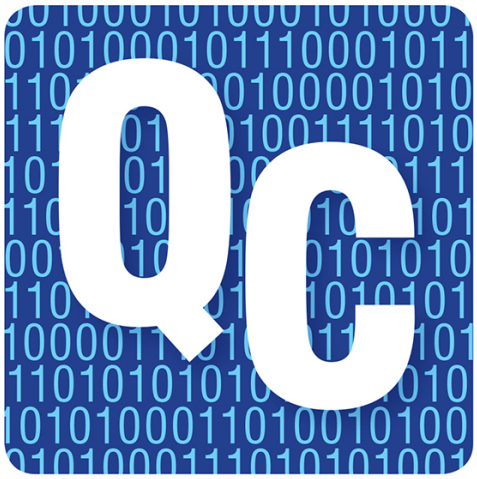

Record linkage and mapping data for UK geographies

Anna Palczewska

QuantiCode's Technical Report, 13 Sep 2016

Abstract

The QuantiCode report provides technical details on a process of record linkage and mapping data for UK geographies. The reader can find here a brief description of methods and algorithms used in each step of record linkage: data standardisations, comparison, classification and evaluation, basing information on the privacy-preserving linkage. This report provides the description of UK statistical geographies, their hierarchies and relationships. This document includes also a short information about methods for data integration from various geographies: lookup tables and GIS.

The QuantiCode project is developing novel data mining and visualization tools and techniques, which will transform people's ability to analyse quantitative and coded longitudinal data. Such data are common in sectors such as health (e.g., electronic health records), local government (e.g., service provision) and retail (e.g., product sales). The project is funded by the Engineering and Physical Sciences Research Council (grant ref. EPN0139801) and, through the Leeds Institute for Data Analytics (LIDA), supported by the Medical Research Council (ESL0118911) and Economic and Social Research Council (ESL0118911).

Contents

Chapter 1

Record linkage

Record linkage is a process that matches records representing the same instance or entity from one or more databases. In this chapter the main terminology for record linkage, the whole process of record linkage and new computational methods for reducing the scalability and privacy presenting problems are described. There is also a list of open source software for record linkage. The list is not exhaustive but the selected software demonstrate variability in terms of functionality.

1.1 Introduction to record linkage

Large amounts of data are generated and collected every day by various organisations, public and private institutions, researchers and individuals. Examples of such data come from shopping transactions, social media, phone records, electronic records for health or census, etc. Integration and analysis of these data can bring benefits for various organisations leading to a better understanding of society. Moreover, linking different sources we can improve data quality, enrich data with additional information, and allow for more sophisticated analysis. At the same time, it is a challenging process due to the lack of unique identifiers, a size of data, quality (typographical errors, variations, different coding), different formats, privacy and confidentiality.

Record linkage is a process of matching records that represent the same instance or entity from one or more databases [1]. In many domains the linkage process is often known as data matching, entity resolution, object identification, identify uncertainty, merge-purge process (for removing duplicates in files). The process of linkage records from one database is also often called duplicate detection, de-duplication or internal data linkage [2]. Traditionally, record linkage was used in statistics (census) and health (epidemiology). Today it is used in many areas that require analysis of big data: immigration, social security, census, fraud, crime, terrorism intelligence, businesses (exchanging customer data), health and social science research. Record linkage is a very powerful tool and can be used for:

- data source cleaning (removing deduplicates)
- merge records into larger dataset
- clean and enrich data for mining and analysis
- create person oriented statistics (longitudinal study)

- geocode matching (match addresses to geographies (spatial analysis of health or geographical information)).

Objects that are subjects for linkage: patients (in health), customers (in business), taxpayers, travellers, business data, consumer product (e.g. for product comparison). One of the main challenges for linkage is that the unique identifiers are not available in all the databases. This is why the attributes that identify entities need to be used for matching. In many cases we have to work on the person (e.g. personal identifiers: names, addresses, dates of birth, etc.) or object description level and we have to apply string comparison to find matching records [3].

The history of data linkage starts in 1950 when the initial techniques were developed. They were ad hoc heuristic methods comparing names, addresses and dates of birth for linking personal information. The first probabilistic methods were introduced in 1962 by Newcombie and Kennedy [4] and theoretical foundations were introduced in 1969 by Fellegi and Sunter [5]. Since then many new methods and approaches have been proposed in fields such as statistics, computer science, databases or information retrieval. Currently, many domains are interested in record linkage to provide new methods for linkage focusing on the scalability and linkage quality (e.g. indexing and blocking techniques) due to a "big data" availability.

1.2 Record linkage process

The process of record linkage consists of three major steps: blocking/indexing, record pair comparison and classification of the compared records. Figure 1.1 presents the full process of record linkage for two datasets. Data cleaning and standardisation is a step to processing the data from datasets into the same format. This will lead to a better linkage quality. In the next step, we apply indexing, searching or filtering techniques to simply not run the pairwise comparison. This allows us to select candidate records with some similarity. Then we compare them based on the similarity of the linked attribute values and we classify them into three classes: **matches** (records that are highly similar to each other), **non-matches** (records that are completely different) and **potential matches** (records for which some linking attribute are similar and some are different). The third class is further manually reviewed to decide if records in that group are matches or non-matches. Such process is often called an active learning because the feedback from the manual review is returned to the classifier. At the end, we evaluate linkage process. The following subsections will describe these steps in detail.

1.2.1 Data cleaning standardisation

Data cleaning and standardisation depend on the quality of data. While cleaning can improve linkage rates, the cleaning process can be quite labor intensive, so researchers should consider the cost-benefit analysis before investing a significant amount of time on cleaning the data. Cleaning has been highly recommended if the data quality is poor and/or only a few identifiers are available [7]. There are the following problems with data:

1. typographical errors (spelling errors, variation of names, different details for the same person in various datasets),

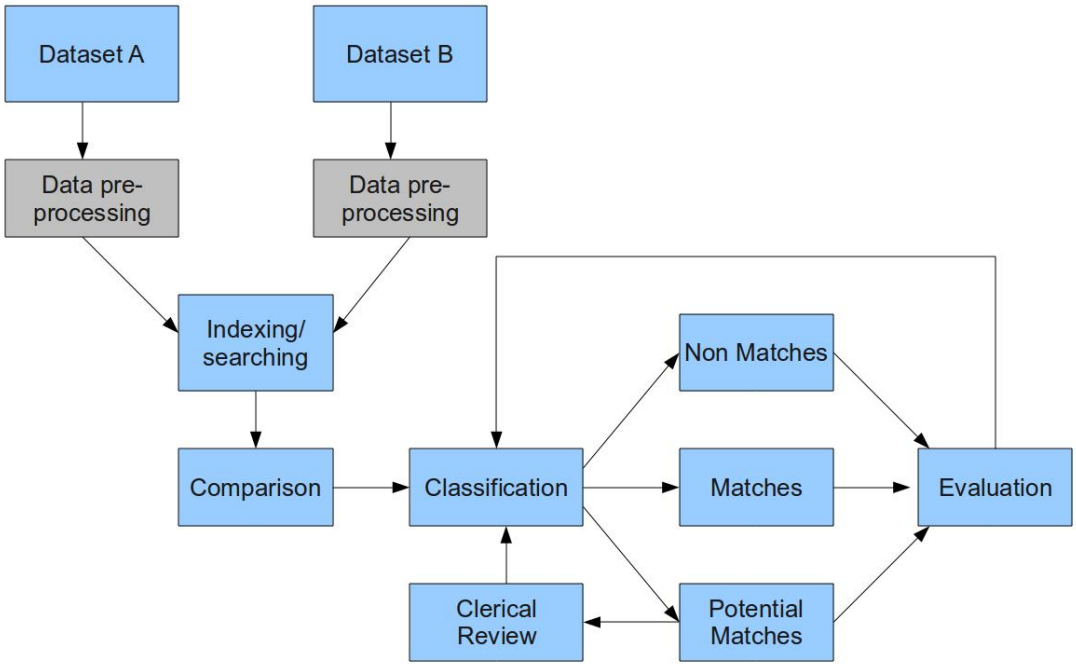

Figure 1.1: Linkage record process [6].

2. different coding schemes (e.g. male/female vs M/F vs 1/2 etc.),
3. missing data,
4. changing data over time.

There is a number of data cleaning techniques that are used in record linkage [6, 7, 8]. Some techniques increase the number of variables by splitting apart strings, where other techniques transform variables into a specific representation:

- reformatting value - ensures that data is in a common standard in all datasets for comparison during linkage process. The data can be easily changed to a new format without creating and removing information. For example, a date's format 01/05/2016 and 1st August 2015 can be reformatted to the same format.
- removing punctuation - unusual characters and punctuation are more likely to be misrepresented, this is why it is important to remove them from the alphabetic values
- removing missing values and not meaningless values - removing values such as NA, 9999 999 for a postcode, "NO ADDRESS", etc. These values are often inserted if there is no information available. It is important because their presence will increase the number of mismatches during the linkage process.
- phonetic encoding - some variables (e.g. surname) they may be inserted with a phonetic error, they can have different spelling but the same sound. Such errors

should be corrected before linkage process. There is a list of available software: Soundex [9], methaphone [10], NYSIIS [11].

- name and address standardisation - name parsing is a process of breaking down a person's full name into individual parts. Also, addresses should be broken down into smaller components such as street number, street name and street type. This can be done applying rule-based methods (e.g. royal mail rules for formatting addresses) or probabilistic methods (hidden Markow models [12]). The first method can be time-consuming, complex to develop and maintain the set of rules. The second methods require a training dataset as it is a learning process.

- nickname/abbreviation lookups - all short names and abbreviation should be expanded. This can be done by creating lookup tables and used them in the data standardisation process.

1.2.2 Blocking, indexing and comparison

Blocking, indexing or filtering was introduced to reduce the number of comparison of records pairs by bringing potentially linkable record pairs together [13]. In traditional blocking [5] a database is split into smaller blocks according to some criteria (known as a blocking key). The records pairs with the same blocking key value from two databases are compared. It is important to find blocking criteria that have an even distribution. For example, postcodes have a similar household count or people population. A phonetic coding for names is another example, similar names should be considered together in a block. Within each block, the records are compared based on string comparison for the linking attributes. There are many techniques for string comparison [6] and the most used for linkage are listed below:

- deterministic linkage: exact comparison of two attribute values

- q-gram: a string is converted into q-grams substrings of length q using a sliding window approach (e.g. "road"={"ro", "oa","ad"} for 2-gram). In the next step, we count q-grams that occur in two strings and we use measure (e.g Dice coefficient) to calculate the similarity between strings.

- edit distance: is based on a string metric for measuring the difference between two sequences. It includes a number of character edits (insert, delete, substitute) needed to convert one string into another [14]. The basic edit distance, is also known as the Levenshtein edit distance. There are a variety of dynamic programming algorithms to calculate this distance.

- probabilistic record linkage: compare records attributes using string comparison functions. These functions are type specific: different for dates, addresses, and strings. The similarity (a matching weight) is calculated to the pair of corresponding attributes). Those weights are sum up. Based on this summarised weight we classify a pair of records as a match, non-match or potential match. This approach requires an estimation of errors for weight calculation, finding optimal thresholds for cutting off matches and non-matches and finally there is a manual clerical review needed for potential matches.

1.2.3 Classification - general linkage techniques

The classification of the candidate record pairs generated in the indexing step is based on the similarity values calculated in the comparison step. More similar two records are, the more likely they refer to the same real-world entity [6]. The classification approaches can be divided into three main groups:

- deterministic algorithms - determine whether record pairs agree or disagree on a given set of attributes, where agreement on a given attribute is assessed as a discrete "all-or-nothing" outcome. There are two methods: **exact linkage** if high quality identifiers (for linking attributes) are present. Identifiers must be precise, robust and stable over time. The second method is rule-based matching. This method is very complex to build and difficult to maintain. The rule methods are data dependent, they should be changed when data changes. There are algorithms that learn rules from data but they require a training dataset which is challenging (many areas do not have gold standard data that can be used for training purposes).

- probabilistic algorithms - classify record on matches, non-matches, possible matches based on the similarity of the linking attributes. Probabilistic methods are often called fuzzy matching and take into account a wider range of potential identifiers and compute weights for each identifier based on its estimated ability to correctly identify a match or a non-match, and using these weights to calculate the probability that two given records refer to the same entity. Record pairs with probabilities above a certain threshold are considered to be matches, while pairs with probabilities below another threshold are considered to be non-matches; pairs that fall between these two thresholds are considered to be "possible matches". Whereas deterministic record linkage requires a series of potentially complex rules to be programmed ahead of time, probabilistic record linkage methods can be "trained" to perform well with much less human intervention.

- computer science approaches - have become more popular recently. They are based on machine learning, data mining, and databases algorithms. A classification approach can be unsupervised or supervised. Unsupervised methods group pairs of records based on the similarity between them without information about the characteristics of true matches and true non-matches. Examples of unsupervised methods: clustering, collective classification (e.g. hierarchical clustering, graph-based approach - linked relationship graph between entities). The supervised approach requires a training dataset with match and non-match characteristics. The accuracy of the built classification model is evaluated using a set of testing data that must be in the same format and structure as the training data. The example of the supervised methods: decision trees, support vector machine.

1.2.4 Record linkage evaluation

Record linkage is evaluated by measuring the linkage complexity and quality [6]. There are two main measures for linkage complexity:

- reduction ratio - how many candidate records pairs were generated by blocking compared to all possible pairs?

$$rr = 1 - \frac{\text{number of candidate pairs}}{\text{number of all pairs}}$$

The smaller number of pairs in blocking the larger the ratio, which means that we reduced the complexity of linking process.

- pair's completeness - how many true matches were generated by blocking divided by all true matches?

$$pc = 1 - \frac{\text{number of true matching candidate pairs}}{\text{number of all true matching pairs}}$$

These two measures tell us how good the blocking method is. To measure a general linkage quality we need to have information about true matches, what can be difficult in many areas. There are two types of errors:

- false non-match - a missed true match (false negative)
- false match - a wrong match (false positive).

A calculation of accuracy as percentage of false matches and false non matches is not meaningful. A classification of all records as non matches can still give a high accuracy. In this case it is better to use two measures that focus on true matches:

- precision: how many true matches are in the set of classified matches? (How many elements are relevant?)

$$prec = \frac{\text{number of true matches pair}}{\text{number of all classified matching pairs}} = \frac{tp}{tp + fp}$$

- recall how many true matches did we find from all known true matches?

$$prec = \frac{\text{number of true matches pair}}{\text{number of all true matching pairs}} = \frac{tp}{tp + fn}$$

1.3 Advanced record linkage techniques

In recent years various indexing techniques for record linkage have been developed [15, 16] In the traditional standard blocking approach, all records that have the same blocking key value are inserted into the same block, and only the records within the same block are compared with each other in detail in the comparison step. Each record is inserted into one block only. To control better a number of comparisons the following techniques have been introduced:

1. sorted neighbourhood approach - a sliding window of fixed size is moved over sorted database tables. Candidate record pairs are generated from the records that are within the current window. This method allows to control the number of comparisons and has a linear complexity [17, 18].
2. Q-gram blocking - convert values to the q-gram lists then generate sublists. Records are inserted into several blocks by generating variations of the record's blocking key value through the use of q-grams (substrings of length q characters) [16, 15].
3. canopy clusters - overlapping clusters - similarity of string record is calculated based on q-grams. Records are inserted into several clusters. Each cluster forms one block from which candidate record pairs are generated [19, 20]

4. string map based blocking - this technique based on mapping block key values into multi-dimensional space such that distances between strings are preserved [21]

5. controlling block size - important for real-time application and privacy-preserving record linkage (iterative split-merge clustering approach) [22]

Advanced classification techniques view record pairs classification as multidimensional binary classification problems. These methods use attribute similarities to classify record pairs as matches or non-matches (there is no summarised similarity for two records). There are three main techniques: machine learning, collective classification and group matching. Machine learning techniques can be split on:

- supervised: require training data (records with true matches and non-matches. The training data has to reflect the real data variation (twins or the same person with changed name and address). For learning decision trees, neural networks, SVM and other known machine learning methods are used.

- active and semi-supervised learning: require training data. Dataset is sampled using bootstrapping, and the initial classifier is built. Then we apply it on the entire dataset. We select "clear/obvious" matches an non-matches and used then as a training data to train a new classifier. Active learning is when we use a human being to help classify difficult cases.

- unsupervised: does not require training dataset. Using clustering methods we group pairs of records based on their similarities.

1.4 Privacy-preserving record linkage

In privacy-preserving data mining (PPDM) the main goal is to perform "data mining" computations on a set of data, in a way that prevents both the computation and the output of the computation from revealing too much sensitive information about the units represented in the data [23]. Similarly to privacy-preserving data mining, privacy-preserving record linkage provides a secure way for record linkage where none can find out about sensitive data [24]. This is used in cases where two or more organisations, governmental agencies or health institutions want to exchange data for integration without violating data privacy. The main challenges are:

- encryption methods cannot be applied directly (e.g two matching record with different name or address will be returned with different identifiers (encryption codes)),

- techniques must not be vulnerable to any kind of attack,

- techniques should be scalable to linking large databases.

In the health domain, the privacy preserving record linkage protocol was proposed in [25] and is presented in Figure 1.2. In the first step, the data source owners send the identifying information to a linkage unit. Linkage unit is a trustee organisation which can see the personal data for linkage. The sent data is encrypted but linkage unit decrypts it and processes linkage on row data following the record linkage process. Linkage unit sends the linked record identifiers back to the database sources. In the next step, the payload information is attached to the linked record identifiers and such data are sent to

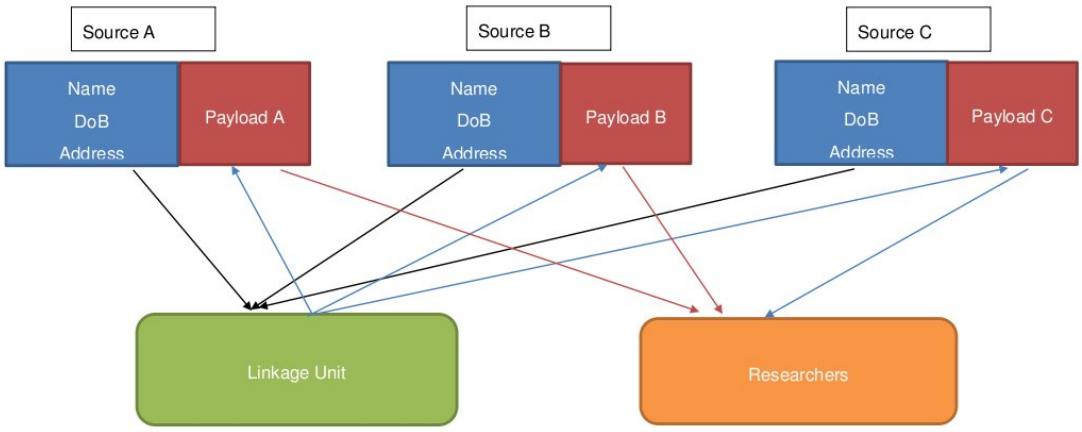

Figure 1.2: Linkage record process.

researchers or third party working on linked data. In such scenario the privacy-preserving record linkage aims:

- to secure that no encrypted data ever leave a data source
- only details about matched records are revealed
- secure against various attacks.

Basic PPRL protocols:

- two-part protocol - only the two database owners participate in the PPRL process. It is more secure, no possibility of collusion and it has a lower communication cost. This method requires more complex techniques to ensure that the two database owners cannot infer any sensitive information from each other during the linkage process.
- three-party protocol - a (trusted) third party (which is called a 'linkage unit') is involved in conducting the linkage, possibility of collusion between linkage unit and one of the data sources.

1.5 Open source linkage software

There is a large number of various software for data linking. They mostly were developed by researchers as part of their work of inventing new and improved data matching algorithms and techniques. Some of them include the graphical user interface and functionality allowing data cleaning and standardisation. Below there is a list of few of them.

- **Febrl** - A Freely Available Record Linkage System with a Graphical User Interface (Febrl) [26] implemented in Python is a free object-oriented programming language that is available on all major computing platforms and operating systems. It contains many recently developed advanced techniques for data cleaning and standardisation, indexing (blocking), field comparison, and record pair classification,

and encapsulates them into a graphical user interface. Febrl can be seen as a training tool suitable for users to learn and experiment with both traditional and new record linkage techniques, as well as for practitioners to conduct linkages with data sets containing up to several hundred thousand records [27].

- **WHIRL** - The Word-Based Heterogeneous Information Representation Language system that allows various similarity string comparison functions (similarity joins) to be applied on textual data system. The system is written in C++ and available here [28].
- **TAILOR** - The Record Linkage Toolbox is a system for data matching including integrated different indexing, comparison, and classification techniques (supervised and unsupervised), as well as various evaluation methods. Standard blocking, sorted neighbourhood approach, comparison and phonetic encoding functions are implemented in this software. TAILOR is written in Java and is available by contacting the developers [29].
- **SimMetrics** - is a system for approximate string comparison. It was developed in Java, currently available at Sourceforge.net [30]
- **R RecordLinkage** - this is the R package that includes functions for standard blocking, and several phonetic encoding and string comparison methods. Both the probabilistic and deterministic matching approachs are included as well. The package, example data sets, and a reference manual are available at [31]
- **FRIL** - The Fine-Grained Records Integration and Linkage system contains several indexing methods (including standard blocking and the sorted neighbourhood approaches), string comparison functions. This system can be run on multi-core systems and it contains a GUI that allows users to easily set-up and customises deduplication or data matching projects. FRIL was developed in Java available at [32]. This system allows the pre-processing of attributes through the use of regular expressions to standardise the input data and to split and merge attributes before they are used for matching.
- **BigMatch** - The BigMatch system has been developed and is being used by the US Census Bureau to match very large census data collections [33]. BigMatch is not a full data matching system, rather it is a program that can be used to extract potential matches from very large files that otherwise could not be processed. These matches are saved into several smaller files so that they can be individually processed with a proper data matching system later on. In contains a standard blocking approach with several blocking criteria. This system was developed in C.

Chapter 2

Linkage and mapping geographically referenced data

Linkage geographically referenced data involve two approaches: address matching using the string comparison methods described in the previous chapter or georeferencing and mapping various geographies for the data aggregation purposes. In this chapter, various English geographies are described together with their mapping methods.

2.1 UK geographies

In the United Kingdom, the Office for National Statistics holds and maintains a number of codes that represent a large range of geographical areas. These codes (ONS codes or GSS codes) refer to the Government Statistical Service of which ONS is a part. Geography provides a structure for collecting, processing, and storing the data. There are many different geographic unit types (administrative, health, electoral, postcode etc) and their boundaries frequently do not align. A range of geographies is frequently revised and geographical boundaries are continuously changing [34]. Figure 2.1 presents the hierarchical representation of UK statistical geographies from October 2015. There is seven main group of geographies: postal, administrative, health, census (statistical block and merged geographies), electoral, Eurostat and other. In this document, only the main English geographies are described in detail.

2.1.1 Postal geography

Royal Mail maintains a UK-wide system of postcodes to identify postal delivery addresses. The Postcode Address File (PAF) is the latest, most accurate UK address database. It contains 1.8 million UK postcodes and over 29 million residential and business addresses [36]. These are constantly updated and verified by ninety thousand postmen and women, making updates to 3,500 records each day. Postcode is used as the main geographic reference when collecting data. This reference can be related to any geographic unit used for statistical production, such as a local authority district or electoral ward. Figure 2.2 presents the structure of the postcode. It is a hierarchical structure supporting four levels of geographic unit see Table 2.1.

There are two types of postcodes:

- Large user postcodes: allocated to single addresses receiving at least 500 mail items per day (e.g. business addresses).

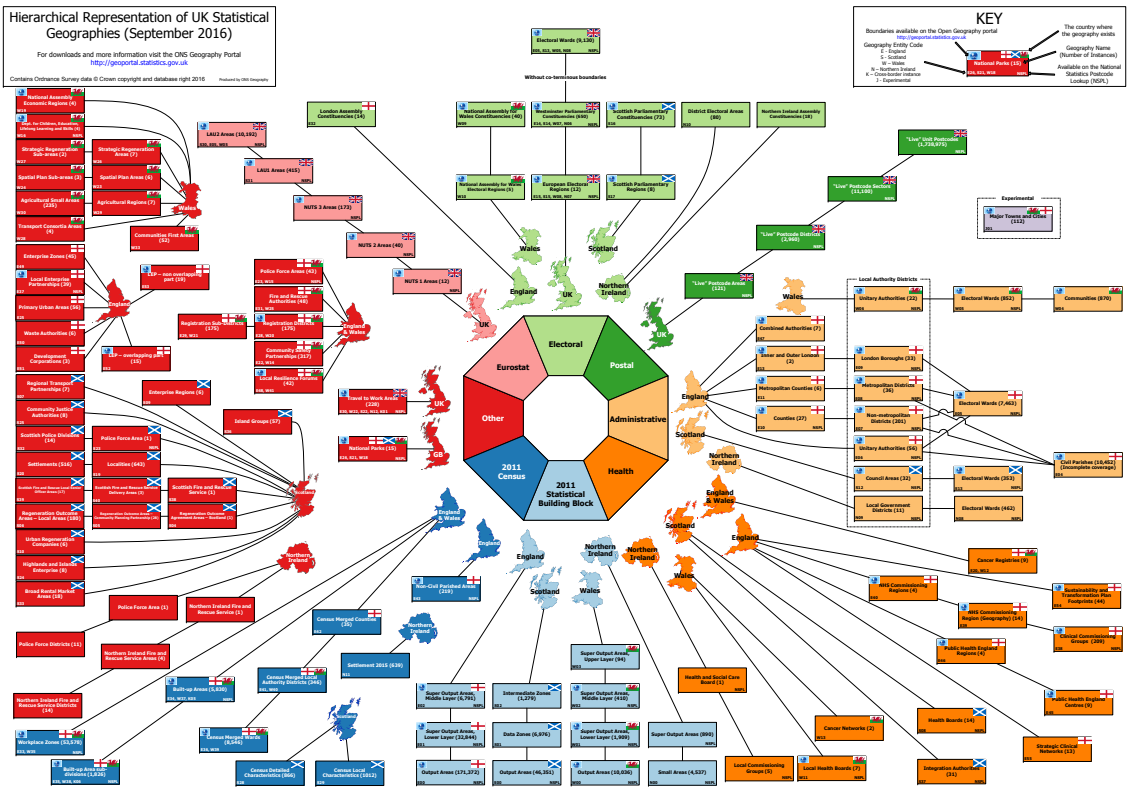

Figure 2.1: Hierarchical Representation of UK Statistical Geographies [35].

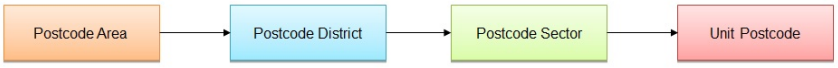

Figure 2.2: Postcode structure.

- Small user postcodes: collections of (usually) adjacent addresses. A single small user postcode may contain up to 100 addresses, but 15 is a more typical number.

Linking postcodes with other geographies is not straightforward:

- Postcode boundaries do not align with other geographic boundaries. A manual assignment of the postcode to the other geography is necessary when the postcode straddles the boundary of a chosen geography (e.g. ward or output area). Usually, it is done by allocating the postcode centroid within the given area boundary [37].

- Postcode boundaries are constantly updated by adding new addresses, removing not used ones.

2.1.2 Census geography

Census provides a detailed snapshot of the population and its characteristics. It is undertaken every 10 years. The most recent was on 27 March 2011. The main geographies directly associated with the Census are Output Areas (OA) and Super Output Areas (SOA) [38]. Output areas are the base unit for Census data releases and the lowest geographical level at which census estimates are provided. Output areas were created for

Geographic	Unit Number in UK	Example
Postcode Area	124	LS
Postcode District	3,114	LS2
Postcode Sector	12,381	LS2 9
Unit Postcode	around 1.75 million	LS2 9JT

Table 2.1: Postcode hierarchical structure.

Census data, specifically for the output of census estimates. Output areas are built from clusters of adjacent unit postcodes. They were designed to have similar population sizes and be as socially homogeneous as possible based on tenure of household and dwelling type [39].

After Census 2011 the total number of output areas is 171,372 for England, they are covered by super output areas. There are 34,753 lower layer super output areas (LSOA) and 7,201 middle layer super output areas (MSOA). Table 2.2 presents the difference between these areas on the number of population and household count.

	Minimum		Maximum		
Area type	People	Household	People	Household	Count
Output Areas	100	40	625	250	171,372
Lower Super Output Areas	1,000	400	3,000	1,200	34,753
Middle Super Output Areas	5,000	2,000	15,000	6,000	7,201

Table 2.2: Lower and upper thresholds for Output Areas

2011 Census estimates for electoral wards/divisions are aggregations of output areas, on a best-fit basis [40]. Boundaries of output areas and super output areas are aligned to local authority district (LAD) boundaries, including those that changed between 2003 and 2011, and also at the border between Scotland and England. For ward and local authority information see the following section.

2.1.3 Administrative geography

Administrative geography represents the hierarchy of areas relating to national and local government in the UK. There is a different structure in each constituent country of the UK. The boundaries of many of the layers in the hierarchy are subject to either periodic or occasional change [41]. Figure 2.3 presents the administrative hierarchical structure for England.

There are nine government offices for the regions (GOR): North East, North West, Yorkshire and The Humber, East Midlands, West Midlands East of England, London, South East, South West. They are split into 322 counties. Metropolitan counties are six heavily built areas (without Great London) divided into metropolitan districts. There are 36 metropolitan district councils as a single-tier authorities [42]. There are 27 counties (shire) split into 201 non-metropolitan districts (LAD - local authority districts) and there are 56 unitary authorities [43]. From 2000 Great London is subdivided into 32 London boroughs with a status similar to metropolitan districts, and also the City of London. The boundaries of all authority districts and London boroughs can be found in the local authority district (LAD) boundary files [44]. There are 7678 wards and electoral division in the UK see table 2.3.

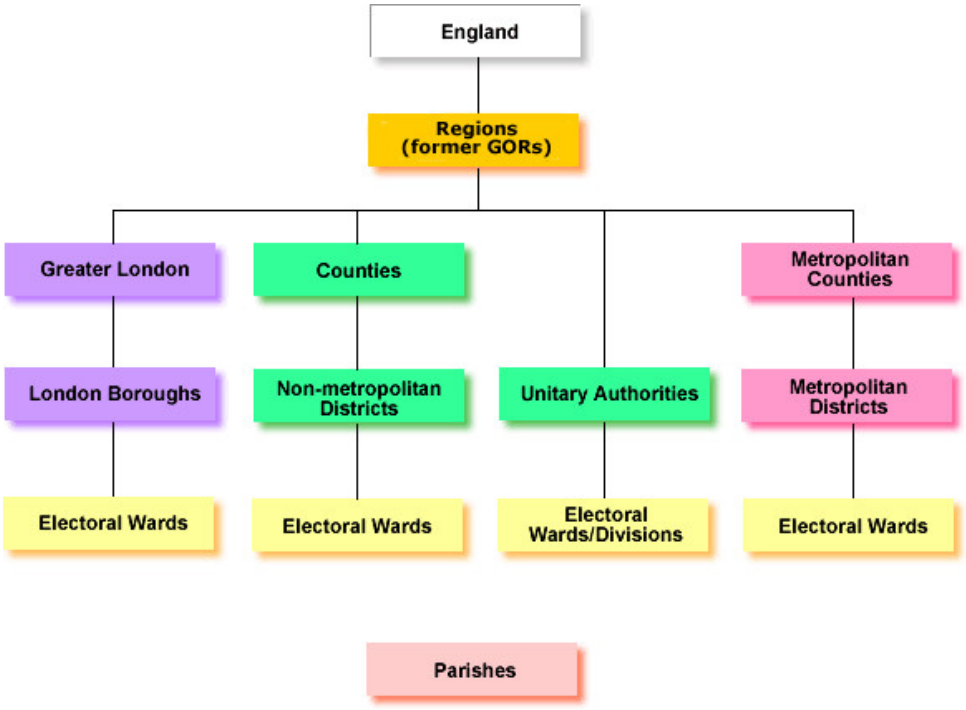

Figure 2.3: Hierarchical representation of England administrative geography [41].

Electoral wards/divisions are the smallest units so called "building block", from which higher units are constituted. They are the spatial units used to elect local government councillors in metropolitan and non-metropolitan districts, unitary authorities and the London boroughs in England. English local authority districts (LAD) (both metropolitan and non-metropolitan), London boroughs and unitary authorities average around 23 electoral wards or divisions each. Electoral ward/division boundary changes are usually enacted on the first Thursday in May each year, to coincide with the local government elections [45].

Area Type	Name	Count
Regions		9
Counties		35
	- Shire (Non-metropolitan)	27
	- Metropolitan	6
	- Great London	2
Local Authorities		322
	- Non-metropolitan districts	201
	- Metropolitan districts	36
	- Unitary authorities	56
	- Landon boroughts	31
Wards and Electoral divisions		7678
	- Census Wards	7218
	- Census Electoral division	453
	- Census Merged Wards	7
Parishes		10,449

Table 2.3: Hierarchy of the Census administrative geography

Parishes are the smallest administrative type in England. They are a very old form of a spatial unit which originally represented areas of both civil and ecclesiastical administration. Many parishes are a similar size to wards, but some can contain several wards, and ward boundaries need not be followed. Parishes are confined within local authority district boundaries but are not contiguous with electoral wards. As at 31 December 2015 there were 10,449 parishes in England [46].

2.1.4 Electoral geography

Electoral geography is a three-tier complex structure. There are three different electoral systems and different areas used for election (see Figure 2.4):

- European and UK parliaments
- Devolved and regional governments
- Local authorities and smaller units

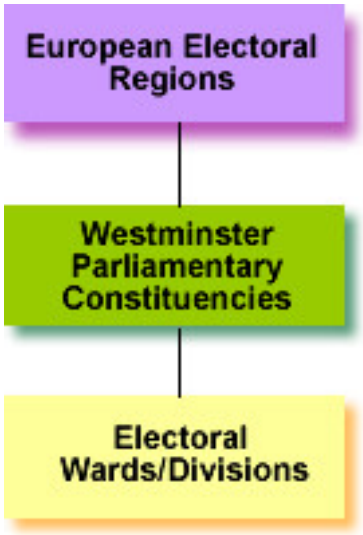

Figure 2.4: Hierarchical representation of electoral geography in UK [47].

The electoral hierarchy is the only electoral structure that covers the whole of the UK. European electoral regions (EER) are used to elect Members of the European Parliament (MEP) to the European Parliament in Strasbourg. England's electoral regions are based on the boundaries of the regions (former government office regions (GOR)) at the start of the year of an election [48]. Westminster parliamentary constituencies are the areas used to elect Members of Parliament (MP) to the House of Commons, which is the primary legislative chamber of the UK and is located in Westminster, London. At the May 2010 General Election, there were 650 constituencies. Councillors in UK districts and unitary administrations are elected to represent the same electoral wards/divisions that are used to constitute Westminster parliamentary constituencies. County councillors, however, represent larger 'county electoral divisions', which are not necessarily based on the electoral wards used at district level [49].

2011 Census estimates for electoral wards/divisions are aggregations of output areas, on a best-fit basis. This is the method used to produce all 2011 Census and national

statistics so that statistics estimates produced on the same geography are consistent, comparable and non-disclosive.

2.1.5 English health geography

A new structure of health geographies in England is valid from 1 April 2013. This structure (Figure 2.5) consists of:

- NHS commissioning regions (NHSCR)
- NHS area teams (NHSAT)
- clinical commissioning groups (CCG)

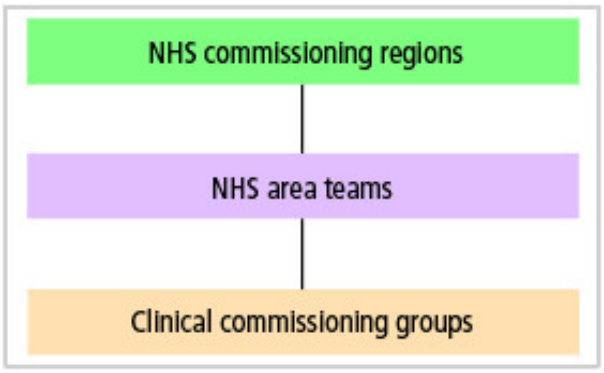

Figure 2.5: Hierarchical representation of English health geography [50].

There are four NHSCRs in England: London, Midlands and East, North and South. These regions cover healthcare commissioning and delivery across their geographies and provide professional leadership on finance, nursing, medical, specialised commissioning, patients and information, human resources, organisational development, assurance and delivery.

There were 27 NHS area teams (NHSAT) that were responsible for GP and dental services, pharmacy services and certain aspects of optical services. Ten of the teams led on specialised commissioning across England and a smaller number of NHSATs carry out the direct commissioning of prison and military health.

There are 211 clinical commissioning groups set up by the Health and Social Care Act 2012 to organise the delivery of NHS services in England. They are clinically led groups that include all of the general practice groups in their geographical area. The aim of this is to give GPs and other clinicians the power to influence commissioning decisions for their patients. CCGs are overseen by NHS England (including its regional offices and area teams). These structures manage primary care commissioning, including holding the NHS contracts for GP practices. CCGs have boundaries that are coterminous with those of lower layer super output areas.

This new structure has replaced the strategic health authorities (SHA) and primary care organisations (PCO) that had been in operation since July 2006 see Figure 2.6. There were 10 strategic health authorities which boundaries aligned with regions. Only South East region comprised two SHAs. There were 152 primary care organisations: 148 primary care trusts (PCT) and four care trusts (CT). They were constituted from groups of local authority districts.

Area Type	Count
NHS Commissioning Regions	4
NHS area teams	27
Clinical Commissioning Groups	211

Table 2.4: Hierarchy of the Census health geography

Figure 2.6: Hierarchical representation of English health geography [50].

The entire NHS structure in England is presented in Figure 2.7. The Secretary of State has overall responsibility for the work of the Department of Health. The Department of Health is responsible for strategic leadership and funding for both health and social care in England. Public Health England provides national leadership and expert services to support public health, and also works with local government and the NHS to respond to emergencies [51] The local council's role is tracking public health problems (e.g. obesity, smoking etc) in their area with the support of health and well-being boards.

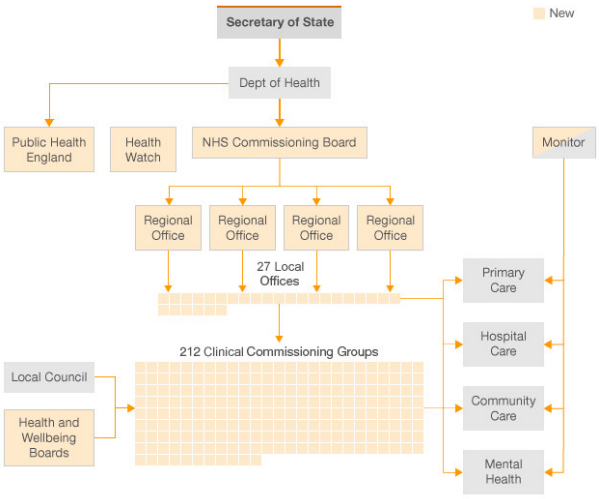

Figure 2.7: NHS structure in England [52].

Despite that the ONS officially publishes on its website two tier for commissioning regions, in April 2015 the area teams were integrated into existing four regional teams (see Figure 2.8) and operate as a single tier [53]. The resent realise of ONS hierarchical representation of UK statistical geographies in September 2016 presents two-tier geographies for NHS Commissioning Regions: NHS Commissioning Regions (4) and NHS Commissioning Regions Geography (14)- see Figure 2.1.

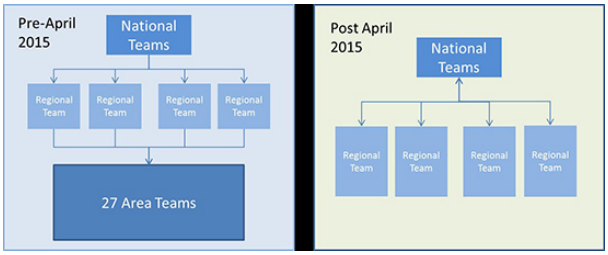

Figure 2.8: NHS Commisioning Regions [51].

2.1.6 Other geographies

In Census 2011, there is a number of geographical units for which statistics are produced with lists of names and codes. For those units we include:

- Build-up areas - a list of the most populated urban areas. Data provides information on the villages, towns and cities where people live, and allows comparisons between people living in urban areas and those living elsewhere.
- Travel to Work Areas - a list of wards for which most of the working residents works in the same area.
- Local Education Authority and Library Board - a list of local councils in England and Wales that are responsible for education within their jurisdiction.
- National Parks - a list of National Park Authority (NPA) that are responsible for conservation, planning, recreation management and fostering the social and economic well-being of local communities.
- Registration Districts - a list of areas for which records of births, deaths and marriages are kept.
- Police Force Areas - a list of territorial police forces.
- Fire and Rescue Authorities - list of local fire authorities.

2.2 Linking geographically referenced data

Geographic location is an element of information that allows defining object position on the earth. Geodata describes the location and characteristic of the real-world object such as houses, roads, boundaries of land parcels, rivers, etc in digital format. A geographically referenced object has two main elements: location and its characteristics. There are two ways to describe a geographical location:

- **Direct georeference:** the information about the location is defined by two- or three-dimensional coordinates in a coordinate reference system. Direct references for geographical objects are generally obtained from physical surveys, remote sensing, digitising of documentary sources or direct capture by Global Positioning System (GPS) receivers

- **Indirect georeference:** the information about the location is defined as administrative areas, postal addresses, postal codes and place names. It does not include explicit coordinates. It is entirely possible to work with geographically referenced data without having direct geographical references, for example by matching two sets of indirect references together (e.g. a postcode and a census ward code) in order to link records from different datasets. However, it is not possible to produce digital maps without using direct references [54].

Spatial Linkage can be understood as a link between direct and indirect referencing. It takes a form of standard reference datasets such as lookup table or digital boundary data. It contains both information: names from indirect referencing and coordinates from direct referencing.

Figure 2.9 presents the example of linkage of asthma rates with area deprivation scores. Assume that we have one dataset with information on asthma rates published for GP practices and second dataset with the deprivation scores for super output areas. To link such data, in the first step we extract postcodes of GP surgeries and their geographical locations and then we map these postcodes to the super output areas. This can be done by using Geographical Information System to geocode addresses and then map them to the super output area's boundaries or using the ONS lookup files.

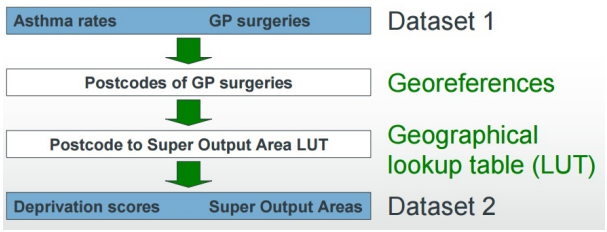

Figure 2.9: Example of linkage two dataset using georeference and geographical lookup table [55].

2.2.1 ONS files

The areas from 2011 census are mostly arranged in hierarchies in which one or more areas from a lower level are combined to form a single area in a higher level. For instance, ward areas can be combined to form local authority levels. The hierarchical relationships of the area sets for three main geographies (health, administration and census) are summarised in the diagram in Figure 2.10. As it was mentioned in Section 2.1, generally the boundaries of various geographies areas do not align with a few exemptions. Figure 2.11 presents the relationship between various geographies. They were extracted from the geography's descriptions. The black lines (within and across geographies) represents the areas with the conterminous boundaries aliments. The dash lines represent areas that split other geography's areas.

Linking data on the postcode level with data on the output area level we may find that one postcode will be linked with two output areas. This can cause problems for statistics because the postcode data will be aggregated twice in each output area statistics. To avoid such situation, ONS introduced the **best-fit** approach. This approaches used output areas as building blocks for any target geography. An estimation is made by aggregating whole output areas of statistics together to form the total estimate for the geography.

Either all the statistics for that output area will be included in the aggregation, or they will be excluded. For each output area, a single point was calculated to represent the spatial distributions and grouping of persons within that output area. This point is known as a population weighted centroid. If the output area's population weighted centroid falls within the boundary of the target geography, the output area will be included in the

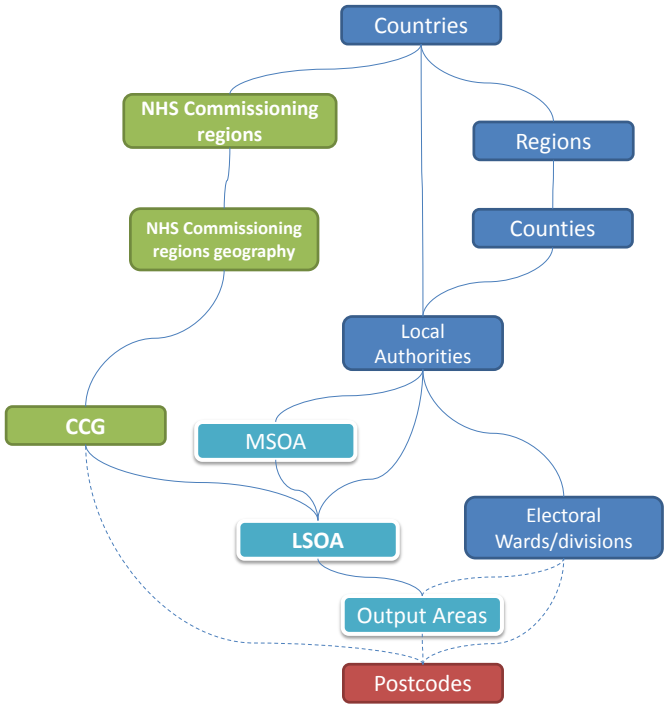

Figure 2.10: Hierarchical structure of English geographies.

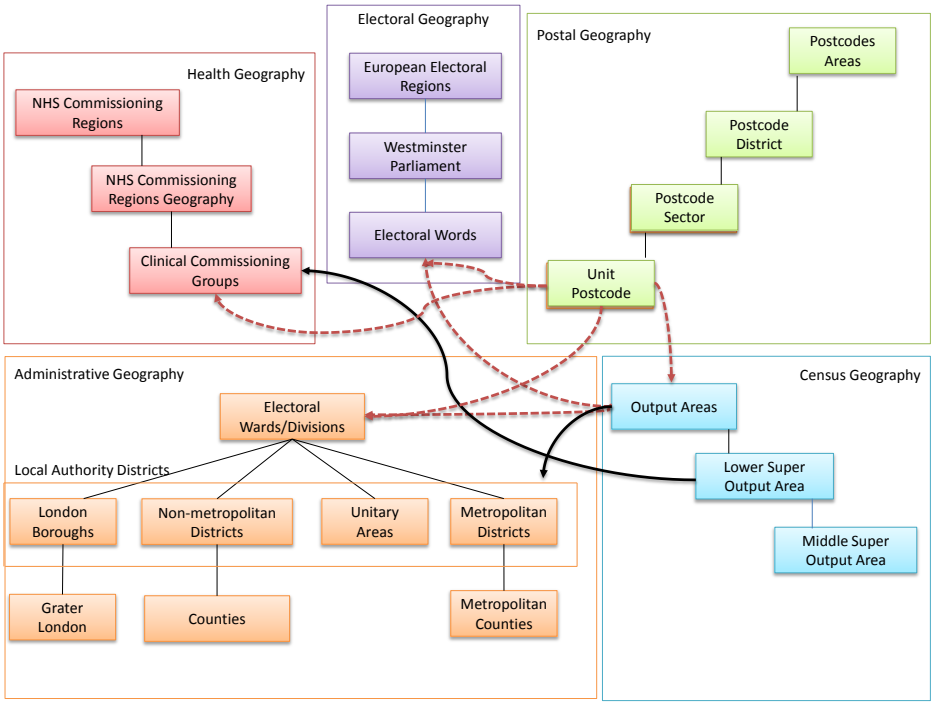

Figure 2.11: Hierarchical structure of English geographies.

best-fit allocation of output area to that target geography. This should ensure that where an output area crosses the boundary of the target geography, it is allocated to the geography that contains the majority of the output area's population. Some instances of geographies, for example parishes, that are smaller than an output area, and which do not contain an output area centroid, will be allocated to the output area with a centroid that is nearest to any part of the target geography's boundary. [56].

The ONS Open Geography Portal [57] provides free and open access to the definitive source of geographic products, web applications, story maps and services. The user can browse data, geography boundaries, maps, documents, postcode products, lookup tables.

UK Geography names and codes database

The Code History Database (CHD) contains the GSS nine-character codes that were allocated for current and new statistical geographies from 1 January 2009. This database includes details of codes, their relationships, hierarchies and archived data. The MS ACCESS database provides multi-functionality and enables user to view or export data for the following options (see Figure 2.12):

- Geography Listings - includes both terminated and live entities, and codes for geographies
- Geography History - contains the change history
- Geography Hierarchies - provides the hierarchies for all entity themes where a one-to-one relationship exists. The file extracted here can be used as lookup tables for geographies described in the Section 2.1
- Geography Constitutions - provided for some geographies in England and Wales. These include parish to ward, electoral and health constitutions
- Geography Equivalents - provides the new codes and previous ONS, CLG, DH and other equivalent names and codes where available
- Geography Information - contains information about the geographies including Statutory Instrument Information

The current database was released on 7 September 2016 and is available as a zip file[1]. This database is updated quarterly.

Lookup files

Using best-fit approach, ONS provide a list of lookup tables with names, codes and relationship between geographies [58]. There are following lookup tables [59]:

1. 2001 to 2011 Census lookups - A range of cross-reference tables to allow linkage between the 2001 and 2011 output areas and super output areas.
2. Lookups for 2011 output areas to other geographies:
    - 2011 output areas to 2011 wards

[1] http://ons.maps.arcgis.com/home/item.html?id=210eb9e8c06e45db8119a42dcedcf8cd

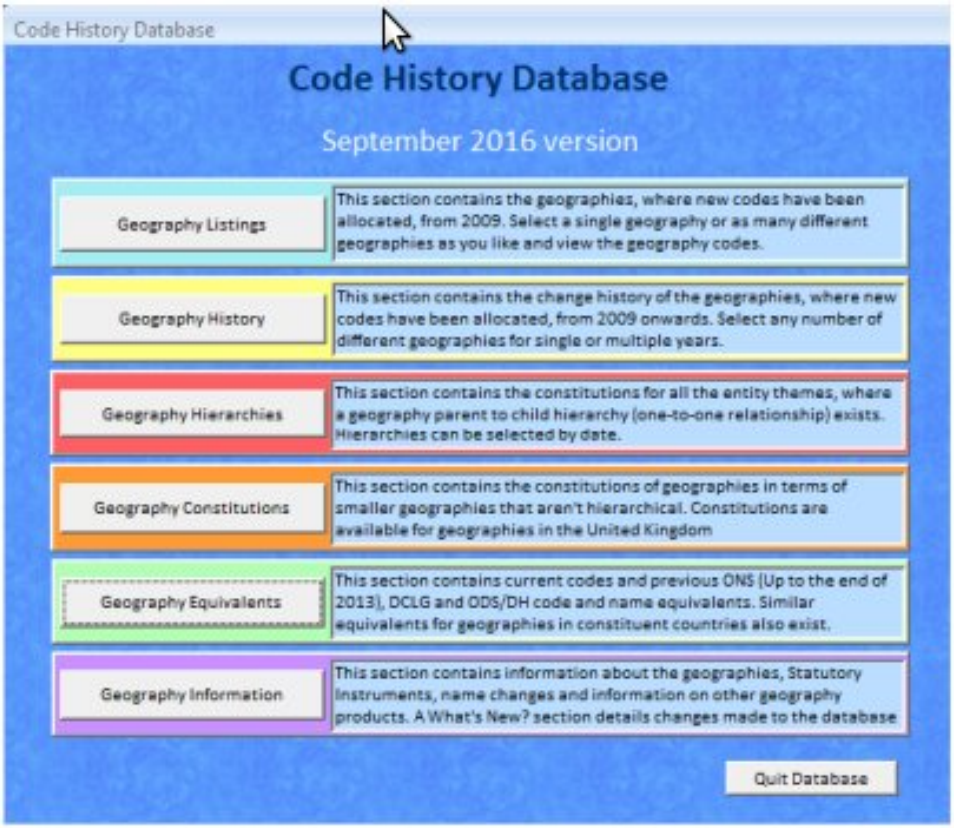

Figure 2.12: The Code History Database

- 2011 output areas to 2011 parishes
- 2011 output areas to current counties in England
- 2011 output areas to regions in England
- 2011 output areas to 2010 Westminster parliamentary constituencies and European electoral regions (with best-fit percentage indicator)
- 2011 output areas to 2011 primary care organisations (PCO) in England/local health boards (LHB) in Wales and strategic health authorities (SHA) in England (with best-fit percentage indicator)
- 2011 output areas to 2011 built-up area sub-divisions (BUASD), built-up areas (BUA), local authority districts (LAD) and regions
- 2011 built-up areas to 2011 local authority districts
- 2011 build-up areas to 2011 regions
- 2011 output areas to 2011 urban/rural definition
- 2011 output areas to 2011 enumeration postcode sectors
- 2011 output areas to 2011 lower layer super output areas (LSOAs), middle layer super output areas (MSOA) and local authority districts - exact fit
- 2011 output areas to 2011 workplace zones (WZ) and local authority districts - exact fit

3. Other geographies

- 2011 enumeration postcodes to 2011 output areas (OA), lower layer super output areas (LSOA), middle layer super output areas (MSOA) and local authority districts (LAD) (with split postcode indicator)
- 2011 wards to 2011 Census merged wards (as used for Census upper-threshold statistics)
- 2011 workplace zones (WZ) to 2011 MSOAs and LADs

More lookup files for various geographies and the newest releases can be found in [57].

Postcode lookup files

The ONS Open Geography Portal provides a number of postcode lookup files. There are four main lookup files:

- National Statistics Postcode Lookup (NSPL) - relates both current and terminated postcodes in the United Kingdom to administrative, electoral, health and other statistical geographies via 'best-fit' allocation from 2011 Census Output Areas (OA). The last release was in August 2016[2]. The file is issued quarterly.
- ONS Postcode Directory (ONSPD) - relates both current and terminated postcodes in the United Kingdom to administrative, electoral, health and other area geographies. It links postcodes to pre-2002 health areas, 1991 Census enumeration districts for England and Wales, 2001 Census Output Areas (OA) and Super Output Areas (SOA) for England and Wales, 2001 Census OAs and SOAs for Northern Ireland and 2001 Census OAs and Data Zones (DZ) for Scotland. The last release was in August 2016[3].The file is issued quarterly.
- NHS Postcode Directory (NHSPD) - relates both current and terminated postcodes in the United Kingdom to administrative, electoral, health and other geographies. It links postcodes to pre-2002 health areas and 2001 Census and 2011 Census Output Areas and Super Output Areas. NHSPD uses information supplied on a monthly basis by Royal Mail. This product contains Royal Mail, Gridlink, LPS (Northern Ireland), Ordnance Survey and ONS Intellectual Property Rights. The last release was in August 2016[4]. The file is issued quarterly.

2.2.2 Linkage using GIS

Geographical Information System (GIS) is designed to capture, store, manipulate, analyse, manage, and present all types of spatial or geographical data. GIS can show many different kinds of data on one map. This enables people to more easily see, analyse, and understand patterns and relationships [60].

The different shapes and symbols are used to illustrate features. There are four main types of symbol used to represent the different feature types:

- Point - for example, a dot symbol to represent a house or a cross to represent a church

[2]http://ons.maps.arcgis.com/home/item.html?id=ad13ce429d9644b88fc1e85af2e6ed8a
[3]http://ons.maps.arcgis.com/home/item.html?id=5a656df5f06b4325aa83f907cf0e8d
[4]http://ons.maps.arcgis.com/home/item.html?id=dc23a64fa2e34e1289901b27d91c335b

- Line - for example, a line to represent a road
- Polygon shape or area - for example, a blue area to represent a lake, boundary of city
- Text - for example, the name of a town or river.

The combination of many different spatial datasets (points, lines, or polygons) creates a new output dataset. Visually, it is done by stacking several maps of the same region. From these overlays, we can extract the features of one data set that fall within the spatial extent of another dataset. For example, visualising the addresses within wards or output area boundaries in the same time linking the postcode areas to the other geography. Ordnance Survey produces many different GIS data products and one of them is the full hierarchy of local government administrative and electoral boundaries in Great Britain. In some cases for data aggregation, the Census "weighted centroid" may not be a good solution. In such situation, using GIS we can define new criteria for data aggregation and spatial relationships between different types of GIS data to integrate and link information together.

Geographical information systems are also used for geocoding. Geocoding is the process of linking an address to a physical location on the earth. From street addresses or any other spatially referenced data such as postcode, parcel and address locations, the GIS calculates geographic coordinates before an address can be displayed on a map. There are two approaches for geocoding [3]. In the first approach, a reference dataset of all known addresses with their geographical location in a certain geographical area is required to geocode individual addresses. The queried address is matched to the addresses in the reference database and their location is returned. The matching is performed using the string comparison and similarity measures described in Chapter 1. The second approach based on using a street centre line database as a reference dataset. This database is made of the geographical locations of small street segments. When an address is matched to such a street segment, its geographical location is extrapolated based on the start and end locations of the street segment and the corresponding start and end street numbers updated by the street offset [6].

The census boundary data can be downloaded from Edina [61]. It is a boundary data selector interface that allows to easily choose the country and geographies of our interest. There is four format for download: CSV file, MAPINFO native format, KML (Keyhole Markup Language) and SHAPE (ESRI Shapefile format).

Glossary

direct georeference information about the location is defined by 23D coordinates in a coordinate reference system.. 18

exact linkage known as exact matching, exact agreement on all linkage columns. 6

indirect georeference iinformation about the location is defined as administrative areas, postal addresses, postal codes and place names.. 19

matches records that are highly similar to each other. 3

non-matches records that are completely different. 3

potential matches records for which some linking attribute are similar and some are different. 3

record linkage is a process of matching records that represent the same instance or entity from one or more databases/datasets. 2

Bibliography

[1] P. Christen and K. Goiser, *Quality Measures in Data Mining*, ch. Quality and Complexity Measures for Data Linkage and Deduplication, pp. 127–151. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[2] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, pp. 1–16, Jan. 2007.

[3] P. Christen, "Privacy, security, and trust in kdd," ch. Geocode Matching and Privacy Preservation, pp. 7–24, Berlin, Heidelberg: Springer-Verlag, 2009.

[4] H. B. Newcombe and J. M. Kennedy, "Record linkage: Making maximum use of the discriminating power of identifying information," *Commun. ACM*, vol. 5, pp. 563–566, Nov. 1962.

[5] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, pp. 1183–1210, Dec. 1969.

[6] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated, 2012.

[7] S. M. Randall, A. M. Ferrante, J. H. Boyd, and J. B. Semmens, "The effect of data cleaning on record linkage quality," *BMC Medical Informatics and Decision Making*, vol. 13, no. 1, pp. 1–10, 2013.

[8] L. Gill, "Methods for automatic record matching and linkage and their use in national statistics," 2001.

[9] The National Archives, "The soundex indexing system." http://www.archives.gov/research/census/soundex.html, 2007. Accessed: 10 June 2016.

[10] L. Philips, "Hanging on the metaphone," in *International Conference on Computer Languages*, 1990.

[11] R. L. Taft, "Name search techniques," Tech. Rep. Special Report No. 1, New York State Identification and Intelligence System, Albany, NY, February 1970.

[12] T. Churches, P. Christen, K. Lim, and J. X. Zhu, "Preparation of name and address data for record linkage using hidden markov models," *BMC Medical Informatics and Decision Making*, vol. 2, no. 1, pp. 1–16, 2002.

[13] L. Gu, R. Baxter, D. Vickers, and C. Rainsford, "Record linkage: Current practice and future directions," tech. rep., CSIRO Mathematical and Information Sciences, 2003.

[14] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, pp. 31–88, Mar. 2001.

[15] R. Baxter, P. Christen, and T. Churches, "A comparison of fast blocking methods for record linkage," in *KDD 2003 WORKSHOPS*, pp. 25–27, 2003.

[16] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, pp. 1537–1555, Sept 2012.

[17] M. A. Hernández and S. J. Stolfo, "The merge/purge problem for large databases," *SIGMOD Rec.*, vol. 24, pp. 127–138, May 1995.

[18] M. A. Hernández and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," *Data Min. Knowl. Discov.*, vol. 2, pp. 9–37, Jan. 1998.

[19] W. W. Cohen and J. Richman, "Learning to match and cluster large high-dimensional data sets for data integration," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, (New York, NY, USA), pp. 475–480, ACM, 2002.

[20] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, (New York, NY, USA), pp. 169–178, ACM, 2000.

[21] L. Jin, C. Li, and S. Mehrotra, "Efficient record linkage in large data sets," in *Proceedings of the Eighth International Conference on Database Systems for Advanced Applications*, DASFAA '03, (Washington, DC, USA), pp. 137–, IEEE Computer Society, 2003.

[22] J. Fisher, P. Christen, Q. Wang, and E. Rahm, "A clustering-based framework to control block sizes for entity resolution," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, (New York, NY, USA), pp. 279–288, ACM, 2015.

[23] R. Hall and S. E. Fienberg, *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2010, Corfu, Greece, September 22-24, 2010. Proceedings*, ch. Privacy-Preserving Record Linkage, pp. 269–283. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[24] D. Vatsalan, P. Christen, and V. S. Verykios, "A taxonomy of privacy-preserving record linkage techniques," *Information Systems*, vol. 38, no. 6, pp. 946–969, 2013.

[25] C. Kelman, A. Bass, and C. Holman, "Research use of linked health data - a best practice protocol," *Australian and New Zealand Journal of Public Health*, vol. 26, no. 3, pp. 251–255, 2002.

[26] P. Christen, "Febrl: A freely available record linkage system with a graphical user interface," in *Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management - Volume 80*, HDKM '08, (Darlinghurst, Australia, Australia), pp. 17–25, Australian Computer Society, Inc., 2008.

[27] Fiebrl, "Freely extensible biomedical record linkage." https://sourceforge.net/projects/febrl/. Accessed: 10 June 2016.

[28] W. based Information Representation Language, "William cohen." http://www.cs.cmu.edu/~wcohen/whirl/, 06 2016. Accessed: 10 June 2016.

[29] M. G. Elfeky, V. S. Verykios, and A. K. Elmagarmid, "Tailor: a record linkage toolbox," in *Data Engineering, 2002. Proceedings. 18th International Conference on*, pp. 17–28, 2002.

[30] SimMetrics. http://sourceforge.net/projects/simmetrics/. Accessed: 10 June 2016.

[31] M. Sariyar and A. Borgs, "Recordlinkage." http://cran.r-project.org/web/packages/RecordLinkage/index.html. Accessed: 10 June 2016.

[32] "Fril." http://fril.sourceforge.net/. Accessed: 10 June 2016.

[33] W. Yancey, "Bigmatch: A program for extracting probable matches from a large file for record linkage," tech. rep., US Bureau of the Census, 2007.

[34] ONS, "UK Geographies." http://www.ons.gov.uk/ons/guide-method/geography/index.html. Accessed: 10 June 2016.

[35] "The hierarchical representation of uk statistical geographies from october 2015." https://data.gov.uk/dataset/hierarchical-representation-of-uk-statistical-geographies-oct-2015.

[36] "Postcode address file." http://www.royalmail.com/business/services/marketing/data-optimisation/paf.

[37] ONS, "Postal geography." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/postal/index.html. Accessed: 10 June 2016.

[38] ONS, "Census geography." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/index.html. Accessed: 10 June 2016.

[39] ONS, "Output area (oa)." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/output-area--oas-/index.html. Accessed: 10 June 2016.

[40] ONS, "The census 2011 bestfit method." http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/new-developments-for-2011-census-results/2011-census-geography/exact-fit-and-best-fit-estimates/BFOverview.pdf. Accessed: 10 June 2016.

[41] ONS, "Administrative geography." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/index.html. Accessed: 10 June 2016.

[42] ONS, "Metropolitan counties and districts." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/metropolitan-counties-and-districts/index.html. Accessed: 10 June 2016.

[43] ONS, "Counties, non-metropolitan districts and unitary authorities." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/counties--non-metropolitan-districts-and-unitary-authorities/index.html. Accessed: 10 June 2016.

[44] ONS, "Greater london and the london boroughs." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/greater-london-and-the-london-boroughs/index.html. Accessed: 10 June 2016.

[45] ONS, "Electoral wards / electoral divisions." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/electoral-wards-divisions/index.html. Accessed: 10 June 2016.

[46] ONS, "Parishes." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/parishes-and-communities/index.html. Accessed: 10 June 2016.

[47] ONS, "Uk electoral geography." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/electoral/index.html. Accessed: 10 June 2016.

[48] ONS, "European electoral region." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/electoral/european-electoral-regions/index.html. Accessed: 10 June 2016.

[49] ONS, "Westminster parliamentary constituency." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/electoral/westminster-parliamentary-constituencies/index.html. Accessed: 10 June 2016.

[50] ONS, "Health geography." http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/health/english-health-geography/index.html. Accessed: 10 June 2016.

[51] NHS, "The NHS in england." http://www.nhs.uk/NHSEngland/thenhs/about/Pages/nhsstructure.aspx. Accessed: 10 Sep 2016.

[52] BBC, "The changing nhs." http://www.bbc.co.uk/news/health-19674838, March 2013. Accessed: 9 Sep 2016.

[53] NHS England, "Nhs england regional teams." https://www.england.nhs.uk/about/regional-area-teams/. Accessed: 12 Sep 2016.

[54] ReStore, "Geographical referencing learning resources." http://www.restore.ac.uk/geo-refer/52620cwors00y00000000.php. Accessed: 10 June 2016.

[55] D. Martin, "Linking and mapping geographically reference data." http://www.restore.ac.uk/geo-refer/files/NCRM\%20Autumn\%20School\%20-\%20Nov\%2008.pdf. Accessed: 10 June 2016.

[56] ONS, "Best-fit." http://www.ons.gov.uk/ons/guide-method/geography/geographic-policy/best-fit-policy/index.html. Accessed: 10 June 2016.

[57] ONS, "Open geography portal." http://geoportal.statistics.gov.uk/. Accessed: 9 Sep 2016.

[58] ONS, "Census lookup tables." http://www.ons.gov.uk/ons/guide-method/geography/products/census/lookup/index.html. Accessed: 10 June 2016.

[59] ONS, "2011 census output areas lookups." http://www.ons.gov.uk/ons/guide-method/geography/products/census/lookup/2011/index.html. Accessed: 10 June 2016.

[60] O. Huisman and R. By, *Principles of Geographic Information Systems: An Introductory Textbook*. ITC educational textbook series, International Institute for Geo-Information Science and Earth Observation (ITC), 2009.

[61] Edina, "Boundary data." https://census.edina.ac.uk/bds.html. Accessed: 10 June 2016.