Office for National Statistics

Better Statistics for English Civil Parishes:

Investigating alternatives to ‘best-fitting’

January 2016

Edition No.: **2016 v 1.0**

Editor: **Bruce Mitchell**

Office for National Statistics

2016 v.1.0

Document Control

Version	Created by	Date last edited	Status	Changes made
1.0	Bruce Mitchell	28/01/2016	First publication version	

About Us

Office for National Statistics
The Office for National Statistics (ONS) is the executive office of the UK Statistics Authority, a non-ministerial department which reports directly to Parliament. ONS is the UK government's single largest statistical producer. It compiles information about the UK's society and economy, and provides the evidence-base for policy and decision-making, the allocation of resources, and public accountability. The Director-General of ONS reports directly to the National Statistician who is the Authority's Chief Executive and the Head of the Government Statistical Service.

Government Statistical Service
The Government Statistical Service (GSS) is a network of professional statisticians and their staff operating both within the ONS and across more than 30 other government departments and agencies.

Contacts

This publication
For information about the content of this publication, contact ONS Geography Customer Services
Tel: 01329 444971 Email: *ons.geography@ons.gov.uk*

EXECUTIVE SUMMARY

The 'best-fit' methodology used to produce statistical estimates left over 1 in 10 civil parishes in England without data in 2011. ONS Geography therefore undertook research to identify a methodology that might improve this outcome.

In order for a new methodology to be judged as successful it had to meet a number of criteria. The method would have to:

a. work within existing disclosure control criteria (e.g. a minimum threshold of 40 households/100 persons);
b. use the definitions of 2013 parish boundaries and reduce the number of these parishes without an estimate;
c. not overly distort the data by providing an acceptable balance between an accurate reflection of the census data and statistical disclosure control;
d. minimise spatial or other bias;
e. be realistically implementable in a technological and methodological context;
f. be transparent to the user community; and
g. be applicable across the whole of England.

Thirty-six options were trialled. Results showed that it was possible to reduce the number of parishes for which census estimates were not available from 1,137 to 715. This outcome was only achieved, however, at the unacceptable expense of seriously distorting the data.

Research proved that an irresolvable trade-off exists between a model's ability to produce precise estimates and its ability to remain close enough to the real spatial distribution of the census data.

We conclude that, while we operate under one or more of the present constraints (chiefly the disclosure threshold limits and the present definition of civil parish boundaries), the problem of 1 in 10 parishes not obtaining census estimates cannot be solved by applying an alternative implementable method to the prevailing best-fitting model.

INTRODUCTION

The Office for National Statistics (ONS) has a legal duty[1], to ensure that all official published statistics protect the identity of individuals or households. No statistics may be released that could reveal any personally identifiable information. To comply with this obligation, the 'best-fit'[2] methodology was developed and written into the official Government Statistical Service (GSS) Geography Policy[3].

Best-fitting is very effective for producing consistent estimates for nearly all geographies. It accommodates high levels of geographic change within administrative geographies, while reducing the risk of statistical disclosure. The best-fit method is built upon the use of the census Output Area population-weighted centroid. Output Areas contain a minimum of 40 Census Households (HH) and 100 Usual Residents (UR) to prevent statistical disclosure. The best-fit process is described in full in 'An Overview of Best-Fitting'[4].

There are two geographies for which the best-fit approach works less well, namely national parks and civil parishes. The national park best-fit problem was resolved by producing actual counts for those geographic areas. The large geographic size of national parks, the disclosure threshold values and the lack of relationship between national parks and other geographies meant this was acceptable. Conversely, for civil parishes, their smaller sizes and the small slivers generated by their previous alignment with OAs prevented an actual count approach being used. Therefore, an alternative solution was required.

[1] Laid down in Section 39 of the Statistics and Registration Service Act 2007
[2] Underlined terms are explained in the glossary at the end of this document.
[3] GSS Geography Policy V.2
[4] https://geoportal.statistics.gov.uk/Docs/An_overview_of_best-fitting.zip

The civil parish is the lowest level of administrative geography in England. The equivalent in Wales is the Community, but best-fit can be successfully applied here[5]. The parish best-fit problem is restricted to England, and so this work applies only to civil parishes in England.

One in 10 of the 10,500 civil parishes in England (2013) does not contain an OA-PWC. Applying the best-fit process to these parishes produced census estimates of zero and yet they actually contained a combined total of about 120,000 residents, representing about 1.4% of the total parish population of England. This issue was aggravated because some parish-level publications that had been available for previous censuses were withdrawn for 2011[6].

To address this issue, ONS Geography has assessed a range of alternative methods for estimating census statistics for parishes. This report provides an overview of the research and recommends a way forward for the 2021 Census.

The origin of the civil parish issue

ONS faces one particular long-standing challenge. The UK has a large number of very different, cross-cutting and persistently changing geographies. Boundary change in the UK is endemic and this has long hampered the comparison of statistics over time. The issue with parishes has arisen as an unintended consequence of the application of best-fit to manage these levels of boundary change and multiple types of geography used for statistical outputs.

A hierarchy of statistical building block geographies, each with a minimum population threshold, was developed for publishing the results of the 2001 Census for England and Wales. These were Output Areas (OA), Lower Layer Super Output Area (LSOA) and Middle Layer Super Output Area (MSOA). OAs are the smallest geography for which census data are released. Information about

[5] There are only two Welsh Communities without an OA-PWC and these are unpopulated and entirely industrial areas of Port Talbot.

[6] Parish Profiles (tables PP01 –PP04) were designed to provide condensed summary results on People; Households; Work and qualifications; and Accommodation and tenure. The profiles were produced for parishes with at least 50 people and 20 households

the OAs, the best-fit process, minimum population thresholds and other disclosure issues has been documented elsewhere[7].

The 'BEST-FITTING' policy and civil parishes

The GSS Geography Policy defines the way that statistics for any 'higher' geography larger than OA should be generated. 'Exact estimates' are calculated for the OAs derived directly from the census households they contain. These estimates are applied to the OA's population-weighted centroid (OA-PWC). Statistics for the individual instances of any higher geography are summed from the respective OA-PWCs they contain. Best-fitting is used regardless of whether the higher (target) geography is made up of exact aggregations of OAs (e.g. LADs) or crosscuts them (wards or postcode sectors).

The best-fit procedure is reliable and simple. It eliminates the risks associated with overlaying incompatible geographies. Statistics calculated by best-fit are guaranteed to be non-disclosive, and remain consistent even if the geography changes, but this method is problematic for civil parishes. Parishes cover 91% of the area of England, but house less than half of the population – the majority of the population of England live in urban unparished areas. In terms of area, parishes may be as small as 400m^2 or as large as 250 km^2 and, in these terms, they are similar to OAs. However, there is a crucial difference: the population thresholds embedded into the OA design means that OAs are small where the population is dense and large where it is thinly spread. In contrast, there is no such correlation between population and area for parishes: in consequence, despite there being over 170,000 OA-PWCs and 10,500 parishes across England, 1 in 10 parishes do not contain any OA-PWCs.

The parishes in this group have a combined census population of nearly 120,000 residents, but they are unevenly distributed. While some of these parishes can have up to 140 households, some are completely unpopulated, yet the entire group is treated similarly by the best-fit method, producing population estimates of zero for all. It is important to stress that area has little to do with it: the largest without an OA-PWC covers 155 km^2. The whole group covers 7,000 km^2. A

[7] https://geoportal.statistics.gov.uk/Docs/An_overview_of_best-fitting.zip

parish may be without an OA-PWC because two or more parishes are contained within a single OA: the OA-PWC can only fall in one of the parishes. The largest wholly unpopulated parish is 13 km^2.

The mix of old and new parishes, densely and thinly populated, small and large, has made the challenge of developing a single methodology that can be applied successfully across the entire parish geography that much harder.

THE PROJECT

Objective

To identify a methodology that would maximise the number of parishes for which publishable counts of Census Households (HH) and Usual Residents (UR) could be produced, while minimising error caused by smoothing the data, and avoiding spatial or other bias.

Assumptions

Any solution would have to:

- work within existing disclosure control criteria (e.g. a minimum threshold of 40 households/100 persons);
- use the definitions of 2013 parish boundaries and reduce the number of these parishes without an estimate;
- not overly distort the data by providing an acceptable balance between an accurate reflection of the census data and statistical disclosure control;
- minimise spatial or other bias;
- be realistically implementable in a technological and methodological context;
- be transparent to the user community; and
- be applicable across the whole of England.

Base data

- The 8.9 million HH in the Census Microdata located within English parishes.

- The x-y coordinates of the Census HH, the HH count and the UR count in each household.

Two different interpretations of a parish boundary:

- The 'actual' civil parishes – the actual geographical boundaries of civil parishes as published by ONS on its Open Geography portal.
- 'Pixellated' civil parishes – approximations to actual civil parishes made up of one or more grid cells.

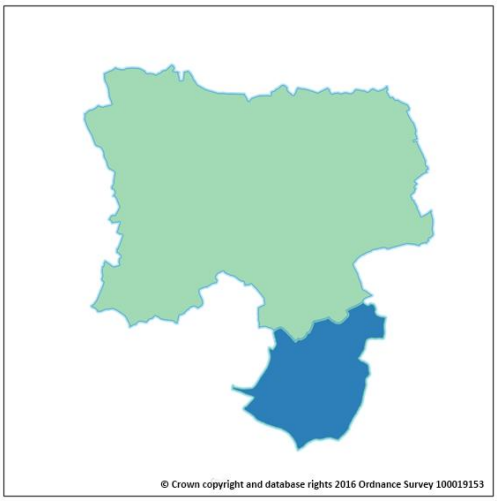

Figure 1: Sample 'actual' parish boundaries

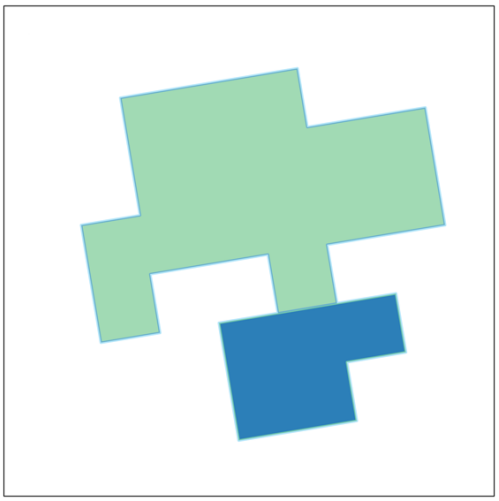

Figure 2: Sample 'pixellated' parish boundaries

Nine statistical models

1: Reference Model

A baseline 'reference model' was built up from the 8.9 million HH and 20.4 million UR resident in English parishes. It therefore represents the finest possible granularity of the census data and therefore the closest possible approach to reality. The reference model supplied benchmarks against which alternative models – including the currently approved best-fit model – could be measured.

The following data were calculated:

- the number of Census HH contained within the individual civil parishes;
- the number and proportion of civil parishes with 40 or more HH;
- the number and proportion of 'sub-threshold' civil parishes with 39 or fewer HH;

2. Eight Comparative Models (A – H)

The same data were then calculated for eight other, 'comparative models'. The results from the comparative models were compared against those from the reference model. A model was initially deemed to perform better than the reference model (termed 'model gain') if it produced a lower number of parishes without estimates.

For the first four models (A-D), the Household-level census data points were aggregated to certain geometric locations relating to administrative or postal geographies. The remaining four methodologies (E-H) were developed from a grid of roughly 120,000 1km by 1km grid cells covering English parishes.

A. Postcode headcounts (counts of HH and UR within postcode units);
B. 'Best-fit' (counts of HH and UR within OAs, allocated to their OA-PWC);
C. Counts of HH and UR within parish, allocated to parish geometric centroids;
D. Counts of HH and UR within parish, allocated to parish population-weighted centroids;

E. Counts of HH and UR within grid cell, allocated to grid cell geometric centroids;
F. Counts of HH and UR within grid cell, allocated to grid cell population-weighted centroids;
G. GRID SPATIAL: HHs within grid cell, allocated to NEAREST of the geometric centroid and the four corners;
H. GRID ARITHMETIC: one-quarter of HH within grid cell allocated to each of the cell corners.

The reference model and comparative models are then run a second time against the alternative 'pixellated' parish boundaries. The results from the comparative models are compared to those from the reference model for the same geography. This gives us eighteen separate data streams.

RESULTS

Actual Parishes

The best results (see Table 1) are produced by the GRID SPATIAL and GRID ARITHMETIC models. However, their gains (38 and 222 parishes respectively) are weakened because these models lose 120,000 and 354,000 households respectively. These are households which are allocated to grid-

cell centroids or corner points located beyond the actual parish boundaries: such points fall either in unparished areas, over the Welsh or Scottish national borders, or in the sea.

Pixellated Parishes

The GRID SPATIAL and GRID ARITHMETIC models again outperforms the reference model. As they were designed with pixellated parishes in mind, they also produce better results when applied to the pixellated geography than when applied to actual parish boundaries.

On the pixellated geography, 76 households are lost with the GRID SPATIAL model, but this is a simple rounding difference. The GRID ARITHMETIC model loses no data at all. But even with a complete dataset, the gains are modest, solving at best 403 of 1,137 problem parishes (see Table 2, 'Model Gain').

There were 21 civil parishes for which no pixellated parish was generated. This is because when overlaid with the 1km^2 grid, these parishes contained neither a grid-cell population-weighted centroid nor a grid-cell geometric centroid, both essential for the creation of pixellated parishes. These parishes have a combined population of 983 HH or 2,329 UR, of which over 80% are concentrated in three. This is an insignificant problem capable of manual adjustment.

The cost of model gain: model error

The solution that resolved most parishes was provided by the combination of the GRID ARITHMETIC approach and pixellated parish boundaries. This was because the method smoothed the data from more populous to less populous areas and increases the household count in sparsely populated parishes. But the gains from smoothing the data come at a cost: the original data are shuffled into a new pattern, which can be significantly different from the original.

If, for each parish, we compare the counts derived from the Comparison Model against those from the Reference Model, we obtain the difference, or 'model error'. This is the difference between what a model calculates the population of a parish to be and what the population actually is expressed as an average of errors across the entire geography.. This should be as close to zero as possible.

Table 1: Results for Actual Parish Boundaries

Model comparisons Geography & methodology		Lost HH[1]	Resolved parishes[2]	Unresolved parishes[3]	Model GAIN[4] (count)	Model GAIN[4] (%)	Error[5]
Actual Civil Parishes (10,500)							
A	Postcode headcounts	-239	9,347	1,153	-35	-0.33%	2.66%
B	OA-PWC ("best-fit")	1,822	9,364	1,136	-18	-0.17%	15.82%
C	PARISH-GEOC	0	9,326	1,174	-56	-0.53%	3.36%
D	PARISH-PWC	0	9,382	1,118	0	0.00%	0.00%
E	GRIDCELL-GEOC	175,228	9,351	1,149	-31	-0.30%	37.20%
F	GRIDCELL-PWC	-103	9,273	1,227	-109	-1.04%	9.11%
G	GRID SPATIAL	120,437	9,420	1,080	38	0.36%	22.11%
H	GRID ARITHMETIC	353,852	9,604	896	222	2.11%	69.98%

Notes:

1. **Lost HH:** Census HH moved to outside of non-parish boundaries (real - model)
2. **Resolved parishes:** parishes which after application of the model have 40 or more HH
3. **Unresolved parishes:** parishes which after application of the model have 39 or fewer HH
4. **Model Gain:** Increase in number of resolved parishes after application of the model (real - model)
5. **Error:** Mean absolute proportional difference between Comparison Model and Reference Model estimates.

Table 2 : Results for Pixellated Parish Boundaries

Model comparisons Geography & methodology		Lost HH[1]	Resolved parishes[2]	Unresolved parishes[3]	Model GAIN[4] (count)	Model GAIN[4] (%)	Error[5]
Pixellated Civil Parishes (10,479)							
A	Postcode headcounts	952	9,371	1,108	10	0.07%	18.91%
B	OA-PWC ("best-fit")	243	9,209	1,270	-152	-1.47%	21.82%
C	PARISH-GEOC	201	9,177	1,302	-184	-1.78%	20.74%
D	PARISH-PWC	0	9,133	1,346	-228	-2.20%	11.42%
E	GRIDCELL-GEOC	175,210	9,255	1,224	-106	-1.03%	1.11%
F	GRIDCELL-PWC	-103	9,273	1,206	-88	-0.86%	0.00%
G	GRID SPATIAL	76	9,416	1,063	55	0.50%	16.76%
H	GRID ARITHMETIC	0	9,764	715	403	3.82%	67.55%

Notes:

1. **Lost HH:** Census HH moved to outside of non-parish boundaries (real - model)
2. **Resolved parishes:** parishes which after application of the model have 40 or more HH
3. **Unresolved parishes:** parishes which after application of the model have 39 or fewer HH
4. **Model Gain:** Increase in number of resolved parishes after application of the model (real - model)
5. **Error:** Mean absolute proportional difference between Comparison Model and Reference Model estimates.

It should be noted that the currently approved best-fit method, using actual parish boundaries, produces a model error of 16% with respect to the reference model. It is this high because the 8.9 million households in the reference model are drawn together and generalised in the best-fit model to around 68,000 OA-PWCs. The GRID ARITHMETIC method might resolve more parishes, but it does so with an error of 68% for pixellated parishes which makes the method unusable. It is clear that the model achieves its success at the expense of severely damaging the data.

SUMMARY

The methodology currently approved for calculating statistics for administrative areas fails to provide statistics for 10% of civil parishes in England: but the parish is a popular geography, and it commonly features in requests for local data. Councils involved with neighbourhood planning for below-threshold parishes can find the lack of local statistics frustrating.

In order to address this issue for the 2021 Census, ONS Geography has researched a range of methodologies. These were tested to see whether – given the disclosure control constraints imposed by the GSS Geography Policy – any might provide a better balance between precision and protection against disclosure. The best model would be the one that maximised the number of parishes for which above-threshold counts of HH and UR could be produced, while minimising model error and avoiding spatial or other bias. Alongside the actual geographic parish boundaries, an alternative 'pixellated' rendering of the civil parish boundaries, made up of collections of grid cells, was developed as an ideal match for the GRID ARITHMETIC and GRID SPATIAL methods.

Model results were somewhat improved when applied to the parish council geography because of the smaller number of more populous areas. However, the legal provision to group sparsely-populated parishes had not been universally adopted across England. Furthermore, there is no central register of parish councils and the civil parishes they represent, so this is not, at present, a practical option.

The research has shown that the parish conundrum is unlikely to be solved with the existing disclosure control measures in place. While some of the Comparative Models resolve hundreds of the problem parishes, those that are most effective do so at the expense of introducing an unacceptable degree of error. The few that do produce more gain than error produce modest gains, while incurring the overheads of generating pixellated parish boundaries.

One potential solution that requires investigation is the aggregation of parishes done on the basis of administrative function. In many parts of the country, individual low-population civil parishes are administered by grouped parish councils. This permitted the ONS to analyses on parish councils, and identify that data grouped at the parish council level has the potential to sharply reduce the number of parishes for which ONS currently has to publish estimates of zero. Grouped parish councils do not currently cover the whole of England and so whether the use of grouped parish councils is a solution for producing parish statistics is not yet fully understood.

GLOSSARY

Actual parishes: the published geographic boundaries of civil parishes in England.

Best-fitting: Calculation of estimates for a target geography based upon the Output Area population-weighted centroids (PWC) that each instance of the target geography contains. Download 'An Overview of Best-Fitting' from the ONS website.

Census microdata: small samples of data for whole households and individuals, which include some associated census characteristics but no information that could identify a household or individual. See definition on ONS website.

Civil parish: a territorial designation in England, normally run by its own parish council.

Exact-fitting: Calculation of estimates for a target geography based directly on the Census Households (HH) that each instance of the target geography contains.

Geometric centroid: a point placed at the centre of gravity of a polygon.

GRID ARITHMETIC method: a grid-based method, whereby the census data located within a grid cell is summed, and a quarter of the total is allocated to each corner of the grid cell.

GRID SPATIAL method: a second grid-based method, whereby each individual census data point located within a grid cell is allocated either to one of the four corner points or to the geometric centroid, dependent upon which one of those five points is closest.

Non-parish boundary: A boundary between an actual or pixellated parish and unparished areas, the coastline or local authority district /national boundaries

Open Geography portal: an ONS Internet resource that allows users to discover, view and download geographical reference data to support National Statistics.

Parish council: An elected corporate body with a legal personality which represents a civil parish. It is the lowest, or first, tier of local government.

Population-weighted centroid (PWC): a point placed at the centre of gravity of a polygon, based on its population.

Pixellated parish: synthetic equivalents to geographical parish boundaries, made up of complete grid cells.

Reference model: A statistical model designed to produce baseline results against which results from 'comparative models' may be compared.