Geocoding ISTV data - methodology

1. Hospital data join

Data from all hospitals is sourced from the secure SafeStats SQL database and individually standardised to ensure each hospital dataset only contains the key information required for geocoding. Once combined into a single dataset, the data is split into two – those records containing location data that can be geocoded, and those that cannot.

2. Clean

The geocodable dataset is 'cleansed' by removing from the location information non-standard characters (eg. @/*/"), as well as ensuring that commonly occurring words/phrases are all spelt the same (eg. station, stn or street, st)

3. Split

Keyword and pattern recognition are used to identify key address components present in each location (schools, stations, landmarks, areas, streets. postcodes etc) and records of each type are flagged accordingly.

4. Prioritise

Based on the likelihood of a component or combination of components being able to place a location at a specific accurate point on a map, priority levels are assigned to each record based on the components present (eg. train station > street > area).

5. Geocode

Different combinations of address components are then geocoded in different ways. If a commonly used location is identified (station, hospital, school, landmark etc), manually created lookup tables containing their point locations are used. If only partial postcode or area components exist, spatial lookup tables are used that match these with a wider electoral ward or borough. For those records that contain full postcode or street-based information, an online map geocoding service is used to locate to a street or postcode area centrepoint. This provides three levels of output – point (actual or centrepoint), ward and borough.

A 'cascade' concept is also utilised to capture those records that cannot be geocoded at each level, and are therefore passed to the next geocoding method down. For example, if a school cannot be matched by the lookup table, it is then passed through the online service; if this service cannot match it, it is checked for partial postcode or area information, and if still no match is found, 'no match' is recorded. These 'fallouts' are captured for ongoing analysis.

6. Re-join and spatial alignment

The geocoded outputs of each method are joined back together, and corresponding ward and borough information added where applicable (eg. to point outputs of the geocoding process).

7. Outputs

Geocoded data is rejoined to the source data that could not be geocoded (stage 1. above) and output into both the ISTV interface and SafeStats Data.

To request access, visit maps.london.gov.uk/safestats/ and select 'Request Access'

For more information, please contact safestats@london.gov.uk

www.london.gov.uk/safestats