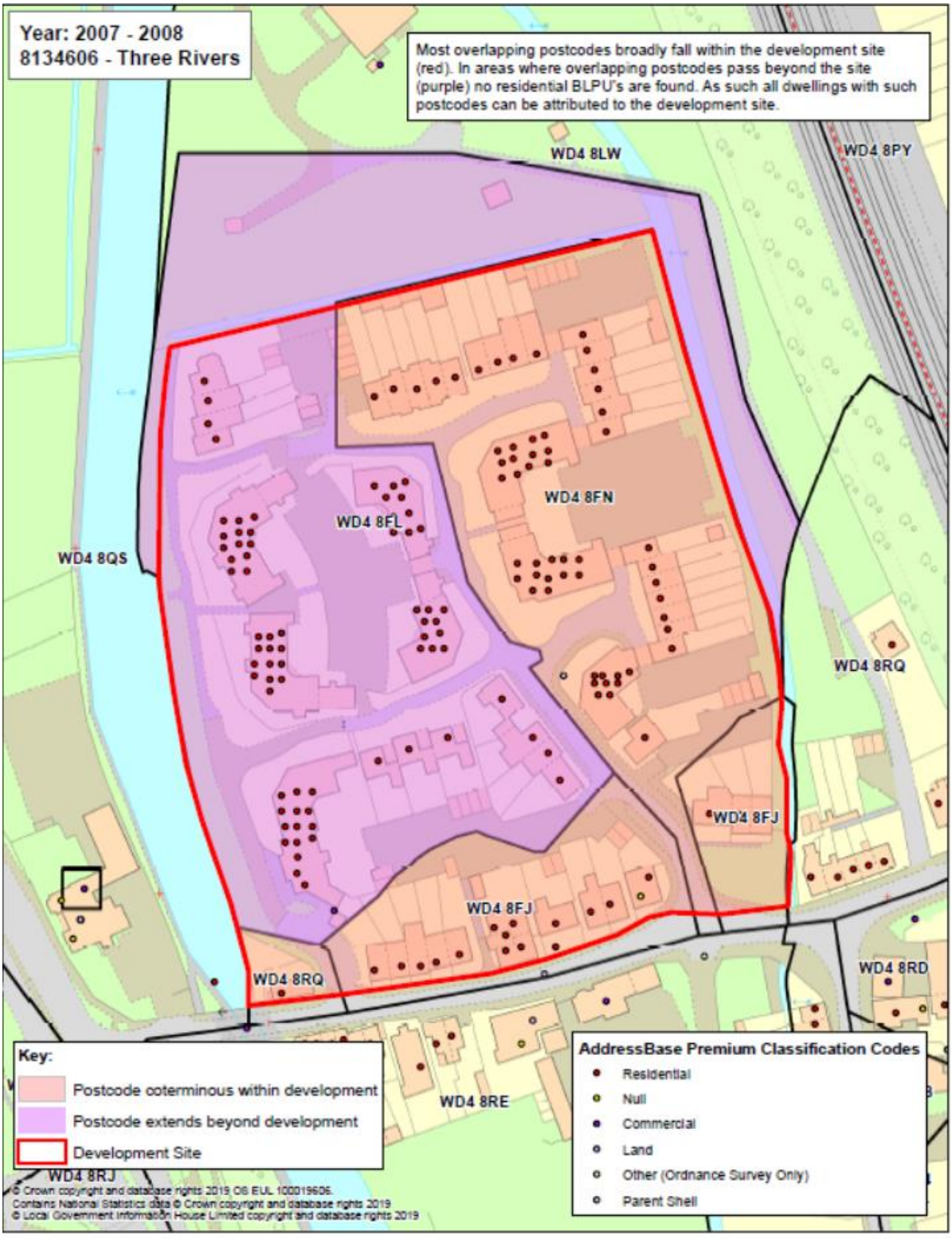


Figure 5. Scenario 2 for determining development coterminus postcodes.

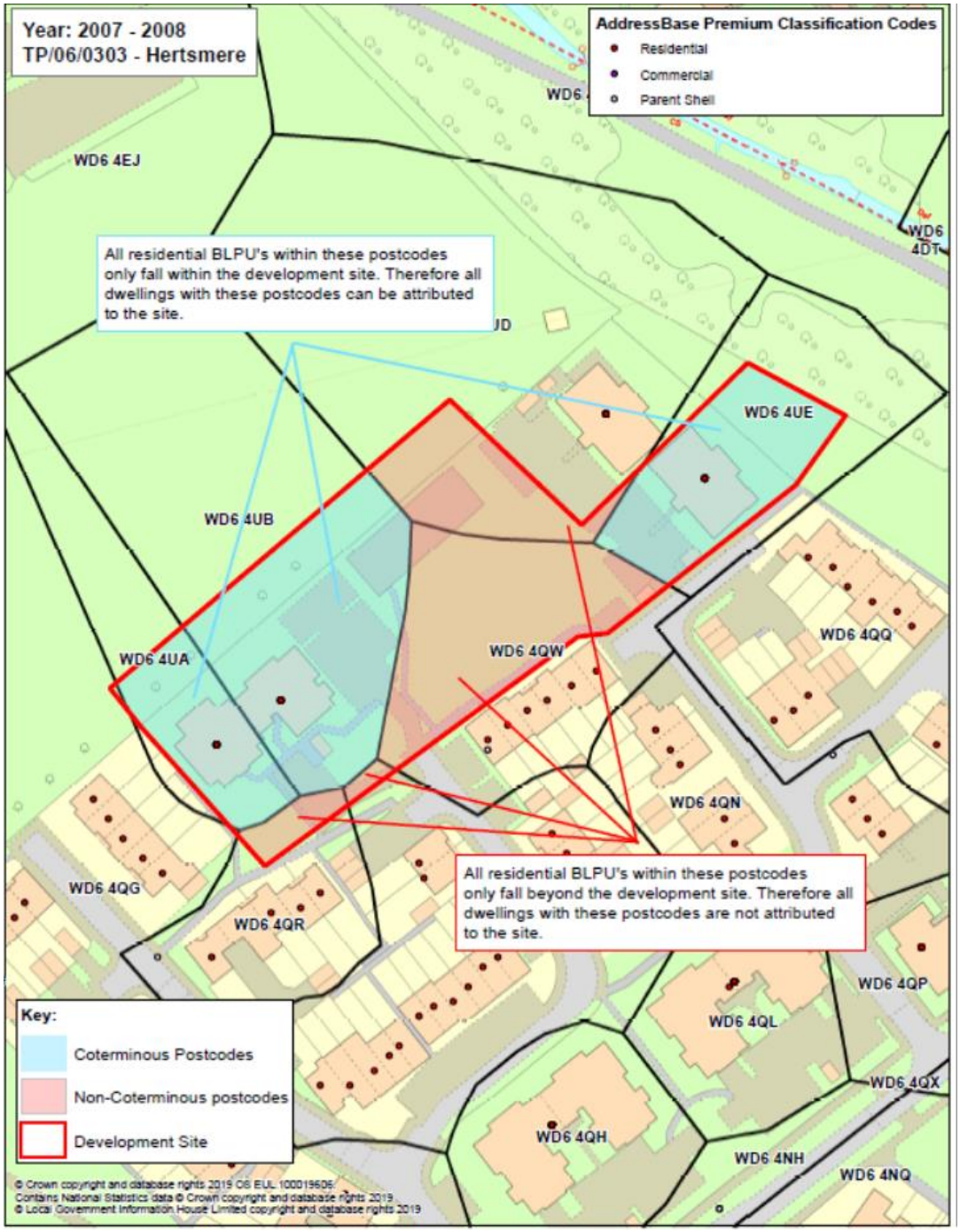


Figure 6. Scenario 3 for determining development coterminus postcodes.

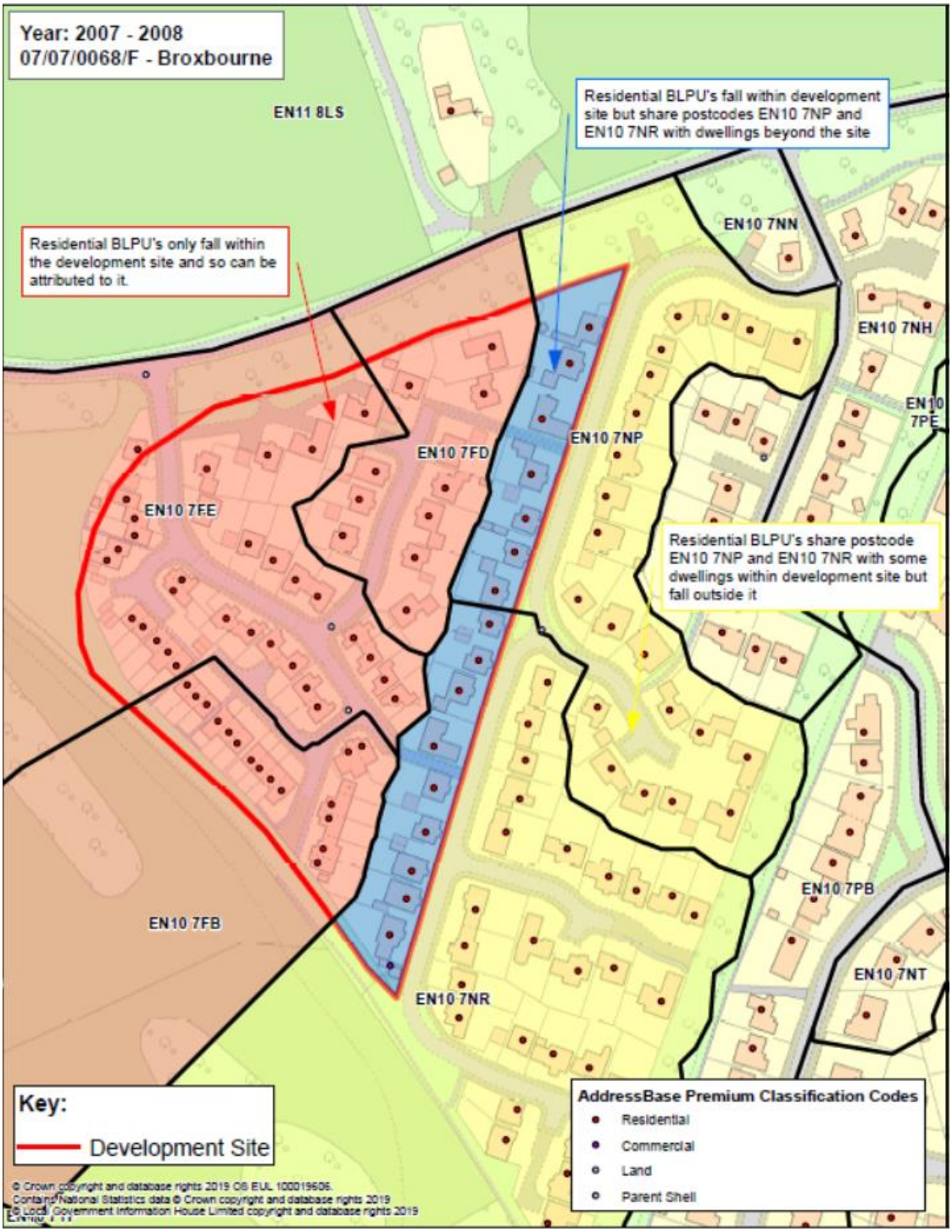


Figure 7. Scenario 4 for determining development coterminus postcodes.

The coterminus postcode files were of relevance to annual development cohorts between the periods 2002 and 2006 and, postcode level GP Registration data sets. In the former, School Census records between 2002 and 2006 required only the pupil postcode to be returned to the DfE, 2007 was the first year that individual pupil address was included as part of the return. Coterminus postcodes within developments, in conjunction with the address files, can be used to determine dwelling counts within the postcode by units only and type. School Census mainstream records associated with these postcodes can be divided by the number of dwellings to calculate yield per 100 dwellings rates. This method would be applied for coterminus postcodes within a specific development, or in aggregate for a cohort, to determine a statistically robust estimate of mainstream yield per 100 dwellings. Such “estimate” yield rates can be used to display patterns in pupil accumulation within developments, and overall annual cohorts, prior to the address specific counts associated with a whole development from 2007 onwards.

A similar premise exists for postcode-based GP registrations counts for children aged 0 to <7 years by individual year of age. Aggregation of coterminus postcodes within an annual cohort of developments permits a statistically robust sample of the included new build population for that year at County level. Individually addresses and associated residential dwelling types from AddressBase Premium allows the count of units only for calculation of rates per 100 dwellings. Where postcodes are wholly houses, or wholly flats, then this permits the estimate of yield per 100 dwellings specific to these dwelling types. Whilst this is included in the overall PYS methodology this element has yet to be progressed, in the majority, in the analytical stage (see Section 10).

8.2 Development address export files

PPREF specific address files determined by GIS processes were exported for matching to the annual cohort master files. The principle address included was usually the Delivery Point Address (DPA), this is used by Royal Mail to deliver packages. These addresses are very spatially accurate as they identify the exact location of the package destination. Until the spatiality of the DPA is concretely determined, a Geographic Address can be used, Local Authorities use these when organising addresses. Geographic Addresses use a combination of Primary Addressable Objects (PAO’s) and Secondary Addressable Objects (SAO’s) to provide either a preliminary address or, a description of an address. Generally, UPRN’s with PAO’s are more accurate than UPRN’s with SAO’s, the latter commonly describe plots of land, such as Plot 238. In most cases the Primary Addressable Object progresses to become the Delivery Point Address. Figure 8 is an extract of a permission in the 2018_2019 cohort which is still under construction.

There are no Delivery Point Addresses for the site, which is expected given the spatiality of the dwellings is not yet determined. However, all the UPRN’s within the site have a PAO address such as, for example, No: X Thorpe Road, Bishops Stortford, Hertfordshire. The AddressBase Premium technical specification recommends using Delivery Point Addresses first and, if these are not available, gaps should be completed with the PAO/SAO. The trade-off however is that Geographic Addresses are less spatially authoritative as Delivery Point Addresses.

In most instances completed development polygons within the Pupil Yield Study were Delivery Point Addresses due to the greater level of spatial accuracy.

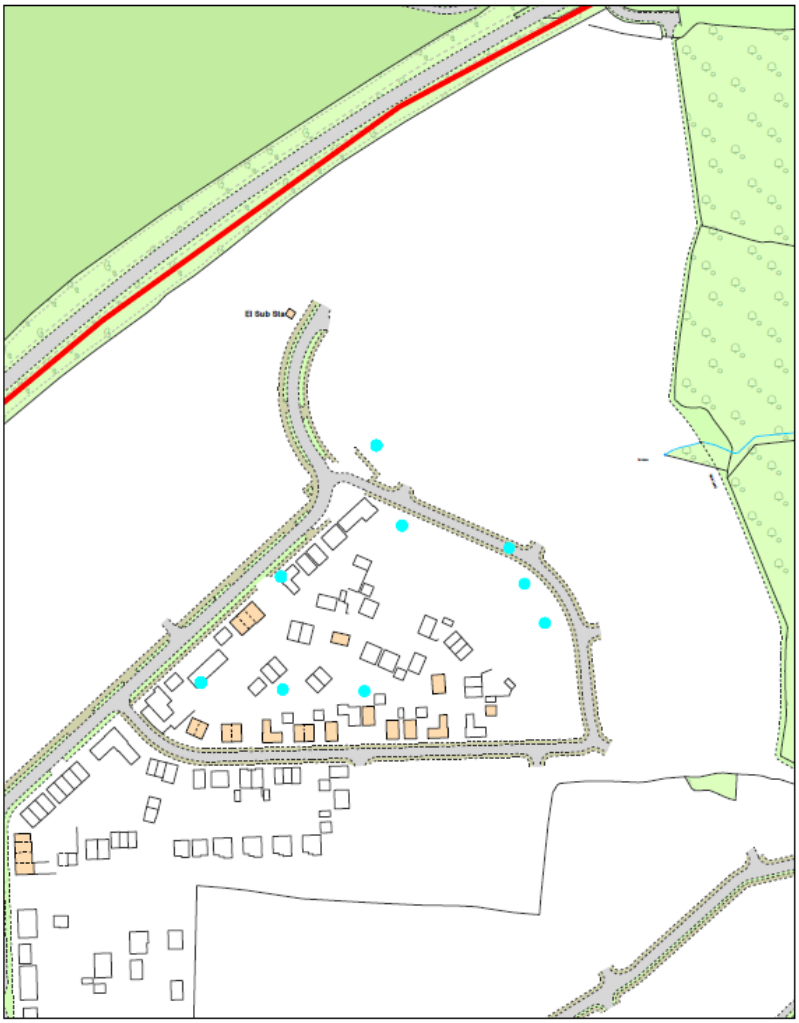


Figure 8. Example map of a development currently under construction with the level of detail in such instances displayed.

The following data fields were included in the finalised address file for each annual cohort:

- Inclusion Year – the financial year to which the PPREF relates
- LAD – Local Authority District Name
- LAD CD - Local Authority District Code (Office for National Statistics)
- PPREF – Unique Planning Permission Reference
- Unique Site Ref - The PPREF prefixed with the LAD CD
- UPRN – Unique Property Reference Number for each dwelling
- Parent UPRN – the parent UPRN for multiple dwellings such as the UPRN associated with a block of flats with each flat in the block also having its own UPRN
- Sub Building – Address data
- Building Name – Address data
- Building Number – Address data
- Thoroughfare – Address data

- Post Town – Address data
- Postcode – Address data
- Classification – RD02, RD03, RD04 & RD06 in the majority of instances
- Class Scheme - AddressBase Premium Classification Scheme
- Class Description – Detached, Semi-Detached, Terraced, Self-Contained Flat
- Primary Description - Residential
- Secondary Description - Dwelling
- Dwelling Classification Type – House_Bungalow – Detached, House_Bungalow – Semi-Detached, House_Bungalow – Terraced, Self-Contained Flat_Apartment_Maisonette.

Pivot table analysis of the address cohort classification determined dwelling counts by units only and type for each unique PPREF. These were cross referenced to the homogenous total permitted dwelling counts arising from the overall permissions, residential completions and size_type data within the cohort master file. This ensured that the GIS determined number of dwellings by type matched that of the SMART Herts data sets. Developments with variance in dwelling counts between addresses “on the ground” versus total completions were subject to further investigation. Most instances where this occurred resulted from an additional permission coming forward for a site. These additional permissions were identified through address and description field searches of the residential completions master data file 2002 through to 2020. Additional permissions were joined to the original permission and overall permission, residential completions and size_type data files merged into one record. The site unique identifier was amended to a concatenate of both permissions which served as an identifier for developments which were merged, for example, [Site 1 ID] & [Site 2 ID]. The analysis process was then repeated to ensure resulting counts tallied between GIS and SMART Herts data sets.

To the annual cohort GIS master address files, a “Concatenate Address” was produced within a single field, this was derived from completed cells within the associated address fields such as building number, thoroughfare, post town and postcode. The single field dwelling address were applied in looking up cleansed school census record addresses to return the master address file UPRN. The UPRN within both files provides a unique identifier for determining specific dwelling mainstream sector counts over time. Additional data from the GIS Address files was also cross referenced to the master data files for each annual cohort derived from the overall permissions, residential completions and size_type data files. These additional data fields were:

- ADPremium Class No: Flats – A count of flats for the unique PPREF observed within the address master file.
- ADPremium Class No: Houses - A count of houses for the unique PPREF observed within the address master file.
- ADPremium Class No: Total - A count of all dwellings (units only) for the unique PPREF observed within the address master file.
- Check ADP Total Flats to Size_Type Data – a Yes/No flag to indicate whether the GIS residential dwelling count of Flats matched the count from the size_type SMART Herts data files for the unique PPREF.

- Check ADP Total Houses to Size_Type Data – a Yes/No flag to indicate whether the GIS residential dwelling count of Houses matched the count from the size_type SMART Herts data files for the unique PPREF.
- Check ADP Total to Total Size_Type Data – a Yes/No flag to indicate whether the GIS residential dwelling count of dwelling units matched the count from the size_type SMART Herts data files for the unique PPREF.
- Development Mix – The percentage contribution of flats and houses to the overall number of units completed in the unique PPREF.
- Dominant Type – The dominant dwelling type associated with the unique PPREF. Where a percentage contribution was >=60% then this became the dominant type (either Houses or Flats). Where the type was within the range 40% to 60% then the dominant type was “Mixed”.

Where GIS address residential classifications matched SMART Herts size-type dwelling type counts then cross reference of the permission tenure and bed size data by type permitted, some instances the determination of individual dwelling bed size and tenure. Where this occurred the UPRN specific bed size and tenure was recorded against the master address record. This was normally the case for developments in the >=10 to <30 dwellings cohorts. For example, a development of 25 flats for which the size-type data indicates all are Open Market and 2-bed dwellings will have “2-Bed Open Market” recorded against each UPRN in the address master file. A proportion of the dwellings included within the overall Pupil Yield Study therefore already have bed size and tenure data associated with UPRNs, the remainder will be determined once all cohorts are finalised.

Finally, postcodes were extracted from the GIS annual cohort master address files and duplicates removed. The resulting list provided the annual cohort postcodes which would be applied to extract School Census records from the relevant databases since annual cohort commencement. For example, the 2007_2008 development cohort would result in individual pupil annual School Census records being extracted, based on a match between GIS master file and pupil address postcode, from 2007 through to 2020.

9.0 The schools census mainstream and special school pupil records

There are three census returns from schools each year termed the Spring, Summer and Autumn returns, the focus of the Pupil Yield Study was the January or Spring return each year 20002 to 2020. Historically this census return has been the dominant one with consideration of the Annual Schools Census and PLASC. Within state schools every pupil has an allocated Unique Pupil Number (UPN) which is retained each school year and acts as a longitudinal unique tracker across a pupils schooling, it is a mandatory field within each Schools Census return**.** Each Schools Census return contains pupil level data for pupils both on and off roll as at the date of the census (Nursery being the exception). Pupil Enrolment Status is however only required for those pupils whom are on roll at the school census date. A count on census day therefore includes all pupils whom are on roll as at census date and whose enrolment status is:

- ‘C’ (current - single registration at this school) [ALL schools]
- ‘M’ (current main - dual registration) [ALL schools]
- S’ (current subsidiary - dual registration) [ALL schools]

- ‘F’ (FE college) where a pupil is registered with the PRU / AP but is taught for most of their teaching time at the FE college [For: PRU / AP only]
- ‘O’ (other provider) where pupil is registered with the PRU / AP but is taught for most of their teaching time by the other alternative provision provider (which is not a school) [For: PRU / AP only]

The count excludes any pupil whose enrolment status is ‘G’/Guest - pupil not registered at the school but attending some sessions or lessons. It therefore follows that Pupil Date of Entry is provided for all pupils both on and off roll as at census date (on roll only for designated Nursery schools) whilst Pupil Date of Leaving will only be provided for those pupils with no enrolment status.

Anonymised child level data from all schools, with no pupil duplication from dual registration codes, as at the schools census date can therefore be extracted on the basis of the pupil UPN and a Pupil enrolment Status like ‘C’ or like ‘M’, where the Pupil Date of Leaving is ‘Null’. Each annual cohort of postcodes defined in Section 7 was imported to each January school census database 2002 to 2020 to provide the reference postcode from which unique anonymised pupil data could be matched and extracted. Development postcode areas are used to refine the number of schools census records extracted for analysis, this both saves analyst time and is also in accordance with the project Information Asset Registration (IAR) entry and requirements of GDPR.

Whilst the main body of data collected via the School’s Census is robust and validated with inbuilt DfE checking processes there remains a known issue with address information. A proportion of HCC schools use British Standard (BS7666) address fields within their Schools Information Management Systems (SIMS) whilst others apply free text Address field 1 through to 5. The DfE does not use address information collected via the Schools Census and therefore has no requirement for a standardised approach as to how address information is stored.

In all instances each pupil has a home address postcode however the quality of associated address field information within the remainder of each pupil’s address is variable. For statistical reporting this is normally not an issue as pupil postcode is assigned to a Census Output Area (OA) based on Office for National Statistics “NSPL Postcode to OA” lookup files. These lookup files relate the centroid of a postcode area to an Output Area, OAs can then be aggregated to larger often bespoke geographies for reporting purposes and is the method recommended by the National Statistics Authority (NSA). Further information on this method is supplied in Appendix 4.

Whilst sub-postcode address quality issues have been known for several years the fact that DfE does not require this data and, the large cost of implementing a standardised address system across all schools versus other HCC priorities, suggests that this will be an ongoing issue. The query structures established within each School Census database to extract unique anonymised pupil records with their associated single row address were subsequently complex and provided in detail within Appendix 5. Resulting extract tables based on matches to annual cohort development cohort postcode records had the following formats for both mainstream and special school cohorts.

- *School Phase* – Nursery, Primary, Middle-Deemed Primary etc.
- *School No:* - The DfE allocated number for the school.
- *School Address* – The school address details including postcode.
- *NCYearActual* - The National Curriculum Year Group in which a pupil is taught for the majority of their time irrespective of their chronological age.
- *Special Educational Need* – Coded as 'E' (Education, Health and Care plan) or 'K' (SEN support) or 'N' (No SEN Support). The Children and Families Act 2014 replaced Statements of Educational Need (SEN statements) with 'Education, Health and Care plans' (EHC plans).
- Pupil Address Type – The pupil address is extracted based on 'C' or 'Current'.
- *Home Address Postcode* - The postcode, mandatory for both BS7666 and address line format, is allocated by the post office to identify a group of postal delivery points. Note that there may be two or more current address for children with divorced/separated parents/in care, in this instance the first address is taken based on the minimum address ID.
- *BS7666 format: SAON* - The Secondary Addressable Object Name (SAON), refers to the flat, apartment name, number, or other sub-division of a dwelling.
- *BS7666 format: PAON* - The primary addressable object name (PAON), refers to the dwelling name and / or number.
- *BS7666 format: Street* - The street name / description.
- *BS7666 format: Locality* - The locality name refers to a neighbourhood, suburb, district, village, estate, settlement, or parish that may form part of a town, or stands in its own right within the context of an administrative area.
- *BS7666 format: Town* - The town name refers to: A city or town that is not an administrative area; A suburb of an administrative area that does not form part of another town or; A London district.
- *BS7666 format: Administrative area* - A geographic area that may be the highest-level local administrative area for example county or a unitary authority.
- *BS7666 format: Post town* - Assigned by the post office, based on the area sorting office.
- *Address line format: line 1* - First line of the address.
- *Address line format: line 2* - Second line of the address.
- *Address line format: line 3* - Third line of the address.
- *Address line format: line 4* - Fourth line of address.
- *Address line format: line 5* - Fifth line of the address.

With respect to Address Fields, returned information was dependent upon whether the schools Information Management System utilises BS7666 or an address line format and therefore all fields were included in order to enable geocoding. Where a child had multiple addresses, such as where a child lives with both parents at different stages of the week, the first address was extracted on the basis of the minimum Address ID. Note that where the DfE had made amendments to codes, or descriptions, since 2002 then the relevant codes were applied as current to the specific census date.

9.1 Cleansing school census address records

Data extract files from each January School Census relevant to the development annual cohort under consideration were appended to a singular data file for that cohort. Mainstream and special school data files were treated separately. A “Year” flag was added to each annual School Census data extract to relate from which census the data was obtained. For example, the 2005_2006 development cohort had a singular workbook titled “2005_2006 SC Data 2005 to 2020” within which were two worksheets “SC Mainstream Raw 2005_2006” and “SC Special Raw 2005_2006”. Within each sheet the extract table from each census was pasted with the relevant year for each census added. Following completion of the extracts for a development postcode cohort the raw data master sheets were copied to create master sheets from which address cleansing could process.

The initial step undertaken was to resolve the Address Line Format 1 through to 5 addresses into BS7666 format. This was processed through cutting the relevant Address Line fields and pasting into the relevant BS7666 fields. Generally, the number of non-BS7666 records in each annual extract is small and this process is not overly resource intensive. Following this the Address Line Format 1 through to 5 fields were deleted to reduce the number of data columns. The BS7666 Administrative Area was also deleted as all postcode extracted records were coterminus to Hertfordshire and it is therefore superfluous. At this point the master data file (for mainstream and special school pupils separately) contained the following fields:

- *School Phase* – Nursery, Primary, Middle-Deemed Primary etc.
- *School No:* - The DfE allocated number for the school.
- *School Address* – The school address details including postcode.
- *NCYearActual* - The National Curriculum Year Group in which a pupil is taught for the majority of their time irrespective of their chronological age.
- *Special Educational Need* – Coded as ‘E’ (Education, Health and Care plan) or ‘K’ (SEN support) or ‘N’ (No SEN Support). The Children and Families Act 2014 replaced Statements of Educational Need (SEN statements) with ‘Education, Health and Care plans’ (EHC plans).
- Pupil Address Type – The pupil address is extracted based on ‘C’ or ‘Current’.
- *Home Address Postcode* - The postcode, mandatory for both BS7666 and address line format, is allocated by the post office to identify a group of postal delivery points. Note that there may be two or more current address for children with divorced/separated parents/in care, in this instance the first address is taken based on the minimum address ID.
- *BS7666 format: SAON* - The Secondary Addressable Object Name (SAON), refers to the flat, apartment name, number, or other sub-division of a dwelling.
- *BS7666 format: PAON* - The primary addressable object name (PAON), refers to the dwelling name and / or number.
- *BS7666 format: Street* - The street name / description.
- *BS7666 format: Locality* - The locality name refers to a neighbourhood, suburb, district, village, estate, settlement, or parish that may form part of a town, or stands in its own right within the context of an administrative area.

- *BS7666 format: Town* - The town name refers to: A city or town that is not an administrative area; A suburb of an administrative area that does not form part of another town or; A London district.

Within the PYS trial school census records were extracted based on the development postcodes and a 200m buffer which were then preliminary cleansed according to the above process. These records were then sent to GIS for passing reiteratively through LocatorHub Transformation Suite, an address-matching application integrated with ArcGIS. LocatorHub cleansed the addresses and defined a Unique Property Reference Number (UPRN) for most address records. Where a specific cohort of mainstream pupils had their addresses cleansed then a combination of UPN and Postcode of a cleansed address joined to other years of records, which have a match on both the concatenate and the underlying original uncleansed address, was observed to significantly speed up the address cleansing process.

Any School Census records which contained insufficient address details to geolocate was checked against any corresponding UPN within the dataset to determine whether it is possible to harvest additional address details from other years census returns. This was required in order to achieve a sufficiently complete address to permit an address match and establish the UPRN and coordinates. Where this was not possible the pupil postcode itself was be analysed to establish if it was feasible to match either to a postcode centroid (by reference to OS Code-Point) or if the postcode is associated with a single structure, such as a block of flats, to an individual building and to a 'Parent UPRN' of that building within a development's boundary. Records that were geocoded to postcode centroid only were retained if all the delivery points were contained within a development polygon and excluded if they were not. In the latter instance it was not possible to prove a mainstream child is located within a specific development "beyond all reasonable doubt" and therefore such records should not be included. Individual pupil records which are located to a development of the basis of a building "Parent UPRN" were flagged as such although these were very small in number.

However, during the trial it was observed that there were several issues associated with this preliminary method:

- LocatorHub processing was heavily dependent on GIS resource which is finite within the organisation and substantial delays could occur.
- It was rare for an address record related to a development to occur in the 200m buffer around a permission, in such instances this related to an incorrect postcode. Removal of the buffer and associated records made no difference to observed yield rates per 100 dwellings in the development outputs nor in the calculated development average yield overall.
- A much more efficient process was not to match school census records to development addresses but rather to match exact spatially defined development addresses to school census extracts exported based on the development specific postcodes only (no buffer applied). This reduced the scale of record cleansing from the tens of thousands per cohort to thousands.

The methodology was therefore refined for development cohorts processed following the trial. The development school census data table, following removal of the

Address Line Format 1 to 5 and Administrative Area fields, was sorted according to the home postcode and Unique Pupil Number (UPN). Whilst some census returns may contain poor quality address data for a particular UPN, other years were BS7666 exact matches and could be replicated. The master address file containing the spatially exact addresses associated with the annual cohort was opened and sorted according to postcode and concatenate address. Visual cross comparison of a postcode between the school census and address master file determined the format of addresses within that area.

This format was replicated within the School Census record extracts for the specific postcode but using the relevant UPN data. For example, the master address file may list a dwelling as: Flat 5, The Dakota Complex, 4 Piggots Lane, Hemel Hempstead whilst the school census data could be: 4 Flat 5, Piggots Lane, Dakota Complex, Hemel. Effectively the School Census address extracts were cleansed to a consistent format in line with the most spatially exacts DPAs within the master address files. Following this a concatenate address was produced for each school census record, this concatenate address was a unique identifier which was looked up against the cohort master address file. Where a match occurred then the Unique Property Reference Number (UPRN) from the master address file was returned against the pupil address record. Where no match occurred then the UPRN against the pupil address was labelled as “Not In Developments”. The pupil records with a returned UPRN of “Not In Developments” were visually compared to the master address records to ensure that they should be excluded.

Whilst, at face value, this process seems resource intensive it requires no training in LocatorHub, nor licence, and provides a faster turnaround in UPRN matched datasets (it is feasible to cleanse 12,000+ records a day by one person). Comparison of the applied method versus LocatorHub cleansed datasets determined no difference in the count of pupil records allocated to an example annual cohort. In some instances, the quality of address information resulting from LocatorHub cleansing was poorer than that of the method applied. This occurred as the method applied herein compares school census records to the most spatially exact and refined addresses determined from GIS analysis for the cohort of interest. A further step undertaken was to create an additional field for “Sector”. The NCYearActual text value for each pupil record was transformed to an education sector category, the values applied are shown in Table 2.

Table 2. The School Census National Curriculum Year Group Code and returned Education Sector.

NC YEAR GROUP	*EDUCATION SECTOR*
N1	Nursery (N1)
N2	Nursery (N2)
R	Primary
1	Primary
2	Primary
3	Primary
4	Primary
5	Primary
6	Primary

7	Secondary
8	Secondary
9	Secondary
10	Secondary
11	Secondary
12	Post-16
13	Post-16
14	Post-16
X	X

The resulting data table, for mainstream and special schools separately, provided a list of all pupils in all School Census returns since the annual development cohort under construction started producing residential completions through to 2020. However, the table includes both UPRN matched records and those which were not matched to the development cohort, it also includes singular UPNs across multiple census years. It was therefore required to create education sector counts by UPRN for each census year.

9.2 Education sector counts by development UPRN

The address cleansed and UPRN matched pupil level data table was pivoted with the following fields included:

- Year – The census year from which a group a pupil records were extracted, applied as a filter.
- UPRN – The Unique Property Reference Number, applied as a row.
- Sector – The Education Sector to which the pupil record is allocated as at the census year. This was applied as both the count and the column header.

Each year was individually selected using the filter and the resulting UPRN list with sector counts copied and pasted into a new workbook titled, for example, “2015_2016 SC Data Sector Counts 2015 to 2020”. Table 3 displays an example output for an individual census year, the specific UPRNs have been replaced however the principle is the same. Outputs were created in a standard format of N2, Primary, Secondary and, Post-16 versus UPRN. The workbook contained a separate worksheet for each school census year for which extracts were made, for example the 2015_2016 cohort had UPRN based sector counts for 2015,2016, 2017, 2018, 2019 and 2020 as separate worksheets. The process was replicated for longitudinal Special school pupil counts by sector.

Table 3. Example standard format list of UPRN versus education sector mainstream pupil counts.

UPRN	Nursery (N2)	Primary	Secondary	Post-16
A	0	1	0	0
B	0	1	0	0
C	0	1	0	1
D	0	1	0	0
E	0	2	0	0
F	2	0	0	0
G	1	1	1	0

H	0	1	0	0
I	0	1	0	0
J	0	1	1	0
K	0	1	0	1

10.0 Processing births data

The method applied herein was used within the PYS trial however it has yet to be run for the 2011 to 2018 annual development cohorts to which the births data relates. This will occur following PYS cohort finalisation and establishing the priority normalised mainstream yields associated with each annual cohort. It is included within the PYS methodology documentation to ensure that coverage is comprehensive.

Live births by financial year (2010/11, 2011/12, 2012/13, 2013/14, 2014/15, 2015/16, 2016/17 and, 2017/18) were selected using the field [DOB] Date of birth with an applied county code of usual residence of mother of child, [COUNTY_MOTHER], being Hertfordshire. The selected records had an applied "Financial Year" and "Month" identifier, for example "2012/13 April" for a live birth occurring 16th April 2012, "2017/18 Sept" for a live birth occurring 18th September 2017 and so forth. This information was used to assist in aiding the identification of completion dates for developments within the specific financial year in additional to temporal birth counts. Based on the applied selection criteria the following data fields were identified as being required for extraction, the fields were identified using the DSA between Public Health and NHS Digital:

- *[ADDR_MOTHER]* - Usual address of mother
- *[PCODE_MOTHER]* - Postcode of usual residence of mother
- *[PCODE_IMP_IND]* - Postcode imputation indicator
- *[CNTY_DIST_MOTHER]* - County district code of usual residence of mother

An aggregate count of the number of births by presence/absence of *[PCODE_MOTHER]* determined the percentage of births records for which postcode information was absent as a quality assurance measure for all financial years. The postcode of usual residence of mother was duplicated as an additional field and, within this duplicated field, any text spaces were removed. This field was labelled as *[PCODE_MOTHER_NoSpace]* and used to match to the list of development postcodes determined in Section 7. Applying the known development postcode overlays reduced the number of records which were required to be extracted from the births database. This was both in requirement of the DSA stipulation that extracts for analysis should be proportional to the scope of a project and, also informed the likely scope of this part of the project in future repeats of the process. County total numbers of live births for the financial periods were provided separately. The individual record extracts for the financial periods which were specific to the finalised development polygons included the following output fields:

- *[UNIQUE ID]* – A created ID which is unique for each record
- *[YEAR]* – Financial and either 2012 or 2013
- *[MONTH]* – Calendar month within which the birth occurred
- *[ADDR_MOTHER]* - Usual address of mother

- *[PCODE_MOTHER]* - Postcode of usual residence of mother
- *[PCODE_IMP_IND]* - Postcode imputation indicator
- *[CNTY_DIST_MOTHER]* - County district code of usual residence of mother

All other identifiers associated with the birth data extract were removed, data was exported to a Microsoft Excel 2010 format file, password protected and passed to the Community Intelligence & Data Science team (CIDS). The data file was both transferred and added to the restricted access project folder in accordance with protocol determined in the Data Protection Impact Assessment. The workbook password was provided separately via internal phone call. Colleagues in Public Health had informed that the births address field was a single column with no separate identifiers such as specified within BS7666 or Address Line 1-5 format, for example such as applied in the Schools Census. Addresses therefore required validation and geolocating using “Locator Hub” in order to identify the Unique Property Reference Number (UPRN) required to match to the finalised development polygon identified address UPRNs.

Following address cleansing, and addition of UPRN to the individual birth’s records, those records which were located outside of the development polygons were deleted based on linking *[Unique ID]* between the GIS dataset and the excel data file. Birth counts by unique development site ID were produced to examine the frequency distribution, by development size (number of dwellings) band, of aggregates for Statistical Closure Control (SDC) in accordance with the DSA. The number of births per 100 dwellings were calculated as: *The number of births in a development polygon / Total number of dwellings in a development polygon*. Frequency distributions were passed to Public Health for a determination of whether appropriate standards had been met at individual development level. The outcomes of the SDC process determined the geographical scale of the analysis of births arising from completed developments.

Following completion of the births aggregation process, and inspection by Public Health colleagues, permission was sought from the working group to delete the individual births records files both as held within GIS software and data files in the project folder. This was in accordance with procedure detailed in the DPIA.

11.0 GP Registrations data and coterminous postcodes

The method applied herein was used within the PYS trial however it has yet to be applied for the majority of the annual development cohorts 2003 to 2020. This will occur following PYS cohort finalisation and establishing the priority normalised mainstream yields associated with each annual cohort. GP Registrations data processing is included within the PYS methodology documentation to ensure that coverage is comprehensive.

The authority produces a School Place Planning Forecast, part of the data which underpins the DfE required, and accepted, forecast is GP registrations data for children aged 0 to <7 years, by individual year, by anonymised counts to postcode area. The Pupil Yield Study will cross match postcode sector counts of children aged 0 to 5 years, and individual year variants into aggregate outputs such as Age 0 to <3 years, to identified development co-terminus postcodes to produce an annual county

wide sample-based assessment of yields in the early years from new build developments. The use of this data for estimating yields from new build developments is in accordance with the entry held within the Information Asset Register.

The use of postcode small area geographies permits the determination of early years yields by new build dwelling at Units Only and Type distinction although, to date, much of this work has been suspended with prioritisation of the mainstream yields study. Further work is required to determine whether bed size and tenure distinctions can be determined dependent on the proposed DfE methodology once it is released. These assessments will be also be useful in the longer term for the accurate location of localised early years services and childcare provision. The process broadly follows that outlined below, applied to the 2002 annual development cohort.

11.1 Determination of coterminous postcodes

Development postcode buffers contain the postcodes associated with a specific permission and those which occur in a 200m range of the boundary. In creating permission specific buffers, postcode layers were superimposed over the development boundaries, where a postcode polygon intersected/overlay a boundary, or was within 200m, then the postcode was extracted. The relevant PPREF was assigned to each postcode for all developments in each annual cohort. To determine whether postcodes were wholly coterminous within development sites a combination of Codepoint Polygons and AddressBase Premium records were applied.

Using both datasets a postcode was deemed coterminous with a development site if it contained residential addresses (determined using BLPU classes within AddressBase Premium) which fell within the site boundary but no residential dwellings beyond the site. Under this methodology a ‘theoretical’ postcode (as defined using Codepoint Polygons) which overlaps a development may go beyond the development significantly, but if there are no other residential dwellings aside from the ones within the site then it can be inferred that all addresses with such postcodes are attributable solely to that development site. Reference should be made to Section 7.1 which outlines the four possible scenarios when determining coterminous postcodes using this methodology.

Figure 9 below displays an example development specific map for a permission within the 2002_2003 annual cohort, the buffer postcodes within a 200m radius of the site are indicated. The coterminus postcode cohort for each development is a subgroup of the buffer postcodes. The coterminus postcodes determined for each development can be aggregated to form a larger, more statistically robust, cohort for all sites included within a specific annual cohort. Coterminus postcodes were more often associated with the larger development sites >=30 dwellings in size, this occurred as, on average, a postcode contains 30 households, or delivery points. Small development sites are therefore most often part of a postcode area as opposed to wholly contained within the site, the exception occurring when the development is close to 30 dwellings in size or, where a postcode area is redrawn by Royal Mail.

11.2 Example of GP and mainstream coterminous postcodes data analysis

The analysis of GP, and mainstream School Census, coterminus postcodes data sets to estimate yields per 100 dwellings, by units only and type, is as outlined within Section 12. The matching of School Census sector counts to individual dwelling permits the extraction of per annum sector counts by postcode. GP registrations data sets are already at postcode level and counts by year group can be cross referenced to included coterminus postcodes.

Figure 9. A development within the 2002_2003 annual cohort with the postcode overlays displayed and the 200m buffer zone shown. The coterminus postcode cohort for each development is a subgroup of the buffer postcodes.

In total 27 developments were included within the 2002_2003 annual cohort for larger sites >=30 dwellings in size. These permissions contained 2,317 dwellings of which the unique addresses were contained within 83 postcodes areas. Nine of the developments contained no coterminus postcodes at all whilst the remaining 18 permissions had 97.5% to 100% of their included dwellings contained in such areas. In total 16 developments had 100% of their dwellings contained within coterminus postcode areas. Overall 1,599 (69%) of the dwellings built were finally contained within 52 coterminus postcode areas in consideration of Units Only. In consideration of dwelling Type, it was observed that 456 flats were contained in 13 coterminus postcodes wherein all units were flats, this represented 50.7% of the total 915 flats included within this annual cohort. Of the 1,402 houses associated with the overall cohort 730 were included in 28 coterminus postcodes wherein the sole dwelling type was houses, this represented 52.1% of the total houses count.

The proportional representation of coterminus postcode cohorts at dwelling units and dwelling type distinction was observed, in this instance, to be robust at 69%, 50.7% and 52.1% respectively. Where coterminus postcode included dwelling count representations, relative to the overall annual cohort sizes, are high then they can be used to undertake robust assessments of:

- How the overall cohort would be expected to behave at Units Only and dwelling Type distinctions for normalised yield per 100 dwellings rates for Early Years cohorts using GP Registrations datasets;
- Provide an estimate of mainstream yields within the years 2002 to 2006 based on postcode level data as individual pupil address information was not required by the DfE, within the School Census returns, until 2007 onwards.

Figure 10 displays a comparison of normalised primary mainstream yields estimated from houses only within the 2002_2003 annual cohort coterminus postcode cohort (n = 730), in comparison to 2007 onwards from the whole cohort individual dwelling address (n = 1,402). It can be observed that the co-terminus postcode data provides a robust measure of yields which would have been observed from the overall houses cohort if individual dwelling address had been included in the School Census returns 2002 through to 2006.

However, there is likely to be a limitation to this position in that the initial years associated with coterminus postcode normalised yields per 100 dwellings rates might be an under-representation of that which occurs in the overall cohort. This results from application of different trajectory annual dwelling count applied as the denominator in calculating such rates. Whilst SMART Herts data sets permit the determination of annual completions associated with each permission, or in aggregate for an overall cohort, disaggregation to postcode level is not possible. In calculating the normalised coterminus postcode rates the overall dwelling count associated with such postcodes is applied. It is likely that not all dwellings within a postcode, and certainly not all postcodes, would have been completed in an annual period until the overall development, or annual cohort, trajectory is fully completed.

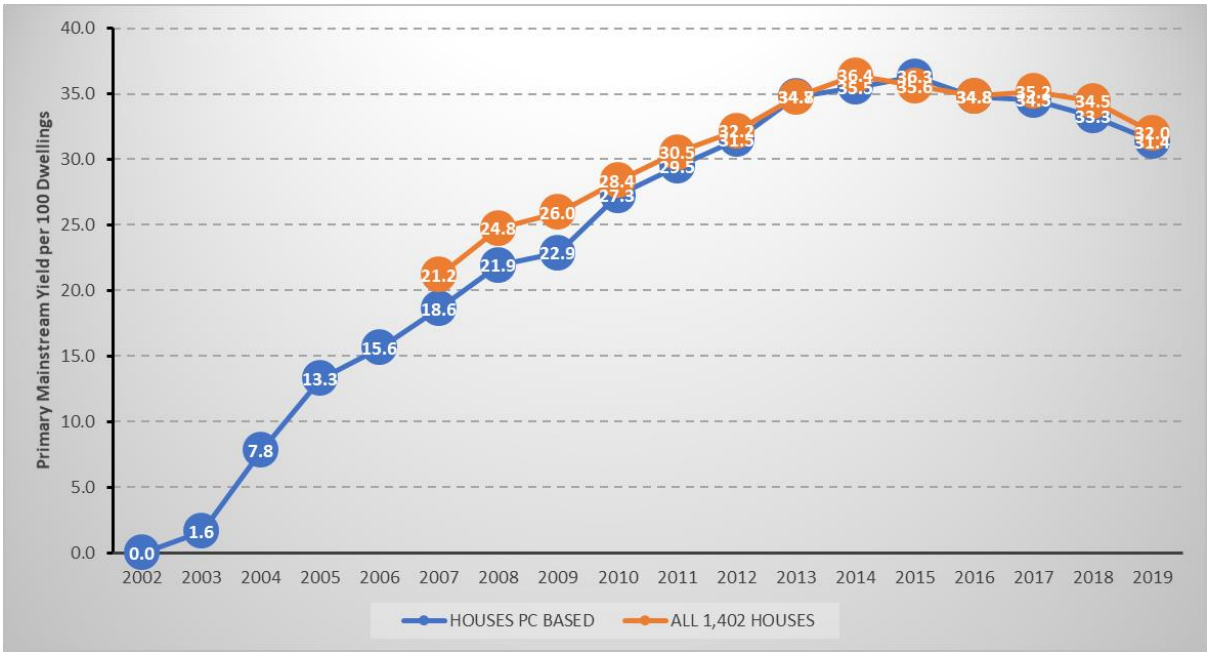


Figure 10. The mainstream primary yield per 100 dwellings for Houses Only from the 2002_2003 annual cohort based on (1) Estimates from the 730 houses within the 28 coterminus postcode areas 2002 to 2019 and, (2) All 1,402 houses from 2007 to 2019.

This increases the size of the denominator applied to aggregate coterminus postcode analysis resulting in smaller yield rates than would otherwise be realised if postcode level trajectory completions data were available. This will not be of relevance following the point where it is determined that the trajectory has completed, either for a single development under consideration or, for an overall annual cohort. This was investigated further using the PYS trial datasets to determine the extent of possible limitations to the method applied.

In total 4,557 dwellings were contained in the 41 developments included within the PYS trial, of these dwellings 3,215 or 70.6% were determined to be contained in coterminus postcodes. Overall 945 flats were contained in a coterminus postcode wherein this was the sole dwelling type which represented 44.6% of the trial 2,117 included flats cohort. Some 835 houses were contained in a coterminus postcode wherein this was the sole dwelling type which represented 34.2% of the trial 2,440 houses cohort. Figure 11 provides a contrast between mainstream primary Houses Only yield between the overall house's cohort and that from the coterminus postcodes only. In this consideration the 2009 start year provides individual dwelling School Census data from commencement (it being later than 2007) and permits comparison to postcode level analysis.

Figure 11. Comparison between the PYS All Houses (n = 2,440) normalised primary mainstream yield per 100 dwellings to that estimated using coterminus postcode data (houses only n = 835).

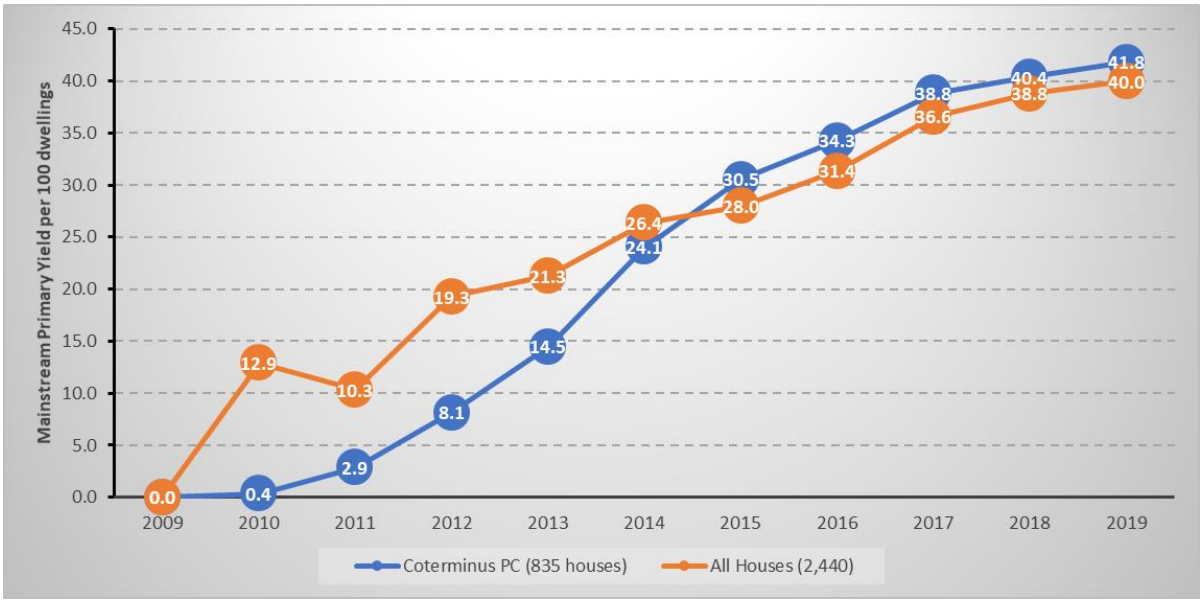
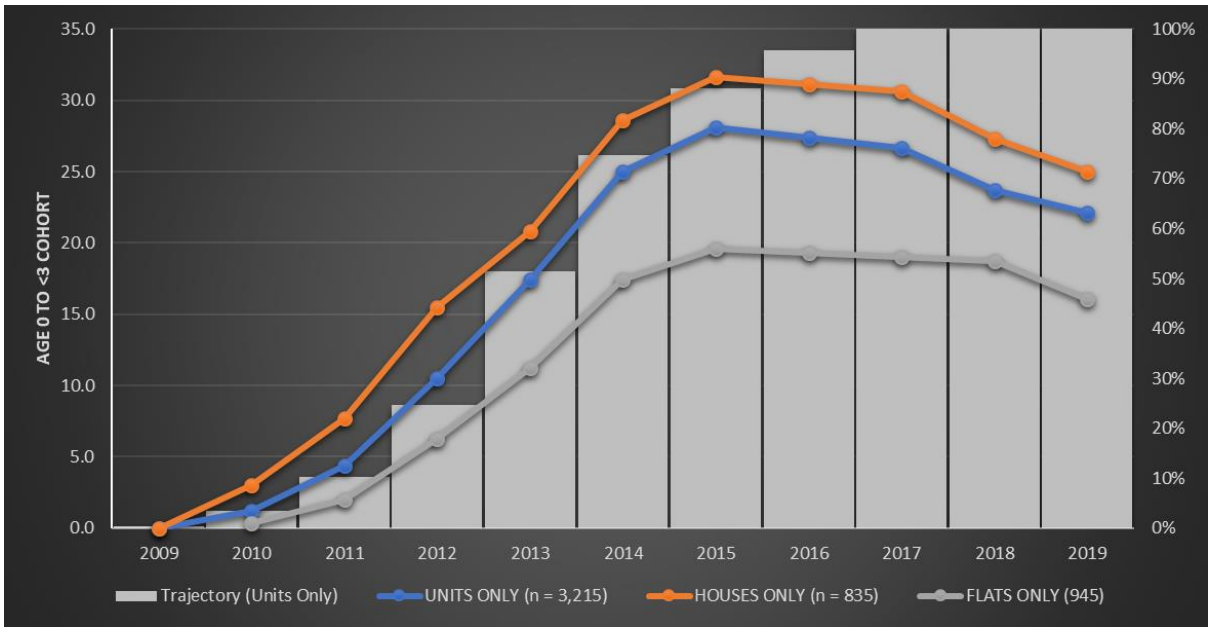


Figure 12. Age 0 to <3 years estimated yield per 100 dwellings (Units Only, Houses Only, Flats Only) using the coterminus postcode method applied to the PYS trial cohort of 41 developments.



Between 2010 and 2013 it can be observed that there are substantial differences in normalised primary yields between all houses in the cohort and those contained in only coterminus postcodes despite that latter being a large sample size. From 2014 whilst there are some differences in yields, they are broadly comparable, as is the rate of accumulation, and therefore a reliable estimate. It was determined that coterminus postcode yields generally only reflect that of the overall cohort when

approximately >85% of the total trajectory has been completed. At such a point differences in trajectory dwelling counts, applied as denominators, in coterminus postcodes versus that overall is relatively low. This was consistent with Units Only and Type distinctions level of analysis considered to date. Observed yields prior to this point should be considered indicative only.

Overall this is not of significance unless subsequent yield analysis of cohorts, particularly in the early years indicates peak yields several years prior to, or before 85%, overall trajectory completion. GP Registrations data sets were applied to the PYS trial coterminus postcode cohort to determine if this would be a limiting factor and Figure 12 displays the results at Units Only, Houses Only and, Flats Only. It can be observed that Age 0 to <3 years cohorts tend to approach peak at circa 90% of completion of the cohort trajectory (Units Only trajectory displayed in Figure 12). Values at, or around peak, were approximated for circa 3 years prior to starting the transition to LTA.

It is likely that Early Years cohort normalised yield rates at this point would be reliable estimates of the overall dwelling cohort given that most of the trajectory has completed. However, the above is without consideration of development typology and further work will be undertaken once this metric can be considered.
Where it is identified that the number of dwellings, at Units Only or in Type distinctions, within coterminus postcodes is small relative to the respective total number of dwellings included in the annual cohort then estimates may not be robust. Comparison between mainstream coterminus postcode rates over the relevant period, to current time, against that derived from the overall annual cohort, will be undertaken to provide supporting evidence for closeness of match. Small population statistical sample size calculations may also be applied to determine confidence interval and confidence level associated with included coterminus postcode cohort sizes versus the overall annual cohort size. Such determinations would be applied at Units Only and Type levels of analysis.

It should be noted that the information presented herein relates to Gross Yields only with no account for local moves such as that required in order to determine net yields. Calculation of net yields will not be possible with GP Registrations aggregate postcode counts by year group as no previous and current dwelling address is provided from which this can be determined. At best, net yields within the Early Years can be estimated by applying observed local move rates from the mainstream sector as a proxy estimate.

12.0 Merging GIS/SMART Herts and school census data sets

GP Registrations and Births data sets are excluded, at current time, from compilation with the overall datasets derived from the PYS. This exclusion occurs as the former are specifically postcode area-based statistics which do not align well with UPRN based record merges and, in the latter case, this is very much an emerging position and the bulk of births data outside of the trial have yet to be sourced via Public Health. Data sets from GIS/SMART Herts and the School Census were merged to create five principle master files: (1) Master Development Cohort (2) Summary Address Outputs (3) Master Address (4) Trajectory (5) School Census UPRN Counts. These five files permitted not only the collation of development specific

information in a standard format for each annual cohort but also for all cohorts 2002 through to 2020 in entirety.

12.1 Master development cohort

This principally involved the merging of data as outlined in Section 6.2 and 6.3 using unique permission ID to create a singular row of data, in Excel, specific to developments included within the PYS or, for an annual cohort, and to which GIS permission specific dwelling counts by type were added. This singular row of planning system data per development contained the following fields:

- Report Year – the annual cohort year to which the permission related.
- LAD – Local Authority District Name.
- LAD CD - Local Authority District Code (Office for National Statistics).
- PPREF – Unique Planning Permission Reference.
- Unique Site Ref – The PPREF prefixed with the LAD CD.
- Address – the site address as recorded in the system capped at 250 characters.
- Description – a description of the site as recorded in the system capped at 250 characters.
- PDL – Previously Developed Land flag (Yes/No).
- Gross Comp in Period – the number of gross dwelling completions in the report year.
- Loss in Year – the number of dwelling losses in the report year.
- Net Comp in Period – the number of net dwelling completions in the report year.
- Total Proposed Gain Check – a formula to check the number of gross dwellings permitted in the overall permissions files versus the sum of the residential completions associated across one or more years for the development.
- Total Gross Completions – the sum of the gross residential completions observed for each specific development across one or more years specific to the PPREF.
- Match – Yes/No error function to determine whether the total proposed gain check equals the total gross residential completions data.
- C2 Development from SMART Herts Report – Yes/No flag as to whether a development was identified as C2 from the specific SMART Herts report.
- C4 Development from SMART Herts Report – Yes/No flag as to whether a development was identified as C4 from the specific SMART Herts report.
- C2 Development from Description – Yes/No flag as to whether a development was identified as C2 from the site-specific description field.
- C4 Development from Description – Yes/No flag as to whether a development was identified as C4 from the site-specific description field.
- Include/Exclude in cohort – Include/Exclude Flag wherein developments indicated as C2/C4 were excluded.
- Reason Exclude – Notes field detailing the reason a site has been excluded for future reference.
- Gross Completions in Year – The number of gross dwelling completions in the start year to which the development is assigned, and for each year onwards

until development completion. This formed the basis of the Units Only trajectory.

- Losses in Year – The number of dwellings losses in the start year to which the development is assigned, and for each year onwards until development completion.
- Net Completions in Year – The number of net dwelling completions in the start year (Gross Completions in Year – Losses in Year) to which the development is assigned, and for each year onwards until development completion.
- Actual number of dwellings – The number of dwellings associated with the specific development polygon, as per residential characteristic inclusion criteria, as determined by GIS analysis.
- Variance in Count – The difference between the Total Proposed Gain to Actual number of dwellings from GIS.
- Note on Variance – where variance in counts above is observed an indication of analysis as to why such variance may exist.
- Final Number of Dwellings – The final total number of dwellings built for the permission in consideration of the included data sets.
- ResLine Provider – The provider type for the development, or part of development to which the size_type data row relates. For example, Private, Housing Association, Local Authority, Unknown.
- Dwelling Type – House, Bungalow, Flat/Apartment/Maisonette.
- ResLine Tenure Type – Tenure of dwellings within the row of dwelling type data.
- Overall Units – Number of dwelling units associated with the tenure, dwelling type and provider row of the dwellings for all/part of the relevant permission.
- 1 Bed Units – Count of 1-bed dwelling units.
- 2 Bed Units – Count of 2-bed dwelling units.
- 3 Bed Units – Count of 3-bed dwelling units.
- 4+ Bed Units – Count of 4+ bed dwelling units.
- Overall Houses – Number of the overall number of houses completed.
- 1 Bed Houses – Count of 1-bed Houses.
- 2 Bed Houses – Count of 2-bed Houses.
- 3 Bed Houses – Count of 3-bed Houses.
- 4+ Bed Houses – Count of 4+ Houses.
- Overall Flats – Number of the overall number of Flats completed.
- 1 Bed Flats – Count of 1-bed Flats.
- 2 Bed Flats – Count of 2-bed Flats.
- 3 Bed Flats – Count of 3-bed Flats.
- 4+ Bed Flats – Count of 4+ Flats.
- ADPremium Class No: Flats – The count of residential classifications of dwelling type “Flats” as observed from AddressBase Premium for the specific polygon.
- ADPremium Class No: Houses – The count of residential classifications of dwelling type “Houses” as observed from AddressBase Premium for the specific polygon.
- ADPremium Class No: Total – The total count of residential classifications of dwelling type “Houses” and “Flats” as observed from AddressBase Premium for the specific polygon.

- Check ADP Total Flats to Permissions Completions – A flag check of whether the SMART Herts size_type permissions data for Flats type matched that of AddressBase Premium.
- Check ADP Total Houses to Permissions Completions – A flag check of whether the SMART Herts size_type permissions data for Houses type matched that of AddressBase Premium.
- Check ADP Total to Total Permissions Completions – A flag check of whether the SMART Herts size_type permissions data for Houses & Flats type matched that of AddressBase Premium.
- Development Mix – The proportional representation of Flats and Houses to the overall proposal specific mix.
- Dominant Type – The dominant dwelling type associated with the permission with the bands of >=60% Flats = FLATS, >=40% <=60% Flats = MIXED, >=40% <=60% Houses = MIXED and, >=60% Houses = HOUSES.

Where any flag check returned an error then further investigations were conducted to resolve discrepancies.

12.2 Summary address outputs

A single row of data per permission which summarises the information determined from GIS analysis of each development polygon using AddressBase Premium. The following fields were included:

- PPREF - Unique Planning Permission Reference.
- Total Dwellings - Based on the total number of UPRNs located to a polygon.
- Address Known – A count of dwellings for which the DPA, PAO or SAO was known – where a difference occurred to Total Dwellings then this was flagged for further investigation.
- Count Flats – A count of UPRNs wherein the residential classification was of type Flats: This data was linked into Section 11.1 above.
- Count Houses - A count of UPRNs wherein the residential classification was of type Houses sub-divided into Detached, Semi-detached, Terraced: This aggregate count was linked into Section 11.1 above.
- Other/Unknown – A count of UPRNs wherein the residential classification was RD (Residential Dwelling) but for which the specific classification for Flats/Houses was unknown.
- CHECK – A check that the count of flats and houses matched that of the total dwellings determined for the specific site polygon.
- No: Flats & No: Houses – The aggregate count of the number of flats and houses as determined from AddressBase Premium.
- Development Mix –The proportional representation of Flats and Houses to the overall polygon specific type mix: This proportion was linked into Section 11.1 above.
- Dominant Type - The dominant dwelling type associated with the permission with the bands of >=60% Flats = FLATS, >=40% <=60% Flats = MIXED, >=40% <=60% Houses = MIXED and, >=60% Houses = HOUSES: This was linked into Section 11.1 above.

- Status: A flag check as to whether the sub-parts of the data file equated to the total dwellings observed for a specific site polygon based on AddressBase Premium residential classification UPRN counts.

Where any flag check returned an error then further investigations were conducted to resolve discrepancies.

12.3 Master address file

This file contains the individual unit addresses by UPRN determined using GIS analysis of AddressBase Premium for each specific and unique development based on the supplied polygon, fields included were:

- Year – the annual cohort year to which the permission related.
- LAD - Local Authority District Name.
- PPREF – Unique Planning Permission Reference.
- UPRN – Unique Property Reference Number.
- Parent UPRN – The unique Parent UPRN assigned to the UPRN e.g. a block of flats would contain many unique UPRNs but all would assigned to an overall parent UPRN associated with the block.
- SubBuilding – An included AddressBase Premium address descriptor.
- BuildingName - An included AddressBase Premium address descriptor.
- BuildingNumber - An included AddressBase Premium address descriptor.
- Thoroughfare - An included AddressBase Premium address descriptor.
- PostTown - An included AddressBase Premium address descriptor.
- Postcode - An included AddressBase Premium address descriptor.
- PC_No_Space – The above postcode but with any spaces removed in order to create a specific string.
- AddressConcatenate – A concatenated address produced in standard format using the supplied AddressBase Premium address fields.
- UPRN for Matching – The UPRN associated with each unique dwelling and replicated adjacent to the AddressConcatenate field.
- Classification – The residential dwelling classification code.
- ClassScheme - AddressBase Premium Classification Scheme.
- ClassDescription – Self Contained Flat, Detached, Semi-Detached, Terraced.
- PrimaryDescription – Residential.
- SecondaryDescription – Dwelling.
- TertiaryDescription – Self Contained Flat, Detached, Semi-Detached, Terraced.
- DwellingClassificationType – “Flat: self-contained (includes maisonette apartment)”, “House_Bungalow: Detached”, “House_Bungalow: Semi-Detached”, “House_Bungalow: Terraced”.
- AggregateType – Flagged as either “FLAT” or “HOUSE”.
- NOTE – Any additional information relating to the specific unique address.

12.4 Trajectory

Trajectory at Units Only was derived from overall number of gross completions associated with each permission, over time, resulting from SMART Herts residential

completions data sets as presented in Section 11.1 above. The Units Only data was replicated into a separate, standard format, file with the following fields:

- PPREF - Unique Planning Permission Reference.
- Completions Year 1 – The year 2002.
- Completions Year 2 – The year 2003
- Completions Year X – Each subsequent year to current period.

Where a permission did not commence producing residential completions until a year after 2002 then each subsequent year, to assigned development reporting start year, had a value of N/A returned. For each subsequent year following permission completion then a value of N/A was returned. For each year which contained counts then such data related to the residential completions reported in that year only. This process was repeated for Flats Only and Houses Only as separate tables using size_type permissions files previously sourced from SMART Herts. This enabled the establishment of annual dwelling completions data sets for each permission at Units Only, House Only and, Flats Only levels of detail. Checks were made between annual aggregate Houses Only and Flats Only counts in comparison to the Units Only counts for that year to ensure unity.

Following collation of the annual completions counts for each level of detail, cumulative completions trajectories were produced in a standard format file equivalent to the above. Where a permission did not commence producing residential completions until a year after 2002 then each subsequent year, to assigned development reporting start year, had a value of N/A returned. Inclusion year completions data, i.e. year 1, replicated the values within the annual completions file. Year 2 counts were the annual completions counts plus the completions in the previous year.

For each subsequent year following development completion then the overall number of completions associated with the permission was returned to current period. This was equivalent to the value returned in the final year of the specific residential dwelling completions associated with the permission. The Units Only, Flats Only and, Houses Only cumulative completions trajectories would be applied to calculate relevant normalised yield rates per 100 dwellings over the longitudinal return of School Census record counts, matched by UPRN to new build completions, and aggregated to PPREF.

12.5 School census UPRN counts

Section 8.2 provides the method applied in determining School Census mainstream and special school pupil counts by N2, Primary, Secondary and, Post-16 to Unique Property Reference Number. The Master Address file (Section 11.3) was replicated into a separate file for N2, Primary, Secondary and, Post-16, this provided for each sector a master list of individual addresses assigned to the annual cohort included permissions. The included UPRN within the new build master address file was linked to each year of School Census cleansed addresses UPRN and per annum counts specific to each sector returned to the relevant file. This permitted the tracking of School Census pupil counts by each sector longitudinally for each specific dwelling included in the study. Pivot table analysis would permit the aggregation of counts, longitudinally, by either unique permission reference or annual cohort overall.

Associated metrics would be dwelling type, bed size and tenure either individually or in any combination thereof.

13.0 Analysing gross and net yields

The analysis of PYS based data sets was based on Units Only, Houses Only and, Flats Only as distinct entities for mainstream and special school normalised yields per 100 dwellings (N2, Primary, Secondary and Post-16) from new build developments. The current method applied relates to Gross Yields only, no discount for localised moves can be applied at this point in order to calculate Net Yields (see Section 12.2). Whilst SMART Herts datasets has enabled the determination of bed size and tenure for approximately 25% of the overall cohort, dominantly in the small development cohorts, HCC is dependent, at current point, on the future publication of the DfE methodology to ascertain such information for the remainder of the permissions.

Currently these metrics are excluded from analysis except in consideration of the overall bed size and tenure associated with both individual permissions and for annual cohorts in entirety. Normalised yields per 100 dwellings were also examined by emerging Typology classification. The two principle data sources applied in the analysis were Section 11.4 (Trajectory) and Section 11.5 (School Census UPRN Counts). The School Census UPRN Counts were subject to pivot table analysis to produce longitudinal aggregate counts by permission reference for each education sector separately, and by Units Only, Houses Only and, Flats Only (Table 4).

Table 4. An example of a Units Only listing of PPREF and the longitudinal sum of primary mainstream pupil counts observed for each permission.

	Primary Mainstream Pupil Counts									
PPREF	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
05/1382/FUL	0	6	23	33	42	51	50	53	46	45
07/01398/FULM	0	8	24	31	39	43	45	43	43	46
08/00485/RM	16	32	40	59	79	97	110	111	113	116
08/00746/FULM	0	0	6	12	14	16	20	20	22	23
09/00445/FULM	0	0	5	7	8	14	16	15	16	16
09/02366/1	0	0	0	5	6	12	14	13	14	14
09/0701/FUL	0	7	13	19	24	26	26	21	18	14
10/00469/FPM	0	0	0	15	19	38	52	69	61	58
10/00470/FPM	0	0	0	0	7	13	22	21	32	34
10/00472/1	0	0	0	7	9	12	13	16	16	14
10/01066/1	0	0	6	14	14	17	22	26	24	24
3/09/1061/FP	0	30	38	49	54	53	63	71	76	77

The sum of the counts, for each permission, in each year relates directly to the number of children of National Curriculum Year Group (NCYG) in that education sector as observed from that specific year's January School Census return. Counts are not cumulative additions rolled forward year on year. The total of each annual

column provides an overall sum of the number of pupils observed, in the new build dwellings included within the cohort, per annum.

The cumulative trajectory provides the count of dwellings within a permission form start date (inclusion year) to current point in time (Table 5), these are dwelling counts from which the pupil counts within Table 4 are observed to arise. Note that in examining Houses Only or Flats Only that the inclusion year (start point) may not have any dwelling counts in some permissions, this occurs where the houses, or flats, commenced construction at a point after the other dwelling type. The total of each annual column provides an overall sum of the number of cumulative completions observed, within the cohort, per annum.

Table 5. An example of the PYS trial Units Only cumulative dwelling completions over time by permission reference number (PPREF).

	Cumulative Dwelling Completions								
PPREF	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
05/1382/FUL	130	130	130	130	130	130	130	130	130
07/01398/FULM	129	129	129	129	129	129	129	129	129
08/00485/RM	250	254	280	370	386	386	386	386	386
08/00746/FULM	62	62	62	62	62	62	62	62	62
09/00445/FULM	71	71	71	71	71	71	71	71	71
09/02366/1		32	32	32	32	32	32	32	32
09/0701/FUL	51	51	51	51	51	51	51	51	51
10/00469/FPM	18	100	100	100	100	100	100	100	100
10/00470/FPM		17	86	88	88	88	88	88	88
10/00472/1	12	38	38	38	38	38	38	38	38
10/01066/1	59	59	59	59	59	59	59	59	59
13/0603/AOD		24	98	99	99	99	99	99	99
13/1117/AOD		50	95	123	126	126	126	126	126
3/09/1061/FP	119	147	147	147	147	147	147	147	147

For each permission dividing the annual count of mainstream pupils, for the relevant sector, by the cumulative number of dwelling completions (by type where relevant) and multiplying by 100 determines the normalised yield gross yield rate. For example, in 2020 the number of primary pupils residing in dwellings within PPREF 05/1382/FUL was 45 whilst the cumulative number of dwelling completions was 130. The normalised yield calculation proceeds as (45/130) * 100 = 34.6 mainstream primary pupils per 100 dwellings. An overall development cohort normalised gross yield is similarly calculated but replacing specific permission counts with overall annual development sums. For example, Figure 13 displays the accumulation of normalised mainstream primary pupils in the 2,440 houses included within the PYS

trial study using this method. The same principles apply when undertaking analysis of Typology data.

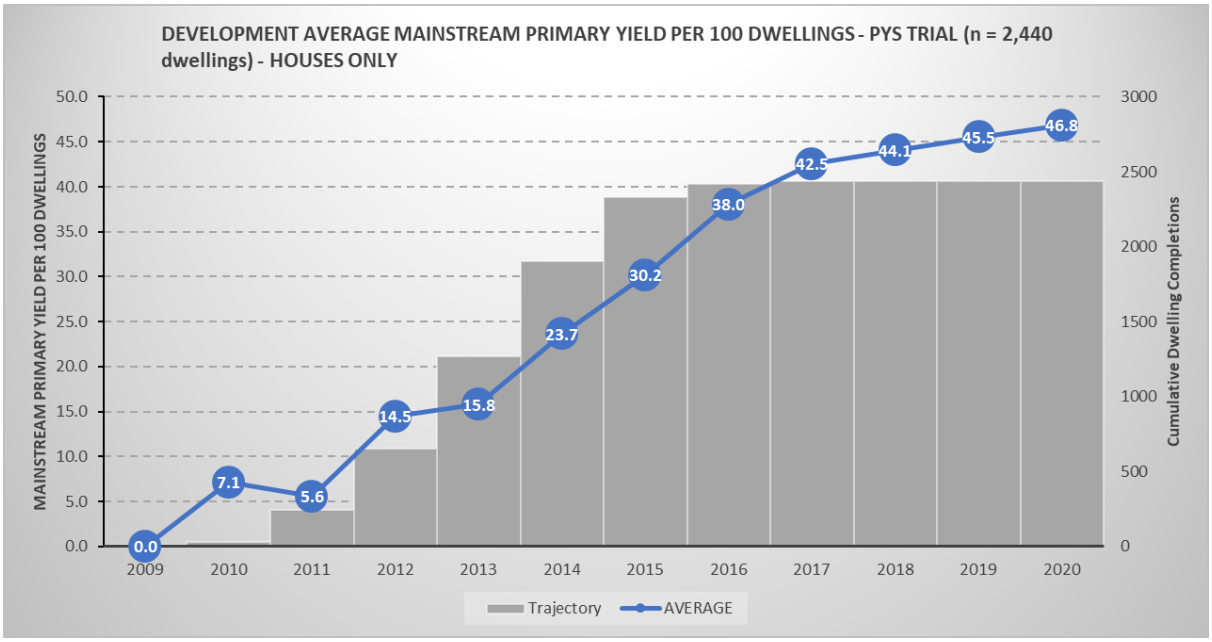


Figure 13. The accumulation of normalised mainstream pupil rates per 100 dwellings for the 2,440 houses included in the PYS trial study.

HCC has applied two calculation methods: the arithmetic mean and the weighted average.

13.1 The arithmetic means yield ad the weighted average yield

The average is a measure of central tendency, it is single value which represents the middle point in a data series such that 50% of the observations are above and 50% below. The Department for Education has indicated within its provisional documentation that summary yields across developments should be expressed as an average. However, to date no further information has been issued with respect to which average is considered the most appropriate. The following presents the two methods currently applied by HCC.

The average currently being applied by the authority is the arithmetic mean yield observed across all developments in an annual cohort, this utilises the normalised yield per 100 dwellings. It can be expressed as:

$$PY_{av\_dev} = \left(\frac{\sum \left(\frac{cp_1}{dw_1} + \frac{cp_2}{dw_2} + \frac{cp_3}{dw_3} + \frac{cp_4}{dw_4} \ldots . \frac{cp_i}{dw_i} \right)}{N_{Dev}} \right) \cdot 100$$

Wherein PY_{av_dev} is the average development yield per 100 dwellings for all developments, cp_i is the count of pupils in development 1, 2, 3 …. *i*, dw_i is the count of dwellings in development 1, 2, 3 …… *i* and, N_{Dev} is the number of developments included in the cohort.

The following example calculation includes three developments for simplicity, developments A, B and C contain 200, 300 and 500 dwellings respectively with a current observed yield of 150, 200 and 350 primary age pupils, the average development yield (PY_{av_dev}) per 100 dwellings can be calculated as:

$$PY_{av\_dev} = \left(\frac{\sum\left(\frac{cp_1}{dw_1} + \frac{cp_2}{dw_2} + \frac{cp_3}{dw_3} + \frac{cp_4}{dw_4} \ldots \cdot \frac{cp_i}{dw_i}\right)}{N_{Dev}}\right) \cdot 100$$

$$PY_{av\_dev} = \left(\frac{\sum\left(\frac{150}{200} + \frac{200}{300} + \frac{350}{500}\right)}{3}\right) \cdot 100$$

$$PY_{av\_dev} = \left(\frac{\sum(0.75 + 0.66 + 0.70)}{3}\right) \cdot 100$$

$$PY_{av\_dev} = \left(\frac{2.11}{3}\right) \cdot 100$$

$$PY_{av\_dev} = (0.70333) \cdot 100$$

$$PY_{av\_dev} = 70.33$$

The arithmetic mean development yield is therefore 70.3 mainstream primary pupils per 100 dwellings. In calculating the arithmetic mean yield equal weight is given to each development such that the smallest development of 200 dwellings has the same weight as the largest at 500 dwellings. In the above example the single dwelling yields from each development are reasonably close at 0.75, 0.66 and 0.70 respectively.

However, for smaller developments, the situation does exist wherein the included number of dwellings at a higher level of granularity, such as dwelling type analysis, will be small whilst the observed pupil count may be high. For example, consider development D for 20 dwellings wherein 15 are flats and 5 are houses. The development is included in the analysis as, at Units Only, the total number of dwellings is >10. Of the 15 flats 2 are Social Rented and 13 Open Market whilst for the houses 2 are Open Market and 3 Social Rented. It was observed that the flats gave rise to 6 primary mainstream pupils whilst the houses had 9 pupils. At Units Only the single dwelling yield is calculated as 15/20 = 0.75 or 75 per 100 dwellings, the arithmetic mean yield including developments A, B and C is therefore calculated as:

$$PY_{av\_dev} = \left(\frac{\sum(0.75 + 0.66 + 0.70 + 0.75)}{4}\right) \cdot 100$$

$$PY_{av\_dev} = \left(\frac{2.86}{4}\right) \cdot 100$$

$$PY_{av\_dev} = (0.715) \cdot 100$$

$$PY_{av\_dev} = 71.5$$

The arithmetic mean yield has increased slightly to 71.5 per 100 dwellings. The following example considers Houses Only, for this dwelling type the observed single dwelling yields from developments A, B and, C are as observed previously at 0.75,

0.66 and 0.70 respectively. The single dwelling yield for development D is 9 pupils/5 dwellings = 9/5 = 1.8, the arithmetic mean is calculated as:

$$PY_{av\_dev} = \left(\frac{\sum(0.75 + 0.66 + 0.70 + 1.8)}{4}\right) \cdot 100$$
$$PY_{av\_dev} = \left(\frac{3.91}{4}\right) \cdot 100$$
$$PY_{av\_dev} = (0.9775) \cdot 100$$
$$PY_{av\_dev} = 97.8$$

The arithmetic mean across all developments has increased to 97.8 per 100 dwellings, the average has risen by 26.3 per 100 dwellings or a relative increase of 39.1% (27.5/70.3), where the houses single dwelling yield is identical to the Units Only single dwelling yield.

Development D had a small number of 5 dwellings which were houses although the single dwelling yield is high at 1.8 due to the tenure. The contribution of development D to the total 1,020 Units Only dwellings (200 + 300 + 500 + 20 = 1,020) is very small at 2.0%, if developments A, B and C were formed of 50% houses then the contribution of development D, at 5 houses, would be even smaller at <1%. However, inclusion of development D in calculating the arithmetic mean has resulted in the development average yield increasing to 97.8 per 100 dwellings, a 39.1% rise. It can be observed that the arithmetic mean is substantially affected by extreme values, this can occur to either increase or decrease the arithmetic mean of an annual cohort and it is not likely that such “extreme” values either size of the central point will balance one another.

The median is less affected by extreme or outlier values however this would only be the case where there was an equivalent number of developments >=30 dwellings as those in the >=10 to <30 dwellings cohort (where extreme values tend to be observed). Observations of the extended PYS have however indicated that the latter category is at least twice the size of the former and therefore it is likely that the median will also be impacted.

An average calculation which considers each developments size relative to the total number of dwellings, commonly referred to as its *weight*, could result in a more accurate measure. This is termed the weighted average, the weight (W_i) given to any one development is the ratio of the number of dwellings (dw_i) in that development divided by the total number of dwellings in all developments (T_{dw}). This is expressed as:

$$W_i = \frac{dw_i}{T_{dw}}$$

The weights for developments A, B and C, in the above example, are therefore:

$$A)\ W_i = \frac{dw_i}{T_{dw}}\ ;\ W_1 = \frac{200}{1{,}000}\ ;\ W_1 = 0.20$$
$$B)\ W_i = \frac{dw_i}{T_{dw}}\ ;\ W_2 = \frac{300}{1{,}000}\ ;\ W_2 = 0.30$$
$$A)\ W_i = \frac{dw_i}{T_{dw}}\ ;\ W_3 = \frac{500}{1{,}000}\ ;\ W_3 = 0.50$$

As the weights are calculated relative to the total number of developments then they will sum to 1 (0.2 + 0.3 + 0.5 = 1.0). The weight for each development is used as a multiplier for the single dwelling yield, $\frac{cp_i}{dw_i}$, the resulting values are summed in order to determine the weighted average pupil yield (*PY_w).* The calculation proceeds as follows:

$$PY_w = \sum \left[\left(\frac{cp_1}{dw_1} \cdot\ w_1\right) + \left(\frac{cp_2}{dw_2} \cdot\ w_2\right) + \left(\frac{cp_3}{dw_3} \cdot\ w_3\right)\right]$$

$$PY_w = \sum \left[\left(\frac{150}{200} \cdot\ 0.2\right) + \left(\frac{200}{300} \cdot\ 0.3\right) + \left(\frac{350}{500} \cdot\ 0.5\right)\right]$$

$$PY_w = \sum [(0.75 \cdot\ 0.2) + (0.66 \cdot\ 0.3) + (0.70 \cdot\ 0.5)]$$

$$PY_w = \sum [0.15 + 0.20 + 0.35]$$

$$PY_w = \ 0.70$$

The mainstream primary weighted average single dwelling yield is 0.70 or 70 per 100 dwellings.

It can be proven that this equals the simpler calculation of the total number of pupils divided by the total number of dwellings, 700/1000 = 0.70 or, 70 per 100 dwellings. In the calculations above each development pupil count is divided by the number of dwellings, the result is multiplied by the number of dwellings in that development divided by the total number of dwellings, for development A this can be expressed as:

$$\left(\frac{cp_1}{dw_1} \cdot\ w_1\right)\ or\ \left(\frac{cp_1}{dw_1} \cdot \frac{dw_1}{T_{dw}}\right)$$

This reduces to:

$$\left(\frac{cp_1 \cdot dw_1}{dw_1 \cdot T_{dw}}\right)$$

However, dw_1 / dw_1 = 1 and can therefore be removed, for development A the residual is:

$$\frac{cp_1}{T_{dw}}$$

Repeating the process for developments B and C then:

$$PY_w = \left(\frac{cp_1}{T_{dw}} + \frac{cp_2}{T_{dw}} + \frac{cp_3}{T_{dw}}\right)$$

Wherein T_{dw} is a common denominator, such that:

$$PY_w = \left(\frac{cp_1 + \ cp_2 + cp_3}{T_{dw}}\right)$$

Given that the total pupil count, cp_{TOTAL}, is the sum of $cp_1 + cp_2 + cp_3$ then:

$$PY_w = \frac{cp_{TOTAL}}{T_{dw}}$$

The division of the total pupil count by the total number of dwellings is a Point Estimate (PE) of pupil yield, multiplied by 100 it is the point estimate yield per 100 dwellings. It can however be observed herein that this is the same as the weighted average yield when considering the proportional representation of each developments number of dwellings relevant to the total number of dwellings in the whole cohort. The weighted average yield of 70 per 100 dwellings in the example is marginally less than the 70.3 per 100 dwellings calculated by the arithmetic mean yield of the developments.

Incorporating development D into the calculation determines (700 + 15) / (1,000 + 20) = 715 / 1,020 = 0.701 or 70.1 per 100 dwellings. This can be checked using the longer form of the equation:

$$PY_w = \sum \left[\left(\frac{cp_1}{dw_1} \cdot w_1\right) + \left(\frac{cp_2}{dw_2} \cdot w_2\right) + \left(\frac{cp_3}{dw_3} \cdot w_3\right) + \left(\frac{cp_4}{dw_4} \cdot w_4\right)\right]$$

$$PY_w = \sum \left[\left(\frac{150}{200} \cdot 0.196\right) + \left(\frac{200}{300} \cdot 0.294\right) + \left(\frac{350}{500} \cdot 0.490\right) + \left(\frac{15}{20} \cdot 0.020\right)\right]$$

$$PY_w = \sum [(0.75 \cdot 0.196) + (0.66 \cdot 0.294) + (0.70 \cdot 0.490) + (0.75 \cdot 0.020)]$$

$$PY_w = \sum [0.14706 + 0.1961 + 0.3431 + 0.0147]$$

$$PY_w = 0.700996$$

$PY_w =$ 0.701 or 70.1 per 100 dwellings.

Including the fourth development, D, into the weighted average, results in a calculated value of 70.1 per 100 dwellings. Including the considerably smaller development of 20 dwellings has increased the weighted average yield by 0.1 per 100 dwellings.
It can be observed that the weighted average has benefits over the arithmetic mean in determining the development average yield: it permits the inclusion of small developments wherein yield per 100 dwelling rates may be many times that observed for larger developments whilst not being excessively affected by possible outliers. It is acknowledged in literature that the arithmetic mean is the most widely understood measure of average wherein it is intuitively understood that it is the centre point of a data set.

HCC investigated whether a hybrid equation could be derived which includes elements of both the arithmetic mean and the weighted average. Such an equation would account for both the sum of the pupil counts, the contribution of each developments number of dwellings to the total dwellings included and, the number of developments included within the study. However, the hybrid approach is equally affected as that of the arithmetic mean when small developments calculated yields per 100 dwellings are disproportionately high. This occurs due to the very small ratio of the number of dwellings in developments like these relative to the total number of dwellings overall. The choice of arithmetic mean or weighted average will be dependent on DfE recommended methodology, currently HCC applies the more commonly accepted arithmetic mean.

13.2 Calculating net yields

The Study will consider whether it is appropriate to apply some form of discount to gross yields to account for localised moves which are suggested to not increase pressure on local school capacity. This will be the case for both Open Market and Affordable Rented dwellings, with a possible exception being where it can be evidenced that older dwelling stock locally vacated homes are backfilled by families with children whom go on to take up a local school place thereby exacerbating demand. The spatial extent of such areas is still under investigation and the authority cannot apply such methods to discount gross yields until further information is forthcoming.

13.3 Statistical tests for normality

Where an average value from a study is applied then it is only an accurate representation of the centre point where the distribution matches, or approximates, that of a normal distribution. The Jarque-Bera and D'Agostino are two tests, of increasing robustness, which can be applied to determine whether the observed distribution from PYS results are within bounds of the normal distribution. The former tests for skewness and kurtosis and, is generally considered as very effective. The D'Agostino tests for skewness, kurtosis and centrality and is considered more of a powerful omnibus. The Jarque-Bera is a goodness-of-fit statistical test for whether data have the skewness and kurtosis which matches that of a normal distribution. It is calculated as:

$$JB = n\left(\frac{(k_3)^2}{6} + \frac{(k_4)^2}{24}\right)$$

Wherein *x* is each observation, $\bar{x}$ is the mean of all observations, *n* is the sample size, *s* is the standard deviation, k_3 is skewness and, k_4 is the kurtosis. The skewness, k_3, is calculated as:

$$k_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

The kurtosis, k_4, is calculated as:

$$k_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

As this is a statistical test a hypothesis, h_o, needs to be established. In this instance the hypothesis, h_o, is that there was no statistically significant difference in the PYS distribution to that arising from a normal distribution. The reverse, or null hypothesis, h_n, is that there was a statistically significant difference in the observed distribution and that arising from a normal distribution. The calculated Jarque-Bera statistic was compared to the Chi-Squared distribution table to determine the critical value at a probability level (alpha or α) of 0.05. In this instance skewness and kurtosis were the variables of interest and therefore there were two degrees of freedom (*df*).

The D'Agostino test is based on the *D* statistic which provides an upper and lower critical value and is calculated as:

$$D = \frac{T}{\sqrt{n^3 SS}} \text{ And } T = \sum \left(i - \frac{n+1}{2}\right) X_i$$

Wherein *D* is the test statistic, *SS* is the sum of squares of the data, *n* is the sample size and, *i* is the order or rank of observation *x*. The degrees of freedom (*df*) for this test is the sample size *n* (thereby *df* = *n*). The data must be ordered from smallest to largest, or vice versa, prior to commencing the test. Within the PYS trial $\frac{n+1}{2} = \frac{41+1}{2} = \frac{42}{2} = 21$, therefore $T = \sum(i - 21) X_i$ which equated to 6021.4 and substituted in the equation. The test statistic *D* was therefore calculated as:

$$D = \frac{T}{\sqrt{n^3 SS}} \text{ or;}$$

$$D = \frac{6021.4}{\sqrt{n^3 SS}}$$

As this is a statistical test a hypothesis, h_o, was established. In this instance the hypothesis, h_o, is that there was no statistically significant difference in the observed PYS distribution to that arising from a normal distribution. The reverse, or null hypothesis, h_n, is that there was a statistically significant difference in the observed distribution and that arising from a normal distribution. If the calculated value of *D* fell within the critical range then h_o was accepted, otherwise h_o was rejected and h_n accepted.

13.4 SEND yields

The number of children whom are resident in the authority, attend an in-county special school and, have been UPRN matched to a new build dwelling is relatively small. Normalised yield rates per permission by education sector are therefore quite variable and do not lend themselves well to determination of average yield via arithmetic mean across an annual cohort. The determination of average yield normalised yields by education sector follows a slightly different approach to that of mainstream yields. In this instance the weighted average yield is determined per annum based on an entire annual cohort i.e. it utilises the overall count of pupils and the overall cumulative number of dwelling completions in that year. The arithmetic mean yield is then taken across all years since cohort trajectory completion to current year.

Following discussion with the DfE the authority has identified further work required in this area. Recent advice received is that those children whom attend SEND bases in mainstream schools should also be included within this cohort and HCC is moving to source this data from the January School Census returns and UPRN match to the new build dwellings. It is likely that such included counts will be small, and it will be more resource efficient to process this element once all annual cohorts have been joined into master lists 2002 through to 2020. However, this will be reviewed once a trial has been completed.

13.5 Public accessibility of results

It is the authority's intent, in the fullness of time, to release permission, annual cohort and typology level results (where relevant by dwelling type, bed size and tenure) from the longitudinal PYS where such results satisfy the requirements of Statistical Disclosure Controls. This ensures that methodology, results and conclusions drawn are within the public domain with relevant consideration of possible statutory restrictions on the release of such data. It is likely that the vehicle for access to such information would be via Herts Insight[6] which is already a repository of information and statistics about Hertfordshire across a wide number of defined and bespoke service geographies. This platform would also permit the integration of PYS data sets with overlays of small area statistics already held and the possible creation of profile reports.

14.0 Calculating long term average (LTA) mainstream yields

The LTA is the overall yield that a development would be expected to attain once enough time has passed post-peak and reflects the wider housing stock yields, it is sometimes referred to as the "All Households" yield. The HDM inclusion of only 2011 census based All Household yields was suggested to not take account of inter-census period changes to overall dwelling stock numbers nor changes in the demographic profile of the authority area. HCC investigated methodologies by which these metrics could be updated, the first is based on official ONS population estimates/projections whilst the second is sample based.

14.1 Applying official population estimates to calculate the LTA

Within the consultation response reference was made to application of ONS Sub-National Population Projections (SNPP) cohort and SMART Herts sourced dwelling stock counts rolled forward from the 2011 census to current period. SNPP are counts for the population as a whole and not just mainstream pupils, a sector relevant countywide mainstream uptake rate would therefore need to be applied prior to calculating yield per 100 dwellings. The benefit of the SNPP is that an official projected primary age cohort can be applied for current point in time whereas the ONS Mid-Year Estimates are normally 12 to 18 months behind current time. An alternative approach is to use the most recent January School Census return refined for in-authority resident children counts.

14.2 Updating dwellings units only LTA values

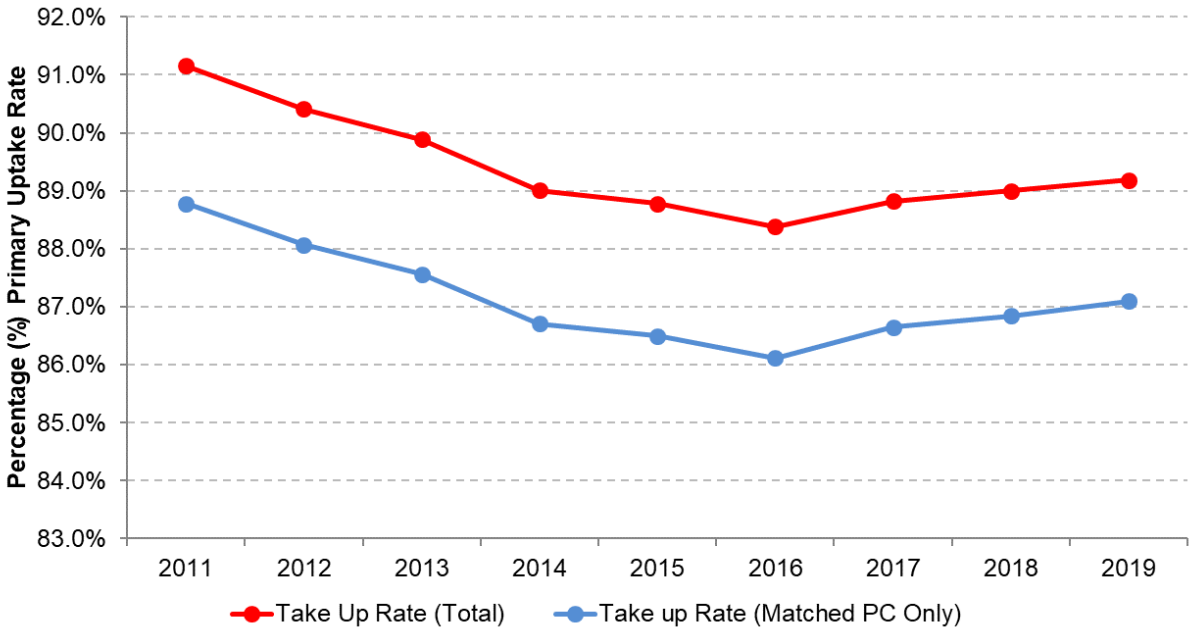
The most recent year SNPP projection for the relevant year, or ONS Mid-Year Estimate, can be used to determine an overall sector count by relevant age band. For example, the 2019 ONS MYE determines a primary age cohort within Hertfordshire of 112,190 children. Data from spatial planning estimates that as at 2019 there were 495,335 dwellings, the Units Only primary LTA value therefore becomes 112,190/495,335 = 22.6 per 100 dwellings. However, this relates to all children of primary age and adjustment must be made to account for mainstream schools only. The current mainstream Hertfordshire resident only uptake by the

[6] https://www.hertfordshire.gov.uk/microsites/herts-insight/home.aspx

primary age population was determined as 87.1%, this was the average of three methods, multiplying the yield per 100 dwelling rates by this percentage determines an updated LTA mainstream only rate. The calculation proceeds as: 22.6 * 0.871 = 19.72 per 100 dwellings. The same principles can be applied to determine updated secondary and Post-16 rates.

The previous paragraph references “Hertfordshire resident only”, there is an important distinction between the number of primary age children whom attend an authority mainstream setting and those whom are resident in the authority itself. The January School Census return can be used to determine cohorts of mainstream children based on the same inclusion query structures as applied within the PYS. It can be observed from Figure 14 that the total take-up rate, i.e. a count of all primary age children in authority mainstream schools, is consistently 3 percentage points higher than that when considering Hertfordshire resident children only. The determination of in-authority residence can be made using the pupil home postcode in conjunction with ONS NSPL data files.

Figure 14. The percentage of Hertfordshire mainstream primary age cohorts relative to the relevant ONS Mid-Year Estimates of Primary age within the authority.



Within Figure 14 the y-axis starts at 83% as opposed to zero and trends in variance in up-take are subsequently exaggerated. Mainstream primary resident only uptake rates decreased from around 89% of the overall primary age population in 2011 to 86% in 2016. Since this point uptake has increased to 87%. In order to account for variability an average of three methods was taken: (1) most recent 3-year average weighted uptake; (2) average of percentage over period 2011 to 2019 and; (3) overall count of mainstream resident pupils 2011 to 2019 divided by the ONS MYE 2011 to 2019 aggregate mainstream primary age population. In practice little

variation exists in either method with 2019 uptake values of 86.97%, 87.15% and, 87.10% respectively.

The application of ONS MYE or ONS SNPP data sets requires use of authority School Census data sets for calculating mainstream resident specific take-up rates. It can be observed that a simpler method is to solely apply the School Census data sets to a current estimate of dwelling stock. For example, the 2019 January School Census indicates that there are 100,058 primary age pupils attending mainstream schools in Hertfordshire. Of these 2,347 were out of county, or not successfully geolocated to authority postcodes. The resulting authority resident mainstream primary age cohort was 97,711 pupils. Data from spatial planning estimates that as at 2019 there were 495,335 dwellings, the Units Only primary mainstream LTA value therefore becomes 97,711/495,335 = 19.73 per 100 dwellings.

There is a negligible 0.01 per 100 dwelling difference to that calculated using the ONS MYE data. In both instances there is a decrease from that observed from the 2011 Census at 20.8 per 100 dwellings for overall primary cohorts. However, considering primary uptake in 2011 at 89% then the census based mainstream resident LTA would have been 20.8 * 0.89 = 18.5 per 100 dwellings, this update effectively increases the development permanent provision costs by 1.2 pupils per 100 dwellings (19.7 – 18.5). A benefit of ONS official population estimates, such as the SNPP or Mid-Year Estimates (MYE) is that they can also be applied to update other age group LTA dwelling unit values whilst School Census data sets are relevant for ages 4 to 18 years only (excluding further education colleges at Post-16 for which the authority does not hold data).

Whilst the authority has uptake rates, over time, readily available by mainstream sector there remains a more substantial barrier to applying official population estimates. This method would be limited to dwelling units only as the SNPP does not provide population projections arising from specific dwelling types. Application of official population estimates would therefore, at face value, be an unsuitable method for LTA calculation of Houses Only and Flats Only rates. Either a divergent approach to dwelling units only versus dwelling type would need to be applied or consideration given to alternative methods of calculation which incorporates both elements.

14.3 Dwelling type LTA census output area based values

Hertfordshire County Council is segmented into 3,516 census outputs areas (OAs), it is likely that a proportion of these would consist solely of houses and some solely flats. Identification of OAs that are of a singular dwelling type permits the allocation of census-based household by bed size counts and, overall household counts. Where OAs consist wholly of unshared households then the household count is in effect the dwelling count. Extraction of current AddressBase Premium residential dwellings for these OAs, and aggregated counts of dwellings in comparison to census estimates, would determine those which have not been subject to development since 2011 and the bed size distribution is most likely as reported at that time.

Where such OAs are identified then the most recent ONS Mid-Year Estimates (MYE) to census output area could be cross referenced and aggregated, this is not

applicable to the ONS SNPP for which District is the lowest geography. This determines most recent ONS based population estimates of areas consisting of solely houses and flats for which the dwelling count and aggregate bed size are known. Multiplying the mainstream ONS MYE population count by county-wide uptake rate determines an estimate of mainstream sector yield.

Ralph (2011) indicates that where the number of aggregate OAs is >=6 then the errors associated with population estimates are likely to be less than that which occurs from small area record swapping applied by the ONS under Statistical Disclosure Control (SDC) measures. Whilst OAs are relatively small in terms of households contained there is a minimum threshold of 40 households, a target of 125 households and an upper limit of 250 households (Figure 15). Where the aggregate OA count exceeds the target rate of 750 households (6 * 125) and 1,872 persons (312 * 6) then estimates are indicated by the ONS to be robust.

The population and household size target and threshold values applied to the Output Area (COA), Lower Super Output Area (LSOA) and, Middle Super Output Area (MSOA) hierarchy for 2011 (ONS November 2012).

	Population thresholds			Household thresholds		
	Lower	Target	Upper	Lower	Target	Upper
OA	100	312	626	40	125	250
LSOA	1,000	1,500	3,000	400	600	1,200
MSOA	5,000	7,500	15,000	2,000	3,000	6,000

Figure 15. The Office for National Statistics Lower, Target and Upper population and household counts by census geography (Source: Office for National Statistics).

Alternatively, as above, mainstream sector counts for the aggregate OAs can be derived from the most recent January School Census either geolocated to point in polygon or through assignment using the ONS NSPL. Division of the most recent population estimates, from either source, by the known number of dwellings by specific type in theory permits the calculation of an updated LTA yield rate related to a specific bed size mix. The match between 2011 census dwelling count and current AddressBase Premium dwelling counts ensures that no development has occurred in the OAs since the census such that new build would be excluded. The approach in detail can be observed as:

- The 3,516 total 2011 Census Output Areas were listed, and census data matched to the following fields: Total Dwellings, Total Houses, Total Flats – sub-groups required for each were counts of 1 bed, 2, bed, 3 bed and 4+ bed unshared dwellings. HMO's and communal establishments were listed against the OAs as separate fields.
- Census table KS401EW or QS418EW provides a total count of dwellings by OA and total count of Unshared dwellings (the difference between the two

being a total count of Shared dwellings). Note that dwelling counts excludes caravans/temporary structures. Within this table are also OA aggregate counts of Household Spaces by Accommodation Type – Houses, Flats and caravans/temporary structures – no bed size data is available from these tables and these Household Space counts are for all Household Spaces and not just those with one or more usual resident. The count of Households Spaces will therefore match the OA Total Dwelling count where there are no Shared Dwellings and there are no Caravans/temporary structures.

- Census Table QS411EW (also LC4405EW) provides OA counts of "All Household Spaces With At Least One Usual Resident by Bed Size" although Accommodation Type is not an included field. These were cross referenced to the OA list.
- From this master list OAs were identified wherein the 2011 census indicates 100% Houses or 100% Flats (unshared dwellings), this cohort was further refined by excluding those OAs where counts of communal establishments or shared dwellings existed.
- The refined cohort was passed to GIS whom extracted the AddressBase Premium residential dwelling addresses and dwelling type characteristic (House or Flat) for each census area. Counts of AddressBase Premium addresses by dwelling type for each OA determined those OAs where development had occurred subsequent to the 2011 census, these were excluded from further analysis.
- The resulting OA cohort was matched to the ONS MYE 2018 most recent OA level population estimates release which enabled calculation of Units Only, Houses Only and Flats Only population sector yield per 100 dwellings for a known aggregate bed size distribution.

Applying the above criteria (Houses only in this example) there are 196 of the 3,516 Hertfordshire OAs where the Total Dwelling Count = Total Unshared Dwellings Count = Household Spaces for Accommodation Type_Houses Count. Census Table QS411EW (also LC4405EW) provides OA counts of "All Household Spaces With At Least One Usual Resident by Bed Size" although Accommodation Type is not an included field. However, the above process determined a 196 OA cohort wherein all the dwellings and unshared household spaces relate to the Accommodation Type - Houses. QS411EW provides bed size data directly applicable to the identified dwelling type of Houses. This is only true where the sum of the OA Household Spaces Bed Size data (QS411EW) matches that of the Total Household Spaces from Table KS401EW. Recall that QS411EW relates to bed size counts of Household Spaces where there are one or more usual residents only. KS401EW reports the total number of Household Spaces irrespective of whether there is a usual resident present.

At this point consideration needs to be given to the ONS MYE and what data they contain. ONS MYE are estimates of the *usually resident population* of a defined geography as at 30th June each year. The Household Spaces counts by bed size (QS411EW) relate to households with one or more usual residents. Where OA based aggregate counts between QS411EW and KS401EW differ then this occurs when either the household space is occupied by a household which does not match the inclusion criteria of usual resident or, the property is vacant. Application of OA observed aggregate counts in a yield per single dwelling calculation wherein the

population estimate is irrespective of vacant dwellings and is for the usually resident population only therefore meets criteria between both data sets. However, it could be argued that this would only be correct where it can be evidenced that the “missing” household spaces continued to be vacant or occupied by a household that did not meet the inclusion criteria for usual resident.

Overall, the above issues can be considered moot if the 196 OA cohort for Houses Only has an overall Household Spaces count (KS411EW) which is sufficiently close to the overall Household Spaces by Bed Size counts from QS411EW for the overall bed size mix to be considered representative. The Total Houses Only Household Spaces (and by inclusion criteria dwelling type of Houses Only) identified in this cohort was 24,408 (KS401EW) and there are no shared dwellings or caravans/temporary structures in any OA. QS411EW determined 23,971 Household spaces with 312 1-Bed (1.3%), 2,974 2-Bed (12.4%), 11,248 3-Bed (46.9%) and 9,437 (39.4%) 4+ Bed.

The difference in the 196 aggregated OA counts for Household Spaces – Houses Only between the data sets was 437 or 1.8% relative to the total observed in KS401EW. The converse of this is that 98.2% of the Houses Only bed size household spaces data set was included in QS411EW. The overall bed size mix can therefore be considered representative of the overall Houses Only dwellings.
If it is assumed that the occurrence of household spaces not meeting the inclusion criteria for QS411EW (vacant or not a usual resident household) is equally distributed across the bed size range then estimates of the actual bed size counts can be determined by multiplying each cohort by 1.018 (101.8%). However it is more likely that the differences would be proportional to the observed bed size distributed – a count of 11,248 3-Bed households spaces in houses would experience a greater number of occurrences of non-inclusion criteria (vacant/not a usual resident household) by virtue of the size of the cohort in comparison to the overall 312 1-Bed houses. Overall there are three determinations of the Houses Only overall Bed Size mix from the census data:

- Apply the bed size mix as observed from available Household Spaces with 1 or more usual residents.
- Examine the overall percentage difference in counts between KS401EW and QS411EW for the identified OAs and assume that the percentage difference is relevant to each bed size
- Examine the count difference between KS401EW and QS411EW and proportion this difference between the bed sizes dependent on the contribution of the bed size cohort to the total overall.

In practice, the observed overall difference of 437 dwellings is sufficiently small relative to the total 24,408 cohort that even where account is taken of differences the resulting estimated percentage bed size mix matches that of the original known data. Where differences are small then the original bed size mix from QS411EW is representative of the overall cohort derived from KS401EW. Applying the criteria, that a match in count is required between KS401EW Household Spaces – Houses and QS411EW Household Spaces with at least One Usual Resident, reduces the cohort from 196 to 43 OAs.

Whilst the overall cohort sizes contained in the 43 OAs specified above is more than enough for a statistically robust sample this excludes GIS analysis to confirm that the total number of dwellings between the census date equals that of the most recently available GIS dwelling data set. Where totals equate then the census observed bed size mix is still relevant, the larger the OA cohort the smaller the errors in the estimates. Of the 43 OAs which matched between census data sets for Houses Only 16 had an exact match to total dwellings, and of residential characteristics indicating a house dwelling type only, as determined by GIS. A further 13 OAs had a match within 1 dwelling count. The exact match OAs had an ONS MYE 2019 primary count of 459 children and, the “match within 1 dwelling” a count of 351 children. The combined count of 810 primary age children equates to a yield of 22.3 per 100 houses.

Applying the Units Only take up rate of 87.1% reduces this to 19.4 per 100 houses mainstream primary yield. The validity of applying a Units Only take up rate is however questionable, there exists the strong possibility that such rates would be higher from specific house type dwellings than that from flats. A similar approach can be taken for other demographic age bands, for example the ONS MYE 2019 indicates a total population of 9,506 persons within the 29 OAs of solely house dwelling types. This equates to a yield of 2.6 persons per house from the included bed size distribution. Whilst the overall bed size mix is known in aggregate it is not possible to calculate at this point individual house bed size LTA yields.
Additional difficulty is presented when considering Flats Only as none of the 3,516 OAs within Hertfordshire are solely of this dwelling type. In considering the percentage contribution of flats to total dwelling stock there are only 19 OAs wherein the representation is >=95%. However, as this dwelling type cannot be considered in isolation via this method then it cannot be applied to determine LTA yield values with a degree of confidence that would likely satisfy requirements of Regulation 122. Consideration of a geography smaller than OA from which aggregate counts can be derived would likely solve this issue. The above process considers census output areas only as the ONS releases MYE to this geography. However, it is equally valid for application to smaller geographies such as postcode area data sets on condition that 2011 census exists for these areas or, it can be proven that included areas specifically exclude new build developments.

14.4 Dwelling type LTA postcode-based values

Postcode areas are specifically implemented for the delivery of post by Royal Mail, whilst they contain clusters of houses (on average 30 dwellings) the boundaries associated with them are somewhat arbitrary. Postcodes are regularly created and terminated and the ONS Geoportal provides quarterly updates via the ONSPD and ONS NSPL. It is the dwellings/households contained within a postcode area, and often its population weighted centroid, which is normally of relevance as opposed to exact spatial extent. Whilst official, or experimental, population estimates are not available for these areas proxy estimates can be created based on, for example, School Census or Electoral Register data sets.
All postcodes within the ONS NSPL which are assigned to Hertfordshire were extracted into a new CSV. AddressBase Premium was used to obtain all UPRNs within Hertfordshire which have a Classification Code which was included in the PYS. The AddressBase Premium points contain the postcode of the UPRN within

the schema. A join was made between the above AddressBase Premium points and the NSPL Hertfordshire postcodes based on the postcode fields, with those UPRNs which matched therefore attributed as belonging to Hertfordshire within the NSPL. This table was then exported for pivot table analysis and comparison of the postcode aggregate proportion of dwellings by Type made to overall Hertfordshire dwelling stock applying the same methodology. Only those postcodes wherein the residential characteristic enabled determination of dwelling type for all units within the area were included.

In total 28,057 postcodes located within the boundary of Hertfordshire were initially included, in aggregate these postcodes contained 110,861 Flats and 386,744 houses (total 497,605 dwelling units). Flats represented 22.3% of the cohort and houses 77.7%. It was observed that 19,744 postcodes were wholly populated by dwellings of House residential characteristic type and consisted of 299,341 units, this represented 77.4% of the total House type dwelling stock in the authority. In total 3,085 postcodes solely contained 62,192 dwellings of Flats residential characteristic type which represented 56.1% of the total Flat type dwelling stock in the authority. Overall 81.4% of the postcodes within Hertfordshire were included within the study at this point, the postcodes contained 72.7% of the total dwelling stock.

The refined cohort of 22,289 postcodes was cross referenced to 2011 census table LC1117EW for obtaining census-based household counts. LC1117EW is for Table 1 postcode counts of 1 or more usual resident occupied household spaces as at census date 2011 and the related total household usually resident population[7]. It excludes Table 2 postcode level occupied households with 1 or more usual resident occupied household spaces and related population counts. These postcode areas straddle OA boundaries and are subsequently split with only overall count and percentage representation data presented. Both data sets exclude postcodes with unoccupied households or where there is no usual resident. As postcodes are included by GIS process only if they wholly contain residential dwellings which meet the inclusion criteria then communal establishments are excluded.

In total 1,158 of the 22,829 postcodes had no match to LC1117EW and were either OA straddling postcodes or, postcodes with no usual residents or, occupied household spaces and/or are postcode areas created or terminated since 2011. These postcodes were excluded from further analysis. Of the remaining 21,671 postcodes a cross check between the GIS dwelling counts to that of census table LC1117EW determined those areas within which development had not occurred[8]. This resulted in a final cohort of 12,954 postcodes of which 1,040 were wholly flatted and 11,914 wholly contained houses only.
In relation to wholly flatted postcodes these areas contained 14,346 units whilst the wholly houses postcodes contained 158,368 units, total included dwelling count was 172,714 units. The number of flats included for further analysis is 12.9% of total authority flatted units dwelling stock, 40.9% of all houses and, 34.7% of total units.

[7] LC1117EW is for total population only - no other census table appears to provide age breakdown by postcode - only postcode sectors or postcode districts.
[8] A natural extension of this would be to cross compare refined postcode lists for inclusion at Units Only and Type analysis to those postcodes observed from the PYS master address list 2002 to 2011. Postcodes should be removed wherein a match is observed as this indicates that new build occurred in the period 2002 to 2011.

Using LC1117EW it was observed that the single dwelling total population yields as at 2011 were 1.72, 2.63 and 2.55 respectively for Flats Only, Houses Only and, Units Only. This methodology results in an included dwelling count by type of such size that sample-based confidence intervals are likely to be less than 0.2% (although possibly 0.5% when considering flats only).

However, a limitation is that whilst dwelling Type yields can be determined this is not available for either bed size or tenure distinction. Once the DfE has provided guidance on how authorities can obtain individual dwelling bed size and tenure data then a natural extension of the postcode analysis will either be to:

- Examine those postcodes of singular dwelling type and singular bed size[9] or;
- Examine individual dwelling level data of all postcodes by bed size and tenure.

Due to resource constraints the authority has yet to proceed with postcode level analysis of mainstream sector yields from the 2020 January School Census by dwelling type.

14.5 Sample based assessment of mainstream LTA yields

The final alternative approach to the estimation of LTA mainstream pupil yields is to conduct a sample from existing dwelling stock. Figure 16 references recommended sample sizes based on the population proportion displaying the characteristic of interest. Whilst the LTA for dwelling units only indicates a percentage value of circa 20 per 100 dwellings, or 20%, the PYS has demonstrated that houses can have a primary pupil yield up to, and in excess of 50 per 100 dwellings, or 50%. The 50% demarcation in sampling is the “worst case” scenario due to the presence of a higher level of uncertainty and hence the sample size required at this mid-point is the largest. Whilst the industry standard is 95% +/- 5% it would be prudent to aim for 95% +/- 2%, based on these criteria the table below indicates a sample size of 2,401 dwellings would be required.

However, experience indicates that some poor-quality School Census addresses will be impossible to geolocate to UPRN level and it was considered prudent, to retain confidence interval, that the sample size of 2,401 be increased by 10% to a total of 2,641 dwellings. The following process was then undertaken:

- All postcodes within the ONS NSPL (2019) which are assigned to Hertfordshire were extracted into a new CSV. AddressBase Premium was used to obtain all UPRNs within Hertfordshire which had a Classification Code which was included in the PYS. The AddressBase Premium points contained the postcode of the UPRN within the schema. A join was made between the AddressBase Premium points and the NSPL Hertfordshire postcodes based on the postcode fields, with those UPRNs which matched therefore attributed as belonging to Hertfordshire within the NSPL (2018). This table was then exported for pivot table analysis[10].

[9] This would be dependent on the determination of a sufficient dwelling count by dwelling type and bed size in order to have a low sample confidence interval such that results are both meaningful and applicable in application to modelling.

[10] Note: OS CodePoint-Polygon cannot be used singularly as the source to identify the UPRN postcodes as flats are more likely to be found within Vertical Streets inside the Code-Point Polygon dataset which are unlikely to

Figure 1: Sample size lookup table

Population Proportion	Precision (at the 95 per cent confidence level)							
	±12%	±10%	±8%	±5%	±4%	±3%	±2%	±1%
50%	66	96	150	384	600	1,067	2,401	9,604
45% or 55%	66	95	148	380	594	1,056	2,376	9,507
40% or 60%	64	92	144	369	576	1,024	2,305	9,220
35% or 65 %	60	87	136	349	546	971	2,184	8,739
30% or 70%	56	81	126	323	504	896	2,017	8,067
25% or 75%	50	72	112	288	450	800	1,800	7,203
20% or 80%	*42*	61	96	246	384	683	1,536	6,147
15% or 85%	*34*	*48*	76	195	306	544	1,224	4,898
10% or 90%	*24*	*35*	54	138	216	384	864	3,457
5% or 95%	*12*	*18*	*28*	72	114	202	456	1,824

If you are expecting non-response or a difficulty in locating your sample selections then it is prudent to over sample to ensure that the sample size achieved provides the required level of precision.

The figures in ***bold and italics*** denote sample sizes of less than the recommended minimum.

Figure 16. Required sample size based on the percentage representation of the characteristic of interest, level of precision and, confidence interval (Source: National Audit Office – Statistical & Technical Team – A practical Guide to Sampling).

- Fields included within the sample extract were UPRN, AddressBase Premium Residential Characteristics for Dwelling Type and, the relevant address fields: ParentUPRN, UDPRN, SubBuilding, BuildingName, BuildingNumber, Thoroughfare, PostTown and Postcode. The provided address was based on the Delivery Point Address, which is the most spatially accurate, where the DPA was missing then the PAO/SAO was provided. Each address was flagged to indicate whether it was DPA/PAO based. Pivot table analysis of the proportion of Houses and Flats in the sample was undertaken and compared to the latest known overall dwelling stock dwelling Type data – this was to ensure that the sample was representative of all dwellings overall.
- A concatenate address for each dwelling, based on the same criteria as applied in the PYS master address files, was created. Addresses were removed where it was determined as new build via either, for example the address supplied was SAO and began with a plot reference e.g. Plot 67, Glebe Street, or cross comparison to the PYS cohort determined it as such.
- Based on the address postcodes from the sample, relevant mainstream and special school pupil records were extracted from the most recent January School Census return. The pupil cohort was included on the same selection criteria as for inclusion within the PYS. Extract fields included the UPN as a

be as accurate as the AddressBase Premium dataset. Using this method only not all the NSPL Hertfordshire postcodes would be successfully merged to the OS Codepoint Polygon layer, resulting in less postcodes being identified within Hertfordshire. Given unsuccessfully matched postcodes (which are contained within the Vertical Streets dataset) tend to contain flats, this would result in a smaller number of flats than expected. The expanded methodology applied resulted in 302 additional postcodes being included which were predominantly flatted areas.

unique identifier, National Curriculum Year Group and, all relevant address fields. An additional flag was provided to determine whether the record was “Mainstream” or “Special School”.

- Pupil address records were cleansed and matched to UPRN. The dwelling sample and pupil records were then cross referenced based on UPRN and mainstream/special school counts allocated to each dwelling where matches occurred. An aggregate Units Only, Houses Only, Flats Only yield per 100 dwellings LTA for N2, Primary, Secondary, Post-16 and Special School cohorts was calculated.

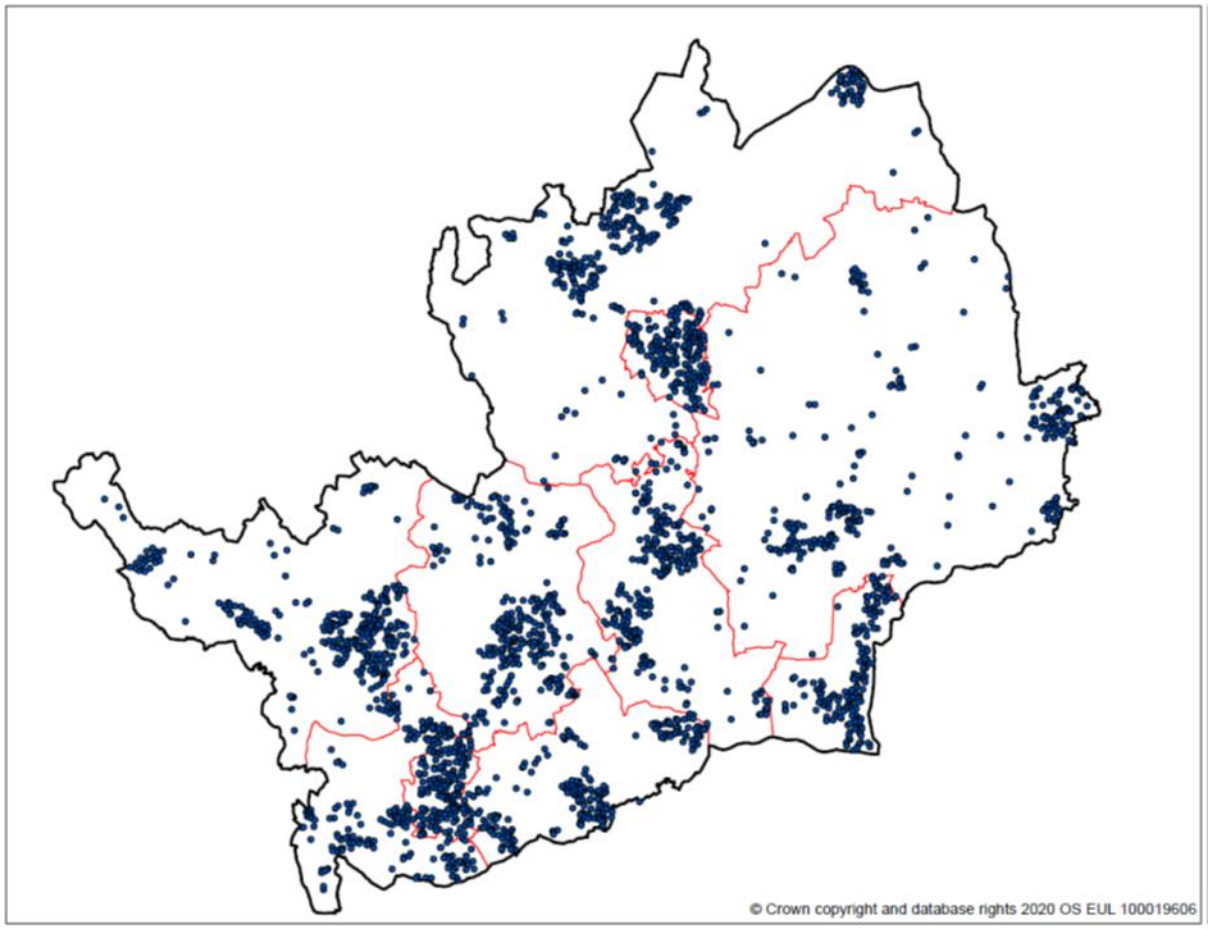


Figure 17. The location of randomly selected dwellings (n = 2,641) within the boundary of the authority based on the applied method.

Table 6. Sample derived LTA mainstream sector yields per 100 dwellings at Units Only, House Only and, Flats Only (2019).

	Dwellings	N2	PRIMARY	SECONDARY	POST-16
UNITS ONLY	2525	36	472	303	71
	Yield per 100	***1.4***	***18.7***	***12.0***	***2.8***
HOUSES	1954	30	425	286	69
	Yield per 100	***1.5***	***21.8***	***14.6***	***3.5***
FLATS	571	6	47	17	2
	Yield per 100	***1.1***	***8.2***	***3.0***	***0.4***

GIS determined that of the overall dwelling stock in the authority 77.7% were houses and 22.3% flats, the sample type mix was 77.6% houses and 22.4% flats and was considered representative of the wider population. Figure 17 displays the scatter of randomly selected dwellings within the boundary of the authority. However, of the 2,641 sample dwellings 7 were removed from the cohort as being identified as new build and a further 109 dwellings removed as the residential dwelling characteristic was “RD” and no dwelling Type distinction was available. The resulting cohort size was 2,525 dwellings. Matching to the January School Census data sets 2019 determined the LTA point estimates shown in Table 6.

As a sense check the Units Only sample observed point estimate 2019 mainstream LTA uptake was 18.7, this is 1 per 100 dwellings less than that determined using the overall authority resident mainstream primary pupil counts and the estimated total number of dwelling units. However, whilst the sample based dwelling units only yield was observed to be 18.7 per 100 dwellings the true population mean will be $\pm$ 2% and therefore within the range 16.7 to 20.7 per 100 dwellings at primary. This encompasses the 19.7 per 100 dwellings determined previously from the overall population and is suggestive that the true population mean may lie toward the upper end of the range.

The sample determined secondary age yield per 100 dwellings (Units Only) was 12.0 (range 10 to 14), this compares to the overall population 60,688 mainstream secondary age pupils / 495,335 dwelling 2019 = 12.3 per 100 dwellings. In all mainstream sectors the LTA yield per 100 dwellings associated with Houses was higher than that for Flats although the differences were observably less for N2 cohorts. For primary, secondary and Post-16 cohorts the differences between LTA values between houses and flats were generally a factor of 3, 5 and 9.

Differences in yield are likely to have occurred due to variance in bed size mix, at Units Only between the sample and that observed applying official population estimates and overall dwelling stock numbers. Currently bed size data to specific dwelling is unknown and the authority is dependent upon DfE clarification of national approach in conjunction with the OS in order to resolve this issue. It is a further complication that this process is specific to the education sector only, there are other contributions sought from developers which require different demographic counts, such as overall population for libraries and waste disposal. Extension would therefore be required to determine availability of dwelling level data sets, such as the electoral role, which the authority can apply to robustly estimate these metrics. However, all variables resulting from sample would have a yield range within which the true population mean would lie.

Narrowing of the yield range will require a substantially larger sample size than included herein. For example, a confidence interval of 1% requires a four-fold increase in sample size from circa 2,600 to 10,000 dwellings. Further increases, or boosting, of sample sizes may be required in order to attain the confidence interval when considering specific dwelling type and bed size counts e.g. 3-bed flats which are generally low in number within the authority. It was also observed that the preliminary dwelling cohort size of 2,600 dwellings was insufficient to reliably

determine special school yields by sector. It is likely that this would be resolved through a larger dwelling cohort.

14.6 Recommended LTA methodology

The preceding examination has determined several methods by which LTA values can be annually updated, each have their strengths and weaknesses. They are all currently limited in relation to known dwelling bed size, in aggregate or individually, and for which the authority would be dependent on DfE recommendations to further resolve. Such information will be required in order to satisfy the requirements of Regulation 122. Once bed size and tenure data are available then it is apparent that the postcode method would have the greatest reliability and usefulness in updating LTA values for dwelling units and type-based analysis. In addition to having the smallest confidence intervals associated with LTA estimates the process also ties in well with the availability of postcode level GP registrations data sets for Early Years cohorts as applied within the DfE approved school place planning forecast. Further work is required on the suitability of electoral register data to create dwelling type specific adult cohort population estimates. This is likely to be a substantial piece of work and will require cross comparison to official population estimates such as the ONS MYE or SNPP to validate results within defined error bands.

15.0 Determining the typology of developments identified within the pupil yield study

Provisional DfE guidance indicates that examination of developments may determine variables which are similar such that permissions could be grouped into clusters of distinct characteristics which typify specific typologies. Distinct typologies could provide a more accurate assessment of development average yield for application to the estimation of likely mainstream yield at the Local Plan stage with consideration of Regulation 122.

Determining the typology of specific developments included within the annual cohorts of the Pupil Yield Study is an emerging methodology currently being trialed with three annual cohorts. The process is dependent upon three stages for larger proposals in the >=30 dwellings cohorts. Smaller developments in the >=30 dwellings cohorts, and generally all developments in the >=10 to <30 dwellings cohorts, were dependent on stages 2 and 3 only. This occurred as smaller developments were generally not included in district local plans and no relevant data would thereby be included for Stage 1 assessment.

The determination of development typology is independent of any observed pupil yield arising from a development and as such is a “blind study” based predominantly on local plan information and data as a proposal comes forward to planning application stage. Prior to conducting a typology assessment, maps specific to each development were produced by GIS and the location of the site, specific to the district within which it occurred, also identified. Whilst the method applied was to determine the typology of development retrospectively it should be noted that it could also be applied to sites in emerging Local Plans. This would provide an indication of the typology of forthcoming developments.

15.1 Stage 1: The Local Plan

The district within which larger sites, contained in an annual cohort, occurred was identified based on data already collated from the PYS master cohort data files. The annual cohort year of inclusion, for example 2009_2010, determines the year in which residential completions began to be produced by each development. This in turn provided an indication as to the Local Plan period within which each development might be found. Sites included within a Local Plan are those which have either been promoted by the District or identified within a “call for sites” and, can be included within Housing & Economic Land Area Availability (HELAA) or SHLAA documentation.

Where developments were identified as being within this documentation then District produced data relevant to these proposals was applied in determining typology. This level of analysis not only provides information relating to a specific site but also contextual data with respect to other developments within the area which may have occurred in a similar time period. Where sites are part of a wider development strategy within a local area then a more accurate assessment of typology may result due to consideration of the broader picture. Some development specific data may also be more accurate than that held elsewhere, for example the proposed dwelling density (dwellings per hectare) and the site area which, spatially, would be highly accurate.

15.2 Stage 2: Key characteristics

Key characteristics for each development are recorded without reference to pupil yield outputs, data items include:

- Number of dwelling units (note that number of dwelling units by type and bed size is unlikely to available at this stage, this would be included in Stage 3).
- Previous land use, of which historical plans (which show the parcel of land), individual planning applications and, Google images (to 2008) can form part of the assessment. There is also a Previously Developed Land (PDL) flag within the PYS data sets.
- ONS Rural Urban Classification (RUC) determined based on assigning the development postcode areas to Census Output Areas using the ONS National Statistics Postcode Locator (NSPL), these are then looked-up against the OA based RUC.
- Housing Density based on the development number of dwelling units divided by the polygon defined site area.
- Build trajectory (provisional assessment of the likely average number of dwelling completions per annum).

This information is used to indicate a provisional typology based on objective, pre-construction stage data.

15.3 Stage 3: Post-development retrospective data

Information within Stage 3 is that derived from SMART Herts during the PYS development data gathering already undertaken prior to the typology process, the latter being the final step in the procedure. Variables determined were:

- The development bed-size mix, particularly the balance/percentage of 3+ bed size dwellings relative to 1/2-bed size dwellings.
- The dominant development dwelling type such as Houses/Flats/Mixed. Where a development is >60% Houses then the dominant type is Houses, where Flats are >60% then the dominant type is Flats, where the type mix is between 60%/40% or 40%/60% then the dominant type is Mixed.

This information may have been available from LPAs at the local plan consultation stage in more general terms. The more specific information determined as a proposal came forward for development, as included within Stage 3, was used to check whether the provisional typology from Stage 2 should be amended. Following the three-stage assessment the final determined typology was recorded against the development unique reference code (PPREF) in conjunction with all data items from each stage.

15.4 Emerging tier classifications

Initial typology classification was derived from the PYS trial cohort. This provisional assessment resulted in the methodology applied herein and the emerging classification of developments as follows:

- Tier 1, 1FE primary per 400 dwellings: *These sites are typically greenfield sites with a dominance of houses (typically 80/20), a higher proportion of 3+ bed properties, and a higher proportion of detached or semi-detached. There tends to be a housing unit density of 22 to 40 per hectare (dph).*
- Tier 2, IFE primary per 500 dwellings: *These sites are typically PDL with a mix of houses and flats, and a higher proportion of terraced, maisonettes or flats. There is generally a 50/50 Split between smaller (1 & 2-bed) and larger (3-bed+) family homes and houses are most likely to be terraced. There tends to be a housing unit density of 40 to 60 per hectare (dph).*
- Tier 3, 1FE per 1,000 dwellings: *These sites are typically PDL with a dominance of 1-2 bed properties and are solely flatted (or at least >75% of) developments. There tends to be a housing unit density of >=60 per hectare (75 to 100 is quite common).*

The assessment of developments within the overall PYS will provide more substantive evidence to support and further refine the initial classifications and structure.

Spatial Planning do not use HELAA information but do classify permissions by their origin, where applicable, to SHLAA sites. Spatial Planning is currently in the process of collating and updating this information as an ongoing piece of work which will be completed at some point in the future. As PYS typology assessment work continues developments will be flagged if they have a SHLAA association, these records will then be cross matched to Spatial Planning datasets once their work is completed.

16.0 Limitations of the PYS methodology

There exist a few limitations in the methodology applied to determining mainstream and special school pupil yields from new build developments in the boundary of Hertfordshire County Council. The principle limitation relates to individual dwelling

bed size & tenure data whilst others relate to the transition to LTA for all sectors and data availability for Post-16 cohorts.

Whilst SMART Herts data sets have enabled the determination of overall type, bed size and tenure for each development it has not been possible to disaggregate to individual dwelling level in most instances. An exception to this was generally the smaller developments in the >=10 to <30 dwellings cohorts. Many of these smaller developments are of singular bed size and tenure and as such this could be cross referenced to the UPRN of the identified dwellings. In aggregate, it is estimated that 25% of the 50,000+ dwellings included in the study currently have specific bed size and tenure data.

The absence of individual dwelling bed size and tenure data does not prevent the determination of mainstream yield from an annual cohort of an overall known bed size and tenure mix, for either dwelling units only or by dwelling type. Such analysis is required by the DfE for determining likely average mainstream yields by development characteristics. However, it does cause substantial difficulties for modelling specific proposals, which may differ substantially to that observed from aggregate cohorts, as they come forward through the planning process.

Where proposals are substantially different to the “average” applied at Local Plan Stage then variance in yields from the norm will occur. For example, the trial PYS indicated an overall 17% contribution of Affordable Rented/Social Rented dwellings to the mix. Proposals currently coming forward within the authority may have a 40% representation of these tenures and as such yields would be anticipated to be substantially higher than “on average”. Specific site modelling based on a proposed type, bed size and tenure mix, using the Pupil Yield Study base data for a homogenous approach, is dependent on knowledge of individual bed size and tenure single dwelling average yields.

Hertfordshire County Council has approached both the ONS Census Team (for dwellings current as at the 2011 census) and HMRC to request access to individual dwelling bed size and tenure data. However, initial requests have been declined. Discussion with senior analysts in the DfE indicated that it has had similar problems in obtaining this data to individual dwelling level. The DfE implemented a project with Ordnance Survey (OS) to resolve these issues but outcomes have yet to be shared. Similarly, HCC approached GeoPlace however it became apparent that this information is not recorded within Local Land and Property Gazetteers. Review of district held council tax registers also determined that this data was not recorded against individual dwelling data by council tax band. Following discussion with the Information Commissioners Office (ICO) the authority is likely to submit a Freedom of Information (FoI) request to HMRC. Whilst it is not expected that the response will differ substantially to that initially received it is a requirement of the FoI Act that HMRC provided a detailed response as to why the data cannot be released. The ICO can then review this response, against the statutory requirements of the authority, and make a judgement as to whether access should be granted.

In relation to the transition to LTA the authority has generally observed that only those developments in the 2002 through to 2011 annual cohorts have attained peak at primary whilst, secondary peaks were only observed for developments in the 2002

to 2006 cohorts. In the latter case these developments peaked some 14 to 16 years following completion. Whilst change from peak to commencement of transition to LTA can be observed in some of the early annual cohort no singular cohort, either at primary or secondary, has been observed to attain LTA. It is likely that at least a further four years of longitudinal study will be required to observed attainment of LTA values for the earliest cohort. Consequently, LTA values applied in modelling will be of a best-estimate basis as outlined within Section 13 as opposed to “on the ground” observation from the PYS. This will also directly affect Post-16 peak and LTA values. An additional factor relating to the Post-16 cohort is that the authority can currently only include mainstream school cohorts. HCC does not have access to data returns from Further Education Colleges, these institutions return direct to the DfE.

. As a result of the current exclusion of this specific cohort then yields observed within the authority PYS at Post-16 will be less than the actual number of children resident in this age group. This will impact not only the accumulation of Post-16 cohorts to peak but also the yields at, and time at, peak in conjunction with transition to, and attainment of, LTA.

17.0 Provisional cohort sizes

In total 1,190 developments >=10 dwellings in size containing 55,470 dwelling units were identified for possible inclusion in the Pupil Yield Study 2002 through to 2020. 114 developments were identified for exclusion as either being C2/C4 or, a permission reference was determined to be part of a larger permission and subsequently concatenated (Table 7).

Table 7. The combined >=10 to <30 and >=30 development cohorts per annum.

Annual Cohort	*No: Devs*	*No: Dwellings*	*No: Devs Excluded*	*No: Dwellings Excluded*	*Devs in Cohort*	*Dwellings in Cohort*
2002_2003	70	3,138	5	148	65	2,990
2003_2004	35	1,901	4	152	31	1,749
2004_2005	75	3,969	6	247	69	3,722
2005_2006	73	3,400	8	258	65	3,142
2006_2007	76	3,288	5	317	71	2,971
2007_2008	85	2,804	2	32	83	2,772
2008_2009	75	3,258	10	171	65	3,087
2009_2010	53	3,472	2	68	51	3,404
2010_2011	44	2,075	3	33	41	2,042
2011_2012	62	3,241	6	225	56	3,016
2012_2013	50	2,024	6	262	44	1,762
2013_2014	46	1,628	6	206	40	1,422
2014_2015	57	1,904	4	171	53	1,733
2015_2016	72	3,580	6	231	66	3,349
2016_2017	77	3,065	7	174	70	2,891
2017_2018	71	3,032	8	221	63	2,811
2018_2019	90	5,868	11	632	79	5,236
2019_2020	79	3,823	15	443	64	3,380

TOTAL	**1,190**	**55,470**	**114**	**3,991**	**1,076**	**51,479**

It can be provisionally indicated that the PYS will therefore include 1,076 developments containing 51,479 dwellings constructed within the boundary of the authority in the period 2002 to 2020. However, the most recent four annual cohorts have yet to be fully finalised either because: SMART Herts residential completions and size_type data sets 2020_2021 onwards will be required to complete the data sets or; there are complex sites for which estate files are being used to resolve. It is likely that the 2016_2017 and 2017_2018 developments will be fully resolved in forthcoming months whereas parts of the 2018_2019 and 2019_2020 may need to be reserved whilst most of their cohorts are processed.

18.0 References

Hertfordshire County Council. January 2008. Planning Obligations Guidance – Toolkit for Hertfordshire: Hertfordshire County Council's requirements.

Hollis, J. 2005. Data Management and Analysis Group – Child Yield. DMAG Briefing 2005/25 August 2005.

Walker, J. 2006. Child Yield Estimates for School Roll Forecasting. BURISA 169, September 2006, pp. 4 – 7.

Bath & North East Somerset. July 2009. Planning Obligations: *Supplementary Planning Document*.

Bracknell Forest Borough Council. July 2007. Limiting the Impact of Development: Supplementary Planning Document.

Buckinghamshire County Council. *Unknown*. Children and Young People's Service: Guidance on Planning Obligations for Education Provision.

Burton, M. 2009. Proposed Development of Land at Dunsfold Aerodrome: Proof of Evidence – Surrey County Council.

Cambridge County Council. 5th March 2009. Cambridge Horizons – Agenda Item No: 7 Revised Pupil Forecasts.

Carmarthenshire County Council. 2007. Appendix 1 – Education and Children's Services Department Planning Guidance Relating to Section 106 Obligations.

Central Bedfordshire Council. November 2009. Planning Obligations Supplementary Planning Document (SPD): Background paper – Methodology used for calculating Standard Charges and Costs. Annex 1.

Conwy County Borough Council. 2010. Educational Facilities.

Cumbria County Council. 2011. Development Contributions to Education Capacity and Other Related Essential Infrastructure: Draft.

Department for Education. 2014. *Area Guidelines for Mainstream Schools: Building Bulletin 103*. Department for Education.

Department for Education. 2017. *Local Authority Interactive Tool – User Guide*. Department for Education. September 2017. Accessed Online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/643132/LAIT_User_Guide_2017.pdf. Accessed On: 2nd January 2019.

EMIE / NfER 2006. Pupil Forecasting One Year On: Report of two EMIE / NfER practitioner workshops – Friday 17th November 2006.

Essex County Council. 2010. EssexWorks: Developers' Guide to Infrastructure Contributions 2010 Edition.

Greater London Authority (GLA). 2005. Child Yield. Data Management and Analysis Group (DMAG) Briefing 2005 / 25 – August 2005.

Greater London Authority (GLA). 2006. Child Occupancy of New Social Housing. Data Management and Analysis Group (DMAG) Update 2006 / 11 – May 2006.

Hampshire County Council. 2010. Building Schools for the Future – Strategy for Change: Phase 1, Havant and Horndean.

Hampshire County Council. September 2010. Developers' Contributions towards Children's Services Facilities.

Hollis, J. 2005. *Data Management and Analysis Group – Child Yield*. DMAG Briefing 2005/25 August 2005.

Lancashire County Council. November 2011. Planning Obligations in Lancashire – Contributions towards education places (Appendix A).

London Borough of Barking and Dagenham. December 2011. Draft Community Infrastructure Plan 2012 – 2025.

Luton Borough Council. 2007. Planning Obligations – Supplementary Planning Document – September 2007.

Milton Keynes County Council. 2008. Current Education Provision Figures (07/03/08) – Contributions Model for 2008: Addendum – Replaces Appendix 2.

Mole Valley District Council. 2006. Surrey Planning Collaboration Project 2006- Mole Valley District Council: Planning Obligations and Infrastructure Provision Code of Practice – Basis for Calculating Formulae and Standard Charges.

Reading Borough Council. 2004. Supplementary Planning Guidance on Planning Obligations - Education Facility Costs and Requirements.

Reigate and Banstead Borough Council. 2008. Planning Obligations and Infrastructure – Supplementary Planning Document.

Rockwell, J., Vargas, N., Vanis, C. and, Chou, S. 2005. Evaluation of Child Yield Within Recently Completed Housing Developments in the Borough of Brent. Worcester Polytechnic Institute.

Surrey County Council. April 2007. Tariff and Surrey Education Formula.

Trafford Council. February 2011. Local Development Framework: Draft SPD1: Planning Obligations. Technical Note 6: Meeting Social Needs.

Walker, J. 2006. Child Yield Estimates for School Roll Forecasting. *BURISA 169*, September 2006, pp. 4 – 7.

West Berkshire Council. May 2010. Delivering Investment from Sustainable Development: Topic Paper 3 – Education.

Wokingham Borough Council. 2010. Planning Advice Note: Infrastructure Impact Mitigation – Contributions for New Development. Revised.