The development of a ‘Postcode Best Fit’ methodology for producing Population Estimates for different geographies

Andy Bates
Office for National Statistics

A ‘Postcode Best Fit’ methodology has been developed by the Office for National Statistics to produce population estimates for a range of different geographies which are entirely consistent with each other, regardless of whether or not the estimates for one geography can be aggregated to produce estimates for another geography. This article describes the Postcode Best Fit methodology, its evaluation, limitations with some of the data sources used with the method, the application of the method for producing population estimates and case studies describing how the method has been used to produce bespoke population estimates to meet specific user requirements.

Introduction

In addition to the population estimates published by the Office for National Statistics (ONS) at national and local authority level, ONS is now publishing population estimates for a wider range of different geographies to meet user requirements, covering Super Output Areas, Wards, Primary Care Organisations, National Parks and Parliamentary Constituencies.

With the published ONS population estimates there is a hierarchy of estimates for different geographies, all consistent with each other. Lower Layer Super Output Area (LSOA) estimates are consistent with Middle Layer Super Output Area (MSOA) estimates; these in turn are consistent with the local authority estimates, which in turn are consistent with the national (England and Wales) estimate. The requirements for additional population estimates for other geographies transcend this hierarchy and cannot therefore always be attained by simply aggregating existing outputs.

Although ONS has previously developed and evaluated different demographic methods for producing population estimates at small area level, the limited geographical availability of some administrative data sources has made it necessary to develop a different method to produce population estimates for geographies which do not fit within the boundaries of existing outputs.

In order to meet the requirements for population estimates for a wider range of geographies, a best fitting method, labelled a ‘Postcode Best Fit’ (PBF) method has been developed. This Postcode Best Fit method provides a mechanism for producing population estimates for different geographical levels which are all entirely consistent. The method is versatile and easy to implement so will help to overcome the difficulty of producing annual population estimates when boundaries are
subject to periodic review and change (as for wards and parliamentary constituencies for example). This method is also capable of producing population estimates for both small areas and large areas.

**Description of the Postcode Best Fit methodology**

In essence the PBF method is an ‘apportionment’ method, apportioning population estimates from the smallest small area geography for which population estimates are published by ONS at LSOA level, to unit postcode level based on age and sex information from patient register postcode level data.

This PBF method uses the population estimates for the 34,378 LSOAs in England and Wales by age and sex (average population 1,560) and apportions these to around 1.32 million residential and communal establishment postcodes in England and Wales (with an average population of around 40) based on the counts of persons by age and sex included on the patient registers. A special allowance is made for population sub-groups not included on the patient registers, covering prisoners, UK armed forces, and foreign armed forces and dependants. The LSOA counts for this special population are removed from the apportioning process and then added back in at unit postcode level, based on postcode information for the special population.

These postcode level population estimates can then be aggregated (or ‘best fitted’) to a range of higher geographies if required using a suitable postcode look-up file, for example the National Statistics Postcode Directory (NSPD) or Geographical Information System (GIS) (Figure 1). Population estimates for some areas for higher geographies may be derivable from existing outputs, and therefore PBF-derived estimates would not necessarily be required for these areas – for example wards which can be derived from aggregation of whole LSOAs.

The assumption is made that any patient register list inflation, whereby patient register counts exceed the mid-year population estimates, is consistent within a LSOA, and so when the LSOA estimates (less special population) are apportioned to unit postcode level, this is done on the basis that the patient register counts by age and sex will closely reflect the actual population distribution (though not necessarily the actual numbers).

It is not intended that derived unit postcode estimates are themselves published or released due to the uncertainty over their accuracy at this fine level of geographical detail. Instead the purpose of these postcode estimates is that they are aggregated to other higher geographies using the NSPD or GIS.

**Evaluation of PBF derived population estimates**

In order to assess whether the PBF method is capable of producing population estimates which are deemed to be sufficiently reliable and accurate, an evaluation of PBF derived population estimates was undertaken.

It was initially proposed than an evaluation of PBF derived estimates could be done by comparing mid-2002 estimates for Census Area Statistics (CAS) wards from the PBF method with the previously published mid-2002 CAS ward estimates. However because of the way in which CAS wards were created from Census Output Areas, for 98.8 per cent of the 8,850 CAS wards, population estimates for these wards could in fact be derived from aggregations of LSOA estimates. In order to make a better judgement of the PBF method, the evaluation later covered the comparison of population estimates for Primary Care Organisations (PCOs, also known as Primary Care Trusts) in England, and because of associated problems with this geography, later extended to cover postcode sectors for the whole of England and Wales.

The key elements to the evaluation of population estimates for wards, PCOs and postcode sectors was a statistical analysis, including for example looking at the maximum and average absolute change and absolute percentage change in the estimates by age and sex, and the correlation of the estimates on a scatter plot.

There are 102 CAS wards where the ward boundaries are non-coterminous with the LSOA boundaries (excluding both the City of London and Isles of Scilly which were each treated as single CAS wards). When comparing the original mid-2003 population estimates for these wards with estimates derived from PBF, for the total population counts, there was a mean absolute percentage difference of 1 per cent, with no wards having an absolute percentage difference of more than 10 per cent, and just 7 wards with an absolute percentage difference between 5 and 10 per cent.

At the time of the evaluation there were 303 PCOs in England for which mid-2002 population estimates had previously been published (as a result of reorganisation there are currently 152 PCOs). 139 PCOs (45.9 per cent) were omitted from the evaluation as estimates for them could be derived by aggregation of the estimates for the component local authorities (as they are coterminous), or were subject to boundary change since 2002. Thus the mid-2002 published estimates for just 164 PCOs were considered alongside mid-2002 estimates generated by PBF for evaluation purposes. The vast majority of the PBF estimates were very close to the published estimates, with only 14 per cent of the quinary age estimates having an absolute percentage difference greater than 1 per cent and only 2 per cent having an absolute percentage difference of more than 3 per cent. There was an extremely high correlation ($r = 0.9999$) between the respective total population counts.

In theory, population estimates for many of the 164 PCOs considered as part of this evaluation could alternatively have been derived by aggregating the estimates for LSOAs where the geographies were coterminous. A more rigorous evaluation would therefore require independent population estimates for a geography where estimates could
not be derived from existing estimates and which ideally had a national coverage.

At the time, ONS was aware of the availability of population estimates by quinary age and sex for postcode sectors within England and Wales produced for the Joint Industry Committee for Population Standards (JICPOPs) by two companies working together, CACI and Experian. JICPOPs kindly agreed to provide mid-2003 population estimates at postcode sector level to assist with the evaluation of the PBF method. These were compared to mid-2003 postcode sector population estimates derived from PBF.

Whilst some ONS data are used in the creation of the JICPOPs postcode sector estimates such as the local authority mid-year population estimates and projections, and 2001 Census data, non-ONS data are also used (for example, estimates of residents in households). The methodology used to produce the estimates is different from that used by ONS. Consequently the JICPOPs estimates are considered to be independent population estimates. Population estimates for postcode sectors spanning both England and Scotland were excluded from the evaluation.

Comparisons were made between the respective mid-2003 JICPOPs and ONS PBF derived postcode sector total population estimates (Figure 2). This indicated a very high correlation ($r = 0.9973$) between the two sets of estimates, though some outliers are noticeable and tend to occur where the ONS estimate is greater than the JICPOPs estimate. The outliers with the greatest absolute differences are shown in Table 1.

These outliers were investigated. Findings of ONS research which reviewed evidence on the 2001 Census estimates indicated that, whilst no single piece of evidence on its own was conclusive, the weight of evidence suggested that the 2001 Census did not cover all people in England and Wales, particularly young adult men. Accordingly the 2001 local authority mid-year estimates, most recently revised in September 2004, reflect this evidence; these include adjustments for missing Census forms, Longitudinal Study adjustments, the Manchester and Westminster Matching Studies and 2004 Local Authority Studies.

We can attribute the biggest differences between the two sets of estimates to revisions to ONS annual population estimates for mid-2001 onwards that were released in September 2004. These have been incorporated in the ONS PBF method, but were not included in the JICPOPs mid-2003 estimates that were released before this date.

It was apparent that for the majority of postcode sectors there were relatively small differences between the PBF and JICPOPs estimates (Figure 3). For over half of the postcode sectors the differences are within ±100, 83 per cent within ±250, and 95 per cent within ±500. Consequently only around 5 per cent are outside the range ±500.

From the evaluation of population estimates produced using the PBF method for CAS wards, PCOs and postcode sectors with other estimates for these areas, the comparisons indicated little difference between the respective PBF and non-PBF derived estimates. Where there were noticeable differences, these could usually be accounted for. For example, adjustments to the mid-2001 estimates for Westminster resulted in noticeable differences between these estimates and the 2001 Census count, contributing to differences between PBF and JICPOPs population estimates for mid-2003 for postcode sectors covering Westminster. As a result of this evaluation no serious concerns associated with the method and the estimates were identified, though it is recognised that there are some limitations with population estimates produced from PBF.

**Limitations with population estimates produced from PBF**

The known limitations can be categorised as follows:

1. Issues with the patient register postcode data
2. Issues with the population estimates for LSOAs and local authorities
3. Issues with the accuracy of the NSPD and postcode changes

### 1. Patient Register postcode data

Through the use of patient register postcode data within the Ratio Change methodology used to produce small area population estimates, we are aware of limitations in the use of patient registers as a proxy indicator

---

**Table 1**

<table>
<thead>
<tr>
<th>Postcode-sector</th>
<th>Local Authority</th>
<th>JICPOPs total</th>
<th>Difference (ONS-JICPOPs)</th>
<th>% Difference (ONS-JICPOPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W9 3</td>
<td>Westminster</td>
<td>10,580</td>
<td>4,810</td>
<td>45.5</td>
</tr>
<tr>
<td>NN8 8</td>
<td>Westminster</td>
<td>6,870</td>
<td>4,660</td>
<td>67.8</td>
</tr>
<tr>
<td>W10 4</td>
<td>Westminster</td>
<td>7,511</td>
<td>3,502</td>
<td>46.6</td>
</tr>
<tr>
<td>W2 5</td>
<td>Kensington &amp; Chelsea/Westminster</td>
<td>9,572</td>
<td>3,201</td>
<td>33.4</td>
</tr>
<tr>
<td>DE23 8</td>
<td>Derby</td>
<td>14,618</td>
<td>2,629</td>
<td>18.0</td>
</tr>
<tr>
<td>W11 1</td>
<td>Kensington &amp; Chelsea</td>
<td>9,966</td>
<td>2,514</td>
<td>25.2</td>
</tr>
<tr>
<td>W10 6</td>
<td>Kensington &amp; Chelsea</td>
<td>10,612</td>
<td>2,471</td>
<td>22.3</td>
</tr>
<tr>
<td>RM16 6</td>
<td>Thurrock</td>
<td>8,738</td>
<td>2,389</td>
<td>27.3</td>
</tr>
<tr>
<td>M14 4</td>
<td>Manchester</td>
<td>5,850</td>
<td>2,257</td>
<td>38.6</td>
</tr>
<tr>
<td>W10 5</td>
<td>Kensington &amp; Chelsea</td>
<td>13,529</td>
<td>2,112</td>
<td>15.6</td>
</tr>
</tbody>
</table>
of population size. Nationally the count of persons recorded on the patient registers exceeds the national population estimate. The mid-2006 count on the patient registers exceeded the mid-year estimate by 2.9 m (5.4 per cent). The excess of persons on the patient registers compared to population estimates, often referred to as list inflation, may occur when some patients have more than one NHS number and are double counted, and patients may be on doctors’ lists after having left the country. List inflation may also be localised, for example in student areas where students do not quickly re-register after finishing their course of study and moving away from an area.

We are also aware of inconsistencies in patient register postcode data over time which could negate the accuracy of any PBF estimates in some areas. An investigation has been undertaken for individual postcodes where there are large patient register counts and instances where the patient register counts fluctuate significantly over time. It emerged that a significant number of these postcodes related to halls of residence at universities. Using student count information from universities and the 2001 Census, for some postcodes a year-specific adjusted patient register postcode count was created to more closely reflect actual numbers of resident students. As patient register data are used in the production of small area population estimates, unless there was information to suggest otherwise, these adjusted patient register postcode counts by age and sex are invariably kept constant over time.

We now have year-specific population-adjusted counts of patient register data to more closely reflect the likely age and sex characteristics of the population usually resident within an area. The fact that in some areas there is significant patient register list inflation should not be an issue as long as this occurs consistently within each LSOA and across all postcodes, as the effect of apportioning the LSOA estimates to the patient register counts at postcode level where widespread list inflation occurs is that all counts by unit postcode would be reduced. The extent to which there is list inflation (or an under recording) which is not consistent within a LSOA (that is, very localised) could, however, impact on the effectiveness of this PBF method when the constrained postcode counts are aggregated to an alternative geography (for example, statistical wards) which may not be coterminous with the LSOA geography, that is, ‘cut across’ the LSOA.

2. Issues with the Population estimates for LSOAs and local authorities

The postcode population estimates are consistent if aggregated to the population estimates for LSOAs, which in turn are consistent with estimates for higher geographies such as the local authority and national population estimates.

Whilst the local authority mid-year estimates have National Statistics status, meeting standards for quality and relevance, it is recognised that some areas can be more difficult to estimate accurately than others. For example areas with high levels of migration could be expected to be more difficult to estimate accurately than areas with very low levels of migration. For this reason any inaccuracies in the local authority mid-year estimates will also be reflected in the LSOA estimates, and in turn in the postcode unit estimates.

It also needs to be recognised that the individual addresses for some postcodes may straddle the target geography (and in some cases the LSOA itself). For this reason some error may result when the unit postcode estimates are aggregated to higher geographies as this will be done on the NSPD grid reference for the postcode. All persons associated with addresses with the same postcode will therefore all be allocated to the same area within target geography, even though some addresses may physically be within a different area of the target geography.

3. Issues with the NSPD and postcode changes

Postcodes with PO Box numbers

Non-geographic postcodes can either be special postcodes assigned to some large users of the postal service or PO Boxes that lie within a (pseudo) postcode sector. These PO Box postcodes will all have been assigned a grid reference, usually the local Royal Mail sorting office.

We are aware that some of the postcodes recorded on the patient registers do, in fact, relate to PO Box postcodes and therefore have an associated population, for example some relate to student halls of residences. Where these patient register PO Box postcodes can be identified as belonging to a particular establishment, they may be allocated on the NSPD to the ‘wrong’ LSOA as the allocation is done on the grid reference (usually the Royal Mail sorting office) rather than the physical location of the establishment. For example the postcode NP18 3YG for PO Box 179 which relates to Caerlon Campus, University of Wales, Newport is allocated on the NSPD to an LSOA some 4 km away from the actual campus, and in a different LSOA. Without a specific grid reference and LSOA adjustment this could result in postcodes within some LSOAs with too low or too high estimated population counts (in this example our LSOA estimates do reflect the actual location of the university campus).

With the mid-2006 patient register adjusted counts, 10,000 people were associated with 4,300 PO Box number postcodes, with 104 people associated with both a PO Box number and a non-geographic postcode (40 postcodes).
There were only seven occurrences of patient register PO Box number postcodes where the counts of persons associated with such postcodes were 50 or greater. In fact, 98 per cent of patient registers PO Box postcode usage related to fewer than 10 persons. Special consideration may need to be given to PO Box postcodes and the allocation of these postcodes to other geographies when applying the PBF method, though nationally this will have very little impact on any PBF derived estimates.

Postcode changes and terminated postcodes
The Royal Mail periodically make changes to postcodes, terminating old postcodes, creating new postcodes (generally when new residential properties or industrial/commercial premises are built) and changing postcodes in particular localities as part of postcode reorganisation. Such changes will impact on the PBF. There may be a time lag between the period when new residential properties are built and occupied, and the time people are included on the patient register. In addition it may take some time for terminated postcodes to be removed from the patient registers and some may remain indefinitely.

Some terminated postcodes will have been renamed and so, for PBF purposes, it is not a problem that counts relate to these terminated postcodes as long as the people (GP patients) to which they relate still exist and reside at the address previously associated with the terminated postcode. There are however procedures in place for individual PCTs to update on a quarterly basis patient register postcode information when new postcodes are introduced as part of a Royal Mail postcode reorganisation.

On the mid-2006 patient registers, there were 382 postcodes which had been terminated which had counts of over 100 persons; of these, however, only two postcodes (0.5 per cent) were terminated before 2000. This suggests that generally postcodes which have been terminated on the patient register are removed (if appropriate) or renamed (if a postcode change is involved).

Given the overall number of patient register counts relating to terminated postcodes (just 0.2 per cent), this is not considered to be a big problem, though it may impact locally on the quality of any PBF-derived estimates produced.

For the PBF method to work as intended, accuracy on the NSPD is essential. It needs to include valid postcodes, but importantly to have accurate grid reference information for these postcodes. Typically there will be around 40 people with the same postcode, grid references in the NSPD are available in one or 100 metre resolution, and the majority are derived from the Ordnance Survey product ADDRESS-POINT containing grid references for each address to 0.1 metre resolution. The addresses for a single postcode may straddle geographies, for example LSOAs or wards, but as these geographies are based on aggregations of Census Output Areas which were themselves based around unit postcodes, this will generally not be the case.

For the majority of postcodes on the NSPD, the grid reference will reflect the mean of matched addresses on ADDRESS-POINT with the same postcode, but allocated to the nearest address (property) of this mean.

Figure 4 gives a visualisation of how ward population estimates can be derived from LSOA population estimates using the PBF method and GIS. The LSOA of Berwick-Upon-Tweed 003C in Northumberland covers the area of two wards, Ford and Lowick. Within the LSOA there are 115 postcodes with associated population estimates for mid-2005, the average population per postcode is 17. Ford ward has 66 postcodes with population estimates, whilst Lowick ward has 49 postcodes with population estimates. These postcode population estimates can be aggregated to their respective wards, this gives mid-2005 population estimates of 1,011 for Ford and 962 for Lowick.

Case Studies highlighting application of PBF
In October 2006 ONS had communication from South Tyneside Council who enquired about the feasibility of ONS producing population estimates by age and sex for their 71 defined Neighbourhood areas covering most of South Tyneside, to assist in the validation of their own population estimates for these neighbourhoods. At the time the PBF
method was still at a development stage and, as South Tyneside Council was able to offer digitised GIS boundaries for their 71 neighbourhoods, this was identified as being a good test of the PBF method, requiring GIS to identify the postcodes and associated estimates to be aggregated for each of the 71 neighbourhoods.

For most of these 71 neighbourhoods there was a broad consistency between the respective population estimates, however there are a few areas where there are differences of >10 per cent (Figure 5). Most of these differences can be explained by local circumstances.

In July 2007 ONS had contact from The Centre for Environment, Fisheries & Aquaculture Science (Cefas) which is an agency of the UK Government’s Department for Environment, Food and Rural Affairs (Defra). Cefas undertakes scientific research, advisory and consultancy work. Cefas approached ONS to enquire about the availability of population estimates for river catchment areas covering England and Wales to assess the impact of human-derived sewage on fisheries in England and Wales, undertaken on behalf of the Food Standards Agency.

No postcode to river catchment area lookup file existed, so GIS was used to overlay the river catchment boundaries onto our postcode population estimates based on their grid reference. In this way the postcodes within each river catchment area could be identified, and the postcode population estimates within them aggregated. In this manner population estimates for all 903 river catchment areas were produced.

Other examples where PBF has been used to produce bespoke population estimates to meet specific requests include the production of updated population estimates for 2001 Census-defined urban areas in England and Wales, and population estimates for parishes and postcode sectors.

ONS is willing to undertake work to produce bespoke population estimates where these areas can be readily defined, either in terms of existing geographies, a postcode lookup file or customer supplied digitised boundary file. Potential customers are reminded that there are some limitations with the estimates produced using the PBF method and that any work undertaken may be chargeable.

Conclusion

The use of a PBF methodology as developed by ONS can go a long way to meet previously unmet demand for population estimates, such as for wards and parishes. The finer the level of geographical detail with any such estimates, the greater the uncertainty of accuracy – as is the case with the standard outputs for population estimates for local authorities, MSOAs and LSOAs.

In addition to producing population estimates for ‘non-standard’ small areas, the method has also been used to produce population estimates for larger areas, such as National Parks and Parliamentary Constituencies, published as experimental statistics. In fact population estimates can be produced for any geography where a postcode to geography lookup or a suitable GIS digitised boundary exists. Greater confidence can be associated with large area estimates derived by PBF because of the greater likelihood that the estimates will contain whole local authorities, MSOAs or LSOAs and, in such cases, only relatively small slither areas may actually have population estimates generated solely from the PBF methodology.

As PBF derived population estimates are a by-product of LSOA population estimates, a population base does not have to be generated for each geography for which estimates are required: the only inputs required are a year-specific patient register file, the associated LSOA and special population counts and, optionally, the NSPD. To generate estimates from PBF for different geographies is not resource (staff) intensive, whereas alternative methods for producing such estimates may be.

From the evaluation undertaken, no serious limitations with the PBF methodology have been identified, but it is recognised that the accuracy of any such estimates is largely dependent upon the accuracy of not only the patient registers, but also the NSPD, and the LSOA and local authority population estimates.

Whilst there are particular limitations with the patient register counts, with appropriate checking and adjustments, the quality of any estimates derived from PBF in this way can be improved. Such checks and adjustments are currently done with the Ratio Change method for producing the LSOA and MSOA estimates.

References

3 The National Statistics Postcode Directory (NSPD) is produced by ONS and lists all unit postcodes in the United Kingdom and assigns them to a range of administrative, health, electoral and other geographies. Postcode grid references are provided and counts of the number of addresses, delivery points and small businesses in each postcode are also available. The product contains both live and terminated postcodes.
4 The postcode sector comprises the first half of a postcode (the outcode) and the first digit of the second half of the postcode (the incode) eg EC1A 1 and WR10 3. There are around 9,000 postcode sectors wholly within England and Wales with an average population of around 5,900, with a very small number in excess of 20,000.
6 The Ordnance Survey product ADDRESS-POINT is a dataset that uniquely defines and locates residential, business and public postal addresses in Great Britain. It is created by matching information from Ordnance Survey digital map databases with more than 27 million addresses recorded by the Royal Mail.
8 The label ‘Experimental Statistics’ basically refers to statistics which are being consciously ‘groomed’ for National Statistics status when, and if, circumstances permit. The label from 1 April 2008 now has no significance in the context of the Statistics and Registration Service Act 2007 or the new UK Statistics Authority’s ‘assessment and designation’ function (unless the UK Statistics Authority decides otherwise).