

Census 2001 Review and Evaluation

Census Geography: Evaluation Report

April 2004

Content	Page
Project Objectives.....	2
Background.....	2
Methodology.....	2
Assessment and Lessons Learnt.....	7
Conclusions.....	9

ONS is carrying out a review and evaluation of the 2001 Census in England and Wales which will culminate in a Data Quality report and a General Report being published.

Plans for individual reports on specific aspects of the Census operation and a timetable for release have been published.

Each report is written in isolation and is subject to amendments as processing progresses and further information comes to light.

Reports will be released on the ONS website in the form of a high level Executive Summary and a more detailed Evaluation Report.

Census Customer Services
ONS
Titchfield
Fareham
Hants PO15 5RR

Telephone: ++44 (0) 1329 813800
Fax: ++44 (0) 1329 813587
Minicom: ++44 (0) 1329 813669
E-mail: census.customerservices@ons.gov.uk
Website: www.statistics.gov.uk/census2001

Census 2001 Review and Evaluation

Project Objective

To provide and maintain an integrated, accurate and timely geographical service to all aspects of the 2001 Census, including geographical information, advice and support.

To minimise data redundancy and duplication of effort by establishing and maintaining a central database of census geography information and ensure, as far as is cost-effective and practical, that strategies for meeting census geography requirements are implemented in an integrated manner.

This review only covers England and Wales but the project took on a coordination role with the General Register Office for Scotland (GROS) and the Northern Ireland Statistical Research Agency (NISRA) to ensure that systems which made use of geography data took account of each country's requirements. A UK Geography Coordination Group was established to integrate each country's census geography data into the UK Geography Database. The group proved to be a successful vehicle for the resolution of any detailed technical issues, quality assurance of specifications, the establishment of UK geography milestones and the integration of Scotland or Northern Ireland only geography requirements.

Background

Conducting a census is a huge and complex task and a key part of the programme of planning and executing the Census is providing the geography that underpins the whole operation. Experience from the 1991 Census together with extensive research and testing during the preceding decade had shown that a more integrated and automated approach to providing the geography was required for 2001.

Census Geography was divided into three main, but closely linked phases, and this review covers these phases, describes how the work was carried out, the outcomes in broad terms and the lessons learnt.

The three phases can be summarised as:

- the creation of enumeration areas (known as EDs) which facilitated the efficient and accurate distribution and collection of census forms by enumerators, whilst attempting to equalise the workloads;
- ensuring a central geography database and systems were in place to assist in the processing of census forms by resolving queries and ensuring that all records had an accurate and valid postcode and grid reference assigned to them; and
- the creation of a separate geography for output, based on grouping postcodes together and using them as the building bricks for the release of census statistics.

Methodology

The main change for the 2001 Census was the introduction of major automation in the design of EDs and subsequently in the creation of Output Areas (OAs). Advances in affordable computer power, Geography Information System (GIS) software and digital data, from the national mapping agency Ordnance Survey (OS) made this possible.

The use of GIS at the front end allowed the EDs to be separate from the areas used for the release of census output. This was a crucial and innovative advance. EDs and output areas serve two very different purposes and previously the same building block had to be used for both purposes, which gave a sub-optimal solution for either enumeration or output.

Enumeration Area Planning

The methods used to plan EDs for the 1991 Census were mainly manual and very labour intensive and provided less than optimum maps showing the areas that enumerators were responsible for. The decision was taken very early in the 1990s to replace the methodology with a more automated system.

An extensive research and development programme was carried out and resulted in the Geography Area Planning System (GAPS) being developed and used successfully to plan the EDs for the Census Test, held in June 1997. The performance of GAPS was fully evaluated after the test, including a number of field follow up visits by HQ staff to assess the accuracy and address coverage of the areas planned. This confirmed

Census 2001 Review and Evaluation

that GAPS had performed well and it was subsequently used to plan the EDs for the Dress Rehearsal in April 1999, and very soon afterwards for the 2001 Census. GAPS effectively automated the majority of the tasks and allowed the operation to be completed in a shorter timescale with a tenth of the staff.

GAPS was developed jointly with a private sector company ESRI (UK) (the GIS software supplier). The supplier was selected after a competitive European Union tendering exercise and contained the novel requirement that the supplier used two Office for National Statistics (ONS) staff as part of the development team. Once the initial development phase was completed the in-house ONS staff carried out further enhancements.

In summary, GAPS used digital representations of 1991 Census EDs (provided by the ED-Line product) as the starting point and various OS products, ADDRESS-POINT (AP) (gazetteer of addresses, each address having a grid reference, postcode and other attributes) and 1:10000 scale raster (maps held as graphic images within a computer), to enable EDs for the 2001 Census to be created. The system ran on a network of UNIX and Windows NT computers running suitably customised ARC/INFO GIS software.

The 1991 Census ED identity was automatically added to each address on the AP database, by a simple computer process of identifying all AP one metre grid references falling within the ED-Line polygons. This allowed revised counts of addresses for 1991 EDs to be established.

Local Authorities (LAs) were approached and asked to provide details about any housing developments of ten or more properties likely to be completed by April 2001 together with details of all demolitions. Every LA provided the information requested and these counts were added to those from AP to arrive at an estimated current number of households for each 1991 ED as at 29 April 2001.

To try and equalise the time needed to carry out the enumeration the number of households were reduced in the more difficult areas. These ED difficulty factors recognised included:

- the number of non residential addresses, sourced from AP;
- the percentage of multi-occupied addresses (such as bedsits), sourced from the 1991 Census;
- the percentage of households whose first language was not English, sourced from the 1991 Census;
- the number of persons in communal establishments, sourced from the 1991 Census and other external sources such as the English Tourist Board (hotels, guesthouses etc.), MoD (armed forces camps); and
- the physical size of the ED calculated by GAPS.

The table below shows the factors together with the 'score' allocated.

Criteria	Factor	Score
Non-residential (No. of addresses per ED)	49 or less	0
	50-99	1
	100-149	2
	150-199	3
	200 or more	4
Multi-Occupancy (Percentage of addresses per ED)	24.99% or less	0
	25-44.99%	2
	45% or more	3
Language (Percentage of heads of household)	9.99% or less	0
	10 - 24.99%	1
	25-49.99%	2
	50% or more	3
Persons in Comm Estabs (Total No. of persons per ED)	99 or less	0
	100-199	1
	200-399	2
	400 or more	3
Area size (Size of ED in Km ²)	<1Km ²	0
	1-5Km ²	3
	5Km ² or more	6

Census 2001 Review and Evaluation

The scores allocated because of any difficulty factors present were then added together to arrive at the final score for the ED. A range of households was deemed acceptable for each score and Enumerators were paid extra for any households enumerated above the standard number.

Score	Standard number	Range
0	210	170 - 250
1	190	160 - 240
2	160	135 - 180
3	145	120 - 160
4	130	80 - 150
5	110	70 - 140
6 or more	100	50 - 120

Acceptable planning ranges etc were built into GAPS and this allowed a first cut of 2001 EDs, based on 1991 ED boundaries, to be produced. Many EDs remained the same but the system highlighted EDs that required amendment.

Redesigning of any EDs necessary to meet 2001 planning criteria was accomplished by identifying 'blocks' of addresses to be moved between EDs. This was achieved by digitising, on screen, a temporary polygon and interrogating the AP database to find the number of addresses present.

The system removed the need to create 1991 Census style map folios and the manual overlaying of statutory boundaries - raster maps and Boundary-Line automatically achieved this. Only changed EDs required new boundaries and these were drawn, interactively on the computer screen. An automatic by-product of the system was the creation of digital boundaries for 2001 Census EDs. Map collation type work was carried out within the system and customised colour maps for Enumerators produced as single sheets (at mainly A4 or A3 size although maps as large as A0 were produced). Larger colour plotters also produced composite maps for use by Census District Managers (CDMs) and Team Leaders.

Enumerators were also provided, from AP, with lists of addresses within their EDs preprinted into their enumeration record books (ERBs) to supplement the 1:

10000 scale maps. In areas of major new development the 1:10000 scale maps were also supplemented by larger 1:1250/2500 scale vector maps.

EDs were created to nest within LAs, electoral wards and parishes (communities in Wales) and were then grouped together to form Census Districts (CDs), a 'management area' which was the responsibility of a CDM. The size of CDs were also reduced where they contained concentrations of difficult EDs and ranged in size from 36 to 66 EDs. In a number of very difficult inner city areas the size of CDs was reduced further and some contained fewer than 20 EDs.

GAPS produced four 'profiles', which described the areas that field staff were responsible for:

- Census District Summary (GEO D 1) listed the area each CDM was responsible for together with the workloads and ED numbers, number of households, non residential properties, persons in communal establishments and a final column which allowed CDMs to add the enumerators names.
- Special Enumeration District Summary (GEO D 2) listed information about large communal establishments (100 or more persons) which were subjected to special enumeration arrangements.
- Census District Check Report (GEO D 3) used by CDMs to report the findings of their pre-enumeration check (covered in more detail later).
- Enumeration District Profile (GEO E 2) listed information for enumerators about their area such as the difficulty factors, the estimated number of households, likely new developments together with space for the CDM to add any additional information thought relevant.

Part of the CDM's duties were to carry out a check of their area using the material generated by GAPS, maps, ERBs and profiles and to report back to HQ any changes thought necessary. These were then assessed, the required changes made and the material re-issued as appropriate.

Census 2001 Review and Evaluation

Processing

A final major by product of GAPS was the creation of a centralised geography database.

The centralised geography database was used extensively in the 2001 Census both during the enumeration period and afterwards during the processing of the data and in the creation of OAs.

During the enumeration phase the database itself or extracts from it were used by the payroll contractor to populate their systems with details of areas and field staff numbers whilst the public enquiry telephone contractor used it in a similar manner. The processing contractor used extracts of the database to control many aspects of their work, as did the in house ONS processing areas.

The database was also used extensively within the Census geography area during the processing of census forms to check the validity of postcodes and allow grid references to be added. Postcode and address queries raised during this process were passed to the geography team for resolution and various automated and manual systems (collectively known as the LOckheed Martin Address System (LOMAS)) were used to ensure the final census database contained accurate and valid postcodes and associated grid references.

The processing contractor supplied three types of addresses to geography for checking:

- Address of enumeration – these related to addresses that were not pre-listed into the enumeration records books at the ED planning stage but had been added by enumerators. The contractor may have found and added a valid postcode but they would not have had a grid reference assigned.
- Address one year ago – these were similar to address of enumeration but often only a partial postcode was present and again no grid reference had been assigned.
- Workplace address – the most difficult type of address to find and often the postcode assigned related to a PO box number.

LOMAS attempted to automatically match (on postcode and address) addresses that had been supplied against addresses held on the geography database on either the original address table (used during the planning of EDs) or an 'updated' address table that had been created from updates to ADDRESS-POINT that had been received from OS, half yearly since April 1999. If a successful match was made then the postcode, address and grid reference were added to the data. If a match was not found then the addresses were flagged accordingly and passed through for manual referencing. Manual referencing of addresses of enumeration entailed physically finding the addresses on large-scale digital OS maps and interactively adding a grid reference and validating, or finding a postcode from the Royal Mail Address Manager product. Manual referencing of the other two address types mainly entailed finding a valid postcode from the Royal Mail Address Manager product and adding the postcode centroid grid reference.

Due to the number of addresses passed to geography for validation (approaching 2 million) it became necessary to amend LOMAS to automate it further and thereby reduce the manual effort. This was done in a systematic way that ensured the quality and integrity of the data was maintained and bias was not introduced. Once the validity of a postcode had been established, the addresses were assigned a grid reference from other valid addresses within the postcode via a routine which ensured an even spread across the whole postcode.

Further manual checks were made on workplace postcodes that contained a large number of workers to ensure they were assigned the correct grid reference and this included, wherever possible, postcodes that related to PO box numbers.

The successful completion of LOMAS processing meant that the geography database then contained an address table which reflected the position at the start of the enumeration period and a 'formid' table that reflected the position at the end of processing. The formid table contained the identities of all the Census forms processed, together with a postcode of enumeration and an associated 'fit for purpose' grid reference. In addition, a series of further geographic codes for other areas, such as health areas, parliamentary constituencies and national parks were added.

Census 2001 Review and Evaluation

The Census form identities present on the formid table allowed direct linkage with the main census database and permitted the updating of postcode and grid reference data. Updated/corrected postcode information for address one year ago and workplaces from LOMAS were also updated on the main census database in readiness for the output phase of the Census.

Output Geography

Another first for 2001 and what many consider to be a world leading operation, was the separation of the collection geography from the output geography, again facilitated by the combined use of a GIS, digital boundaries and the fully grid referenced census database. An automated zoning system for output purposes (Output Area Production System (OAPS)) was developed jointly with Professor David Martin, from Southampton University, whose expertise was invaluable in developing theory into a working system.

The separation of the collection and output geographies was a major advance over previous censuses as it meant that the output building blocks could be optimised for output purposes and therefore better able to meet the requirements of users.

In summary OAPS used the data, postcodes and grid references, present on the formid table to create postcode polygons by automatically drawing lines midway between pairs of adjacent points and then dissolving the internal lines to create 'thiessen' polygons around each postcode. This process was repeated for each ward and parish.

As the boundaries were computer generated they were fairly random in nature and, to make them more useable, they were subjected to a tidying up exercise by being snapped to 2001 Census ED boundaries and road centre lines present in the OS product OSCAR. These tidied up boundaries were then used to automatically build the Output Areas (OAs).

The creation of OAs then involved two further main processes of grouping the postcodes together and then making adjustments to improve them. Built into OAPS were a series of constraints that were used to fine tune the postcode groupings to produce a consistent set of

areas across the whole of England and Wales. Some of these constraints were hard (must follow) and some were soft (try to follow).

The hard constraints were:

- contiguity – all postcodes must be adjacent to each other and must not cross ward or parish boundaries; and
- thresholds – to preserve the confidentiality of the data all OAs had to contain at least 100 persons and 40 households.

The soft constraints were:

- size – ideally each OA should contain 125 households;
- shape – avoid boundaries that were too irregularly shaped;
- homogeneity – degree of sameness, for which tenure of household was used after research showed this was the best indicator; and
- urban/rural – each postcode was assigned an urban or rural indicator based on whether it fell wholly or partially within the Office of the Deputy Prime Minister/OS defined urban areas and like types were kept together if at all possible.

Actual 2001 Census postcode population and household counts were used to decide how many OAs to create, and an initial grouping was produced. After the initial grouping of postcodes was completed a series of iterative postcode re-combinations took place until a 'best' solution was arrived at. This process happened thousands of times, improving OAs and then starting again somewhere else until it had done its best. No solution produced was perfect as the constraints often pulled in opposite directions.

Once the OAs had been finalised they were allocated a unique code, which consisted of a county code, district code, ward code and sequence number within the ward. This code together with other attribute data such as, geometric and population weighted centroids and area size formed a further database. An automatic by product was a set of digital OA boundaries, which were made available to users.

Census 2001 Review and Evaluation

The OA code was added to the formid table and subsequently to the Census database, which then allowed individual records to be aggregated to form OAs.

As other area codes had already been added to individual records, on the formid table, it was then possible to create a database of 'best fitted' OAs to these other geographies and provide census statistics for them whilst maintaining confidentiality by avoiding differencing problems. OAs that were split by other geographies were assigned ('best fitted') to the area containing the largest population.

Digital OA boundaries together with look up files, indexes and metadata were subsequently provided 'free at point of use' to users via Census Access.

Assessment and Lessons Learnt

Geography is crucial to the Census operation, being present in all stages. The successful development and implementation of the 2001 geographical system is all the more impressive given that it was a key innovation. Innovation always carries risks; in this case the risks were well managed and contained. The detailed lessons learnt during that process are described below.

Although in live running the geography systems worked well, overall, the geographical aspects of each stage of the Census process were not always considered sufficiently at the design and development stage. Geography therefore needs to assume a higher profile and become more integrated in the overall conduct and management of a future census. Geography 'experts' need to be involved, from the beginning, in the design of any systems which will make use of geography data.

System development was helped greatly by the use of integrated teams, with generalists and technical programming staff under the same management and this approach should be repeated. The more widespread use of the GIS and database software within ONS should make this easier to achieve.

Enumeration Area Planning

Overall this can be considered an overwhelming success with EDs being planned in an extremely cost effective way. Enumerators were also provided with better, more up to date maps and the labour intensive task of writing addresses into their record books was largely removed. GAPS was also flexible enough to allow late changes, identified by CDMs during their check, to be incorporated and revised materials supplied.

The planning of EDs started in April 1999 and was completed on 28 September 2000 with a total of 116,918 EDs planned. Running in parallel with the planning of EDs was the creation of the address lists (completed on 29 September 2000), the production of maps (completed on 16 October 2000) and the production of profiles (completed on 5 October 2000). Although completion was slightly later than originally planned it did not cause any knock on problems with later phases of the enumeration.

The principle lessons learnt can be summarised as follows:

- the specialised nature of GIS systems meant that difficulties were encountered in finding sufficient, suitable qualified staff to develop the systems needed and this must be addressed in any future exercise;
- the task of ensuring that boundary data sets were consistent was very time consuming and sufficient resources should be made available in future at an early stage; and
- the printing of field staff maps and documentation was very time consuming and contracts with external printers should be set up well in advance to ensure that the materials are available when needed.

Processing

Although far more records than originally planned were passed to geography for query resolution, the flexible nature of the systems meant that they could be adapted to cope and ensure that crucial end dates were still met.

The resolution of the queries started on 24 September 2001 and was completed on 26 June 2002. In all 1,930,614 queries were resolved, 1,629,643 automatically within LOMAS and the remaining 300,371 clerically.

Census 2001 Review and Evaluation

The principle lessons learnt can be summarised as follows:

- experience gained this time will prove invaluable in designing systems for any future census and should enable an even quicker turn around time for query resolution; and
- careful consideration must be given to the balance between the need for timeliness and the need for total accuracy against a 'fit for purpose' solution especially in regard to grid reference allocation, as this has a significant impact on time and resource.

Output Geography

The creation of OAs was another resounding success which has created building blocks ideally suited to the presentation of census statistics as they are smaller than enumeration areas and, being constructed from postcodes allow other data to be easily referenced to them. Initial reaction from customers has been favourable, seeming to meet most of their needs.

A few users have commented that they would have liked more input into the physical design of the OAs but this was not possible due to time constraints and the need to retain uniformity over the whole of England and Wales. Users were consulted about the principles governing the creation of output areas and were provided with some examples to illustrate the implementation of those principles. If development work was completed earlier in a future census, it would be possible to build on this aspect and allow a greater degree of involvement

Another indicator of success is the fact that OAs have been adopted as the single small area building brick for National Statistics and are the basis for "Super Output Areas" to be used in Neighbourhood Statistics (NeSS).

The planning of OAs started on 10 September 2002 and was completed on 9 December 2002. 175,434 OAs were created with some 37.5% lying between 120 and 129 households and 79.6% between 110 and 139 households. Some 5% lie between 40 households - the confidentiality threshold - and 99 households, and many of these are single small parishes. The distribution

of size by households has a single sharp peak, showing a consistency in size compared with the number of peaks which appeared in the distribution of 1991 Census EDs by size.

As the creation of OAs was an almost wholly automatic process the actual time taken to create them was relatively short but an equal amount of time was set aside to manually check for oddities and inconsistencies, especially those associated with spuriously split postcodes. A fully cleaned set of OA boundaries and associated attribute data was completed in late February 2003.

The principle lessons learnt can be summarised as follows:

- OAs were created in a very cost effective manner over a short space of time to be fit for the purpose of disseminating 2001 Census Statistics; it may be possible to refine the boundaries to integrate them with the Ordnance Survey Code-Point with Polygons product * and/or the likely replacement MasterMap, and it will be necessary to consider the impact changing topography on the boundaries in future;

*The Code-Point with Polygons product contains digital postcode boundaries created in a similar manner to those created by ONS but they were not used because of incompatibility with 2001 Census postcodes of enumeration and possible copyright issues.

- OAPS was basically developed within ONS by one person, despite strenuous efforts to find additional resources and this over-reliance on particular individuals must not be allowed to happen again; and
- changing to follow the new ONS Boundary Compliance Policy meant that OAs had to be created within wards in existence at the end of 2002, which caused major problems and delayed the production of OAs by 3-4 months. The timing of the introduction of new geography policies and the impact on outputs agreed with census users should in future be considered carefully at an early stage in the policy formulation.

Census 2001 Review and Evaluation

Conclusions

The moves from the predominantly manual, labour intensive work carried out in the 1991 Census to the automated methods used in the 2001 Census were completed successfully. The development partnership with Professor David Martin resulted in an output geography system that is viewed as a world leader amongst countries with a complex geographical structure. However, technology moves on and the development work required for any future census should not be under estimated.

Current systems should be retained and if at all possible ported across to the latest ONS corporate environment of Oracle 9i and ARC/INFO 8.3 to ensure that all, or as is more likely parts of them will be available for use during the intercensal period and form the starting point for the systems required for the 2011 Census.

The importance of geography throughout the Census operation needs to be recognised, and sufficient and suitable resources put in place at an early stage. To help achieve this there is a need to ensure a strategy is in place to keep key specialist/technical staff throughout the life of any future project.

Census 2001 Review and Evaluation

Census Topics	Target Dates for Release
Legislation	Published
Non-Compliance (Executive Summary Only)	Published
Data Needs	Published
Geography	Published (Executive Summary)
Publicity	Published
Data Collection Development	Published
Data Collection Support	Published
Census Coverage Survey	Published
Processing	Published
Annex: Quality of Data Capture and Coding	Published
Downstream Processing	Published (Executive Summary)
Data Quality	
- Question non-response rates	Published
- Disclosure Control (Executive Summary only)	Published
- Data Validation (Executive Summary only)	Published
Edit & Imputation	Published
One Number Census	
- Quality Assurance	Published
- Lessons learnt (Executive Summary only)	Published
Output Policy	Published (Executive Summary)
Output Production	
- Part 1:Review of Output Released to date	Published (Executive Summary)
- Part 2:including Sample of Anonymised Records (SARs)/Origin Destination Matrices	Published
Census Access	Published
Programme Management	Published (Executive Summary)
Quality Report	Published
General Report	Published

Please note that the dates for release of individual evaluation reports noted above are target dates, and therefore subject to change. For the latest information please visit www.statistics.gov.uk/census2001/reviewevaluation.asp