HSE
Health & Safety Executive

A National Population Data Base for Major Accident Hazard Modelling

Prepared by **Staffordshire University** for the Health and Safety Executive 2005

RESEARCH REPORT 297

HSE
Health & Safety Executive

A National Population Data Base for Major Accident Hazard Modelling

Graham Smith, Charlie Arnot, John Fairburn, Gordon Walker
Institute for Environment and Sustainability Research
School of Sciences
Staffordshire University
College Road
Stoke on Trent
ST4 2DE

In order to undertake the modelling of major accident hazard events the HSE needs to estimate the numbers of people potentially at risk from such events, particularly when societal rather than individual risk is of concern. To-date, however, the data used on the spatial distribution of populations at risk has been highly generalised. This project has developed a sophisticated methodology for producing a national population database, drawing on multiple data sets and including populations located within residential, workplace, retail, transport and leisure land uses and within communal establishments involving particularly sensitive populations (such as schools and hospitals). The final database has a greater coverage of population types and a better level of spatial resolution than any others that currently exist. It has a flexible and user-friendly interface, which provides for many different potential uses by HSE and other Government Bodies and Departments.

There were major challenges involved in producing the database including the need for national coverage, for accuracy at small spatial scales and for representing highly variable patterns of population concentration over time (for example in retail areas). For these reasons, the population database has to be approached as a representation of patterns of potential occupation, rather than a precise measure. Throughout this report, and the user documentation which accompanies the database, the constraints on the data for different population categories and therefore on how it should be used are emphasised.

This report and the work it describes were funded by the Health and Safety Executive (HSE). Its contents, including any opinions and/or conclusions expressed, are those of the authors alone and do not necessarily reflect HSE policy.

HSE BOOKS

First published 2005

ISBN 0 7176 2941 4

Acknowledgements

We would like to thank the following people and organisation for their help with this project:
John Mooney for continued advice and problem solving
Adam Knight for work on some of the datasets.
Christine Dover for providing administrative support.

The project management team provided invaluable guidance and feedback:
Andrew Fowler and Shaun Welsh at the Health and Safety Executive
Helen Balmforth and George Brownless at the Health and Safety Laboratory

We would like to thank the following for locating and or providing data.
Teresa Groves, Welsh Assembly Government
Jenny Cole and Lillian Goswell, Department for Education and Skills
Colin Houston and Dougie Ferguson, NHS, Scotland
Richard Morrison, HM Inspectorate for Education, Scotland
John Tidball, Welsh Health Estates, NHS
Brian Penlington, Stoke on Trent City Council

CONTENTS

EXECUTIVE SUMMARY

Context

1. Major hazard accident modelling is undertaken by the HSE in relation to installations and pipelines defined under EU and domestic legislation as presenting a potentially significant off-site risk to people. This modelling work is used to inform the carrying out of regulatory functions by the HSE, and the provision of advice to land use planning authorities.

2. Part of this modelling work involves estimating the numbers of people potentially at risk from such accident events, particularly when societal rather than individual risk is of concern. To-date, however, the data used by the HSE on the spatial distribution of populations at risk has been highly generalised and of poor quality.

3. This project constitutes the final stage of the development of a more sophisticated approach to deriving population data for major accident hazard modelling, drawing on multiple data sets to include populations located within residential, workplace, retail, transport and leisure land uses and within communal establishments involving particularly sensitive populations (such as schools and hospitals). Two previous projects undertaken by Staffordshire University had evaluated the feasibility of constructing such as database within the key constraints of 'reasonable' cost and potential for application at a national level.

Aims

4. The aims of the project were to:

- refine the method and process for deriving a national population database for major accident hazard modelling to take account of newly available datasets.
- develop and apply generic population multipliers to be attached to buildings, transport routes and land uses. These multipliers should, where appropriate and possible, provide differentiation between population levels at different times of the day and between populations of different sensitivities.
- produce a database for estimated populations associated with buildings, transport routes and land uses across England, Scotland and Wales
- develop a user interface to enable simple use of and interaction with the database by personnel within the HSE.

Key Challenges

5. Whilst the production of a population database may superficially appear a fairly straightforward task, in reality there are many complex challenges involved:

- the need is for a database which can represent where people are likely to be at any time when an accident could take place. Major concentrations of people can build up at certain times of the day, in places which at other times are almost entirely empty and patterns of occupation can be highly variable through the week and over longer seasonal timescales
- national scale coverage is required within the database. It is not therefore possible just to rely upon locally sourced or intensively collected local data.
- the size of accidents the HSE needs to model is also highly variable, extending up to 10-20km in some cases, but in others to only 50-100 metres. This means that whilst for larger scale analysis some reduction in spatial accuracy may be acceptable, for smaller hazard ranges the precise location of individual buildings may be crucially important.

- there are significant constraints on the availability of digital national scale data on land uses and population, although less than in the past
- patterns of population distribution do not remain static over the longer term.

6. For all of these reasons, the population database has to be approached as a representation of patterns of potential occupation, rather than a precise measure. Throughout this report, and the user documentation which accompanies the database, we emphasise the constraints on the data for different population categories and therefore on how it should be used.

The Database

7. The completed National Population Database has been provided in the ArcGIS Personal Geodatabase format. A summary of the final contents and structure of the database is shown in the Table below.

Feature Dataset	***Layer***	***Differentiation***
Residential	Residential	Usual or Night Time Daytime Term Time / Non-Term Time
Transport	Roads (major)	Average Daily Flow; Peak Flow; Maximum Capacity
	Railway Stations	Location only
	Ports	Location only
	Airports	Location only
Sensitive and Communal Establishments	Schools	Daytime
	Boarding Schools	Night time
	Care Homes	Maximum Capacity
	Hospitals	Maximum Capacity
	Prisons	Maximum Capacity
Workplace	Workplace Populations	Total Workplace Population
Retail	Retail Populations	Core Retail Centre; Town Centre; Retail Park
Leisure Facilities	Stadia	Maximum Capacity
	Camp Sites	(Location only) Camp Sites; Caravan Sites
	Public Attractions	(Location only) Aquarium; Historic House; Motor Racing Circuit; Racecourse; Theme Park; Wildlife Centre; Zoo

8. Two versions of the final database have been developed:

- *Individual point locations*. This version represents features/populations at their actual location, usually to 1 metre accuracy depending on the source data used. The features in this database have all been spatially located by combining address information with the Ordnance Survey (OS) data sets AddressPoint and CodePoint that list the spatial location of addresses.
- *100 metre by 100 metre grid*. This is the most comprehensive version of the database. Populations are generalised to a 100m by 100m grid.

9. There are a number of issues to be considered in using and combining the data layers, as problems could arise due to incompatible spatial scales and to temporal differentiation. The potential problem of double counting may also occur where different data layers are combined and decisions need to take account of this on a case by case basis.

10. In some source data sets large buildings or collections of buildings with a common function (such as hospitals) are represented as only single points. To account for the unknown positional error in such situations, a flag system was used. The flag functions as an error

buffer increasing the size of the feature. No population is attached to the flag itself, it is an indication that the feature in question may be present. This might be particularly important where the edge of a hazard area narrowly fails to capture the grid point representing a particular geographic entity such as a school or hospital. The flag will help ensure in this situation that the presence of the school or hospital is still registered within the hazard area.

11. Some of the key limitations of the database include:

    - data accuracy will decrease with the passing of time. However accuracy over time will also vary between datasets as some data is more static than others.
    - the 100m grid resolution is relatively coarse if working at a very local scale and adds to positional uncertainty for smaller geographic entities such as households.
    - dynamic populations, such as the transport layer, are invariably generalisations and may be relatively inaccurate when considered at a local scale.

12. Some population categories could not be included in the database largely due to an absence of data at a national level. These include night time populations in retail and leisure areas, populations in major leisure facilities and populations in railway stations.

13. The National Population Database (NPD) is delivered in the ESRI Geodatabase format, designed to be accessed through the ESRI ArcGIS platform. A customised interface has been created to facilitate user functionality. The interface consists of a selection of ArcGIS and bespoke tools to enable the selection, extraction and exportation of data from the NPD.

Conclusions

14. This project has successfully produced a national population database, which has a greater coverage of population types and a better level of spatial resolution than any others that currently exist. It has a flexible and user-friendly interface, which provides for many different potential uses. As such, it represents a major step forward for the HSE in their representation of populations at risk within accident modelling. Moving from a crude method of visual inspection of OS maps, to a multi-layered database constructed from good quality national data available within a GIS environment is a very significant advance, improving the quality of current work and opening up new opportunities for analysis.

15. There are number of ways in which the use of a national population database could be extended within the HSE. These include for other forms of point source risk (explosive sites and nuclear facilities); transport and pipeline routes; macro-level risk studies and performance indicators. Outside of the HSE the database could be used by departments and agencies concerned with other forms of risk, such as flooding, extreme weather events and terrorism; various organisations concerned with issues of resilience, emergency preparedness and emergency response in the event of disasters; private industry including both those companies operating hazardous sites and pipelines and the insurance industry assessing premium risks.

16. There are various issues to be considered in deciding when and how to update the database over time. These are complicated by ongoing changes in the availability of data and a dynamic environment for potential applications within the HSE. We therefore recommend that 18 months after the first version of the population database is delivered that the need for updating is thoroughly and carefully evaluated.

1 AIMS AND OBJECTIVES

1.1 Rationale and Context

Major hazard accident modelling is undertaken by the HSE in relation to fixed sites and pipelines defined under EU and domestic legislation as presenting a potentially significant off-site hazard and risk to people and the environment. This modelling work is used to inform the carrying out of regulatory functions by the HSE, and the provision of advice to land use planning authorities on planning applications for new hazardous installations and development in their vicinity (HSE 1989, Walker 2000).

A range of software packages are used by the HSE to model the behaviour of toxic, flammable and explosive substances under a range of accident scenarios. Part of this modelling work involves estimating the numbers of people potentially at risk from such accident events, particularly when societal rather than individual risk is of concern. To-date, however, the data used by the HSE on the spatial distribution of populations at risk has been highly generalised and of poor quality (Walker, Mooney and Pratts 2000).

This project constitutes the final stage of the development of a more sophisticated approach to deriving population data for major accident hazard modelling. Previous projects undertaken by Staffordshire University have (i) evaluated different sources of data for producing a residential population database (Walker and Mooney 1998) and (ii) produced a methodology for mapping buildings and a range of land uses to which population estimates can be attached extending beyond residential uses (Mooney and Walker 2000). This methodology was developed within the key constraints of 'reasonable' cost and potential for application at a national level.

This project aimed to further refine the methodology for deriving a population database to take account of newly accessible data, evaluate and agree a range of population multipliers and then produce a population database for use within a GIS environment for all of Great Britain (England, Wales and Scotland).

This database will provide a key source of information for the modelling of societal risk, as well as inform other aspects of the HSE's work, including the assessment of COMAH safety reports, the analysis of the impact of land use planning controls and the monitoring patterns of change in populations at risk over time.

1.2 Aims

The aims of the project were to:

1. refine the method and process for deriving a population database for major accident hazard modelling to take account of newly available datasets.

2. develop and apply generic population multipliers to be attached to buildings, transport routes and land uses. These multipliers should, where appropriate and possible, provide differentiation between population levels at different times of the day and between populations of different sensitivities.

3. produce a database for estimated populations associated with buildings, transport routes and land uses across England, Scotland and Wales

4. develop a user interface to enable simple use of and interaction with the database by personnel within the HSE.

1.3 Previous Research

The second of the two projects referred to above took the form of a jointly funded PhD studentship between HSE and Staffordshire University. An HSE research report was completed and published on this work (Mooney and Walker 2002), which includes a literature review of the state of the art in population mapping for risk analysis at that time.

The aim of the project was to evaluate and derive sources of population data to be utilised in major accident hazard modelling and quantified risk assessment (QRA). A number of key criteria were laid down to guide the process of selecting, manipulating and integrating population data sets. The population data needed to provide national cover, at high levels of detail, without incurring excessive cost and to take account of diurnal changes in population patterns associated with a range of different land uses. The solution developed involved using postcode geography as a foundation and then adding and combining further datasets to provide a "richer" set of population data that encompasses activity away from the home. The basic features which were derived and mapped are buildings, transport routes and land uses. Datasets drawn on included cartographic data, remotely sensed land use data, postcode data, commercial directories and other list data and area based socio-economic data.

This research concluded that there was an affordable method and process available for deriving data that extends beyond the simple distribution of residential populations. This method and process could not provide a 'perfect' view of the distribution of people potentially at risk from accident events (as discussed in the next section this is in any case an impossible goal). It could however provide reasonably robust data on the distribution of buildings, transport routes and land uses to which population estimates can then be attached.

Since this research was completed there have been changes both in the specific and wider context for producing a national population database. New, improved and more up to date data sets have become available and some of the cost constraints referred to above have been removed or lessened in significance. The ability of the HSE to make use of the capabilities of Geographical Information Systems (GIS) has been enhanced, particularly through the services provided by the Health and Safety Laboratory. Furthermore, the need for the HSE and others to have access to a good quality, inclusive population database has strengthened due to regulatory developments, the lessons learned from accident events and the escalation in threat posed by terrorist activities.

1.4 Key Challenges

Whilst the production of a population database may superficially appear a fairly straightforward task, in reality there are many complex challenges involved. Most importantly, for the purposes of accident modelling, the need is for a population database which can represent where people are likely to be at any time when an accident could take place. For this reason simply knowing where people live, from census or other household data is insufficient. People clearly do not stay at home all of the time, but rather move around to work, recreate, shop, go to school, stay in hospital etc.. This means that major concentrations of people can build up at certain times of the day, in places which at other times are almost entirely empty (e.g. shopping malls). Patterns of occupation can also be highly variable through the week and over longer seasonal timescales (e.g. retail areas, sports stadia, camp sites etc..). Attempting to predict in detail how these fluctuations of population are likely to take place is a deeply uncertain and, arguably, impossible task. Whilst some

patterns are more predictable than others, simplifying assumptions have universally to be used to remove some degree of real-world complexity.

These problems are exacerbated by the levels of scale at which the HSE need to be able to undertake accident modelling. National scale coverage is required within the database, as new hazardous installations could potentially be proposed anywhere in the country. If accidents from hazardous substances in transport are involved the spatial coverage also needs to be extensive. It is not therefore possible just to rely upon locally sourced or intensively collected local data around existing hazardous sites in constructing the database. The size of accidents the HSE needs to model is also highly variable, extending up to 10-20km in some cases, but in others to only 50-100 metres. This means that whilst for larger scale analysis some generalisation and reduction in spatial accuracy may be acceptable, for smaller hazard ranges the precise location of individual households and occupied buildings may be crucially important.

These needs then have to be matched against the spatial resolution, accuracy and usability of available and affordable data sets. Whilst, in general, the availability of digital national scale data on land uses and population has significantly improved and the HSE now benefits from access to Ordnance Survey data through a pan-governmental licence, there have still been major issues to consider in evaluating what data is available and how applicable it may be to the needs of the project. In some cases whilst good quality information on the spatial location of particular types of land uses/buildings may exist, information on the numbers of people that occupy these land uses/buildings may not. In others, information on numbers of people over an area may be obtained; but precisely distributing these people across the occupied buildings in the area may be impossible to do.

Finally, it is obvious that patterns of population distribution do not remain static over the longer term. A population database produced today will become gradually less reliable, although some categories of population may stay more stable than others (e.g. school populations compared to workplace populations).

For all of these reasons, **the population database has to be approached as a representation of patterns of potential occupation, rather than a precise measure**. We have been careful to avoid as much as possible the problem of spurious accuracy - where population figures appear more precise than they really are - and throughout this report, and the user documentation which accompanies the database, **we emphasise the constraints on the data for different population categories and therefore on how it should be used.** We have also provided for additional data input by the user, so that local information can be used to supplement that which can be derived from national scale databases.

2 PROJECT METHODOLOGY

In this section the overall programme of work and common tasks that were involved is discussed. This is then supplemented in following sections by more detailed methodological discussion of the processing and manipulation required for each of the population types.

2.1 Refinement of method and multipliers

This initial phase of work involved updating, rethinking in some cases, and finalising the method and process to be used for producing the population database. It had two elements:

- *review of method in light of newly available datasets*.

  Since completion of the previous project some important changes had taken place in availability and access to data sets. These included:

  - the 2001 census provided a source of data which is both more up to date and produced to a new geographical framework based on postcode geography, and also including employment data.

  - HSE now had access to a wide range of OS data under a service level agreement. Most importantly this included Address Point, a high accuracy spatially referenced dataset of addresses. This was not used in the previous research due to its high cost. Access to national coverage digital mapping under this agreement also provides new avenues for improving accuracy and deriving multipliers

  The first stage of the project therefore involved a review of these changes in access and availability, to decide if they could significantly enhance the method and process for deriving population data of various forms.

- *development and agreement of population multipliers*. The previous project had derived data on buildings, transport routes and land uses. It had, in some cases, also derived a method for attaching population to these features, but this was not fully resolved for all categories of population, such as those in transport systems and in retail areas.

The objective at the beginning of the project was to assess the feasibility of now producing a database that could provide the main categories of population data shown in the first column of Table 2.1 below, with the differentiation within each of these categories noted in the second column. As will be discussed later, whilst it was decided that most of these categories and differentiation could be included in the database, in a few cases this did not prove possible.

Table 2.1: Categories of population and differentiation within the initial project objectives

Category	***Differentiation***
Residential population	Day Time Night Time Weekday Weekend
Population within the road system (for major roads)	Day Time Night Time with ability to set peak or typical flows
Population at railway stations	Day Time Night Time
Children at schools (daytime)	Term time Holidays
Populations in hospitals, care homes etc..	Differentiation between patients and workforce
Populations in major retail and leisure areas	Day Time Night Time Weekday Weekend
Working population (within areas or specific buildings)	Day Time Night Time Weekday Weekend
Major Leisure Facilities (sports stadia, theme parks, caravan sites etc…)	Capacity when occupied Period Occupied
Background populations (populations estimated to be present within different general categories of land use e.g. parks and recreational land)	Day Time Night Time

2.2 Implementation for a sample area

A sample area covering 40km x 40km was selected in order to experiment with data sets and methods of processing, organisation and representation. An area covering from Stoke-on-Trent in the south and extending up to Chester in the North was selected; i) due to the diversity of land uses and sources of hazard this included and, ii) the need to use local knowledge in refining aspects of the methodology. The operational database and an initial interface were also produced for this sample area before then scaling up to the full national data set. In order to work on the sample area it was necessary to:

- *obtain or purchase up to date source data sets.* New versions of datasets need to be purchased in some cases and arrangements for access to others put in place.

- *derive population data for the sample area.* This followed the methods initially developed for each population category, enabling evaluation of any scaling and practical and implementation issues. If necessary the method and process was further refined to take account of this experience

- *check for accuracy and errors.* a verification routine was used on a sample basis to check for problems with the derived data

- *develop a user interface within ArcGIS.* This involved programming using Visual Basic for Applications within ArcGIS software environment to provide a simple means of

accessing and interacting with the data. A discussion of the interface and the capabilities it includes can be found in section 3.3.

2.3 Scaling up to National Level

Finally, a database was established with coverage for England, Wales and Scotland and provided to the HSE for accessing through the user interface. Specific tasks were to:

- *derive population data for a sequence of tiles covering the entire area*
- *where possible, check for accuracy as appropriate to the data source, the quality of the database being used and the availability of comparison data sets*
- *evaluate and assess each layer indicating its strengths and weaknesses*
- *resolve any issues for the GIS database in the integration of the population database with outputs from MSDU accident modelling tools*
- *production of user guide and associated documentation*
- *deliver and demonstrate to HSE personnel*

3 THE DATABASE AND USER INTERFACE

This section outlines the contents and characteristics of the final database, including a summary of the source datasets used to produce the various layers. In addition to this there is an overview of the functionality of the user interface.

3.1 Profile of the Database

The completed National Population Database has been provided in the ArcGIS Personal Geodatabase format.

Each unique population is stored in a Feature Class layer e.g. schools, hospitals. These layers are grouped into Feature Datasets, each representing a set of similar populations e.g. populations within the transport system, sensitive populations.

A summary of the final contents and structure of the database is provided in Table 3.1. All of the feature datasets within the database are fully useable within the ArcGIS environment.

Each layer within the database is stored as point data. Point data is an efficient way of organising spatial data and allows for large volumes of spatial information to be archived, accessed and processed with minimum computational overheads. Two versions of the final database are provided:

- *Individual point locations*. This version represents features/populations at their actual location, usually to 1 metre accuracy depending on the source data used.

- *100 metre by 100 metre grid*. This is the most comprehensive version of the database. Populations are generalised to a 100m by 100m grid. Each point in the database represents the centre point of an area which is 100m by 100m in size. This is detailed further in the next section.

Table 3.1 Structure and content of the National Population Database

Feature Dataset	*Layer*	*Differentiation*	*Source Data*	
			Location	*Population*
Residential	Residential	Usual or Night Time Daytime Term Time / Non-Term Time	AddressPoint[1] CodePoint with Polygons	2001 Census Table: ks02, ks16, ks19, t10
Transport	Roads (major)	Average Daily Flow Peak Flow Maximum Capacity (Bumper to Bumper)	OSCAR Traffic Manager	Department for Transport
	Railway Stations	Location only	Strategi	
	Ports	Location only		
	Airports	Location only	Strategi	
Sensitive and Communal Establishments	Schools	Daytime	AddressPoint[1]	DoE Schools Census
	Boarding Schools	Night time	CodePoint[1]	Care Home Registry
	Care Homes	Maximum Capacity	CodePoint[1]	Care Home Registry
	Hospitals	Maximum Capacity	AddressPoint[1]	NHS Hospital Census
	Prisons	Maximum Capacity	AddressPoint[1]	HM Prisons directory
Workplace	Workplace Populations	Total Workplace Population	AddressPoint[1] Census Output Areas	2001 Census Table: uv75
Retail	Retail Populations	Core Retail Centre Town Centre Retail Park	ODPM Areas of Town Centre Activity CACI Retail Footprint CACI Retail Locations	
Leisure Facilities	Stadia	Maximum Capacity	Internet directories	
	Camp Sites	(Location only) Camp Sites Caravan Sites	Strategi	
	Public Attractions	(Location only) Aquarium Historic House Motor Racing Circuit Racecourse Theme Park Wildlife Centre Zoo	Strategi	

Notes:
1. See Appendix 2

The two versions of the database exist to allow for user requirements at a range of spatial scales. This is particularly useful for the residential layer. The levels of positional uncertainty associated with some source datasets preclude their inclusion as individual points. Table 3.2 lists the layers in each database.

Table 3.2 Comparison of the two databases

Feature Dataset	***Layer***	***Database***	
		Individual Point	***100m by 100m***
Residential	Residential	✓	✓
Transport	Roads (major)		✓
	Railway Stations		✓
	Ports		✓
	Airports		✓
Sensitive and Communal Establishments	Schools	✓	✓
	Boarding Schools	✓	✓
	Care Homes	✓	✓
	Hospitals	✓	✓
	Prisons	✓	✓
Workplace[1]	Workplace Populations	✓	✓
Retail	Retail Populations		✓
Leisure Facilities	Stadia		✓
	Camp Sites		✓
	Public Attractions		✓

Notes:
1. The workplace population layer has a third version in the form of a point layer that attaches populations to Census Output Area Centroids.

3.2 Combining the Data Layers and Double Counting Issues

There are a number of rules that need to be followed when using and combining the data layers. Problems could arise due to incompatible spatial scales and to temporal differentiation. The rules are as follows:

- *Do not combine individual point data and 100m by 100m grid data*. These two types of layer report populations at different spatial scales and should not be summed together.

- *Do not mix up daytime figures and night time figures*. Some of the differentiation within layers reports populations at different times of the day. Combining these could give a false indication of population. For example, combining night time residential populations with daytime school populations would result in the user double counting school children.

The potential problem of double counting may also occur where different data layers are combined. For example, when combining day-time residential, retail and transport populations, it theoretically could be the case, in some situations, that all of the people in the retail area or transport system are those that are already being counted as at home – therefore it might be appropriate to *subtract* the retail and transport populations from the residential rather than add them together. This might be appropriate, for example, if an entire town was included in a hazard study area.

However, we have purposefully sought to focus the populated non-residential layers on populations which should usually be largely *additional* to the local residential population. So that:

- local traffic is not counted, only major motorways and A roads where cars are more likely to be moving through rather than just within the local area

- only retail centres are included, not smaller areas of local shops servicing essentially a local population. It is reasonable to presume that retail centres will be drawing people from a wider vicinity, beyond the residential population that maybe already included within the hazard area

- only major leisure facilities are included which again would be expected to be drawing population from a wider area

It follows that the extent to which double counting issues arise will depend on the characteristics (size, land uses, layout) of the particular area that is being examined in hazard modelling. A judgement therefore needs to be exercised by the user as to whether or not, in any one situation, it is appropriate to entirely add layers, entirely subtract layers, or add or subtract in some proportion. This judgement will be influenced by the geography of the area being examined, and also whether or not 'worst case' assumptions are being used. For example, in an area with a stretch of motorway, a major retail centre and housing, a 'worst case' assumption would be to assume that the motorway is full and entirely made up of through traffic; and that the retail centre is at capacity and contains shoppers who live entirely outside of the study area rather than in local housing. Moving away from the 'worst case' could involve assuming both lower capacities for the motorway and retail centre, and that a proportion of the people living in local housing are amongst those in cars on the motorway or out shopping in the retail centre.

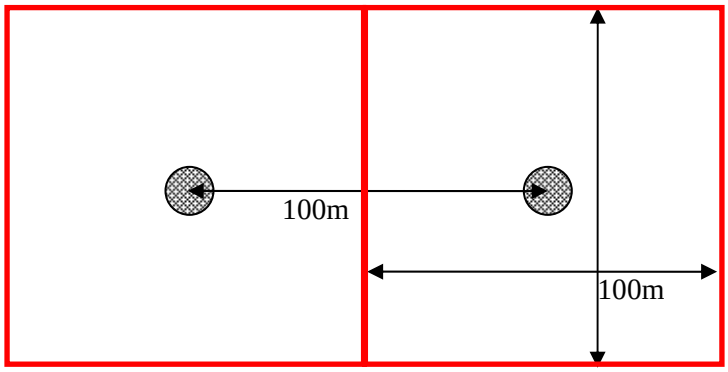
3.3 Individual Point Database

The individual point database provides a high degree of positional accuracy. The features in this database have all been spatially located by combining address information with the OS data sets AddressPoint and CodePoint that list the spatial location of addresses. Further details of these data sets are provided in Appendix II. Each point in the database is the address location for the feature in question plus the associated data. The primary use of this layer is to provide augmentative information to be used in conjunction with the Grid Point Database when a local scale analysis is undertaken.

3.4 100m by 100m Grid Point Database

The 100m by 100m grid point database is conceptually different to the individual point database, and should be conceptualised as a raster grid rather than a point location data set. Each point within the grid represents the centroid of a 100m by 100m square area (Figure 3.1). All data falling within the square is assigned to the square's centroid point. This creates a potentially confusing situation since the 100m by 100m square is represented by a single point in the ArcView GIS. An illustration of edge effects that result from this discrepancy is discussed in section 3.4.2.

Figure 3.1 Two grid points and corresponding spatial extents



The grid point database differs from the address point database because a single grid point may represent multiple locations, e.g. urban residential households; a single location e.g. a school; or be part of a location e.g. part of a hospital core area.

The way the grid point functions can be divided into three categories:

i) Multiple location – the grid point represents multiple geographic entities i.e. residential properties
ii) Single location - the grid point represents a single geographic entity, i.e. small communal establishments
iii) Part location – the grid point represents part of a geographic entity, i.e. hospitals

In the second and third functions the grid points are modelling a geographic entity. In both cases there is a degree of spatial uncertainty introduced by three unknown factors:

i) The accuracy of the original location
ii) The positional error introduced by joining the original location to the grid point
iii) The true spatial extent of the geographic entity.

To account for the unknown positional error, a flag system was used. The flag functions as an error buffer increasing the size of the feature. No population is attached to the flag itself. It is an indication that the feature in question may be present. This might be particularly important where the edge of a hazard area narrowly fails to capture the grid point representing a particular geographic entity such as a school or hospital. The flag will help ensure in this situation that the presence of the school or hospital is still registered within the hazard area.

3.4.1 Flag Creation

Flags were created using either a one point rule or a two point rule depending on the type and size of population being reported.

One point rule: Once a core point or core area has been assigned, all neighbouring grid points are labelled as flag points (figure 3.2).

Two point flag rule: All points neighbouring the core point or core area that are two points away are designated flagged areas (figure 3.3).

Figure 3.2: One point flag example

Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

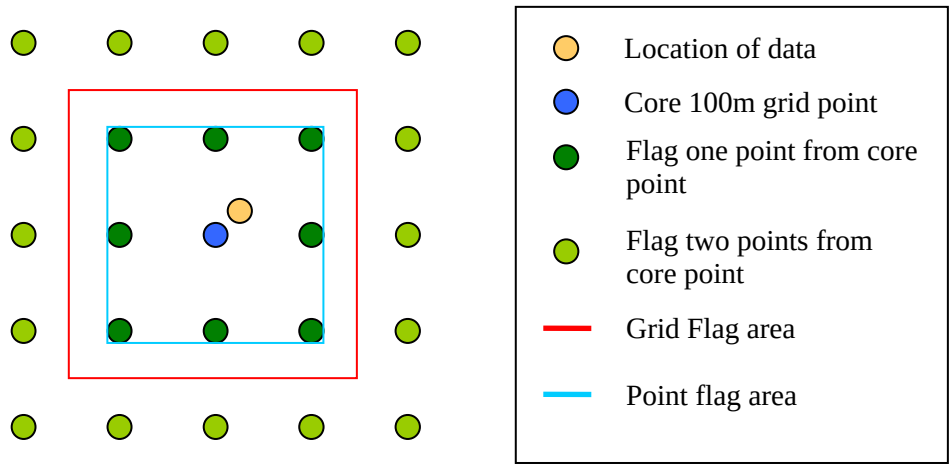
Figure 3.3: An example of the two point flag rule

Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

3.4.2 Edge Effects

There is a significant edge effect possible due to the discrepancy between the spatial area that the grid points represent and the actual form they have in GIS software ArcView. Figure 3.4 provides an illustration of this. The yellow point indicates the original data location, whilst the blue centre point indicates the nearest 100m grid point to which the data would be joined. When creating a flag, this point is referred to as the core point, and is not included as a flag point. The flag is also prone to an edge effect as indicated by the two squares, the red square indicating the true spatial extent of the flag, whilst the smaller blue square represents the actual size of the flag within the GIS system

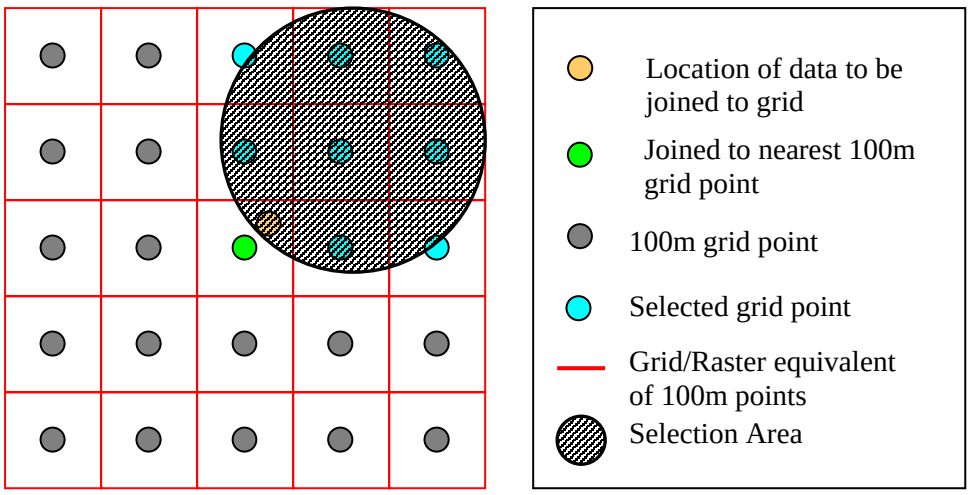
Figure 3.4 The data transformation process for creating a flag.



When selecting an area, i.e. defining an area in which a potential hazard might occur, it is possible to select the addressed location of a feature, and miss the core area. Figure 3.5 gives an illustration of this.

The yellow point shows the address point for a feature. In turn this is attached to the nearest grid point (coloured green). The bright blue points indicate selected points defined by the greyed out selection area circle i.e. a hazard area. The original data point lies within the area of interest (selection area), however the grid point to which the original data point has be joined, lies outside the area of interest and therefore would not be included in any analysis.

Figure 3.5 100m grid with edge effect.



To overcome the edge effects illustrated in Figure 3.5, certain data locations, in particular sensitive communal establishments have been flagged. Flags act as buffers around core data locations. In some instances the core area is given a spatial extent derived from the attached data (i.e. hospitals) whilst in other cases the core area is a single point.

3.4.3 Over lapping flags

In certain circumstances, particularly with respect to schools in the sensitive layer, the flag of one feature may overlap the core and flagged area of another. In this situation a simple logic is applied. This is better illustrated by considering two neighbouring features 'A' and 'B'.

- If a point is a core area it cannot be a flag
- If a flag core area (feature A) occurs to the left of and above the conflicting flag core area (feature B), feature A core area takes precedence and the point in question is labelled with the feature A flag.

3.5 Limitations of the Data Base

Some of the key limitations of the database have already been discussed with respect to positional uncertainty and edge effects. Other limitations include:

- Data accuracy will decrease with the passing of time. However accuracy over time will also vary between datasets as some data is more static than others.
- The 100m grid resolution is relatively coarse if working at a very local scale and adds to positional uncertainty for smaller geographic entities such as households.
- Dynamic populations such as the transport layer are invariably generalisations and may be relatively inaccurate when considered at a local scale.

A number of populations or differentiations of a particular population that were identified as desirable in the initial project objectives (see Table 2.1) have not been included in the final population database. These elements are listed in Table 3.3. The decision not to include each of these elements is due largely to limitations in source data sets, but also to a need to focus on the most important types of populations from the HSE's point of view. In some cases the commitment of further time and resource could potentially enable further population categories to be included, but this is only merited if further enhancement of the database is useful for the HSE.

Table 3.3 Populations included in the initial project objectives, but omitted from the final database

Category	***Differentiation***
Residential	Weekend
Railway Stations	Daytime / Night Time (location only)
Retail and leisure areas	Night Time / Weekend
Workplace	Night Time / Weekend
Major Leisure Facilities	Period Occupied
Background Populations	Daytime / Night Time

3.6 User Functionality

The National Population Database (NPD) is delivered in the ESRI Geodatabase format, designed to be accessed through the ESRI ArcGIS 8.x platform. A customised interface has been created to facilitate user functionality. The interface consists of a selection of ArcGIS and bespoke tools to enable the selection, extraction and exportation of data from the NPD. These tools are grouped into six different groups based on their function. These groups are:

- Selection drawing tools: these provide various methods for defining an area of any shape i.e. a circle, rectangle or polygon.
- Selection option tools: these provide a number of selection options, i.e. create a new

selection, add to current selection, and remove from current selection.

- View tools: provide the ability to navigate the database
- Information tools: allow the user to find features or areas, and investigate data represented by particular points.
- Export tools: provide a number of features and allow the user to export shape files, text files, jpegs, and to create raster images of the point data sets which can in turn be converted into text files.
- Raster Catalog tool: provides the means to create a background layer of multiple rasters i.e. the OS 1:10,000 data set.

The interface also provides a Raster Factory to enable the creation and export of rasters from the NPD. The user is able to define and select an area using a number of different methods. The user can then analyse the data, creating graphs, screen grabs and simple summaries of the included variables. Finally the user can export the data in a number of formats, from an ASCII grid, to a user defined text file listing the variable present for each point selected.

4 RESIDENTIAL POPULATION

4.1 Population Characteristics and Variability

Residential populations are the most readily mapped and estimated of all. There are established data sets available covering the whole UK which provide information both on the location of households and the people they contain (although not down to numbers of people per specific house).

Whilst residential populations are fairly static compared to some of the others included in this project, there are still patterns of variability to take account of. It is reasonable to presume that most people are at their home overnight, so a night time residential population can be fairly directly estimated. Estimating day time population is more involved, as patterns of occupation will vary between weekends and weekdays and between term time and school holidays.

The 'Residential Layer' of the National Population Database locates and reports populations at their usual place of residence. The following population figures have been produced for each household:

- *Usual Resident population.* This assumes that all people are in their homes.

- *Weekday Day-time population*. This figure accounts for people being away from home at a place of employment or school. This is reported in two forms, weekday term-time and weekday non term-time.

The final database contains two versions of the Residential Layer:

- *Household level data*. This layer reports population for each household/place of residence. The purpose of this layer is to aid a user who is looking at localised problems that have relatively small risk areas.

- *Aggregated 100m accurate data*. This layer reports the population of the household data aggregated to a 100m by 100m grid with the population assigned to the centre point of each grid cell. This layer is more suitable when looking at large risk areas. This will dramatically reduce data volume, increase the speed of processing data for areas with a large spatial extent and make it easier to spot the pattern of population.

4.2 Source Data Sets

Various source data sets have been used to produce the residential layer. These can be split into two main types, locational data and population data.

4.2.1 Locational Data

In order to produce an accurate residential population database it is essential to find source data that locates households as precisely as possible. Two datasets have been used to locate residential addresses:

- *AddressPoint*. This is the primary source of household location data. AddressPoint gives a 1 metre accurate grid reference to every postal address in Royal Mail's Postal Address

File (PAF). This is a point dataset and therefore does not give an indication of the spatial extent of properties. This is not a significant problem for normal residential households with relatively small spatial extents.

- *CodePoint with Polygons – Vertical Streets Polygons.* CodePoint with Polygons is a polygon dataset that gives the spatial extent of every unit postcode in Great Britain. Within this dataset are Vertical Streets Polygons. These polygons are very small in area but they flag up locations that have a number of postcodes (or AddressPoints) in one place, i.e. a block of flats.

More detailed information about these datasets can be found in Appendix 2.

4.2.2 Population Data

The most comprehensive source of population data available is the 2001 Census. The smallest spatial area that census data is reported at is Output Area level. These Output Areas are made up of aggregations of unit postcodes and have an approximate size of 125 households.

The following census tables have been used to produce the residential layer:

- *Key Statistics Table KS02 – Age Structure.* This table gives the age breakdown for each output area. It was used as a factor to help calculate daytime residential populations.

- *Key Statistics Table KS16 - Household spaces and accommodation type.* This table gives a breakdown of the housing type in each output area. This data was used in the verification of the residential layer. This is discussed in section 4.4.

- *Key Statistics Table KS19 - Rooms, amenities, central heating and lowest floor level.* This table gives the *average household size* for each output area. This was used to populate addresses based on their location. The table also gives a breakdown of the lowest floor level of households within each output area. This data was used in the verification of the residential layer. This is discussed in section 4.4.

- *Theme Table T10 - Resident, workplace and daytime population.* This data is only available down to Ward Level. This table provides data on the number of people that are of working age but do not work within each ward. This data was used when calculating daytime residential populations.

4.3 Data Transformation and Processing

Two major processes are involved in producing the residential layer. Firstly, *locating residential populations* and secondly, accurately *assigning a suitable population to each location.*

4.3.1 Locating residential populations

AddressPoint is a very detailed dataset and it is updated every 3 months. As a result, a number of quality indicators and flags are used in the data. These flags evolve as data is verified on the ground by OS. Each indicator is used differently to produce the residential population layer. Below, each indicator is discussed and details are given about how they were used.

- *Change Type*. This specifies if the address is a new address, a changed address or a

deleted address. The action performed with this field was to *remove all of the addresses classified as deleted.*

- *Status Flag – physical state.* The status flag code gives an indication of the building type, the positional accuracy, the physical state of the building and a measure of how the building description compares to the Royal Mail description. At this stage the status flag was used to *remove all addresses classified as demolished.*

- *Status Flag – positional accuracy*. Addresses classified as having *temporary coordinates will not be removed* from the database. It is considered more important to identify such addresses than to remove them completely.

- *PO-Boxes*. AddressPoint includes addresses that are classified as PO-Boxes. In general such addresses are commercial and large receivers of post. These addresses are not positioned in their actual spatial location. They are positioned at their delivery location which is the nearest Royal Mail Depot. As a result it is necessary to *remove all addresses classified as PO Boxes.*

More detailed descriptions and statistics about these indicators can be found in Appendix 2.

AddressPoint includes all addressed buildings, both residential and commercial. In order to produce the residential layer it was necessary to select the addresses that were classified as domestic delivery points only. Residential and non residential addresses were separated using OS CodePoint definitions. These include:

- *Domestic delivery points*: Non PO-box delivery points that have no PAF organisation name. These points will be classified as residential and will be included in the residential population layer.

- *Non-domestic delivery*: Non PO-box delivery points that have a PAF organisation name. These points will be classified as commercial and will not be included in the residential population layer.

The result of these processes is a point dataset of current residential addresses. Due to potential errors in the descriptions of addresses it is possible that some addresses have not been assigned an organisation name. This would result in the database classifying a small number of addresses as residential when they are in fact commercial. However, this will be a very small problem in the database and the potential effect on the residential layer will be very marginally to over-estimate residential populations. This is deemed to be an acceptable level of error and much less of a concern than potentially under-estimating residential population.

4.3.2 Assigning a population to residential locations

Usual Resident Population

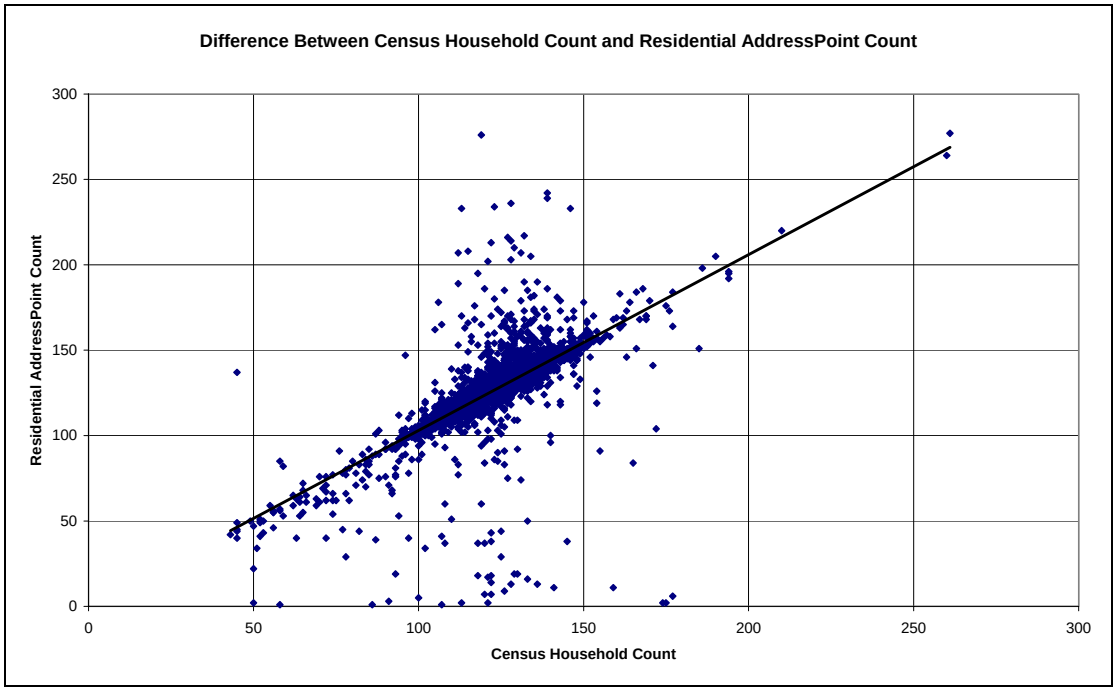
Each AddressPoint within an Output Area has been assigned a population based on the average household size reported in the census for that area. This statistic excludes people living in communal establishments because they would bias the reported size and are dealt with in another part of the database. This household size value is the *usual / night-time population* and assumes that all residents are at home.

Using this method the characteristics of an area in the 2001 census are used to model the population for 2003. In order for this method to be valid the 2003 AddressPoint count needs

to reflect the number of households in the census. Figure 4.1 shows the number of residential AddressPoints in an Output Area compared to the number of households reported in the census. It shows that AddressPoints are a good approximation of the number of households reported by the census.

Of the 2824 Output Areas in the Sample Area there are a number that demonstrate a large difference between the two counts, due to two factors. Firstly, new households have been built and old households demolished since the 2001 census. This is an advantage of using AddressPoint because it gives a more up-to-date picture. The method assumes that new households in an Output Area are of the same average size as existing ones. Secondly, large clusters of inaccurately located AddressPoints (i.e. a block of flats) could be in the wrong Output Area. This will have an effect on both the Output Area the addresses are reported to be in and the Output Area the addresses should be in. This type of error will be discussed further in section 4.4.

Figure 4.1: Comparison of AddressPoint and Census Household Counts within the Sample Area



Weekday Day-time population

Using census figures, usual / night time populations were adjusted to take account of people being at a place of work or in the case of children, at school. This process was used to calculate weekday daytime residential population.

This method assumes that people of working age go to work all year round. This is due to the fact that available data does not indicate precisely when people work. However, in the case of school children it is assumed that during term time they are away from home (at school) during the day, and out of term time they are at home during the day. Therefore the weekday daytime population was reported in the database as two values:

- *Term-time, weekday, day-time population.* School age children are away from home.

- *Non term-time, weekday, day-time population*. School age children are at home.

The most comprehensive set of daytime population figures reported in the 2001 census are available at Census Ward level and above. The census reports the *number of people of working age (16 – 74) who live in an area but do not work*. This method assumes that *these people will be at home during the day*. For each Ward the proportion of people of working age but who are not working was calculated. This value acts as a daytime multiplier (DM) for people of working age in each ward.

At Output Area level the number of people falling into different age groups is reported. This method classifies people into four age groups and different assumptions are made about their daytime location. This is shown in Table 4.1.

Table 4.1: Daytime behaviour of four age groups

Age group	***Description***	***Daytime location assumptions***
0 to 4	Pre School Age	At home.
5 - 15	School Age	At school in term time. At home out of term time.
16 - 74	Working Age	At place of work. Unless classified as not working in the census.
75+	Retired	At home.

For each Output Area the *proportion of people of School Age (S) and Working Age (W)* was calculated. Each AddressPoint in the database has a usual population (P). These values are then used to calculate the two daytime population values for each AddressPoint. This is achieved in the steps shown in Figure 4.2.

Figure 4.2 Daytime population calculation

Where:

S	proportion of people of School Age (i.e. 80% = 0.8)
W	proportion of people of Working Age
DM	proportion of working age population at home during the day
P	usual / night time population of the AddressPoint
ATW	proportion of total population at work
ATH	proportion of total population at home

Each AddressPoint in the database will be given a value for S and W based on the Output area it falls within. Each AddressPoint will be given a value for DM based on the ward it falls within.

Term time daytime population (TTDP) is calculated using the following formula:

$$ATW = (1 - DM) * W$$
$$ATH_{TT} = 1 – (S + ATW)$$
$$\underline{TTDP = ATH * P}$$

Non term time daytime population (NTDP) is calculated using the following formula:

$$ATW = (1 - DM) * W$$
$$ATH_{NT} = 1 – ATW$$
$$\underline{NTDP = ATH * P}$$

This method reduces the usual population of AddressPoints using the characteristics of the area taken from the census. The populations are reduced because of daytime migration but this layer does not attempt to relocate daytime migrants. These populations will be represented in the Schools layer and in the Workplace layer.

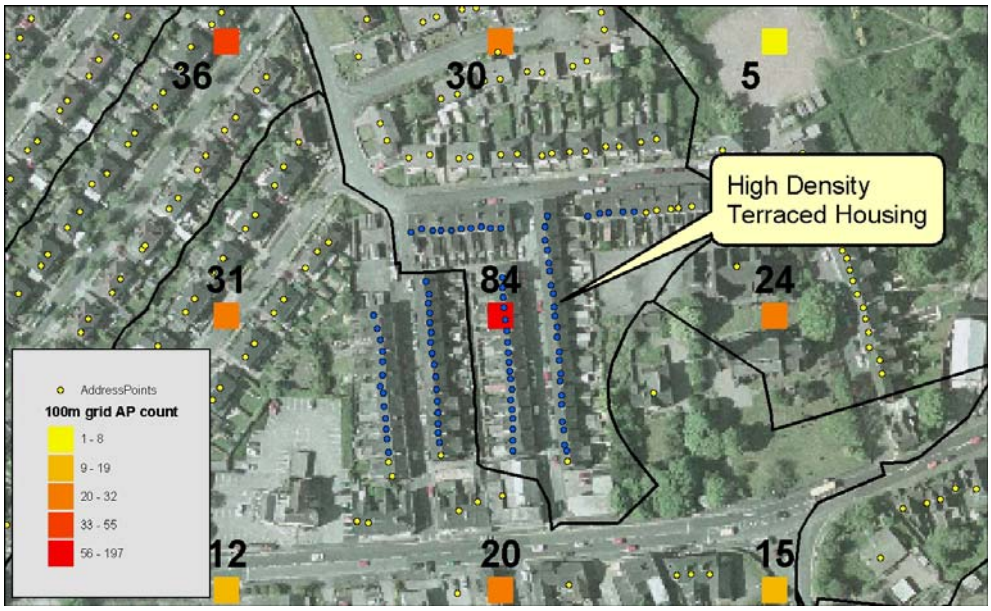
4.4 Verification of the Data

4.4.1 Population Density

The spatial accuracy and characteristics of the residential layer are largely dependent on the spatial accuracy and characteristics of AddressPoint. AddressPoint itself is verified by the Ordnance Survey and is a very accurate dataset. However, a number of additional steps have been developed to examine how well the residential layer represents the real world.

The main characteristic used for checking the data is to look at the density of households. This is necessary because the density within the database is influenced by the positional accuracy of AddressPoints. The density of addresses in the real world is influenced by the size of building and whether they are multi-storey. Therefore the characteristics of buildings in an area were derived from the Census and from Vertical Postcode Lookup files (taken from OS CodePoint with polygons). By examining these characteristics for each 100m grid point a judgement was made about whether the reported AddressPoint count seemed sensible. The typical maximum density of normal housing with ground floor addresses is approximately 80 households per hectare (i.e. one 100m by 100m grid point). This is demonstrated in Figure 4.3.

Figure 4.3 Typical maximum density of households with ground floor addresses.



As a result of this finding all of the grid points with counts of 80 households or more were flagged for further investigation.

Vertical Postcodes

The OS product, CodePoint with Polygons contains small polygons that flag the locations of AddressPoints that are vertical postcodes. This means that addresses are above each other or

spread around the point. In such areas the density of buildings could be much more than 80. As a result of this all of the grid points that have vertical postcodes within them were flagged up.

If a grid point had an AddressPoint count of greater than 80 but had vertical postcodes within it then it was deemed to be acceptable. However, grid points with counts greater than 80 but with no vertical postcodes needed to be looked at more closely.

Household Type

The next step was to look at the characteristics of households within each Output Area. These characteristics were then used to assess the grid cells flagged up for investigation. After examining the sample area it was decided that if a cell was in an Output Area(s) made up of 80% Flats and Terraced Housing then it is possible to have a density greater than 80 addresses. An example of a grid cell in which this occurs is given in Figure 4.4.

As a result, grid points with address counts greater than 80 but with no vertical postcodes are deemed to be acceptable if they are located in an Output Area(s) made up of 80% Flats and Terraced Housing.

Figure 4.4 Higher densities produced in areas of Flats and Terraced housing.

Table 4.2 An example of the data verification process

ID	*Address Count*	*Addresses with temporary coordinates*	*Addresses with a Vertical Streets Polygon*	*% of Household (Terraced or Flats)*	*Address Count is greater than 80*	*There are no Vertical Street Postcodes*	*The % of Terraced houses and Flats is less than 80%*	*This point should be investigated further*
1	137	81		79.78	1	1	1	1
2	91			77.88	1	1	1	1
3	84			62.52	1	1	1	1
4	81			78.16	1	1	1	1
5	81			49.47	1	1	1	1
6	80			73.21	1	1	1	1
7	80			62.54	1	1	1	1
8	118			97.57	1	1	0	0
9	197	58	176	86.06	1	0	0	0
10	148		148	87.43	1	0	0	0
11	136		94	94.67	1	0	0	0

Notes: A value of 1 in the final 4 columns indicates that the statement made in the column title is true.

Table 4.2 gives a number of examples of this process. After completing the process for the sample area a small number (7 out of 27,000) of populated grid cells required further explanation (point IDs 1-7 in Table 4.2). Examination of these 7 grid points helped to highlight two limitations of the residential layer. The first is caused by temporary coordinates and the second is localised under-estimation caused by large clusters of addresses being generalised to a single point.

Temporary Coordinates

One of the characteristics of AddressPoint is that some addresses are given temporary spatial coordinates. These coordinates are updated in a later release of the product after which time the OS has verified the true location of the property on the ground.

Large clusters of inaccurately located AddressPoints (e.g. a block of flats) will have a significant localised impact on the address count of a grid cell. It is also likely that the temporary coordinates may be in the wrong Output Area. This is why some grid cells have a large address count but the characteristics of the output area suggest that it should be lower. Figure 4.5 shows the example of a block of flats with temporary coordinates. These flats have been located in the middle of a row of terraces and all 81 addresses are located at one point. Figure 4.6 shows the same block of flats, however, they are now in their final location in a later release of AddressPoint. This demonstrates the potential error in the location of temporary coordinates and highlights the need to be aware of their occurrence.

As detailed earlier, AddressPoints classified as having temporary coordinates (by way of their status flag) were not removed from the database. It is considered more important to identify such addresses than to remove them completely.

As a result AddressPoints with temporary coordinates were highlighted in the final database using an indicator. In addition the number of AddressPoints with temporary coordinates within each grid cell was reported in the 100m grid layer of the database.

The problem of temporary coordinates is important but its occurrence in the database is not large. In the sample area only 4,963 AddressPoints out of 358,112 have temporary coordinates. This is *1.4% of residential AddressPoints in the sample area.*

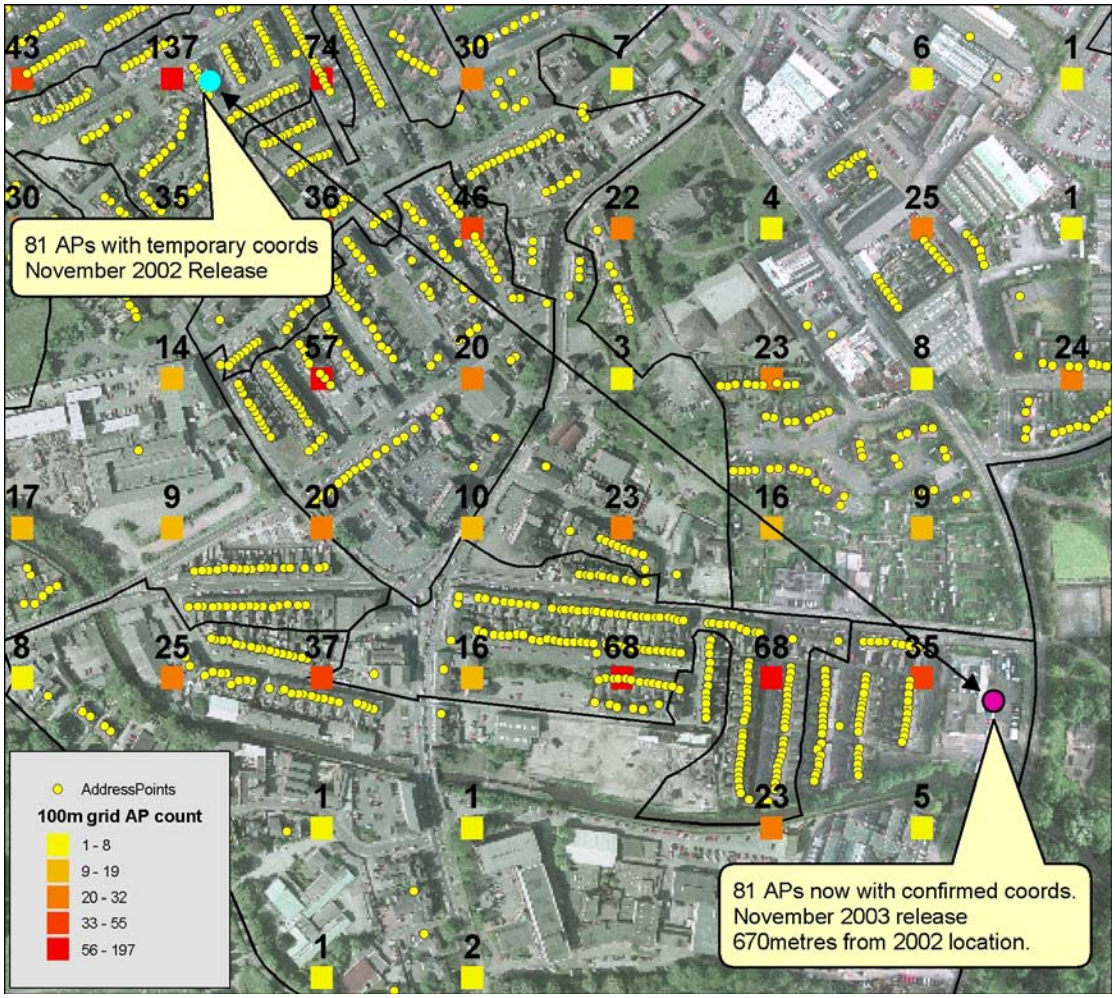
The 100m grid layer indicates that 8.67% of grid cells have some number of AddressPoints with temporary coordinates. However the occurrence of just one or two of these AddressPoints in a cell will not have a large impact on its population. Table 4.3 illustrates the level of the problem and shows that *only 0.23% of populated 100m grid cells contain 10 or more AddressPoints with temporary coordinates.*

Table 4.3 100m grid cells containing AddressPoints with temporary coordinates

There are 26970 populated 100m grid cells in the Sample Area.	***100m grid cells with temporary AddressPoints***	
	count	***%***
Grid cells with 1 or more temporary AddressPoints	2337	8.67
Grid cells with 2 or more temporary AddressPoints	821	3.04
Grid cells with 5 or more temporary AddressPoints	163	0.60
Grid cells with 10 or more temporary AddressPoints	62	0.23

Figure 4.5 The effect of clusters of AddressPoints located with temporary coordinates.

Figure 4.6 Levels of spatial inaccuracy observed in temporary coordinates.



Localised under-estimation

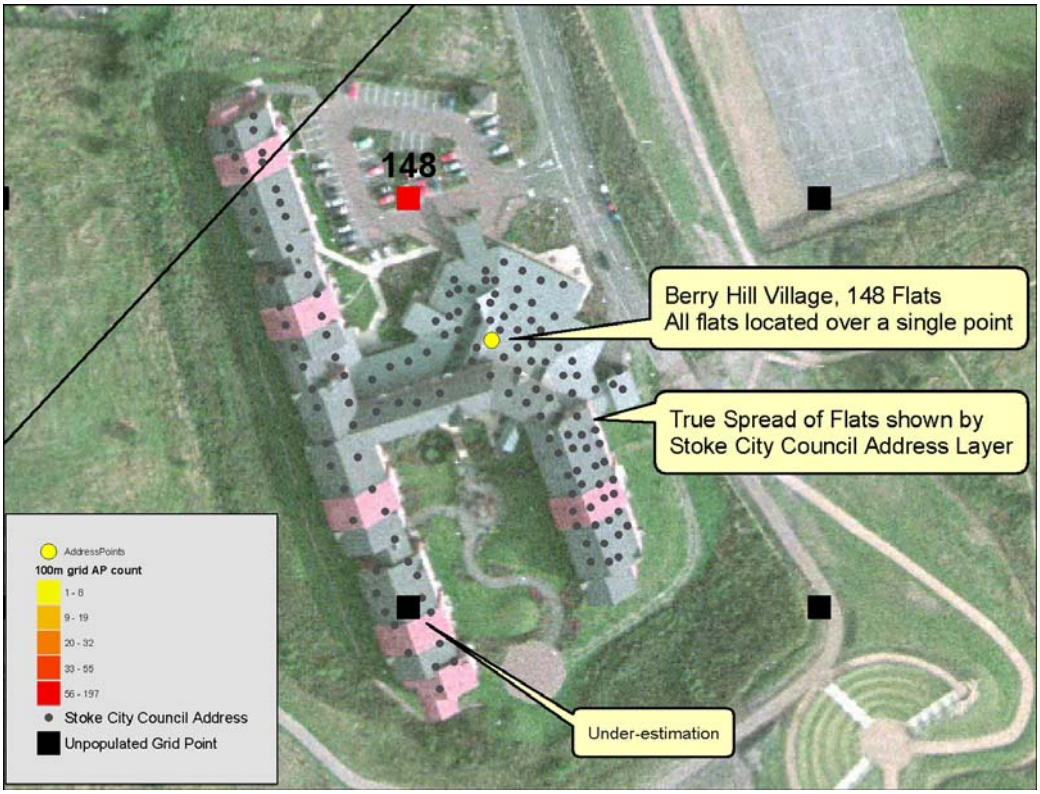
Another feature of AddressPoint is that large clusters of addresses that have a common location, i.e. a block of flats, are often assigned to the same precise location. In other words there will be a large number of Addresses on top of each other in the database. This characteristic is not a large problem in the database but is one that the user needs to be aware of because it produces two problems.

Firstly, Figure 4.7 demonstrates that clustering points together can lead to a 100m grid cells population being underestimated because the addresses have not been spread to their true spatial pattern. A sample address database provided by Stoke-on-Trent City Council has been added to the image to demonstrate how the addresses *should* be spread. Secondly, Figure 4.8 demonstrates the same problem. However, in this case grid cells are being reported as unpopulated.

As with the temporary coordinate issue, AddressPoints and grid cells where this clustering occurs are highlighted in the database by an indicator.

Figure 4.7 Under-estimation of population in grid points as a result of large buildings being located around a single point.

Figure 4.8 Unpopulated grid points as a result of large buildings being located around a single point.



4.4.2 Comparison with the 2001 Census

A major part of the verification of the residential layer is a comparison of data with the 2001 census. Table 4.4 shows a comparison between England and Wales of the 'population in households' reported in the 2001 census and the population reported in the residential layer of the National population database. The 'population in households' figure was used because the majority of the remaining population is that 'in communal establishments'. A large number of these communal establishments are care establishments and these are not present in the residential layer but are contained in the sensitive populations layer. The comparison reveals that the residential layer reports a population figure for England and Wales that is 4.4% greater than that reported as 'in households' in the 2001 census.

Table 4.4 Comparison of household counts between the 2001 Census and the Residential Layer

Region	***Population in Households (Census 2001)***	***Population in the Residential Layer (NPD 2003)***	***Difference +/-***	***Difference %***
North East	2,472,884	2,618,145	145,261	5.87
North West	6,615,672	6,993,907	378,235	5.72
Yorkshire and The Humber	4,880,728	5,147,739	267,011	5.47
East Midlands	4,095,557	4,316,272	220,715	5.39
West Midlands	5,186,248	5,419,168	232,920	4.49
East of England	5,296,534	5,564,878	268,344	5.07
London	7,078,632	7,121,461	42,829	0.61
South East	7,809,823	8,129,561	319,738	4.09
South West	4,812,072	5,006,976	194,904	4.05
Wales	2,859,489	3,030,967	171,478	6.00
England	48,248,150	50,318,107	2,069,957	4.29
England and Wales	51,107,639	53,349,074	2,241,435	4.39

There are a number of reasons for this difference in population. These include:

- *Vacant Properties*. The method that was used to construct the residential layer populated every address and this is therefore a cause of potential over estimation. At a local level, areas that have larger numbers of vacant properties and/or greater numbers of holiday/second homes will experience greater levels of over estimation.

- *Population Growth.* Mid-year population estimates for 2003 released by the Office of National Statistics (Autumn 2004) reveal that the population of England and Wales has grown by 1.4% since the 2001 Census.

- *Communal Establishments.* Some properties classified as 'communal establishments' in the 2001 census may be represented differently in AddressPoint and are therefore included in the residential layer. This is the case for a number of student flats. This produces a discrepancy between the two population figures.

The discoveries made concerning reasons for over-estimation in the database have not suggested that there is a need to review the population figures. However, any user of the data should be aware of each of the issues highlighted above.

4.5 Evaluation of the Data

The residential layer reports weekday night time and weekday daytime populations. In addition to this it reports weekday populations for school term time and non term time. An initial aim of the project was to produce a weekend residential population. This aim has not been achieved because the available source data gives no indication of weekend behaviour. One solution would have been to apply a general multiplier to all weekday populations, but it was decided that this was not feasible. Weekend behaviour varies enormously and therefore for the purpose of this database it is recommended that all populations are seen to be at home during the weekend. This is a worst case assumption and is all that can be achieved with available population data.

There are a number of communal populations that are not consistently represented in the residential layer. These include some university halls of residence and other large establishments such as armed forces sites. Some local user knowledge will be required with regard to locating and populating these sites.

The other major omissions from the residential layer are more transient or temporary populations, in particular those in hotels.

Population trends, particularly internal migration patterns, vary enormously throughout England, Wales and Scotland. Many areas are consistently losing population whereas others are consistently gaining population. In addition there are areas that will have more or less consistent levels of population.

These trends mean that as the residential population layer gets more out of date there will be more potential for errors in the data.

The user of the data needs to be aware that potential errors in the data layer, such as addresses with temporary coordinates or populated vacant addresses, become more problematic when looking at small areas of only a few hundred metres. If the user is looking at areas of a few kilometres then the potential impact of errors is less important.

Overall and within the known limitations the residential layer is a complete and well scrutinised layer.

5 POPULATIONS IN THE TRANSPORT SYSTEM

5.1 Population Characteristics and Variability

The transport system is usually characterised by dynamic population flow that can vary from virtually stationary and very dense, to sparse and free flowing, whether the population refers to passengers waiting at a train station, or travelling on a motorway in a car. As such any population figures given in the database represent a snap shot of a particular scenario. To account for the large population variations possible, three different situations (daily average, peak flow and bumper to bumper) were considered and are described in more detail in section 5.4.

The Transport System layer contains two primary data sets.

a) **Terminal locations**, the location of train stations, international airports and maritime ports. Populating these locations proved to be problematic and so a population has not been attached. This data set therefore provides location information only.

b) **Road network population**, selected road types (single carriageway A-roads, dual carriageway A-roads and motorways) were populated based on flow rates and average vehicle speeds derived from ONS transport surveys (Statistics Bulletin - Road Traffic Statistics 2002; Transport Statistics - Vehicle Speeds in Great Britain 2002; Statistics Bulletin - Traffic Speeds on English Trunk Roads:2001)

5.2 Source Data Sets

5.2.1 Data sets

a) OS Oscar Asset Manager (2003). This is a vector data set representing the road network as a series of lines and nodes. The data set indicates the category of road, i.e. motorway, dual carriageway or A-road. A roads are represented by a single line unless split by a physical barrier, as is the case for motorways and dual carriageway A-roads, when it is represented by two lines. Unusual road configurations e.g. more than 3 lanes per motorway, are not indicated in this data set and thus not accounted for in the database.

b) OS Strategi 2003. This is also a vector data set. Strategi data is organised according to a hierarchy in which features high up the hierarchy have a higher degree of positional accuracy than features at lower levels. The features included in the database were not features high on the hierarchy, and therefore liable to positional inaccuracy however steps were taken to overcome this problem. The features used were,
   - urban polygons, and
   - terminal locations (railway stations, international airports and ports)
   - Government Office Regions

c) Transport data from the ONS transport statistics bulletins. This data includes
   - motor vehicle flows by road class, country and Government Office Region (GOR) in units of ‘thousand vehicles per day’, and

- motor vehicle flow for major sections of motorway network by maximum and average flow (thousand vehicles per day).

5.3 Data Transformation and Processing

5.3.1 Terminal Locations

The point locations of stations, international airports and ports were joined to the nearest 100m grid point. Based on the typical spatial extent of stations, airports and ports, a flagged area was created. For stations a 1 point flag was created. For airports and ports a two point flag was created. Please see chapter 3 for more information on flags.

5.3.2 Road Network Populations

Three road forms were considered for creating this population, these were

a) **motorways**,
b) **dual carriageway A-roads,** and
c) **single carriageway A-roads**.

Other road categories (such as B-roads and private roads) were not populated as they were assumed to be carrying largely light, local traffic (see section 3.2). The three scenarios considered were:

a) **a bumper to bumper** population associated with stationary traffic i.e. as a result of a road blockage,
b) **average population**, representing an average flow derived from a daily average over 24 hours, and
c) **peak population**, representing the peak flows associated with rush hours, i.e. typically between the hours of 8.00am to 10.00am, and 4.00pm to 6.00pm.

In all cases there were **key assumptions** made about the traffic flows and road form in order to overcome practical limitations in data and calculations, these were:

a) The average number of people per vehicle is 1.5.
b) Traffic was deemed to consist entirely of cars. This will tend to produce an over estimation of the road population density that provides worst case scenario figures.
c) Single carriageway roads (represented by the original data set as a single line) contain two lanes per line, one in either direction.
d) Dual carriageway roads (represented by the original data set as two lines) contain two lanes per line.
e) Motorways contain three lanes per line (represented by the original data set as two lines).

5.3.3 Calculation of Road Populations

The three scenarios for road populations are explained below.

5.3.3.1 Bumper to Bumper Road Population

This assumes the roads are full with stationary vehicles that are bumper to bumper. The average space of road occupied by individual vehicles was taken to be 4m as specified in the HSE Safety and Reliability report "The implications of major hazard sites in close proximity to major transport routes" (WS Atkins 163/1998).

The population for a given length of road (l) is calculated by

$$\left(\frac{l * n}{c}\right) * 1.5$$

Equation 5.1: Bumper to Bumper Population

where c = length of road occupied by 1 car (metres)
l = length of road (metres)
n = number of lanes

5.3.3.2 Average Road Population

Average population figures are derived from the AADF (**Annual Average Daily Flow**) with the use of equation 5.2, as specified in the Atkins report (WS Atkins 1998, p.79). AADF refers to how many vehicles pass a point over a 24 hour period. The AADF figures are given in Tables 5.1 and 5.2.

$$\text{Average density} = \frac{\text{AADF}}{V \times 24 \times 3600} \text{ vehicles per metre}$$

Equation 5.2: Average Density

Where: AADF = Annual average daily flow (vehicles per metre)
V = Average traffic speed (m/s)

Table 5.1: Motor Vehicle flows by road class, country and Government Office Region: 2002

		Major roads		Minor roads	
	Motorway	Rural	Urban	Rural	Urban
North East	50.7	12.9	20.8	0.7	2.7
North West	69.9	10.3	17.9	0.9	2.1
Yorkshire & the Humber	65.7	12	18.5	0.9	2
East Midlands	89.6	13	19	0.9	2.1
West Midlands	80.4	11.2	20	0.9	2.8
East of England	83.6	17.5	18.2	1.2	2.5
London	100.8	29.8	28.8	1.5	2.7
South East	91.8	17.7	19.4	1.4	2.5
South West	64.7	10.6	19.7	0.7	2.2
England	77.8	13.4	20.7	1	2.4
Wales	59.5	7.6	16.7	0.6	2.1
Scotland	39.8	4.7	15.9	0.5	1.8
Great Britain	72.9	10.5	20.1	0.8	2.3

Thousand vehicles per day - source ONS Transport Statistics Bulletin, Road Traffic Statistics 2002 (table 2.2 p.15)

Table 5.2: Motor vehicle flows for major sections of motorway network; 2002

Motorways	Max flow	Average Flow
M1 - North of M6 Junction	134	99
M1 - South of M6 Junction	162	100
M2	63	53
M3	124	91
M4 - England	146	93
M5	109	74
M6 - North of M62 Junction	121	59
M6 - South of M62 Junction	147	98
M11	84	61
M20	125	65
M23	111	92
M25 - Eastern links from a1(M) to M23	142	122
M25 - West links from a1(M) to M23	194	147
M27	119	100
M40	114	87
M42	176	91
M56	149	90
M60	174	112
M62 - East of the Pennines (junc 22)	130	74
M62 - West of the Pennines (junc 22)	135	96
A1M	96	49
M4 - Wales	103	66
M73	74	44
M74	85	35
M77	62	47
M8	151	68
M9	55	31

Thousand vehicles per day - source ONS Transport Statistics Bulletin, Road Traffic Statistics 2002 (table 2.3 p.17)

Average density incorporates spatially sensitive variables, AADF and V (vehicle traffic speed). For motorways both variables are assumed to remain constant across a particular Government Office Region (GOR) unless indicated to the contrary by Table 5.2. For both dual and single carriageway A-roads the variables vary between GORs and within GORs between urban and non urban areas. The methods for calculating the spatial variations in AADF and V were similar and involved further ONS data on traffic speeds (Table 5.3) and the production of multiple buffering regions around urban areas.

Table 5.3: Weekday comparisons: Great Britain: 2002

Vehicle type	Road type	Average (miles per hour)
Cars	Motorway	70
	DC	69
	SC	47
LGVs	Motorway	68
	DC	66
	SC	46
Buses/coaches	Motorway	59
	DC	57
	SC	43
Rigid 3/4 axle	Motorway	53
	DC	52
	SC	41
Articulated	Motorway	54
	DC	53
	SC	43

Source: ONS Transport Statistics – Vehicle Speeds in Great Britain 2002, p.8

The ONS Transport Statistics Bulletin "Traffic Speeds on English Trunk Roads:2001" provides information with respect to "average traffic speed by road type and time period", and "average traffic speed by region and time period", which is of potential use and is presented in Tables 5.4 and 5.5.

Table 5.4: Average traffic speed (mph) by road type and time period, 2001

		AM peak	Off-peak	PM peak
Motorway	Built-up	45.7	58.8	50.9
	Non built-up	55.7	61.3	60.7
	All	54.5	61.0	59.5
A roads - DC	Built-up	30.1	35.2	31.9
	Non built-up	54.9	59.3	54.9
	All	47.2	52.1	47.9
A roads - SC	Built-up	24.7	24.4	24.3
	Non built-up	41.0	42.3	42.6
	All	36.6	37.3	37.5
A roads - all	Built-up	28.2	30.8	29.0
	Non built-up	49.3	51.9	50.0
	All	43.1	45.8	43.8
All trunk roads	Built-up	33.3	37.7	34.9
	Non built-up	52.9	57.0	55.6
	All	48.8	53.1	51.3

Source: ONS Transport Statistics Bulletin "Traffic Speeds on English Trunk Roads:2001" p.6

Table 5.5: Average traffic speed (mph) by region and time period, 2001

	AM peak	Off-peak	PM peak
North East	53.1	57.6	52.2
North West	50.2	52.4	52.2
Yorkshire & the Humber	51.5	55.6	54.9
East Midlands	52.1	53.4	52.5
West Midlands	47.2	52.9	49.9
Eastern	50.1	55.0	52.5
South West	58.8	55.5	57.4
South East London	43.3	49.5	47.4
South East	49.9	56.6	54.2
London	24.9	30.0	28.2
All of England	48.8	53.1	51.3

Source: ONS Transport Statistics Bulletin "Traffic Speeds on English Trunk Roads:2001" p.7

However incorporating this information within the current study is problematic. The database being constructed identifies only a single peak time, i.e. night versus day, rather than two peak times (morning and evening rush hours). Secondly regional differentiated data is not categorized by road type (Table 5.5) and urban/non urban speed data is not given at a regional level. An alternative approach was therefore necessary to describe spatial variations in speed. Table 5.6 lists the data used as derived from the ONS 2002 Transport Statistics Report and listed in Table 5.3.

Table 5.6: Traffic speeds: by class of road (mph)

	Non-Urban	Urban
Motorways	70	70
A-roads DC	69	50
A-roads SC	47	35

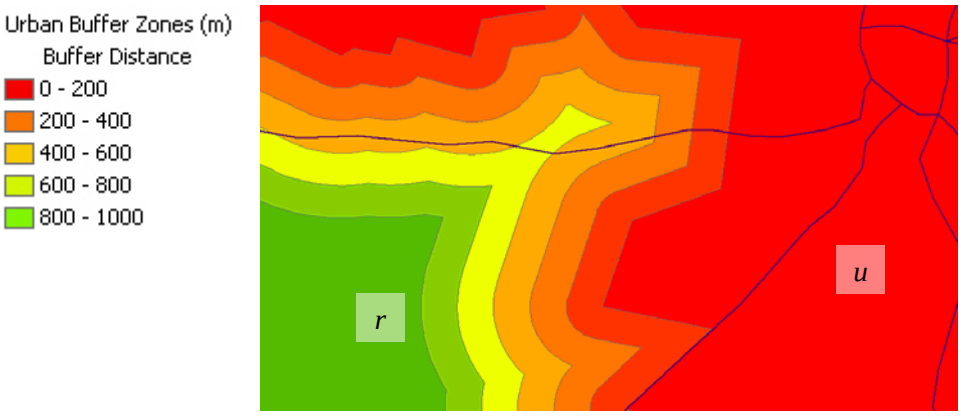
The non urban figures in Table 5.6 are from the ONS Transport Statistics – Vehicle Speeds in Great Britain 2002, whilst the urban figures are based on the speed restrictions usually imposed within urban areas. It should be noted that the urban figures are probably greater than would actually be recorded. For example Table 5.4 gives the national road speeds for urban and non-urban areas in which motorways generally show reduced traffic speeds in urban areas (45.7 to 58.8mph) as compared to non-urban areas (55.7 to 61.3mph). The current study, however, uses 70mph for both motorway conditions on the basis that in some conditions the motorway speed will be unaffected by the presence of built up areas and that errors of over estimation are preferable to errors of under estimation.

5.3.3.3 Creating urban buffer zones

Both speed and flow data were graduated between urban and non-urban areas using the same process. The ONS AADF traffic flow data provides differentiation between urban and non-urban traffic flows for A-roads by GOR. By a series of spatial joins, the data was spatially located to urban polygons. The urban polygon data consists of urban polygons from the OS Strategi data set. There were three issues to consider with respect to this data set. These were:

a) the urban polygons are often inaccurately positioned in the Strategi data set,
b) the data set represents the division between urban and non urban as abrupt, however traffic flows between urban and non-urban would not change abruptly, and
c) there were numerous small urban polygons within the data that would have little impact on traffic flow figures.

Figure 5.1: Urban buffer zones from Non-urban (r) to Urban (u) and A-roads.



Buffer Zones *1 2 3 4 5 6*

To overcome these issues the following steps were taken.

a) only urban polygons with an area larger than 1km^2 were used, and
b) the remaining polygons were buffered with 5 buffers spaced at 200m (Figures 5.1 and 5.2).

This produced a graduated boundary of 1km in width around polygons, reducing the error associated with spatial inaccuracy and providing a more realistic model of traffic flow from non-urban to urban environments.

The AADF was calculated with the following formula (Equation 5.3).

$$AADF = r + \left(\left(\frac{u - r}{6} \right) * z \right)$$

Equation 5.3: AADF Calculation

where *AADF* = *Annual average daily flow (car/metre)*
r = *Non Urban daily flow*
u = *Urban daily flow*
z = *buffer zone*

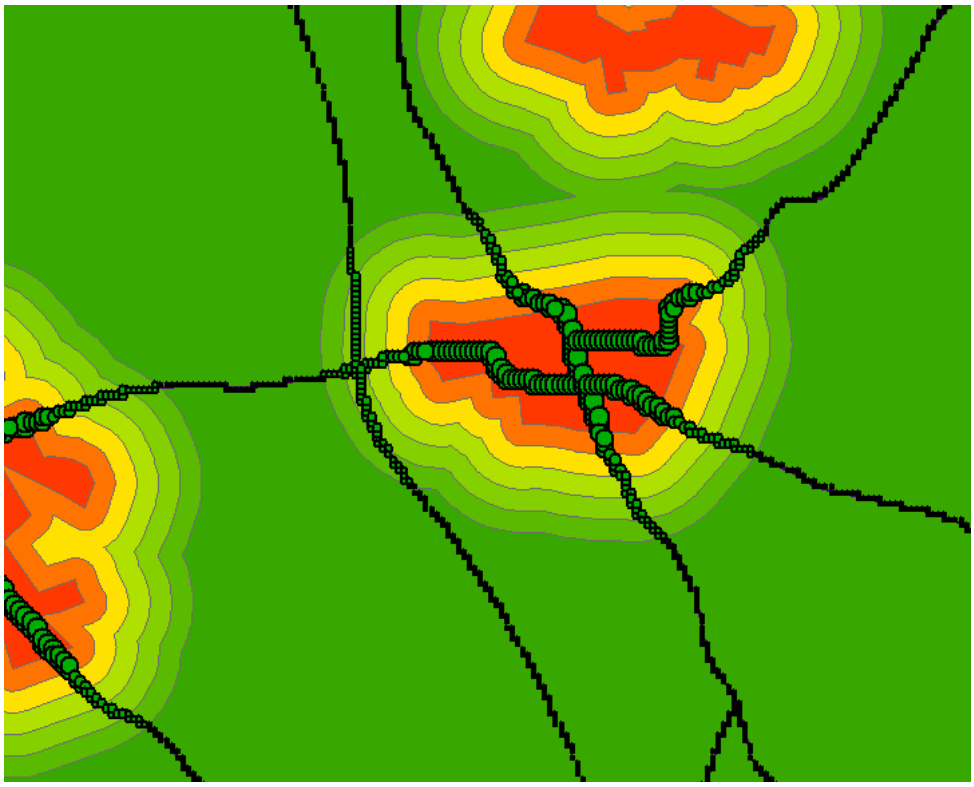
Vehicle speed (V) was assumed to be constant for motorways unless otherwise indicated by the figures in table 5.2. For dual carriageway and A-roads the information in tables 5.4 and 5.6 was combined to give vehicle speeds for each GOR and road type for both urban and rural categories. The resulting figures were used in Equation 5.4 to calculate vehicle speed for any given point.

$$V = r + \left(\left(\frac{u - r}{6} \right) * z \right)$$

Equation 5.4: AADF Calculation

where *V* = *Vehicle Speed (m/s)*
r = *Non Urban vehicle speed (m/s)*
u = *Urban vehicle speed (m/s)*
z = *buffer zone*

Figure 5.2: Urban polygons over 1km^2 with buffers and A-road 10m points. The size of each point is proportional to the traffic density (car/m) represented.



5.3.3.4 Peak Population

Peak population figures were calculated with the assumption that flow rates at peak times are approximately 1.5 times more than flow rates at non peak times. This assumption is derived from data in the 2002 ONS traffic bulletin. The peak population figure was thus a simple multiplication of the Average Road Population by 1.5

Peak Road Population = Average Road Population * 1.5

Equation 5.5: Peak Road Population formula

5.4 Verification of Data

Verification of road traffic data was problematic. The positional accuracy of roads when generalised to the 100m grid was examined for the study area and found to be good. Figures 5.3, 5.4 and 5.5 provide illustrations of this. Calculations of traffic densities were based on previously used methodologies and government data and therefore are presumed to be correct.

Figure 5.3: Point Data. The roads split into 10m points and categorised by road type, with the 100m grid overlaid on top.

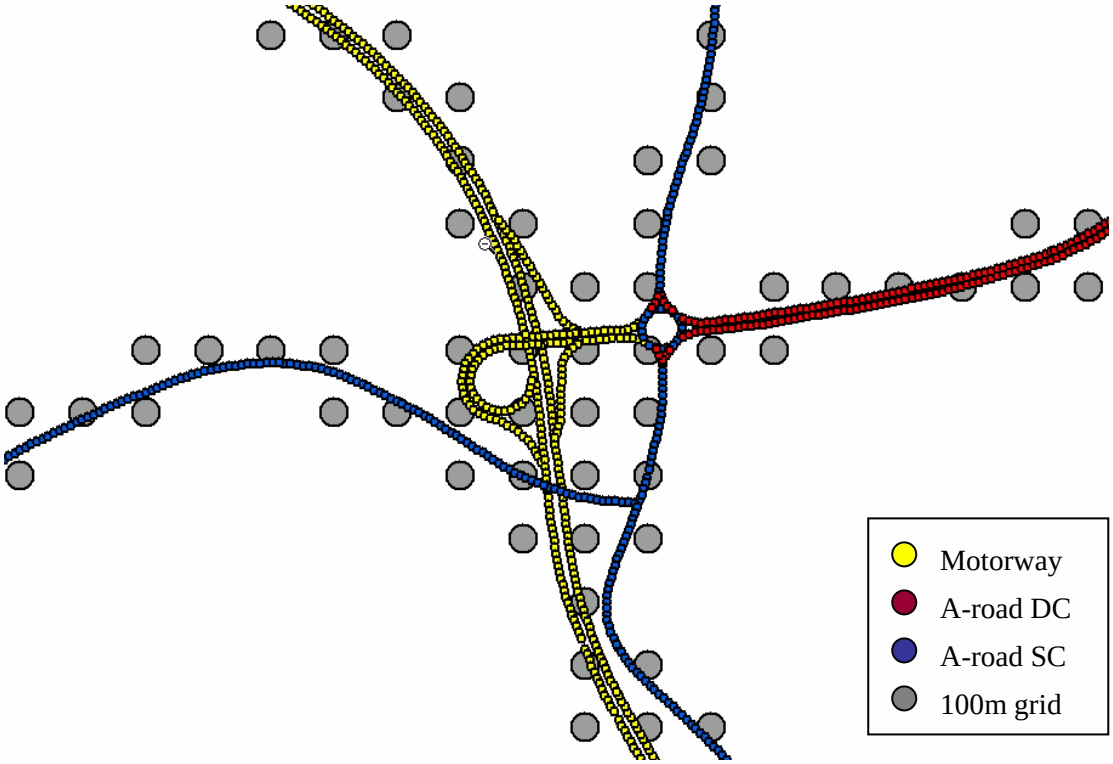
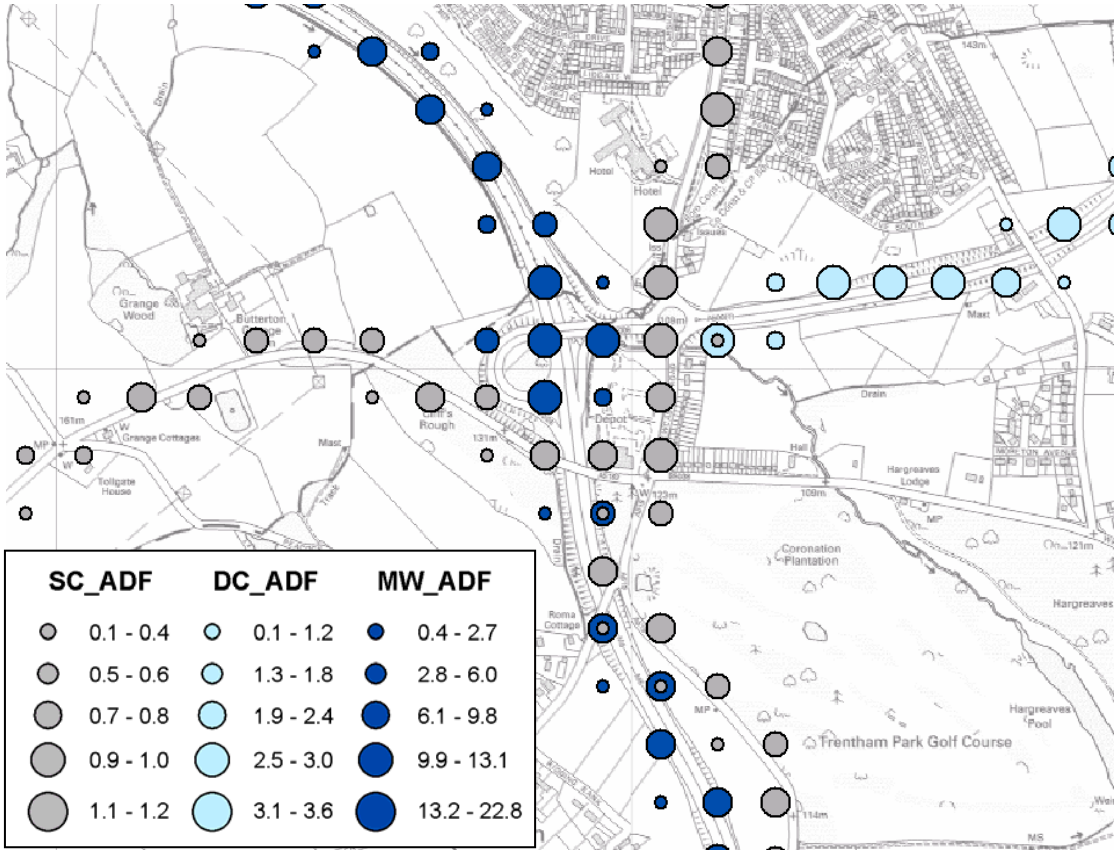
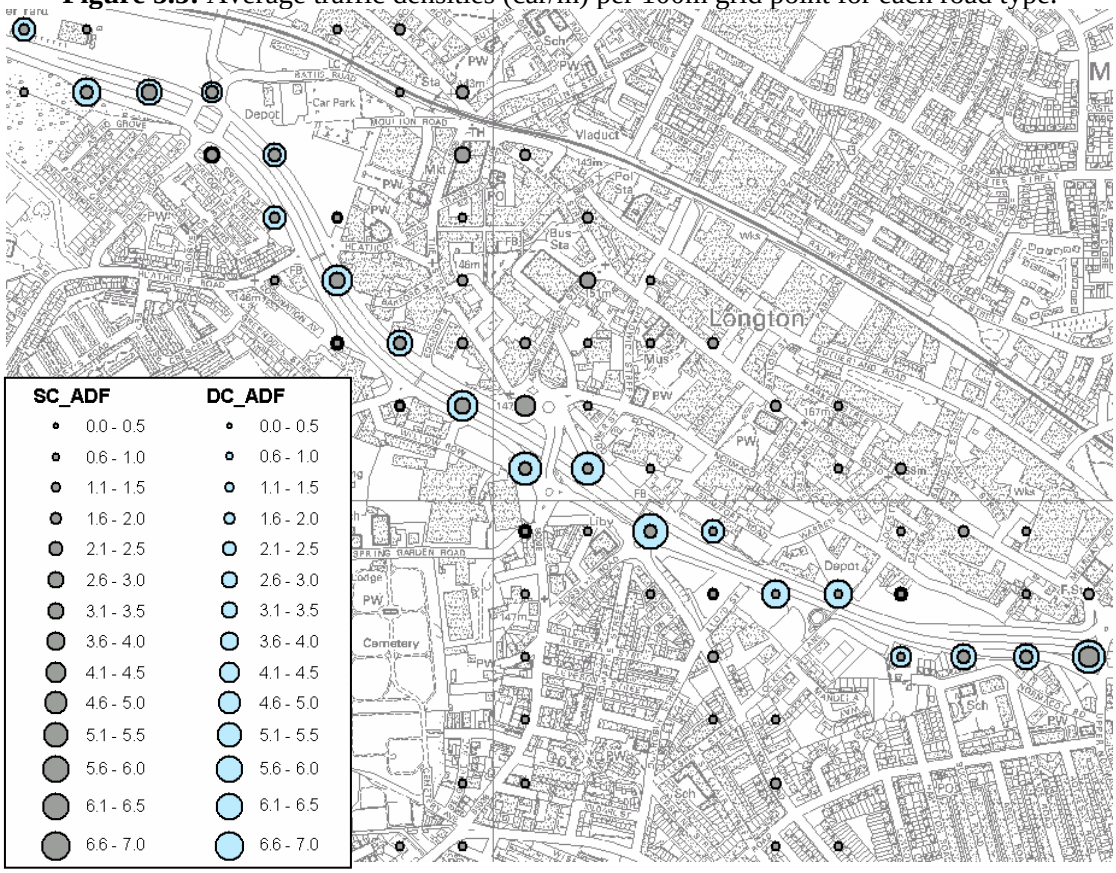


Figure 5.4: The average traffic densities (car/m) associated with each road type (Single and Dual carriageway, and Motorway) as captured by the 100m grid. Note that different road types have different scales to enhance the visualisation of the given road type and that therefore the symbols are not comparable.



Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

Figure 5.5: Average traffic densities (car/m) per 100m grid point for each road type.



Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

5.5 Evaluation of the Data

The data produced is very detailed and builds on existing work by HSE in this area. The approach taken in buffering urban polygons to give graduated population densities provides a comparatively realistic model compared to previous traffic population models. The underlying road transport data has a high degree of positional accuracy. The three scenarios provided allow for a range of traffic populations to be calculated.

5.5.1 Primary Limitations

The primary limitation of this data set is a result of the inherent difficulty in measuring a dynamic fluctuating population. The three scenarios used generate figures that cover medium to large road traffic populations. Scenarios in which lighter traffic is present are not provided. The inevitable generalisations made in the process of generating the populations provide figures that are, at best, a rough snapshot of what the true population may be. Local level variations will lead to local road populations often being significantly different from those estimated in this database.

Further limitations are associated with the key assumptions made about the traffic flows, road form, and terminal locations these are:

a) The average number of people per vehicle is 1.5.

b) Traffic was deemed to consist entirely of cars. This will tend to produce an over estimation of the road population density that provides worst case scenario figures.
c) Road forms are consistent i.e. Single carriageways always contain two lanes, dual carriageways four lanes and motorways six lanes.
d) The spatial extent of terminals

5.5.2 Missing Populations

There are a number of populations for which finding appropriate data was problematic, but which may make up a significant population within the transport network.

a) Railway network, stations and trains
b) Bus network, bus stations and buses
c) B-roads and minor roads
d) Motorbikes and cyclists.

6 SENSITIVE AND COMMUNAL ESTABLISHMENTS

6.1 Population Characteristics and Variability

Communal establishments are defined as geographic entities with a single address point and a population significantly greater than the average residential property. They can vary in composition from a single building (such as a care home) to a large area with several buildings (for example large hospitals). The HSE assesses some populations to be particularly sensitive in respect of the potential impact of accident events, and this layer defines the communal establishments which are associated with these sensitive populations - schools, hospitals and care homes. Non-sensitive communal establishments (prisons) are also included in this layer because the methodology applied to add prisons to the database is similar to the other communal establishments.

With the exception of schools, communal establishment populations tend to be one of the most static populations represented in this database, with population levels remaining essentially the same throughout a 24 hour period. Schools are characterised by a different population pattern, with sensitive populations being present during term-time working hours. Boarding schools are the exception here adding a further complication to the pattern, with a permanent term time sensitive population that changes from day to night time depending on the number of boarding students and day students. Boarding schools have therefore been placed in a separate communal establishments layer.

Although communal establishment populations are relatively static, care homes in particular, change relatively frequently over time and the database will need to be updated here on a regular basis if this is deemed to be important.

6.2 Source Data Sets

Source data sets are divided into locational data sets for geocoding (Table 6.1) and population data sets, for providing addresses and populations (Table 6.2)

Table 6.1: Source data sets for locating sensitive and communal establishments

Data set	*Date*	*Source and further information*
OS AddressPoint*	11/2003	Available through the OS pan governmental agreement.
OS CodePoint*	11/2003	Available through the OS pan governmental agreement.
Streetmap.co.uk	05/2004	Internet mapping site capable of locating postcodes and generating National Grid XY coordinates. Used only to validate problematic postcodes.

*See Appendix 2 for further details of source data sets.

Table 6.2 Source Data Sets for populating Sensitive and Communal Establishments

Data set	*Establishment Type*	*Date*	*Source and further information*
English Boarding Schools	Sensitive	2003	National Care Standards Commission (NCSC) - Boarding pupils only
Care homes	Sensitive	2003	Data from English (NCSC), Welsh and Scottish Care authorities.
Primary Schools	Sensitive	2003	Data from English, Welsh and Scottish Education bodies
English Secondary Schools	Sensitive	2001-2002	Data from English bodies
Secondary Schools (Scotland and Wales)	Sensitive	2003	Data from Scottish and Welsh bodies.
Hospitals (England and Wales)	Sensitive	2003	Data from English, Welsh health authorities- contains bed numbers and internal floor space
Hospitals (Scotland)	Sensitive	2002-2003	Data from Scottish health authorities- contains bed numbers only
Prisons	Non sensitive	Current prisoner numbers and recommended capacity	HM Prison services http://www.hmprisonservice.gov.uk annual report 2002-2003 and the Scottish Prison Services http://www.sps.gov.uk/ annual report 2002-2003.

6.3 Data Transformation and Processing

Data transformation for communal establishments poses three challenges

a) Geo-coding the communal establishment
b) Defining the spatial extent of the communal establishment.
c) Describing the distribution of the associated population within the communal establishment.

Different methodologies were used to define the spatial extents and population distributions of each type of communal and sensitive establishment. The different establishments are therefore discussed separately in the following sections.

6.3.1 Care Homes

Care homes with less than 10 beds available were not included in the database because they were not considered a large enough population to be sensitive. Care homes comprise of two layers.

1. A ‘Care_homes’ layer containing care home data attached to the 100m grid.
2. A ‘Care_homes_ap’ layer indicating the CodePoint location of the care home post code and respective care home data.

Methodology: Care homes were geo-coded initially using AddressPoint. The large number of erroneous matches returned though precluded the possibility of using AddressPoint because filtering the correct address matches manually was not possible. Instead CodePoint was used resulting in a single match per postcode. Care homes without postcodes or with incorrect postcodes, were researched on the internet and, where possible, postcodes were added or corrected. For a small number of care homes it was not possible to resolve the postcode errors or find any details on the internet and these care homes have been omitted. This initial geo-coding generated the ‘Care_homes_ap’ layer.

Once care homes were geo-coded, the points were joined to the nearest grid point. A flag was then created using the one point flag rule to allow for locational uncertainty introduced by the following reasons,

1. generalising to the 100m grid and the use of OS CodePoint,
2. the spatial extent of the care home is unknown.

Figure 6.1 provides an example of two care homes and flagged areas. Care home A is positioned almost equidistant between four grid points, and provides a worst case scenario for positional error when generalising to the grid. The flag and core areas are indicated by both points and 100m raster grid cells. The flag is necessary because it is possible that some of the care home spatial extent may lie outside the core area, and in the flagged area. The flag is therefore indicates possible locations for part of the care home.

Figure 6.1: Care homes with flags

Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

6.3.2 Schools

Schools comprise four layers. There are two types of school layer, 'schools' and 'boarding schools England' for which the methodology was the same. Data for boarding schools in Scotland and Wales was not available and has therefore been omitted. For each of these layers there is also a 100m grid layer and AddressPoint layer. The boarding school layer gives boarding pupil numbers only, i.e. excludes non-boarding numbers and is therefore used in calculating night time populations.

Table 6.3 School layers.

	Database Type	
School Type	Boarding School AP School AP	Boarding School 100m Grid School 100m Grid

Methodology

The schools were geo-coded with OS AddressPoint and filtered manually to remove erroneous matches. This yielded the two AddressPoint layers, 'Schools_ap' and 'Boarding_Schools_ap' which provide the location of the address point of an individual school.

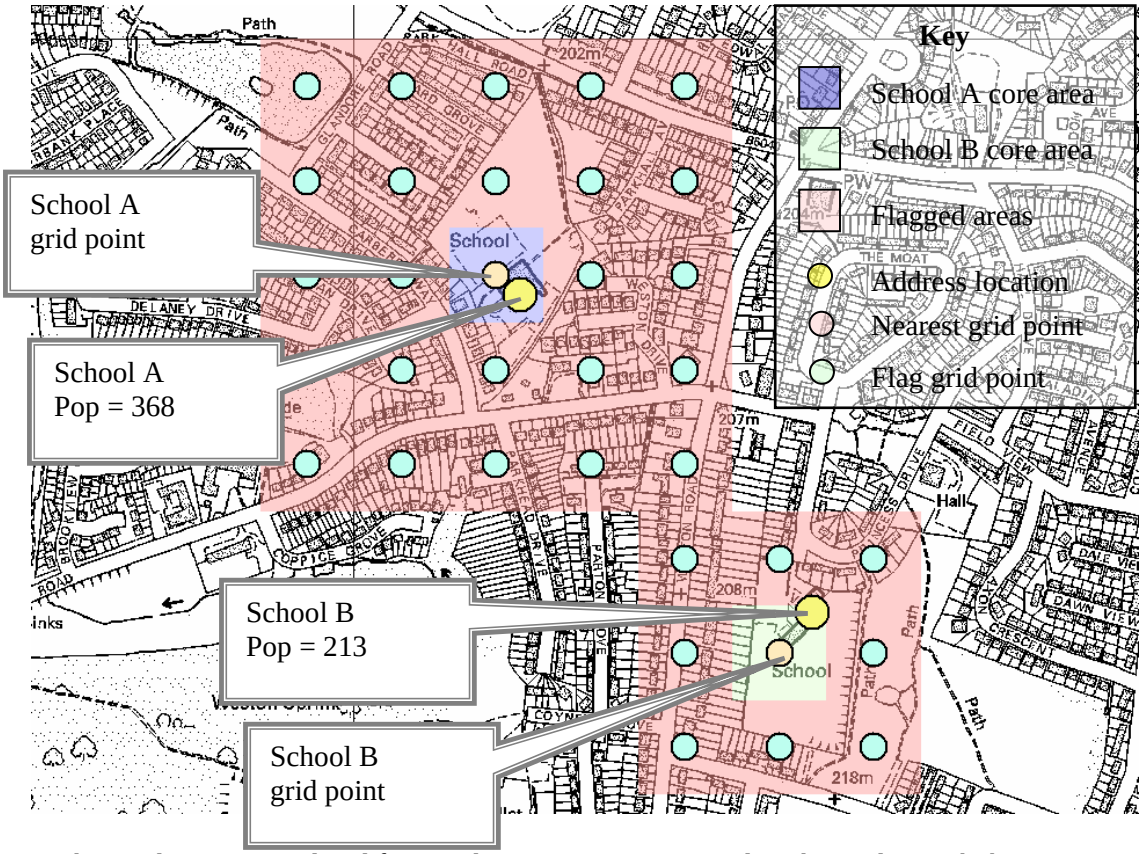
The school locations were then joined to the nearest 100m grid point, to give a school core area. Flags were then created for the following reasons:

1. the spatial extent of the school is unknown and varies considerably between schools,
2. generalising to the 100m grid introduces positional uncertainty.

Two types of flag were used depending on the number of pupils attending a school. A 1 point rule was used to create flags if the number of pupils was less than or equal to 300, a 2 point rule was used to create flags if the number of pupils was greater than 300.

Figure 6.2 illustrates two schools, A and B, with the two types of flag. School A has a flag created using the 2 point rule in which all points that fall within two points of the core point are classified as flag points. Further details of the flag methodology are given in chapter 3. School A illustrates why the use of a 2 point flag is useful. The open area to the north of the school, which forms part of the school grounds, extends to the edge of the flagged area. School B shows the spatial extent of the 1 point flag.

Figure 6.2 Examples of Schools with flagged areas



Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

6.3.3 Hospitals

The hospital layer consists of hospitals that have in-patients and bed counts. Hospitals and other care facilities with out-patients only were not included in the database. Hospitals represent the largest communal establishment represented in the database. Due to the large spatial extents and dispersed populations of some hospitals, a different approach to other communal establishment layers was taken. A core area for each hospital, derived from the **Gross Internal Floor Space** (GIFS – m^2) was created and the population dispersed within the core area. GIFS data was not available for Scotland and another technique was therefore used.

Methodology

Hospitals were geo-coded using AddressPoint to give the hospital address point layer. For England and Wales the hospital core area was then modelled on GIFS using following formula

$$r = \sqrt{\frac{GIFS}{\pi}}$$

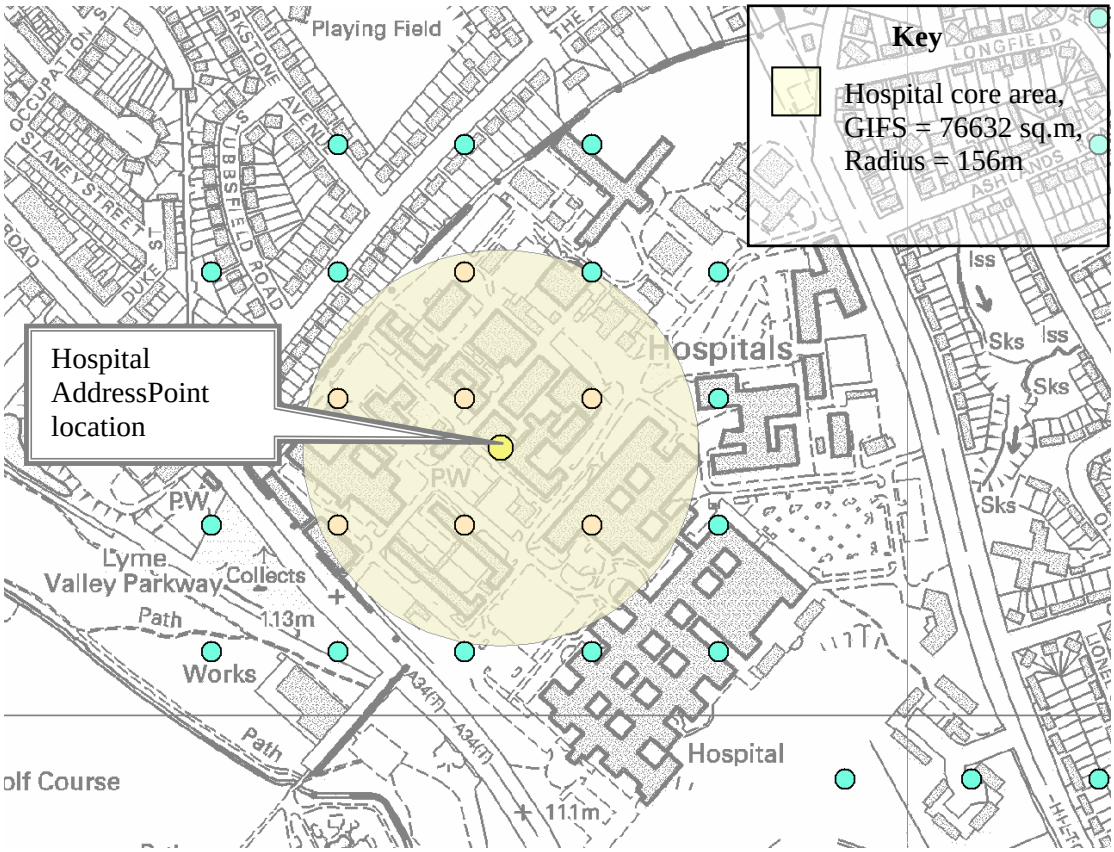
Equation 6.1 Core Area Formula

where r = *radius of core area*

$GIFS$ = *Gross Internal Floor Space*

From this radius a circular core area was generated (figure 6.3) using the address point as the circle centroid. 100m grid points falling within the circle were defined as hospital core area grid points. The population associated with the hospital was then distributed over the core area grid points to give a *distributed* population, an approach which has not been used for other communal establishments.

Figure 6.3: Hospital Core Area



Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

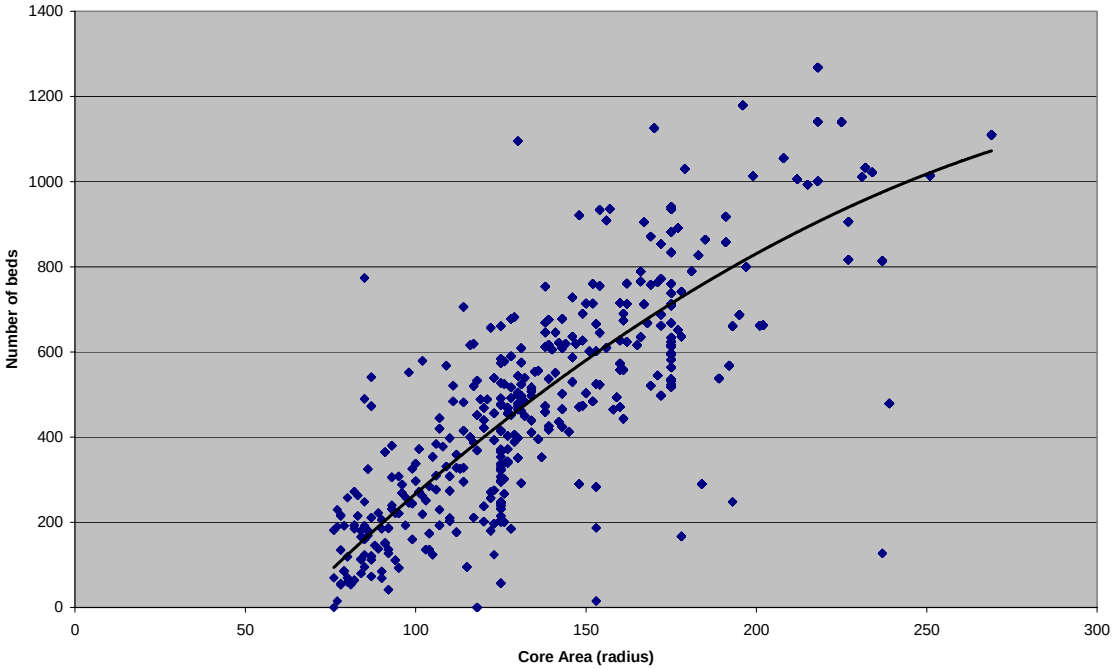
Figure 6.3 demonstrates a working example of this process for a large hospital. The hospital address point location is indicated by the yellow point. The radius derived from the GIFS is 156m and gives rise to the circle defining the core area. Figure 6.4 illustrates the next step in the process. The 100m grid points within the core area form the basis of a flag created using the 1 point rule i.e. all points that neighbour a core area become flag points. The flagged area does not provide any information about the spread of population within the flagged zone, instead providing information about the total hospital population only. The resulting model is moderately successful at capturing the hospitals spatial extent, with the core area in which the population is spread, positioned over hospital buildings.

Figure 6.4: Hospital Core and Flag areas. The flagged area to the bottom right hand side is from a different hospital.

Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

GIFS data for Scottish hospitals was not available, however 'beds available' data was. In order to ascertain the core area for Scottish hospitals, the relationship between beds available and floor space for English and Welsh hospitals was calculated. This allowed a core area to be approximated for Scottish hospitals based on beds available. Figure 6.5 is a scatter graph of the population and radius of core area. The best explanation of the relationship was given by a quadratic equation.

Figure 6.5 Relationship between beds available and radius of core area for English and Welsh hospitals.



The resulting quadratic equation was then used to classify Scottish hospital populations into four bands with different sized core areas (table 6.4)

Table 6.4: Scottish hospital core area bands

Population	*Core Area (radius)*
<200	75 m
200 – 500	125m
500 – 1000	175m
>1000	200m

6.3.4 Prisons

Prisons are characterised by a compact spatial form. Since they are not considered a sensitive population and GIFS data was not available, no attempt was made to model a core area spatial extent. In part this was because there is no clear relationship between prison capacity and spatial form. This is discussed further detail later in this section. The methodology is therefore similar to that used for care homes, and is detailed below.

Methodology

Prisons were geo-coded using AddressPoint and attached to the nearest grid point. A one point flag was then created. Figure 6.6 illustrates the resulting representation in the database. In this example the spatial extent of the prison has been successfully modelled by the core and flag areas.

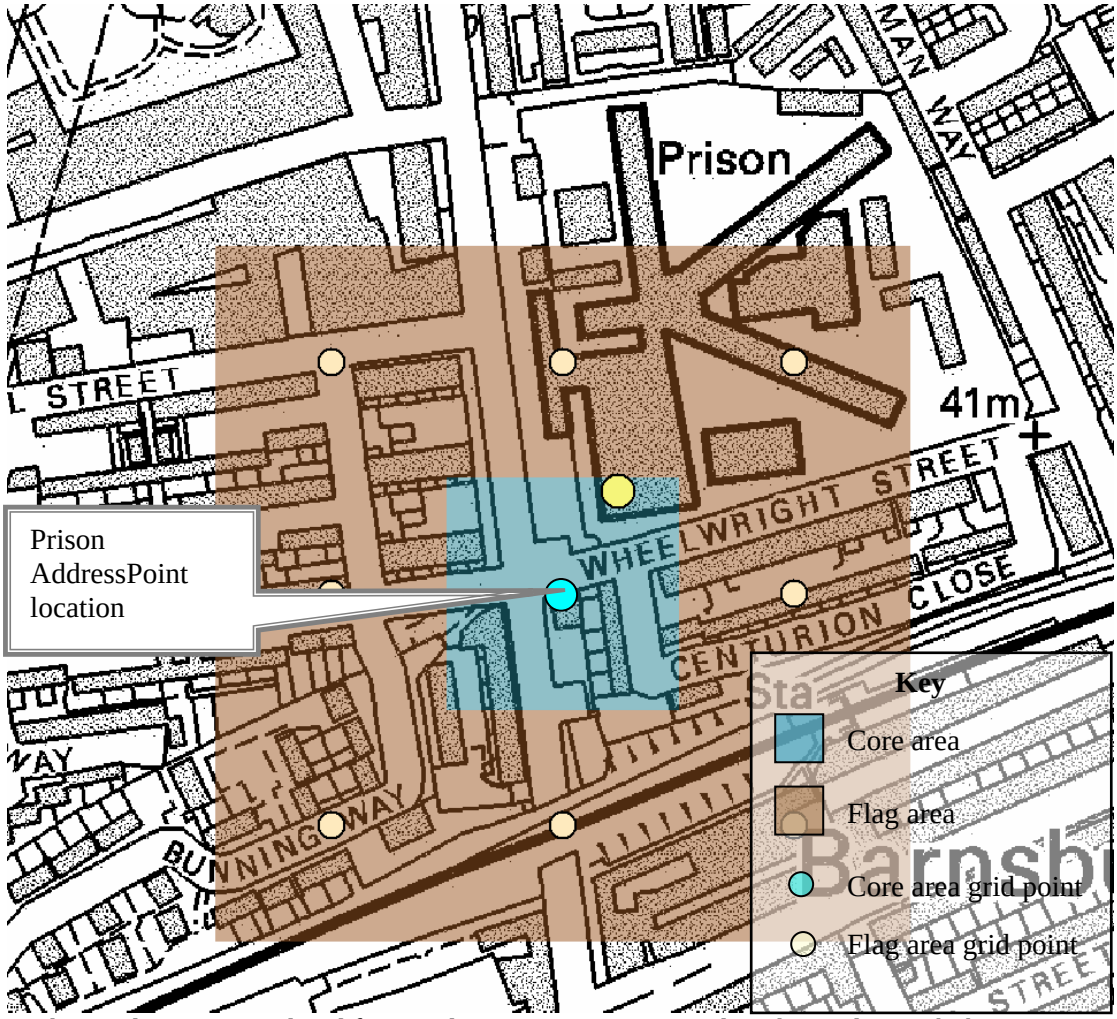
Figure 6.6: Prison example A. Prison population = 480

Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

Figure 6.7 is an example in which the spatial extent of the prison is not as successfully captured by the flag because the address location occurs at the bottom left corner of the prison spatial extent. The nearest grid point happens to be the furthest away from the prison resulting in some of the prison lying outside the flagged area. For the majority of prisons and other communal establishments the first prison example (Figure 6.6) characterises the most common scenario. The second example is a worst case scenario.

The populations of the two prisons also illustrate a characteristic that makes the prisons spatial extent difficult to model from the prison numbers alone and is a justification for not attempting core area modelling. The spatial extent of prison A appears to be larger than prison B, however the population of prison B is over twice that of prison A. This illustrates a difficulty in characterising and modelling prisons.

Figure 6.7: Prison example B. Prison population 1167



Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

6.4 Verification of the Data

The data was verified by visual inspection and comparison of modelled spatial extents to actual spatial extents interpreted either from aerial photography of the study site or from the 1:10,000 raster for those layers that contained no examples within the study site.

6.5 Evaluation of the Data

The primary limitations of the data are the positional uncertainty associated with the core areas. To help mitigate this limitation a flag approach has been used. The flag marks a buffer zone into which part of the communal establishment is likely to extend, however virtually no parts of the communal establishment should lie outside of the flag area. This both alerts the user to a potential problem and also prevents a spurious population being presented that may mislead a user.

Whilst this rule is adhered to for the sensitive communal establishments, it is relaxed for prisons, in part because they are not considered a sensitive population. Scottish boarding schools have been omitted due to lack of data, as have care homes with a population of less than 10.

7 WORKPLACE POPULATIONS

7.1 Population Characteristics and Variability

Workplace populations are arguably the most dynamic and problematic populations that have been included in the database. Unlike populations such as residential households, workplaces vary enormously. They vary in both physical size and workforce size, but these two variables do not share a common relationship. A relatively small office building can contain a large and high density population; whereas a very large warehouse can contain a very small, very low density population.

These characteristics, combined with the limitations of the available source data, mean that the layer reporting workplace populations has the most limitations and provides the most problems when looking at small spatial areas. It is also important to note that workplace populations can change very rapidly and substantially over time as companies expand, restructure their workforce, or shut down.

7.2 Source Data Sets

- *AddressPoint.* AddressPoint gives a 1 metre accurate grid reference to every postal address in Royal Mail's Postal Address File (PAF). Addresses classified as commercial were extracted from AddressPoint. This was based on the CodePoint definition, *"Non PO-Box addresses that have a PAF Organisation Name."*

The following data was used from the 2001 Census at Output Area Level:

- *Univariate Statistics Table UV75 Age (Workplace Population).* This was the primary source of workplace population figures. These are aggregated figures at Output Area Level.

- *Census Output Area Boundaries.* The Census Output Area boundaries were also used as locational data. The aggregated population figures were attached to the geographical centroids. The reasons for this are explained in section 7.3

Source data on workplace populations has a number of limitations:

- Output area level data is not sufficiently precise to identify the population of a specific workplace.

- AddressPoint gives an accurate location but does not give an indication of workplace size (physical or population).

- Although the additional use of a product such as OS MasterMap would help to determine the physical size of a property, this would not be a suitable indicator of workplace population.

- Many workplaces have PO-Box postcodes. AddressPoint locates these at the nearest Royal Mail depot, not at the location of the workplace itself.

7.3 Data Transformation and Processing

Due to the limitations of source data it is not possible to produce a population layer that gives populations that could be seen as 'site' specific. A number of alternative options were therefore considered:

- *Option 1 - Spread Workplace Populations Evenly.* Spread workplace populations across commercial addresses and then assign to the 100m by 100m grid. *Option 1 was considered inappropriate because there is no advantage in spreading the population when it is known that this is wrong, as the variability is so great.*

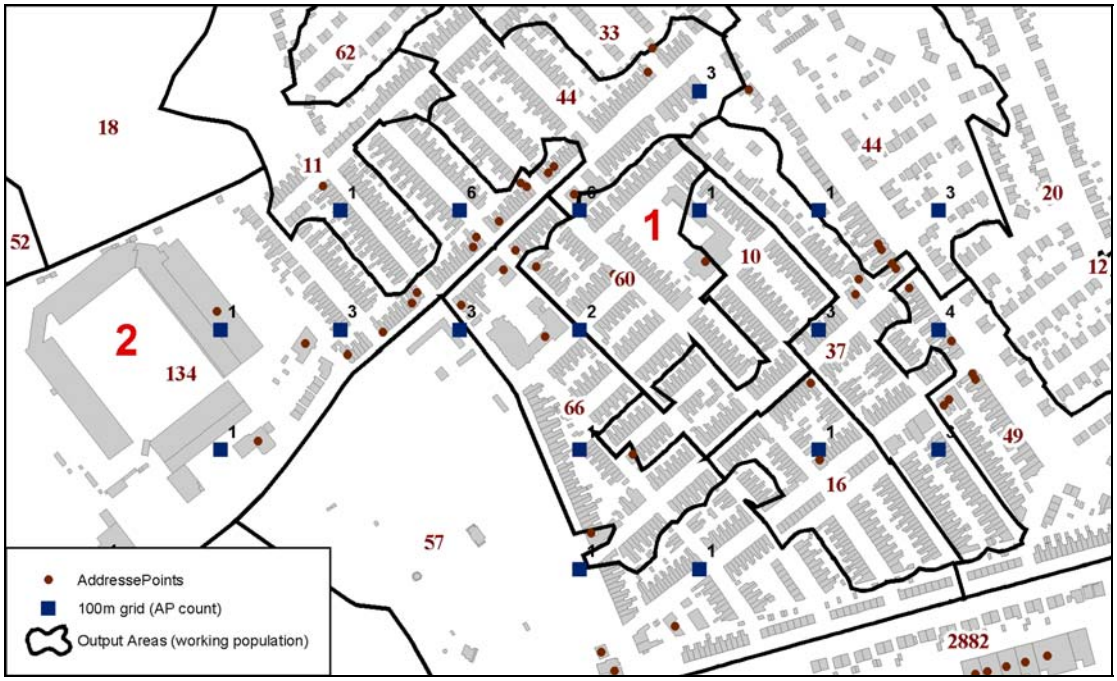
- *Option 2 - Assign Workplace Populations to Multiple Centroids.* This method produced multiple centroids within an output area based on clusters of AddressPoints. The population assigned was based on the number of AddressPoints in the cluster. *Option 2 was considered a good option as far as locating clusters of workplaces. The method has the same problem as option 1 when it comes to assigning populations.*

- *Option 3 - Standard Densities.* This option could start as option 1 or 2 but then the populations assigned would be based on approximate underlying land use. Ward based data of floor space (from the ODPM) differentiated by retail, office and industry would be used to approximate 3 different population densities. AddressPoints are assigned populations based on the average composition of the area with regard to proportion of retail, office and industry. *Option 3 is attempting to be more accurate but the populations would only be generalised and indicate a spurious level of accuracy.*

- *Option 4 - Indicator Values (or flags).* This option would assign indicators to grid cells that report the workplace population of the entire output area it lies within. This approach could also be used in conjunction with option 2. *Option 4 was considered a good alternative because it gives the total workplace population of the area a workplace lies within, but does not attempt to assign site populations.*

After examining these options the initial approach was to highlight clusters (locations) of workplaces on a 100m by 100m grid using the AddressPoints. Then for each grid point give a **count** *of the number of workplaces* and flag up the census output area that it falls within as well as the *total workplace population of the output area.*

Figure 7.1 Results of the initial approach to assigning workplace populations to a 100m by 100m grid



Background image reproduced from Ordnance Survey MasterMap data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

Figure 7.1 gives an example from this initial approach and highlights a number of problems. Looking to the right of point 1 (in red) in the above figure it can be seen that an AddressPoint is actually in a different Output Area than the 100m grid point that it is assigned to. This was a common problem in areas that have mixed residential and commercial land use and have relatively small output areas. The AddressPoint is in an Output Area with a working population of 60 but the grid point is in an OA with a population of 10. This resulted in workplaces being incorrectly flagged.

- *Solution:* the population flag was attached to the AddressPoint first and then transferred to the grid point. However, looking to the left of point 1 it can be seen that there is a grid point that has AddressPoints from 3 OAs. In this instance the above solution did not work because of difficulties in determining which Output Area should be used for the flag.

Point 2 (in red) highlights the location of a potentially large workplace in the same Output Area as a number of small workplaces.

- This example is actually a football stadium. This is potentially quite a large employer and sits in an OA with a working population of 134. In the database it was simply flagged as a grid point with a count of 1 address.
- The Output Area is relatively large and a very strange shape (as a lot of them are) and at the other end there is a more concentrated set of addresses which are all small workplaces such as a barbers and corner shops etc. These sites were assigned more population than the football stadium and this was known to be inaccurate.

Following discussions with the HSE regarding the problems and limitations shown above a

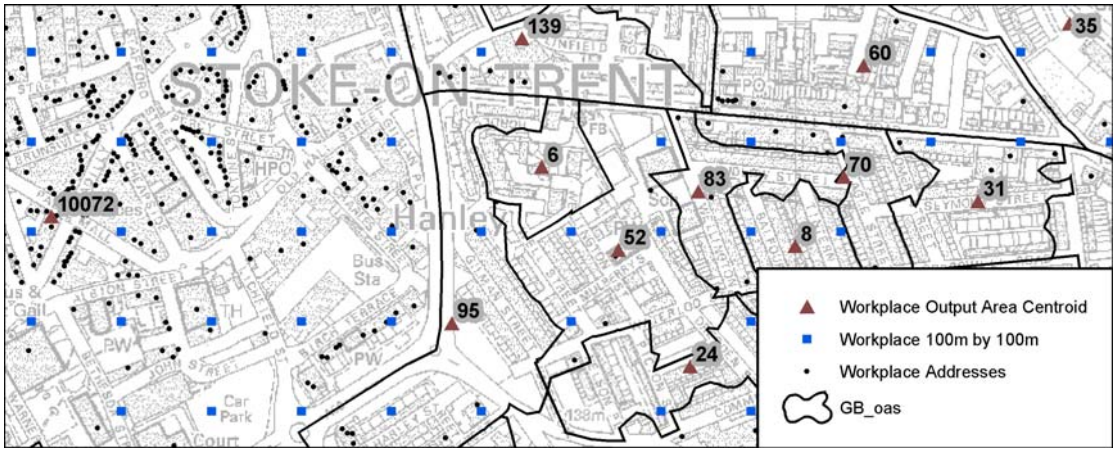
simplified approach was agreed. Three versions of the workplace layer were produced:

- *Workplace Address Layer.* This layer contains all commercial non PO-Box AddressPoints. The purpose of the layer is simply to locate workplaces. In addition, each location has address information and indicates the census Output Area that it is located in. No population data is attached to the addresses.

- *Workplace 100m by 100m Grid Layer.* This layer is a 100m by 100m generalisation of the Workplace Address layer. Each grid point contains a count of the number of commercial addresses. The purpose of the layer is to highlight clusters of addresses. No population data is attached to the grid points.

- *Workplace Output Area Centroid Layer.* This layer has a point to represent the geographical centroid of every Output Area that has a workplace population. This population is attached to each point. The purpose of this layer is to report the workplace population data. Figure 7.2 gives an example of these three workplace layers.

Figure 7.2 An example of the three workplace population layers



Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

7.4 Verification of the Data

The workplace layers contained in the population database are simply reporting the locations of workplaces and give workplace populations at Output Area level. This is non site specific data with a number of known limitations and therefore no specific verification of the data was carried out.

7.5 Evaluation of the Data

The original aim of the workplace population layer was to produce site specific (or as close to it as possible) populations. The source datasets available had a number of limitations and have been discussed above. These limitations made the original aims difficult to meet. In addition to this the release of workplace figures in the census was delayed significantly because of concerns within the Office for National Statistics regarding disclosure of individual workplaces. When the census data was finally released there were two tables missing that could have been useful. These contained a breakdown of the workplace populations by occupation and type of industry. If these tables were available it would allow a method that predicts workplace size to be evaluated. The final result is three layers that simply report the

locations of workplaces and give workplace populations at Output Area level. This restricts the user of the data in a number of ways:

- *Individual Workplaces*. The user does not have the populations of individual workplaces. However, the user does have accurate locational data for workplaces as well as a generalised version of this data to help pick out workplace clusters.

- *PO-Box Workplaces*. Workplaces that are PO-Box addresses are not located in the database. These addresses are usually large user postcodes which mean that they receive a high volume of mail. It is therefore sensible to conclude that these are potentially large employers.

- *Generalised Area Data*. Population data is limited to generalisations of Output Areas in the form of centroids. This is a bigger problem when the user is looking at small risk areas because Output Areas vary enormously in both size and shape and a major effect of this will be the varying degree of edge effects experienced when selecting and analysing data. A major benefit of including this area data is that the data contains populations from the missing PO-Box workplaces and therefore the user could observe a centroid indicating a workplace population in an Output Area where there is no site locational data.

If alternative workplace datasets become available in the future then this population database could be enhanced considerably. One possible alternative could be the use the Interdepartmental Business Register available within government departments, but only if it can be significantly enhanced and incorporated within a GIS environment. However, at the moment source data on site specific workplace populations is simply not available, particularly at a national level.

8 RETAIL POPULATIONS

8.1 Population Characteristics and Variability

'Retail Populations' refer to people who visit areas of retail land use to shop or recreate. An individual's retail habits and behaviour vary enormously as they are influenced by a number of factors, including, the type of retail activity and the size of a retail area.

Retail areas vary in size from small clusters of shops serving a local community to large regional city centres or retail parks. For a number of reasons, particularly those related to potential double counting of populations (see section 3.3), the database populates only larger retail areas expected to draw in population from a wide area that is over and above the neighbourhood population that would shop there. A number of small town centres will be included as locations only but this excludes smaller local clusters of shops. In addition to the physical size of a retail area the level of population is influenced by the type of retail activity, for example grocery retailing has very different patterns to comparison shopping. The type of retail activity also has an influence on the frequency and duration of an individual's shopping activity.

Temporal factors also play a major role in observed patterns of retail activity, from seasonal variations, to weekday / weekend variations, to daytime / night time variations.

All of these factors have a varying influence on specific retail areas and not only do they produce differences between retail areas but within them too. Producing an accurate picture of how populations differ within a retail area usually takes a detailed site specific study. Once a pattern of retail activity has been produced, turning this into a population that is within a retail area at any given time is extremely difficult. With this in mind, as well as the factors mentioned above, the task of accurately producing a retail population layer at a national scale becomes extremely difficult.

8.2 Source Data Sets

Source datasets were evaluated and chosen with a number of goals in mind. The source data needed to:

1. Locate a wide range of retail areas.
2. Give an indication of the type and size of the retail area.
3. Give an indication of the type of retail activity within the area.

The dataset used were:.

- *AddressPoint* - Addresses classified as commercial can be extracted from AddressPoint. This will be based on the CodePoint definition, *"Non PO-Box addresses that have a PAF Organisation Name"*

- *Retail Footprint* - Locates 2600 retail centres. In addition the data includes a classification of the type of centre i.e. Major Regional Centre, Small District Centre, Out of Town etc. Also provided is a Weighted Population figure which relates to the catchments of retail centres. This data was purchased from CACI Limited.

- *Retail Locations* - Locates 38,000 retail stores defined as multiples **(chain stores)**. The stores included fall into 20 retail categories. The full database included approx 70

categories but cost constraints limited the choice to 20. Categories were chosen based on their usefulness in helping to locate different types of retail areas. This data was purchased from CACI Limited.

- *Statistical Areas of Town Centre Activity (ATCA)* - Boundaries and statistics for consistently defined Areas of Town Centre Activity in England and Wales relating to 2000 data on employment, net internal floor space and rateable value for 1029 Areas of Town Centre Activity and 46 Retail Cores (concentrations of retail activity in large town centres) produced for the Office of the Deputy Prime Minister. This dataset is only available for England and Wales. This data is available through the Office of the Deputy Prime Minister and can be accessed at: http://www.iggi.gov.uk/towncent/index.htm.

8.3 Data Transformation and Processing

There are two tasks involved in producing the retail population layer. The first is to correctly locate and classify retail areas and the second is to assign populations to these areas.

8.3.1 Locating and Defining Retail Areas

During the initial stages of the project it was thought to be desirable that the retail layer be on a 50m by 50m grid. Following examination of the source data and formulation of the method it became clear that a 100m by 100m grid was more suitable.

A 50m by 50m grid can produce 'holes' within a retail area and this can give a false sense of accuracy in the data. The 100m by 100m grid gives a more continuous pattern within retail areas which is more appropriate given the source data and methods used to produce the layer.

Grid cells were classified based on a number of factors, some more important that others. Retail areas fall into six categories., Tthese are:

- *Town Centres* – Medium to large town or city centre areas with average levels of retail activity.

- *Retail Cores* – Large city centres have core retail areas where the level of retail activity is higher than the rest of the centre.

- *Small Town Centres* – Smaller town centres typically less than 4 hectares in size. They would also be expected to serve much smaller catchments of population.

- *Large Out of Town Centres* – Large 'regional' out of town centres, such as, The Trafford Centre or Bluewater. Only 9 of these centres exist in the UK.

- *Retail Parks* – Small to medium sized 'district' out-of-town retail parks.

- *Other Non Town Centre Retail* – Other non town centre retail activity believed to be attracting more than just local populations.

The classification listed above was produced using the source datasets. This process followed the hierarchical structure listed below:

Step 1. Is a 100m by 100m grid point within an Area of Town Centre Activity?

Digital Boundaries for ATCAs were obtained through the ODPM. This dataset includes town centre retail areas greater than 4 hectares in size and also classifies retail cores within large city centres. The dataset also includes very large out of town centres. These were picked out using the Retail Footprint dataset.

Therefore, grid points falling within the boundaries of an ATCA were classified as either:

a. Town Centres
b. Retail Cores
c. Large Out of Town Centres

At this stage small town centres and non town centre retail areas, other than large out of town centres have not been classified. This is achieved in the next steps.

Step 2. Is a 100m by 100m grid point within 125m of a Retail Footprint point?

The CACI retail footprint data includes a classification hierarchy and this includes smaller retail areas that were not defined by the ATCA data. The classification was used to pick out the smaller town centre areas and these points were buffered by 125m. This buffer distance was chosen because it will pick up most of the neighbouring 100m by 100m grid points and this will act as a good indicator of the location of small town centres. These centres will be typically less than 4 hectares, thus making a rough approximation of the spatial extent of the centre more acceptable. The retail footprint data also includes retail park classifications. These points were also buffered by 125m and then used to flag up areas likely to be a retail park. The spatial extent of larger retail parks will be accounted for in the next step using the retail locations dataset.

Therefore, grid points falling within 125m of chosen retail footprint points were classified as either:

d. Small Town Centres
e. Retail Parks

At this stage some smaller retail parks and stand alone non town centre retail areas have not been classified. This is achieved in the next step.

Step 3. Is a 100m by 100m grid point near to a Retail Locations point?

The retail locations data contains 20 categories of retail multiples. These categories were chosen to help in the classification of non town centre retail areas. Many of these points will be part of retail parks but many will be stand alone. A stand alone retail location could be a large site such as a supermarket or a DIY warehouse or it could be a small site such as a public house or restaurant. Therefore, retail locations were used differently depending on whether they were classified as a small or large site. This classification was based on the retail location category of each site. The details of this classification can be found in Table 8.1.

Table 8.1 The classification of large and small retail location sites.

Stand alone size: Small / Large	*Retail Locations Categories*
Large	General Furniture, DIY Superstores, Bedroom Furniture, Kitchen Furniture, Cash & Carry, Carpets, Department Stores, Cinemas, Bingo Halls, Computer Centres, Warehouse Clubs, Bowling Alleys.
Small	Pubs / Restaurants, Hotels, Restaurants, Builders Merchants, Health Clubs/Gyms, Swimming Pools (Public), Roadside Eateries.

Step 3.1. Is a 100m by 100m grid point within 125m of a Retail Locations point that is classified as large?

Retail locations classified as large were buffered by 125m. This method picks up some of the neighbouring grid points and acts as a flag for the location of each site. These points were buffered to take account of their large spatial extent as well as areas such as car parking around the site.

Step 3.2. Does a 100m by 100m grid point contain a Retail Locations point that is classified as small?

Retail locations classified as small were joined to the nearest grid point. The sites were deemed unlikely to have a spatial extent much larger than 100m by 100m and would not have the same levels of car parking.

Therefore, grid points within 125m of large retail locations sites or containing small retail locations sites were classified as:

f. Other Non Town Centre Retail

The three steps discussed above have provide the user with a good overall picture of the retail land use in both town centre and out of town areas. The user also has a good indication of the size and type of retail area in question.

8.3.2 Assigning visitor populations to retail areas.

The final stage of the retail method was to assign suitable retail populations to the grid cells defined as retail. There is no specific source data that will provide this data ready for use. Therefore estimates of visiting populations at any one time have been produced from a number of sources.

The major factor used in estimating populations within the retail layer has been floorspace. The total retail floor space of a town is included within the Statistical Areas of Town Centre Activity data. This figure was used to multiply up a standard retail population density figure. This total population was then spread throughout the town centre. Core retail areas have their own floorspace figures and can be expected to have higher densities of population than the rest of the town centre area.

As a result of floorspace being the determining factor in calculations of population only those areas classified using the ATCA data will be populated. This includes:

a. Town Centres
b. Retail Cores

c. Large Out of Town Centres

Areas classified as Small Town Centres, Retail Parks and Other Non Town Centre Retail do not have floorspace figures available and therefore they will not be populated.

A number of sources were investigated in order to provide a suitable population density figure for retail areas. These included:

- Fire and Safety Building regulations
- Pedestrian Flow data
- Crowd modelling literature
- Retail Levels of Service literature
- Car Parking Standards and Statistics

The final density figures were produced from a combination of building regulations and car parking statistics and testing in the study sample area. Using building regulations as a guide, a maximum population density figure of 10 sq.m per person was produced. In shop floor areas alone a figure as high as 2 sq.m per person could be used. However, the retail floorspace figures include all areas of retail outlets, including storage areas. In addition, it would be incorrect to assume such a density across all of a retail space because a high proportion of space is occupied by the 'goods' being sold.

Using very detailed car park occupancy figures for Stoke-on-Trent, a number of assumptions were made which are key to the population figures. In addition to people driving into a town centre area, much of the population is made up of people using public transport or walking in. Reliable data on these people could not be found and therefore, drive-in populations were increased by 10% to compensate. The key assumptions derived from the car parking data were:

1. Average vehicle occupancy at peak retail periods is 2.4 persons.
2. Retail Car Parks are full at peak retail periods.
3. The average occupancy of car parks at peak weekend periods is 2.2 times greater than peak weekday periods.

These assumptions were used to predict visiting retail populations and these were then applied to the known retail floorspace figures to produce population densities. These standard densities were produced to apply to town centre areas, including retail cores. These densities were too low for out of town centres and were recalculated separately using car park spaces data for the Trafford Centre but using the same underlying assumptions as town centres. Table 8.2 gives the details of these standard population densities which were applied to known floorspace figures.

Table 8.2 Standard retail population density figures

Population description	***Retail Population Density (sq. m per person)***	
	Town Centres / Retail Cores	***Out of Town Areas***
Maximum population	10	4
Weekend Peak Population	12	5
Weekday Peak Population	26	11

These figures were combined with the floorspace figures to give total populations for each town centre or out of town area. These populations were then spread evenly across all of the 100m by 100m grid points that classify each centre.

8.4 Town Centres in Scotland

In England and Wales large town centre areas were defined using the Statistical Areas of Town Centre Activity dataset accessed through the ODPM. This data was not available for Scotland, making the modelling of a town centre more problematic. There are 292 centres (as defined by the Retail footprint dataset) in Scotland but only 24 fall into 'larger' classification categories.

As a result a combination of looking at the CACI Retail Footprint points, Retail Locations points and commercial AddressPoints was used. In combination with 1:10,000 raster background mapping the edges of the 24 larger town centres were defined manually.

Only these 24 centres were populated for Scotland. In order to do this the retail floorspace needed to be estimated for each centre. This was achieved by looking at the relationship between the amount of retail floorspace and the weighted catchment population (included in the Retail Footprint data) for a large sample town centres in England and Wales. The floorspace figures were then used to populate the centres in the same way and with the same assumptions as town centres in England and Wales.

This means that large town centres and large out of town centres in Scotland need to treated with caution when using the data. The method for classifying small town centres, retail parks and other non town centre retail in Scotland remains unchanged.

8.5 Evaluation of the Data

Producing a national population database of retail populations has proven very problematic. The major source of the problem is the lack of good quality consistent data regarding retail populations.

The classification of retail areas has largely been a success. This has been achieved through the use of good location data. The user can determine the type of retail area being looked at as well as its physical size.

It is extremely important to note that populating these areas has been based on a small set of very general assumptions and the figures should be used with a great deal of caution. The figures are attempting to reflect high density scenarios. With this in mind, if the user would like to look at more low level density scenarios it is recommended that some local verification of the data is carried out.

The use of very general assumptions will lead to potentially large errors in some town centres. Retail centres can be very unique. In reality there is huge variation between centres and temporal factors, such as, seasonal variations, weekday/weekend variations and daytime/night time variations all have different effects on individual retail areas.

In addition, the populations contained in this database have been spread evenly throughout a town centre. In reality there are big variations between levels of retail activity/population within a town centre area and these variations can be expected to change during different times of the day, week or year.

9 LEISURE FACILITIES

9.1 Population Characteristics and Variability

The leisure facilities layer contains two primary data sets,

a) Stadia locations and maximum capacity,
b) Recreational facilities, locations only.

The population associated with stadia vary significantly over time and are, for the majority of time, a negligible population. However for brief time periods they can hold significantly high and dense populations. Obtaining data for the other recreational facilities proved problematic and has not been possible to include in this release of the database.

9.2 Source Data Sets

A list of stadia was produced by internet research to find a number of different sources that were compiled to produce the current list. The remaining recreational facilities were clipped from OS Strategi 2003.

Table 9.1 Source Data sets for recreational facilities

Facility	*Facility sub type*	*Count*	*Source*	*Additional information*
Stadia	Athletics	1	Internet directory	Maximum capacity
	Basketball	2		
	Cricket	18		
	Football	121		
	Ice Hockey	2		
	Rugby	14		
	Rugby League	6		
	Rugby Union	7		
	Speedway	1		
	Tennis	1		
Camp Sites	(Location only)		OS Strategi 2003	Location Only
	Camp Sites	239		
	Caravan Sites	587		
Public Attractions	(Location only)		OS Strategi 2003	Location Only
	Aquarium	36		
	Historic House	456		
	Motor Racing Circuit	32		
	Racecourse	61		
	Theme Park	52		
	Wildlife Centre	72		
	Zoo	40		

Table 9.2 Location Data sets for Recreational Facilities

Data Set	*Details*
OS Code Point	November 2003 release
OS Strategi	2003 release

9.3 Data Transformation and Processing

9.3.1 Stadia Data Transformation and Processing

Stadia data was geo-coded using the stadium address and OS Code Point. Code Point was used in preference to OS Address Point because features the size of stadia are inevitably associated with a single postcode area. Since OS Code Point gives the centroid of the postcode area, this would provide a more accurate spatial reference to the stadium than using address point which would indicate the location of the addressed building. The resulting point was joined to the nearest 100m grid point. In turn the grid point was given a two point flag (see chapter 3 for details on the creation of flags).

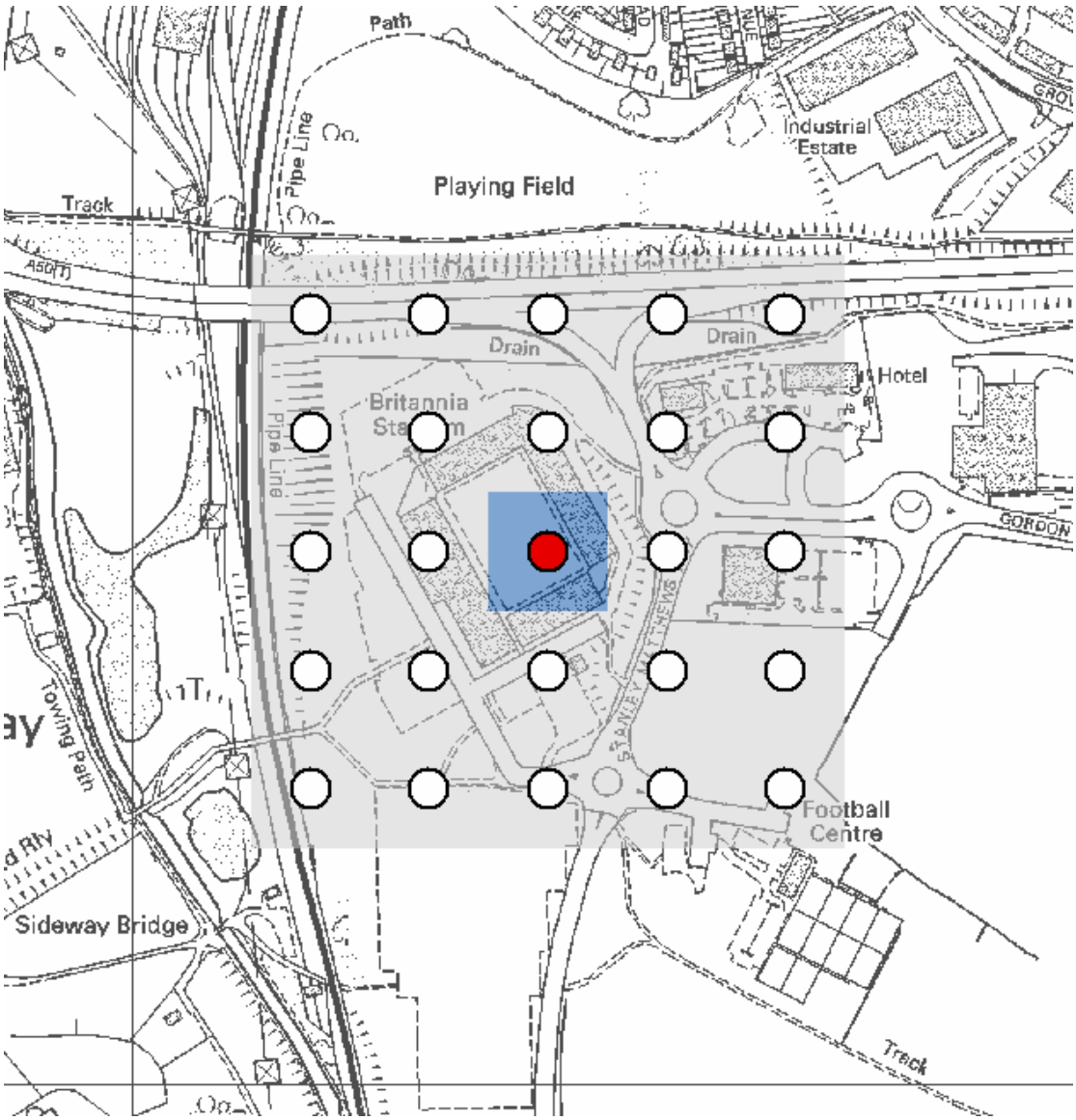
9.3.2 Recreational Facilities Data Transformation and Processing

The remaining recreational facilities were already geo-coded and therefore did not require any further geo-coding. Unnecessary data was stripped from the Strategi tables and the resulting points were not generalised to the grid. Since the recreation points are therefore the original address point, they were stored in the Address point database as opposed the 100m grid database..

9.4 Verification of the Data

The data for this layer was verified by examining sample areas. Figure 9.1 and 9.2 are two such examples.

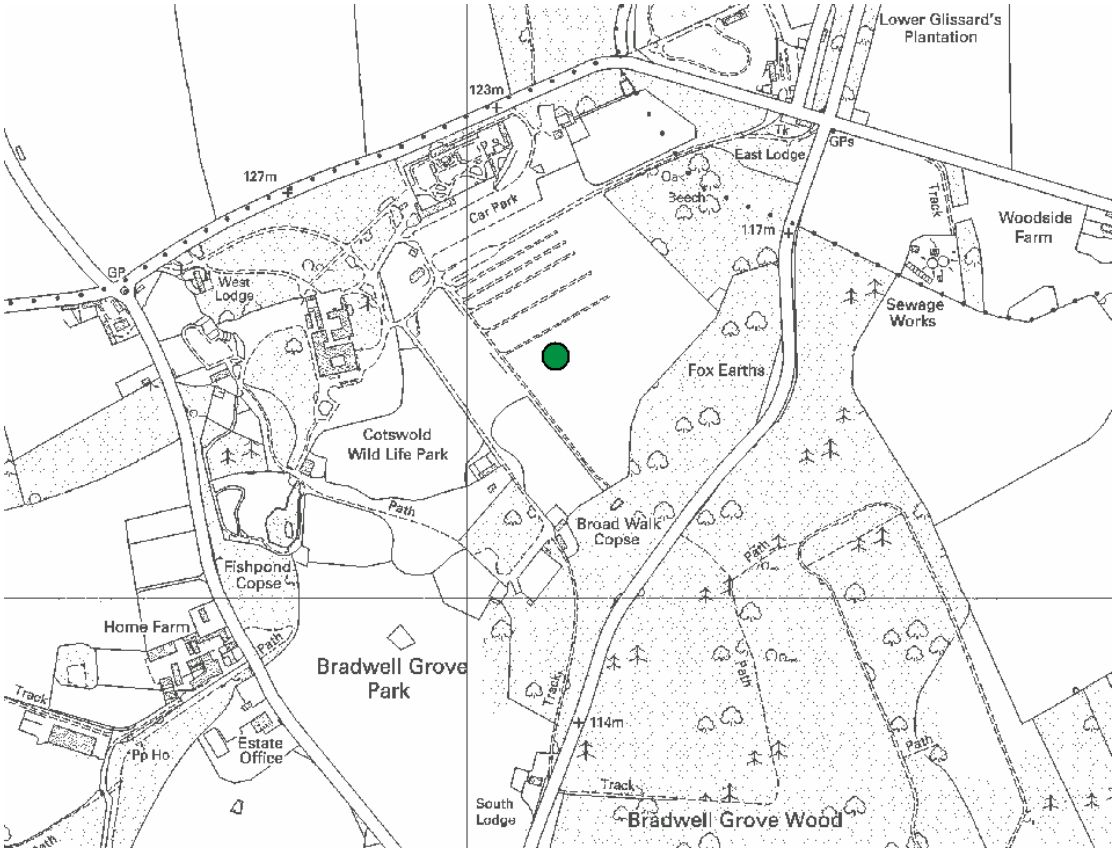
Figure 9.1 Stadia example: Stoke FC football club



Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

Figure 9.1 illustrates a football ground. The ground is completely contained within the flagged area, as are the parking facilities to the top left corner of the ground. Figure 9.2 illustrates a country park in the leisure layer. Because of the large variations in size associated with leisure facilities, neither the spatial extents were modelled nor flagged zones created. Therefore local knowledge and other augmentative data such as the 1:10,000 raster background may be needed to identify the spatial extent of most leisure facilities.

Figure 9.2: Leisure Facility example: Cotswold Wild Life Park.



Background image reproduced from Ordnance Survey 1:10,000 digital map data with the permission of the Controller of Her Majesty's Stationary Office © Crown Copyright

9.5 Evaluation of the Data

Leisure facilities are largely included and flagged up to alert the user to the potential for occasional large increases in population within the area. They would also alert the user to potential sites that could be used during a hazardous event for evacuation purposes.

Limitations to stadia and leisure facility layers are:

- The list of stadia is not exhaustive,
- The list of leisure facilities is derived from Strategi, there might be other facilities that warrant inclusion but are not recorded by Strategi and therefore not included in the current database.
- Flags are not included for the leisure facilities
- Populations are not included for leisure facilities. This may be particularly relevant for camping and caravan sites.

10 FUTURE DEVELOPMENT AND APPLICATIONS

This section considers some of the issues involved with updating and maintaining the different layers. It also identifies ways in which the database can be used by the HSE and other potential users of the database within government.

10.1 Updating the data layers

Following discussions with members of the HSE and HSL it is recommended that the National Population Database update requirements are evaluated 18 months after the first NPD is delivered.

10.1.1 Residential population

Key datasets for the production of this layer have been the 2001 census (which occurs once every ten years) and Address Point (a product from Ordnance Survey licensed under the PAN government OS agreement).

The Address Point data is particularly important, however local authorities are currently compiling Local Land and Property Gazetteers (LLPG) which when finished will create a National Land and Property Gazetteer (NLPG) in December 2005. The National Land and Property Gazetteer is being co-ordinated through a group of public and private sector bodies (see www.nlpg.org.uk for details).

The NLPG will be continually maintained and should provide an even more accurate set of data than address point. However acquiring access to it would cost money and negotiations are still going on between the parties as to what the level of charge and licensing should be. Currently it is possible to acquire a “development and demonstration” licence, which includes access to the full dataset and updates for £4750 per annum. The NLPG will also include information on houses that are demolished as well as new additions through use of the Unique Property Reference Number (UPRN). The intention is also to have a flag to distinguish residential and commercial properties. This should be a more accurate dataset than Address Point, although Address Point would still be available under the PAN OS agreement.

Population shifts could be worked into the model using the annual estimates from the ONS to adjust overall population, although this will not be at any level lower than local authority. It is debatable how useful it would be to update the database every year given the amount of work needed to do so and the extent of change in the population. An update every five years may be enough. The key period to tie into would be with the release of the census, which is due in 2011, but results usually take another two years to be released.

10.1.2 Schools

Information about primary and secondary schools is collected on an annual basis by various government bodies. However the data was not geo-coded until we carried it out for this project. By allocating all schools a unique identifier it should be reasonably straightforward to maintain an accurate list of schools, by updating the data on an annual basis for additions and subtractions if locating the school was the primary concern. The updating of the school population should also be relatively straightforward.

10.1.3 Hospitals

Information about hospitals is also collected on an annual basis. The issues that apply to schools are similar, although as floor space is the main determinant in the modelling there would be more work involved for any updates.

10.1.4 Care homes

There are many care homes and the situation changes fairly frequently. The data is collected on an annual basis by the English, Scottish and Welsh care home authorities.

10.1.5 Retail

There are no plans as yet to update the ODPM dataset concerning town centres. AddressPoint may still be needed to locate commercial areas. CACI maintain the data we have used to locate some of the commercial areas.

10.1.6 Transport

The transport layer uses a variety of data sources and would require a complete rebuild to update. Traffic flow characteristics and the transport network itself are relatively static therefore an update cycle of 5 years is recommended.

10.2 The use of the database by the HSE and other organisations

This project has successfully produced a national population database which has a greater coverage of population types and a better level of spatial resolution than any others that currently exist. It has a flexible and user-friendly interface which provides for many different potential uses. As such, it represents a major step forward for the HSE in their representation of populations at risk within accident modelling. Moving from a crude method of visual inspection of OS maps, to a multi-layered database constructed from good quality national data available within a GIS environment is a very significant advance, improving the quality of current work and opening up new opportunities for analysis.

In our last report (Mooney and Walker 2002) we identified a number of ways in which the use of a national population database could be extended within the HSE. Having now produced the database these applications remain valid and bear brief repetition:

- Other forms of point source risk – the key applications for the database have focused on major accident hazard installations identified under COMAH and associated legislation. However, the database could also be used in other aspects of the HSE's work including risk assessment and regulation related to nuclear and explosive risks.

- Transport and pipeline routes – the data could be used in the assessment and management of these linear forms of risk. For example, the road data already in the database could be used to identify the route of proposed or current transfers of hazardous materials and provide estimates of populations potentially affected both within and alongside transport routes.

- Macro-level risk studies and indicators – there are a range of possibilities for the data to be used at a macro level to examine policy issues and derive performance indicators. For example, monitoring over time the changes in population levels near to hazardous sites;

the targeting of regulatory assessment and inspection resources on sites where the greatest numbers of people are at risk; analysis of the social characteristics of populations living at risk under an environmental justice agenda

Outside of the HSE there are many other potential users of the database. These include:

- departments and agencies concerned with other forms of risk, such as flooding, extreme weather events and terrorism.

- Various organisations concerned with issues of resilience, emergency preparedness and emergency response in the event of disasters.

- Private industry including both those companies operating hazardous sites and pipelines and the insurance industry assessed premium risks.

- Regional Development Agencies and Government Offices for the regions needing to locate populations and have a better understanding of migration patterns.

- Large scale regeneration schemes such as the proposed Sustainable Communities in the South East and Housing Market Renewal Pathfinders in the North.

APPENDICES

APPENDIX 1: Bibliography

Health and Safety Executive (1989) Risk criteria for land use planning in the vicinity of major industrial hazards, London: HMSO.

Mooney J and Walker G P (2002) The Derivation and Use of Population Data for Major Accident Hazard Modelling, HSE Research Report Series, 410/2002, HSE Books, Sudbury.

Walker G P and Mooney J (1998) Spatially Referenced Population Data for Land Use Planning Advice, HSE Research Report Series 189/1998, HSE Books, Sudbury.

Walker G P, Mooney J and Pratts D (2000) The people and the hazard: the spatial context of major accident hazard management in Britain, Applied Geography, Vol. 20, pp 119-135

Walker G.P. (2000) Urban Planning, hazardous installations and blight: an evaluation of responses to hazard-development conflict, Environment and Planning C, Vol. 18, No 2, pp127-143

APPENDIX 2: Address Datasets

This appendix outlines the major locational datasets used in the database, made available for use through the pan governmental data agreement with the Ordnance Survey (OS). These are:

1. AddressPoint
2. CodePoint
3. CodePoint with Polygons

The reader is referred to the original manuals if further details and in-depth discussion of the datasets are required. These user guides can be accessed online through the following addresses:

http://www.ordnancesurvey.co.uk/products/addresspoint/pdf/apuserguide.pdf
http://www.ordnancesurvey.co.uk/products/codepoint/pdf/cpuserguide.pdf
http://www.ordnancesurvey.co.uk/products/codepointpolygons/pdf/cppolygonsuserguide.pdf
http://www.ordnancesurvey.co.uk/products/strategi/pdf/StrategiUserGuide.pdf
http://www.ordnancesurvey.co.uk/products/oscarasset/pdf/OSCAR_user_guide_web.pdf

AddressPoint

ADDRESS-POINT is an Ordnance Survey data product that provides a National Grid coordinate and a unique reference for each postal address in Great Britain (this includes England, Scotland and Wales). The creation process for ADDRESS-POINT, entails the addition of Ordnance Survey National Grid references, a unique reference and other metadata, to Royal Mail's Postal Address File (PAF).

Further details are given in the AddressPoint user guide. The user should consult this guide for detailed discussions regarding the following key issues:

- Temporary coordinates
- The status flag
- Change type
- How AddressPoint represents a variety of complex structures, such as, blocks of flats.

CodePoint

Code Point provides a National Grid reference for each unit postcode in Great Britain. Multiple postcodes in a single block of flats or offices will share one National Grid reference. With each co-ordinated point there is information about the postal delivery points within the postcode unit and codes for a number of administrative boundaries, which coincide with the postcode unit.

Further details are given in the CodePoint user guide. The user should consult this guide for detailed discussions regarding the following key issues:

- Coordinate Accuracy
- Domestic and non-domestic delivery point definitions

CodePoint with Polygons

Code-Point polygons represent postcode unit boundaries in Great Britain. The Code-Point polygons are derived from ADDRESS-POINT® National Grid reference (NG ref) coordinates for each postal delivery address in Great Britain. PO boxes are not included, but their postcodes are supplied in a separate file. A vertical streets lookup file is also included. This highlights delivery points that share the same location.

Further details are given in the CodePoint with Polygons user guide. The user should consult this guide for detailed discussions regarding the following key issues:

- Vertical Streets Polygons
- How postcode polygons are produced

APPENDIX 3: Glossary

Glossary, terms and acronyms

AddressPoint	– An Ordnance Survey data set listing all addressed locations in the UK. Described in appendix II in more detail.
Address point	– An addressed point usually originating from either AddressPoint or CodePoint.
ArcGIS	– The GIS software suit produced by ESRI. This consists of a number of packages all contained within ArcGIS. The project primarily makes use of ArcView 8 which can also be called ArcEditor and ArcInfo depending on the software license installed.
Area of Interest	– See Selection Area
CodePoint	– An OS data set listing all postcode units in the UK.
COMAH	– Control of Major Accident Hazard EU Directive and UK Regulations
Core area	– (hospitals only) the collection of contiguous points that represent the core area. A population is distributed over these areas.
Core point	– The central point representing the location of a geographical entity such as a school. Core points are surrounded by one or two point flags.
Edge effect	– edge effects are illustrated in chapter 3.3.2 and refer to artificial results produced sometimes when selecting an area.
ESRI	– Environmental Systems Research Institute, the software company that produces and markets ArcGIS software.
EU	– European Union
Feature class	– An ESRI term referring to a type of feature. There are three primary types of feature, a point, a line and a polygon.
Feature dataset	– A group of feature classes.
Flag	– An error buffer used to account for positional uncertainty. Flags are generated according to a one point rule or a two point rule. The one point rule specifies all points (including diagonals) surrounding a core area are designated as flags. For a two point rule the logic is the same, the width of the flag is increased by two points. Examples of these rules are found in chapter 6.
Geo–coding	– attaching coordinates (i.e. a spatial location) to data.
Geographic entity	– any real world object or feature that has a spatial extent, for example a hospital.

Grid – In the context of this dataset a grid refers to a regular matrix of points spaced at 100m. Grid is also the name for an ESRI raster file format.

Grid point – A point that makes up part of the grid.

Household – A household comprises one person living alone, or a group of people (not necessarily related) living at the same address with common housekeeping – that is, sharing either a living room or sitting room or at least one meal a day. As defined by the Census 2001.

HSE – Health and Safety Executive

Layer – A collection of similar geographic features – such as households, hospitals, stadia.

Major accident hazard – the potential for an accident involving toxic, explosive or flammable substances at installations identified under EU and UK legislation

OS – Ordnance Survey

Output area (OA) – Smallest spatial unit in the census with an approximate size of 125 households.

Selection Area – An area selected by a user in a GIS to identify features relating to a particular area.

Sensitive population – sensitive populations are those judged by the HSE to be particularly vulnerable in the event of an accident event. Sensitivity levels are defined by the HSE which combine factors relating to age, ill health and numbers of people.

Unit postcode – An abbreviated form of address made up of combinations of between five and seven alphanumeric characters. A postcode may cover between 1 and 100 addresses. The average number of addresses per postcode is 15.

Vertical postcode – Vertical postcodes are taken from the OS CodePoint with Polygons. Where two or more postcodes are associated with a single building seed, a single distinctive square polygon will represent all the postcodes attached to the seed.

Printed and published by the Health and Safety Executive
C1.10 01/05

MAIL ORDER
HSE priced and free
publications are
available from:
HSE Books
PO Box 1999
Sudbury
Suffolk CO10 2WA
Tel: 01787 881165
Fax: 01787 313995
Website: www.hsebooks.co.uk

RETAIL
HSE priced publications
are available from booksellers

HEALTH AND SAFETY INFORMATION
HSE Infoline
Tel: 08701 545500
Fax: 02920 859260
e-mail: hseinformationservices@natbrit.com
or write to:
HSE Information Services
Caerphilly Business Park
Caerphilly CF83 3GG

HSE website: www.hse.gov.uk

RR 297

£25.00

ISBN 0-7176-2941-4

9 780717 629411

A National Population Data Base for Major Accident Hazard Modelling

HSE BOOKS