Minimising Fieldwork Costs In National Random Door-To-Door Surveys
Presented at AAPOR, Chicago, 14 May 2010

Andrew Zelin, Director - Sampling & RMC
Ipsos MORI, London
andrew.zelin@ipsos.com

Abstract

Door-to-door interviewing is still very much a mainstay of sampling in the UK and developed countries for studies where a high degree of robustness is required. This paper describes a fast and practical method for batching-up randomly-selected sampling points into "clusters" of three; such that the points in each cluster are closely-positioned. The sample remains technically unclustered, and therefore benefits from the statistical robustness and power of pure random samples, yet to fieldwork departments, these can be seen as "clustered" and they can thus benefit from cost savings in terms of reduced costs of travelling within a cluster. The work centers around a large-scale national door-to-door Media survey and the programming is carried out in SAS. The aim is to position as high a proportion of sampling points within a survey region as possible into "triples" (ie clusters of three) such that no two points within a triple exceed a certain distance from each other. Described are the methods for allocating points to these triples on nearest-neighbour bases, whilst making suitable adjustments to the distance measures in order to avoid being left with the challenging task of trying to group the most peripheral points. Further methods are employed to optimise the chance of positioning the residual singles and pairs into triples. A further advantage of this method is its ability to be carried out based on just the northings and eastings of the original sample point locations, obviating the needs for slow and expensive specialist software to be purchased / incorporated. The method described in this paper is flexible and can easily be bolted on to in-house sampling programs.

Key Words: Centroids, Sampling Area (SA), Primary Sampling Unit (PSUs), Triple, Cluster Sampling.

1. Background and Introduction

Random Sampling techniques are considered by many to be superior in terms of accuracy to producing Quota Samples (Marsh & Scarborough 1990)[1], [2], Smith (2008)[3]. However, it is very rare in practice for a nationwide door-to-door or face-to-face (F2F) survey to be carried out as a simple random sample. Usually, an element of clustering is applied. This has the effect of making the interviewing task practicable by reducing travel distances, although it has the undesirable effect of reducing the precision of the estimates that are derived from the research (Zelin 2005)[4].

In this paper, we describe a method of sampling and collecting-up of sample points which was required for a national (UK) survey of TV viewing. The requirement was for a pure random sample, in terms of the fact that the actual sampling of primary sampling units (PSUs) were unclustered. However, in order to minimise costs and optimise efficiency, the fieldwork department required the PSUs to be delivered to them in groups of three, such that each member of a "triple" was closely located to each other – ie that no two pairs

1 Marsh, C. and Scarbrough, E. (1990). Testing nine hypotheses about quota sampling. JMRS, vol. 32 no.4.

2 Report of the First Cathie Marsh Memorial Seminar: November 1994. Quota vs Probability Sampling (SCPR)

3 Patten Smith 2008. Is Random Probability Sampling really much better than quota sampling?

4 Cluster sampling: a false economy? Andrew Zelin and Roger Stubbs, International Journal of Market Research, Vol. 47, No. 5, 2005, pp.501-522

within a "triple" were more than 10 miles apart (straight line distance). The fieldwork department could then treat each group of three PSUs as a sampling point from an operational viewpoints and work each of the three PSUs simultaneously, allocating all of it to the same interviewer. This effectively meant that from a fieldwork point of view, the sample was "Clustered", but statistically, they were "Un-clustered". This would deliver the benefit of allowing the survey to run and the interviewing to take place at reduced cost, whilst ensuring that no precision is lost through clustering-sourced design effects.

The core sampling process involved selecting PSUs in a random, systematic and stratified way from all possible PSUs in UK. These PSUs were based on the 2001 Census Output Areas (OA); of which there are about 180,000 in the UK universe and each typically contained 125 households. In England and Wales, these were the actual Census OAs, whilst in Scotland, they were pre-created aggregates of 2-5 Census OAs. The 2010 wave of the survey required 11,336 PSUs to be drawn. This was stratified by the project-specific "Sampling Area" (SA) initially, of which there are 48 in UK and the number needed per SA was pre-determined, ranging from 12 to 1,050. Within that, it was then stratified by other factors, such as whether or not the PSU covered an area where Cable TV was available, the penetration of ethnic minority groups and the CACI ACORN category. In each PSU, four addresses were selected for interview, thus the overall aim of the clustering exercise was to back up the PSUs into assignments of 12 closely-positioned interviews.

Having selected the 11,336 PSUs in this way, the task of tripling them up took place, and it is this tripling task, which is the main topic described in this paper.

2. The Core Matching Process

This effectively takes place on the basis of "minimum distance" between the centroids, akin to the Agglomerative Hierarchical method of clustering of data used in statistical software (eg SPSS) on carrying out Cluster Analysis (Wuensch 2007)[5], except for the fact that we have limited the cluster sizes to three PSUs and we start with data on only two dimensions (ie North and East grid locations).

A program in SAS v9.1 was set up to carry out an overall matching routine for each SA in turn. The reason for the latter was two-fold. Firstly, from a practical point of view, the time taken to estimate every combination of distances between each 11,336 PSUs would render the task highly computationally intensive and would take an amount of time which is not compliant with the operational needs of this process. Secondly, it was not possible to have points from more than one SA within the same cluster as the results of the survey need to be reported at SA level.
A SAS program ran the following routine to carry out the initial tripling of the points:

a) Start by creating a file of original sampling points for each SA. These points will each initially be composed of a single PSU. This is referred to as "File 1" and will contain

5 Karl L. Wuensch, Cluster Analysis with SPSS, 2007 http://core.ecu.edu/psyc/wuenschk/SPSS/ClusterAnalysis-SPSS.doc

the northings and eastings for each point, along with the number of PSUs they contain (the latter will always be 1 at the start of the process);

b) Using the current list of points ("File 1"), create a file ("File 2") that contains a records for each and every possible pair of points within an SA (eg SA #15, which contained 72 points – and hence 72 PSUs originally - would contain [72*71/2 = 2,556 records]);

c) Calculate the straight line (Euclidean) distance between the pair members for each point. This was calculated as:

$$d_{12} = \sqrt{\left[(n_1 - n_2)^2 + (e_1 - e_2)^2\right]}$$

(where n1 and n2 are the northing co-ordinates of the two points and e1 and e2 are the eastings).

d) Take the pair with the smallest distance (eg within SA #15, this will be PSUs #35 and #58), checking that this distance does not exceed 10 miles and that the total number of PSUs between the 2 points does not exceed 3 (in the first round of this program the total number of PSUs across the points in the pair will only ever be 2);

e) If the conditions in d) are both satisfied then remove these from File 1 and create a single replacement point which is located mid-way between its two component original points. The fact that this new point is composed of 2 PSUs (where all the others would still, at this stage, be composed only of 1), is to be flagged.

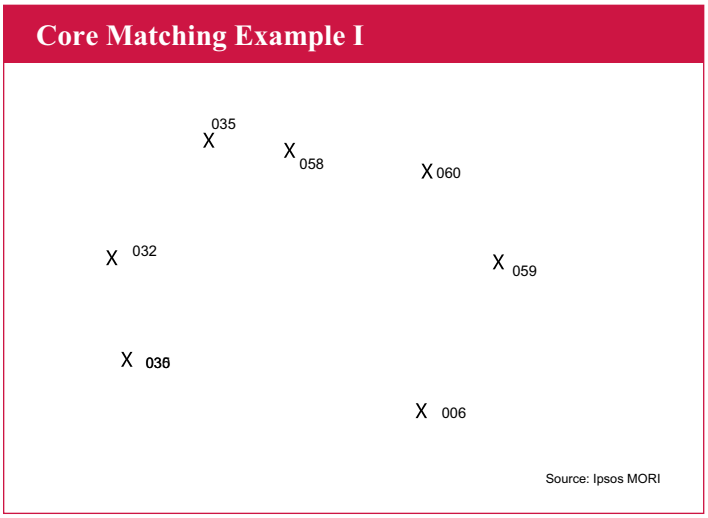
f) Transfer and put aside any points which contain 3 PSUs into a separate file of completed "triples". These points and PSUs will play no further part in this routine.

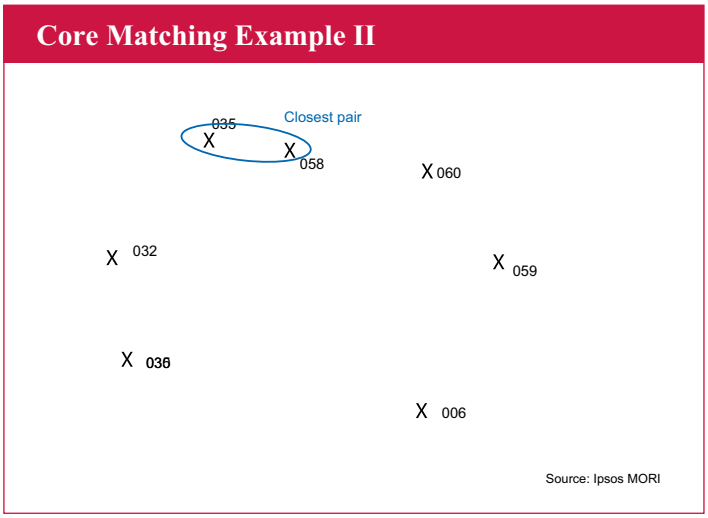
g) Return to b) to re-create Files 1 and 2 using the newly-combined points of single and paired PSUs;

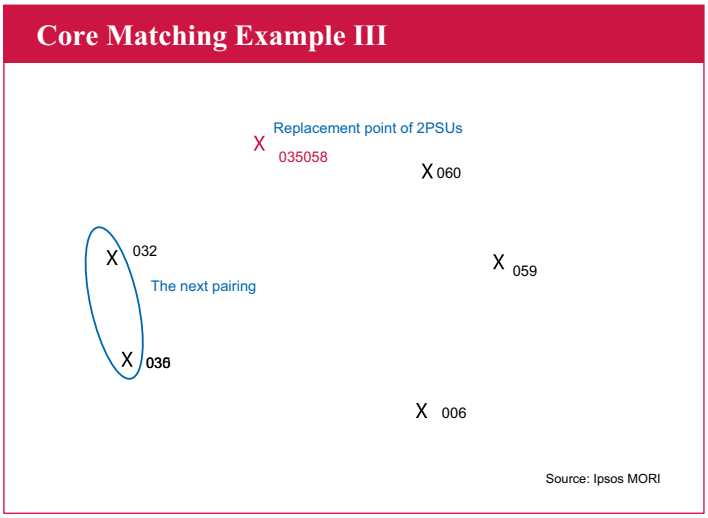
h) Repeat b) to g) until no further combining of points that satisfies the two conditions set out in d) can be carried out;

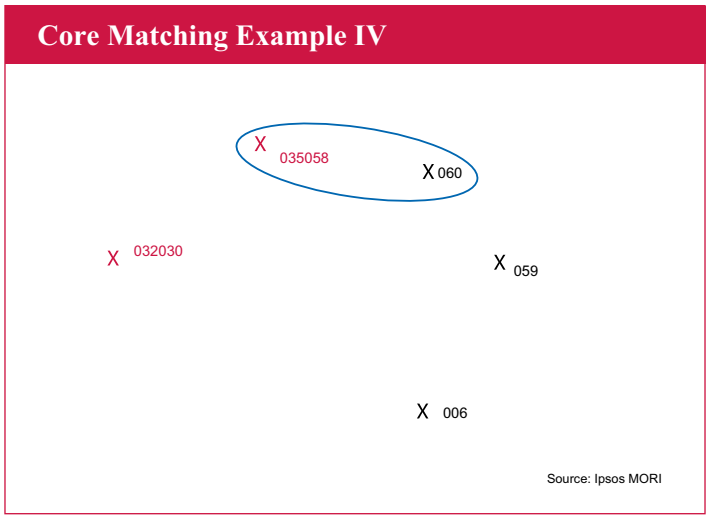
i) End the process with a list of completed triple PSUs ("File 3") and a residual list of pairs and single PSUs ("File 4"). Verify that the total number of PSUs in Files 3 and 4 equate to the original number of PSUs in the SA, and that there is no duplication.

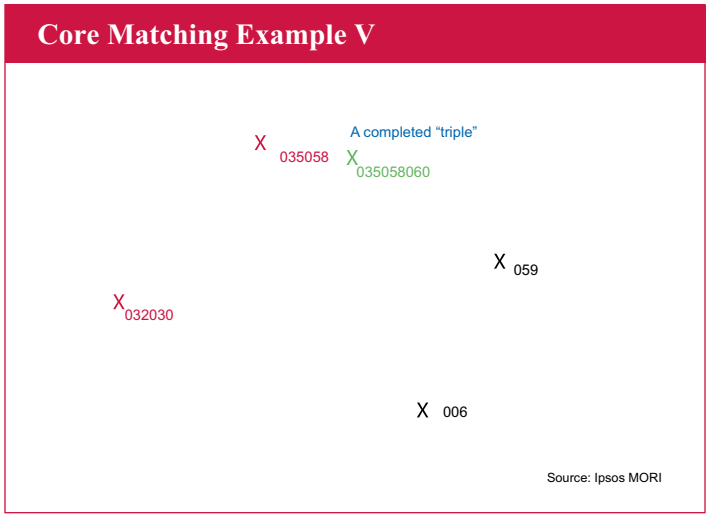
The following sequence shows a worked example:











3. Definition of The Distance Function

The above section describes the core matching process. Having run this for a number of Sampling Areas (SAs), it was being found that large numbers of PSUs were not able to be adequately placed into a triple and that a certain number of singles and pairs were remaining. This was not an ideal situation, as the overall fieldwork costs ran approximately (although this is covered in more detail in Section 6) proportional to the total number of clusters, irrespective of their size; a "single" consisting of four interviews costing virtually as much to cover as a "triple" consisting of 12.

At the random sampling stage, prior to the tripling, the selecting of an over-sample of PSUs in each SA was considered, as followed by interviewing at only the PSUs which were successfully clustered into triples and discarding the remaining points. However, this was

soon rejected as it would lead to overall biasing of the sample in favour of the more easily "cluster-able" PSUs, such as those in the more urban parts of an SA and / or those not near to the borders with other SAs.

Focus then moved towards looking at an alternative way of creating the clusters, sampling and selecting only the exact number of PSUs where interviewing would take place. The core of the matching program described in Step 2 was the calculation of the Euclidean distance.

$$d_{12} = \sqrt{\left[\left(n_1 - n_2\right)^2 + \left(e_1 - e_2\right)^2\right]} \quad \ldots(3.1)$$

From previous work carried out and from intuition, those PSUs which were furthest from the population centre (ie the "centriod") of the SA were the most likely to be left as singles or pairs. Therefore, the opportunity to modify the distance formula so as to give priority to the more peripheral points was explored. This gave rise to the "distance function" (f), which took into account the distance between the SA centroid and the position of the newly-combined point, as well as the distance between the PSUs prior to combining them – ie:

$$f_{12} = d_{12} \Big/ \left[\left(n_c - n_{12}\right)^2 + \left(e_c - e_{12}\right)^2\right]^k \quad \ldots(3.2)$$

Where:

- fij = distance function between points i and j;
- dij = distance between points i and j;
- ni and ei = northing and easting of point i;
- nc and ec = northing and easting of the SA population centroid;
- nij and eij = northing and easting of the mean position of points i and j;
- k = calibration indicator (where k>=0).

Essentially, the distance function equates to the Euclidean distance between the two points, scaled-down by the extent to which both points are away from the centre of the SA. If "pair of points" A and has its members as closely-positioned to each other as does "pair of points" B, but pair A is closer to the SA centre, then pair B will have the smaller distance function and will be paired-up first. However, the relative importance of the distance from the SA centroid will be determined and controlled by the variable "k".

As an example, we ran the tripling program for one of the (smaller) SAs – ie a mixed rural and urban part of Yorkshire consisting of 72 selected PSUs (Table 1).

Table 1: Relationship between k factor and triangle distance

Performance	Dist tri
75%	4.06
83%	5.22
83%	5.22
83%	5.22
83%	5.22
83%	5.43
83%	5.43
92%	5.70
92%	5.71
92%	5.75
92%	5.56
100%	6.30

Using the crude formula 3.1, which does not accounts for distance from the centre of the SA, 75% of the 72 PSUs are placed successfully into triples (ie the "performance"). This is equivalent to the situation where k=0 in formula 3.2. Even giving k a small value, and hence the distance from the centre of the SA a small influence, increases the performance to 83%. Increasing this influence further – ie to k=1.6 is necessary to see a further step-change in tripling performance, although the cost of this is an increase in the mean straight-line triangle distance[6] between the centres of the 3 contained PSUs. At the extreme end, k=5 will triple all points for this particular SA, but at the expense of making the overall travel distances and costs prohibitively high, offsetting the benefits of having no points left as incomplete pairs.

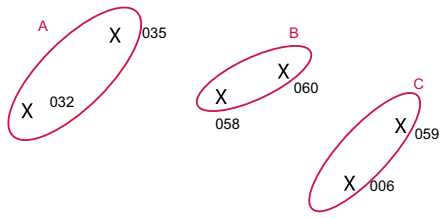
Intuitively, it makes sense not to allow k to be too large. Ideally this should not exceed 0.5 as if it does, the implication would be that the distance of the points from the SA centre would become more "important" than the distance between the individual PSUs that are being paired. From the outset, the latter was to be the key criterion for matching, whilst the former was intended to play the part of acting as little more than an "adjustment factor". It was found that the optimum value of k for most SAs was 0.5. If unlimited time were available to carry out this tripling exercise, a specific optimum k value could be applied for each SA. However, given that this was to be a "productionalised" process whereby a complete sample of points needs to be carried out on each wave, the only practical way forward would involve setting a general optimum value to apply to all SAs.

6 This is the sum of the three Euclidean distance between each pair of points within a triple.

4. Re-Construction of Triples From Pairs And Singles

Being able to allow for distance away from the centre of the SA and adjust it accordingly has been shown as one way to improve the tripling performance, although this benefit of this may be limited where it starts to prioritise pairs that can be a long way apart if this adjustment is set too high.

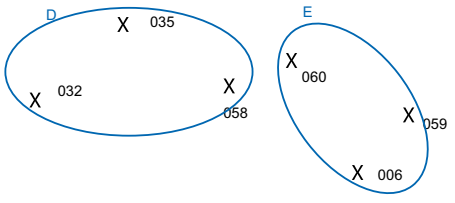
Re-construction Example I



Source: Ipsos MORI

However, there is a way of improving the tripling further. If we consider the example (Reconstruction I, above), where there are 3-reasonably closely-located pairs of PSUs. We might find early on in the core process, that members of these pairs are easily joined as they are so close together. However, later on in the process, when the distance criteria for joining (ignoring for now, the adjustment due to proximity to SA margins) has increased, it would theoretically have been possible to join PSUs #60 to #59 and #6 to successfully form a triple. However, as #60 had already been joined to #58 to form a pair, this tripling would not be possible. By the same token in this example, the potential tripling of #58 to #35 and #32 has been prevented.

Re-construction Example II



Source: Ipsos MORI

Therefore, having taken the process outlined in Section 2 and 3 as far as the initial tripling of the PSUs, could further triples have been created had the formation of pairs never taken place? In practice, it would not be possible to prevent the formation of pairs, as this is a necessary step to forming the triples, given the underlying mechanism of the program. However, this effect could be simulated by "turning the clock back" and preventing the members of the "unsuccessful pairs" (ie pairs which can not be extended further) from joining with each other and instead, making them available to join with other close PSUs; unions which might be more likely to lead to a successful triple formation.

One method by which this could be made possible and was amenable to programming involved allowing the PSU pairs **and** the individual PSU members of these pairs to co-exist; both alongside the un-joined single PSUs. Allowing this to happen meant that, for example, PSU #58 could be available at the same time as and combine with the pair of #32 and #35 to form a triple (as in Reconstruction II). This could only be possible if there was some mechanism of preventing #58 and #60 from combining, or indeed the members of any pair that has been generated in the earlier matching stages. This was achieved by flagging each PSU with a code that related to the identity of the single or pair that had been formed in the earlier stages. If attempts are made in this re-construction stage, to combine two singles, or a single and a pair with the same identity, this combination would not be allowed to proceed. This rejection would operate as a "filter" in a similar way to Stage d) of the program described in Section 2.

Table 2 below shows that for four example SAs, tripling performance is substantially enhanced by re-constructing the singles and the pairs.

Table 2: Effect of re-construction of pairs and singles

			Before reconstruction				After reconstruction			
SA	Location	no_oas	triples	pairs	singles	performance	triples	pairs	singles	performance
15	Yorkshire	72	20	6	0	83%	23	1	1	96%
12.2	North West	276	74	27	0	80%	91	1	1	99%
25	South West	135	40	7	1	89%	42	4	1	93%
32.2	Wales	171	44	19	1	77%	55	2	2	96%

5. Removal of Residual Singles

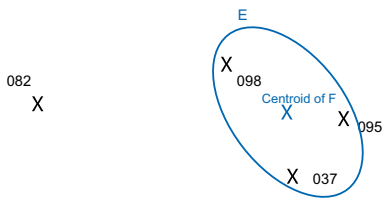
The processes described thus far converted the 11,336 original PSUs into 3,665 successful triples, 116 pairs and 109 singles. This equated to an overall performance of 97.0% (ie in terms of the proportion of PSUs that are part of a triple). This, we felt to be a good outcome, especially as the previous incumbent for this survey had managed a performance of 96%.

There was one final task required in that it was necessary to "cluster" all of the single PSUs. Fieldwork costs for this largely depended on the total number of assignments that required to be covered, almost irrespective of the number of interviews. Therefore, the cost of interviewing at 3 pairs was the same as the cost of interviewing at 3 triples, although the

latter would deliver 50% more actual interviews. Whilst it was accepted that the geography of the UK could not allow for all PSUs to be accommodated into triples, the presence of singles was deemed unacceptable and therefore it was necessary to incorporate a final stage to cover this, even at the expense to reducing the number of triples and hence the performance measure.

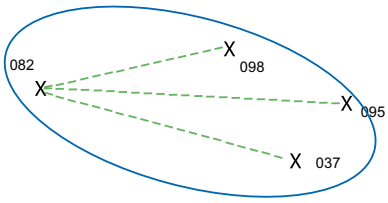
This stage was relatively simple in principle and involved converting a single and a nearby triple into two pairs and is summarised in the following diagrams.

Remove Singles I



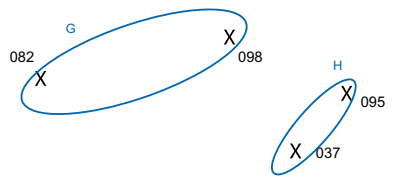
Source: Ipsos MORI

Remove Singles II



Source: Ipsos MORI

Remove Singles III



Source: Ipsos MORI

a) The first stage was to find the most closely-positioned triple for each unmatched single, in terms of the distance between the single PSU and the centroid of the triple, to temporarily form a cluster of 4 PSUs;

b) The distance between the single and each member of the triple would be calculated and the member of the triple which was closest to the single would form and split-off as a pair with the single;

c) The remaining two members of the original triple would form a second pair;

d) A final check was carried out to ensure that the distance between any two pairs of points (within a pair formed in this stage) did not exceed 10 miles. In the small number of instances where this was exceeded (ie 5 – see Table 3 below*), the PSUs were returned to their previous format – ie a single and a triple.

Table 3: Effect of removal of singles

	Before singles removal	After singles removal
No. of singles	109	5*
No. of pairs	116	324
No. of triples	3,665	3,561
Performance	97.0%	94.2%
No. of assignments	3,890	3,890

6. Summary, Conclusions and Further Work

As the quest to provide increasingly robust research at lower margins becomes more challenging, pressure increases on researchers to make the most cost-effective usage of the data collected, which means collecting more data at lower cost. In most practical situations where door-to-door fieldwork is involved, this would be through clustering of the sampling locations in order to minimise travel costs. However, in this paper, we show how this can be done from a fieldwork point of view, whilst still keeping the points statistically unclustered, and therefore not having to introduce a loss of effective sample through clustering-based design effects.

We start by demonstrating an intuitive, but thorough and exhaustive way of aggregating closely-positioned sampling PSUs, by calculating the distances between every possible pair of points at any stage in the process. This is carried out separately for each Sampling Area (SA) in order to allow this to work effectively within the capacity of most business PCs. We then show how the distance function, underlying this can be modified to improve overall performance by giving a little more priority to the PSUs that may be less "accessible" in the terms of this exercise, such as being close to a border with another SA. In this work so far, only the "border" issue has been investigated. It may be of interest, going forward, to explore how this modification might vary between the more rural as opposed to the more urban areas. One could attempt to calculate the mean population density of the output area or postcode sector in which each PSU lies and build that into the distance function. After all, one would expect the rural PSUs to be harder to cluster. Although the 10-mile rule would still need to be adhered to, giving a small amount of additional priority to clustering the rural PSUs may prevent a just-sufficient number of the urban PSUs from being clustered at an early stage, making more of them available later of for grouping with rural PSUs.

We also showed that many of the remaining PSUs that are still in singles or pairs may be successfully tripled as part of a second stage in the exercise. Here, we adapt and flex the "time dimension" of the tripling process to allow pairs and singles from the same point / cluster to all co-exist. The integrity of the process is maintained by disallowing any PSUs from the same original cluster to "re-pair" with themselves. Finally any residual singles were removed by allowing them to pair with one of the members of an existing triple. Although doing this reduces the overall performance measure of the process, in terms of reducing the final number of triples, it does alleviate the unacceptably costly situation of having spare single PSUs. (Admittedly, we did end up with 5 unresolvable single PSUs and this very small number was accepted by our fieldwork department who agreed to relax the 10-mile rule in these situations and pay the interviewers for these areas, a small additional fee for covering these points).

This analysis can be very much considered as "work in progress", which is evolving. We are very soon to generate another full wave of points for the 2010/11 wave of the survey, and will need to triple this new set too. All of the techniques described above would certainly be applied here, although we will be able to exploit the fact that a freshly-drawn sample has become available to further validate them, and also allow further exploration and fine-tuning of the rules and formulae to be carried out.

However, at this stage, we can say that we are satisfied in having the solid foundations of a methodology that enables a sample to be produced which combines the advantages in terms of being clustered to optimise the task for interviewers, whilst ensuring the sample generates robust and precise results, without being subject to clustering. Although at Ipsos MORI, we have been using SAS and employing straight-line distances to develop and run the process described in this paper, the techniques are transferable to most programming languages. With the appropriate IT skills, it may be possible to combine this overall scheme with appropriate topographical mapping software in order to base the clustering on actual road-based travel times and distances. In such situations, the Euclidean distance calculation would be replaced by appropriate functions embedded within the mapping software.

References

Marsh, C. and Scarbrough, E. (1990). Testing nine hypotheses about quota sampling. JMRS, vol. 32 no.4.

Report of the First Cathie Marsh Memorial Seminar: November 1994. Quota vs Probability Sampling (SCPR)

Patten Smith 2008. Is Random Probability Sampling really much better than quota sampling?

Cluster sampling: a false economy? Andrew Zelin and Roger Stubbs, International Journal of Market Research, Vol. 47, No. 5, 2005, pp.501-522

Karl L. Wuensch, Cluster Analysis with SPSS, 2007 http://core.ecu.edu/psyc/wuenschk/SPSS/ClusterAnalysis-SPSS.doc