

Census Coverage Survey

Design and Estimation

CCS WORKING PAPER 1

FOR STEERING COMMITTEE 12.6.97

1) Introduction

This paper describes an outline of the design of a Census Coverage Survey (CCS) with the aim (a) of estimating underenumeration at a County Level (by age and sex); and (b) of allocating this underenumeration to a small area level (say electoral ward) in the form of a few basic counts. A model-based approach is adopted for the design and estimation. It is assumed that the CCS is a postcode based survey and *all* people within a sampled postcode are counted without error. Other papers will discuss what to do when this assumption is not tenable. The sample of postcodes needs to be sufficiently large to estimate the total population by 24 age-sex groups for each county in England and Wales. At the last Census there were 47 counties in England and 8 in Wales. The postcodes can be stratified into groups by a 'Hard to Count' Index and size within each county. In this paper hard to count is defined in terms of increasing deprivation. For 2001 the ONS are working on a categorisation which takes practical details into account as well. A sensible number of hard to count categories is probably five. The problem is to estimate the 24 age-sex totals for each county, all with a Relative Standard Error (RSE) less than $\alpha\%$. (See Appendix I for a definition of RSE.)

2) Direct Estimation from the CCS

Let us consider a population with the following notation:

$c = 1 \dots 53$ counties in England & Wales.

$d = 1 \dots 5$ hard to count categories of postcodes.

$a = 1 \dots 24$ age-sex groups (0-4, 5-9, ..., 40-44, 45-79, 80-84, 85+).

$i = 1 \dots N_{dc}$ postcodes in hard to count group d of county c of which n_{dc} are in the sample s , the rest are in the non-sampler.

The quantities of interest are:

$Z_{aids} = 1991$ adjusted Census count for age-sex group a of postcode i , etc.

$X_{aids} = 2001$ unadjusted Census count.

$Y_{aids} = \text{True } 2001 \text{ count (given by the CCS for those postcodes in sample).}$

2.1) The Estimation Problem

For direct estimation from the CCS it is required that the total populations T_{ac} be estimated to a certain degree of accuracy (ie estimated $RSE < \alpha\%$). This can be treated as 24 similar estimations within each county. For this reason the subscripts a and c are dropped and estimation for one age-sex by county group is considered where the same model framework applies for all other age-sex groups. As each county is being treated as an independent estimation problem higher level totals follow from direct summation of county totals.

2.2) The Ratio Super-Population Model for a Hard to count by County Group

It seems sensible to assume that the 2001 Census count and the CCS count will be related. If this is not true then one really should be suspicious of one of the counts. Further, within sub-groups of the population a linear relationship may well be appropriate. Given that we know from the 1991 Census that age and sex are crucial to undercount, as well as local characteristics, it is sensible to consider a model within age-sex groups for each hard to count by county group where the hard to count index allows for different local characteristics. Therefore the simple ratio model stratified by the hard to count index for an age-sex by county group is:

$$E_{\xi}\{Y_{id} | X_{id}\} = \beta_d X_{id} \quad i \in d$$

$$\text{Var}_{\xi}\{Y_{id} | X_{id}\} = \sigma_d^2 X_{id}$$

$$\text{Cov}_{\xi}\{Y_i, Y_j | X_i, X_j\} = 0 \text{ for all } i \neq j.$$

Under this model the BLUP (Brewer 1963, Royall 1970) for the total T is:

$$T_{\xi} = \sum_d (T_{Sd} + \beta_d \sum_{Rd} X_{id}) \quad (\text{Substitute in the OLSE } \beta_d \text{ of } \beta_d)$$

$$T_{\xi} = \sum_d \{(\sum_{Sd} Y_{id} / \sum_{Sd} X_{id}) \cdot \sum_{Pd} X_{id}\} \quad (1)$$

where \sum_{Sd} is the summation over the sample postcodes in hard to count index d, \sum_{Pd} is the summation over all postcodes in hard to count index d, and \sum_d is the summation over all the hard to count groups in county c. There are standard model-based formulae for estimating the variance of $T_{\xi} - T$, the estimation error, under the model. Therefore an estimated RSE can be calculated for each age-sex group total in county c. In general, when the ratio model is appropriate, the estimator in (1) is more efficient than a simple stratum by stratum expansion estimator for a given sample size.

2.3) The Design Problem

The question that needs to be answered is how many postcodes, and therefore approximately how many people, are required to achieve a specified level of accuracy, or RSE, for each total T_{ac} . It is not possible to design using the ratio estimator given in (1), as there is not sufficient information. The main problem is that no sensible value for σ_d^2 (the variance around the expected value from the regression line) is known from a previous survey. There is another problem. In general, postcode level information, beyond number of addresses, is not known. This leads one to suggest a two-stage design, selecting enumeration districts and then sampling postcodes within selected enumeration districts. The rest of this paper mainly concentrates on the first stage and assumes that the enumeration district values are known without error. In practice sampling postcodes will give a loss of efficiency in the design and this is briefly discussed later.

For selecting enumeration districts the adjusted 1991 Census counts, Z_{ai} , can be used in the design (bearing in mind, of course, that they give out-of-date information and that some enumeration districts will have changed significantly since 1991). It is proposed that the adjusted counts be used if possible as these are the ‘best’ guess for what the Y_{ai} ’s might be (ideally one would like to use the X_{ai} ’s as a design variable but this is NOT possible as the CCS must be ready to go at the same time as the Census). It may be sensible to use a simple Estimating with Confidence approved strategy to update the 1991 Census counts.

2.4) Super-Population Model for the CCS Design

Within a county the enumeration districts can be stratified by hard to count and then size using the 1991 adjusted Census counts. As the intention for prediction is to use a ratio estimator for each hard to count group it is necessary to have these as strata in the design to ensure enough units are selected from each level of hard to count. (Let us ignore, at this stage, the issue of how to choose boundaries when there are several, in this case 24, size variables.) This leads to a stratified homogeneous super-population model of enumeration districts with $h = 1 \dots H_d$ size strata within each hard to count group. Using the same model framework for each age-sex group in county c the model for a given age-sex group can be written as:

$$\begin{aligned}
 E_{\xi}\{Y_{ihd}\} &= \mu_{hd} && i \in h \text{ within } d \\
 \text{Var}_{\xi}\{Y_{ihd}\} &= \sigma_{hd}^2 \\
 \text{Cov}_{\xi}\{Y_i, Y_j\} &= 0 && \text{for all } i \neq j.
 \end{aligned}$$

The BLUP for the county total for a given age-sex group is in this case the simple stratum by stratum expansion estimator, details of this and its estimation error variance formula are in Appendix I.

2.5) Multivariate Stratification and Design

As within a hard to count by county group, estimation is required for each age-sex group there are 24 potential size variables, each of the Z_{ai} ’s, to stratify on. The simplest approach would therefore be to ignore size stratification and just use a simple homogeneous super-population model within a hard to count stratum, calculate each sample size, choose the largest for each hard to count stratum. One would expect that any stratification based on a single variable will give different boundaries to any other variable and the strata will only be efficient for that age-sex group. This suggests a method that uses only one age-sex group with which to stratify may not be very efficient. However, it still may be a simple solution to choose an age-sex group, such as males aged 20-24, which is known to suffer from a potentially high undercount and use this variable. Then for the given set of strata calculate the sample size required for each age-sex group to achieve the required RSE and then choose the largest. This will give the best gains in efficiency for the age-sex group with potentially the worst undercount problem.

Another solution would be to construct a design variable Z_i' based on all the Z_{ai} 's, stratify on this, and then calculate the sample size for each stratum based on the design variable. This is the strategy adopted in this paper. Principal Component Analysis, a standard multivariate technique for reducing the dimensions of a data set, is used to reduce the number of variables to the smallest number of components needed to get over 80 percent (an arbitrary boundary) of the original variability. Cluster Analysis on these components then finds a specified number of clusters to use as size strata in each hard to count group. The actual size variable used in the sample size calculations is then given by:

$$Z_i' = \frac{|Z| \times \sum_j (P_{ji} + \delta_j)}{\{\sum_j \text{var}(P_{ji})\}^{1/2}} \quad (2)$$

where \sum_j is over the principal components chosen, P_{ji} is the j^{th} component score for the i^{th} enumeration district, δ_j is a constant that makes the j^{th} component scores positive, and Z is the variance-covariance matrix of the original size variables. As the components represent measures of size it makes sense that they are positive and adding a constant does not alter the variance. Using the determinant of this matrix as a measure of size, and bearing in mind that principal components are orthogonal, the variance of design variable in (2) is scaled to something which represents the original variability in all the variables.

2.6) Sample Size Calculations

The important feature of stratification is that, regardless of the stratification scheme, for a given set of strata and super-population model you can write-down the BLUP for the total and its variance. Then using the appropriate Z_i' 's you can estimate the number of enumeration districts needed to achieve a specified RSE. Appendix I sets out the theory in detail giving the appropriate formula for the super-population model specified in section 2.4. The formula for the approximate sample size is given by (7) in Appendix I. The strategy uses optimal allocation of the total sample to the strata and has extra built in protection against designing on a variable which is only a proxy for the actual survey variables. This extra protection comes from the simplification of the variance in (6) of Appendix I.

2.7) Second Stage Sampling

The previous three sections have ignored the fact that really we want to enumerate completely a sample of postcodes within selected enumeration districts but not ALL. This introduces more sampling error into the model as the sampled enumeration district totals, Y_{ai} in Appendix I, are not known but would have to be estimated. Consequently there will be a loss in efficiency and an increase of the RSE. As pointed out in section 2.3 there is not detailed postcode information for all postcodes in England and Wales. However, the expectation is that within an enumeration district postcodes will be reasonably homogeneous. This can be investigated using the anonymised unit level 1991 Census data for seven counties which has been made available for the project. The key is to estimate the loss in efficiency from taking a fixed sized simple random sample of postcodes from the selected enumeration districts. This is a simple second stage sample which would not require any postcode level information when implemented across all the counties.

2.8) Discussion of the Design and Estimation Strategy

As with any survey the design stage relies on out-of-date data. If the actual values existed one would not be doing a survey. The question is how much does this affect calculations of sample sizes needed for the required RSE? In areas where there has been large change in the last decade this could be significant and that is why Estimating with Confidence strategies may be useful. One advantage of fine size stratification is it tends to spread the sample through-out the enumeration districts. This is important for estimation. The proposal is to use the ratio estimator and while this is not the 'optimal sample' under the model, a reasonable spread of data will help determine if this is a sensible model. It will also give a near to balanced sample which is a robust sample for the ratio estimator. A regression estimator may fit the data better and indeed make sense. A positive intercept would represent people missed when the Census finds none. In that case a balanced sample is 'optimal' also making the spread-out sample sensible.

Another advantage of designing using a simpler model is that the more efficient model used for estimation will protect against efficiency loss from using the out-of-date design information. It will also protect against the expected loss of efficiency from the second stage of sampling the postcodes from within selected enumeration districts. There is also the added protection of having a sensible estimation alternative if a ratio/regression estimator is NOT appropriate. As the design model is a non-parametric model it is robust to model mis-specification. While it may not be the best model this robustness, given the sensitive nature of the survey, is considered a highly desirable property.

3) Estimating Totals for Small Areas (*Districts By Age and Sex*)

This section looks at the problem of estimates for small areas. The approach proposed is a very simple one stimulated by an example given by Hansen et al (1953) using the US 1945 FCC Radio Survey. In the survey a large- scale mail survey was conducted in 500 county areas. In 85 of those county areas a more accurate interview survey was also carried-out. A simple regression model was then fitted for the 85 counties and used to adjust all the mail survey returns to improve overall accuracy. In the example the results were not as good as they expected because of poor correlation between the two surveys for the 85 areas with both.

This idea was extended by Ericksen (1973, 1974) and is the basis of the regression approach to small area estimation. Ericksen's model was for updating the last Census using a current population survey for some areas and other information on all the areas such as births, deaths and school registration. This extends the model suggested by Hansen et al (1953) to a multiple regression. In the case of adjusting the Census the CCS represents the accurate count for a unit, the postcode. This is available for the sampled units while the less accurate Census count is available for all units. Using this method a regression model can be fitted for those postcodes with both counts using the Census count as a predictor along with postcode characteristic information, such as its hard to count index (which may be specified the same for all postcodes in a given enumeration district). The model is then used to predict the more accurate returns for the non-sampled postcodes in the districts. There has been much work extending these ideas to multilevel populations which more accurately reflect the hierarchical nature of a considerable amount of survey data. They also allow for more small area variation although the independence assumption between the random errors of adjoining areas is not completely realistic and current research is investigating spatial correlation.

3.1) A Model for Estimating District Undercount

Consider the same population structure as before with the additional fact that postcodes are within districts $j = 1 \dots J$ which are themselves within counties. For the districts containing sampled postcodes you can fit a different regression model for each age-sex group given by:

$$Y_{aidjc} = \beta_{1adj}X_{1aidjc} + \beta_{2aj}X_{2aidjc} + v_{ac} + u_{ajc} + e_{aidjc}$$

where Y_{aidjc} is the age-sex CCS count for postcode i from hard to count group d in district j of county c .

X_{1aidjc} is the age-sex 2001 Census count for postcode i in district j of county c .

X_{2aidjc} is a matrix of p characteristic variables at a postcode, district, and county level.

β_{1adj} is a random slope parameter (varies at the district level j by the hard to count group of the postcode) for the adjustment to the Census count.

β_{2aj} is a vector of p random slope parameters (varies at the district level j) for the p other predictors.

v_{ac} , u_{ajc} , and e_{aidjc} are random effects at the county, district, and postcode level.

Another clearer way to write the above model is:

$$Y_{ajc} = \beta_{1ad}X_{1aidjc} + \beta_{2a}X_{2aidjc} \\ + \gamma_{1aj}X_{1aidjc} + \gamma_{2aj}X_{2aidjc} + v_{ac} + u_{ajc} + e_{aidjc}$$

where the β 's are now fixed parameters in the model and the rest are random parameters with:

$$\gamma_{1aj} \sim N(0, \sigma_{\gamma 1}^2), \gamma_{2aj} \sim N(0, \sigma_{\gamma 2}^2), v_{ac} \sim N(0, \sigma_v^2), u_{ajc} \sim N(0, \sigma_u^2), e_{ajc} \sim N(0, \sigma_e^2)$$

all being iid random variables. Independence is also assumed between different levels but covariance terms are estimated between the three district level random effects. This is a standard multilevel model set-up for estimation using MLn.

The model appears rather complicated but it is both an attempt to reflect what is known from 1991 as well as to write-down a fairly general model. The model is fitted for each age-sex group. This is because experience says that certain groups suffer far more from undercount than others. In practice it will probably be necessary to combine some age-sex groups before estimation as it is expected that only a selected few age-sex groups will be really different. The fixed parameters on the Census count also depends on the hard to count group of the postcode, this reflects the expectation that poorer areas suffer more from undercount. The random slope parameters, the γ 's, allow for small area effects on the fixed effects. This means that some districts may suffer from higher (or lower) undercount than the rest even when the socio-economic characteristics of the postcodes are already accounted for. The random intercepts are then the remaining unexplained local variation in the model not captured by the predictor variables.

3.2) Prediction from the Model

The model given in Section 3.1 can be fitted in MLn. Strictly, the response is a count and so the model will be fitted using a log transformation. For a non-sampled postcode from a sampled district Y_{aij} , the predicted survey total, follows directly from using the fitted model. If a district has no sampled postcodes fitting the model gives no predicted district random effects. Under the model assumptions of the distributions of the random effects the BLUP for the district random effects is zero. While for model fitting this is a convenient assumption this is where the independence assumption between ‘neighbouring’ districts may not be sensible. It seems intuitive that a better estimate for the district random effects can be made by ‘borrowing strength’ from surrounding districts which have a predicted value from the model.

A simple solution is just a weighted average of the random effects for neighbouring districts, contiguous districts being given more weight with the weight decreasing as distance increases. There are several standard spatial functions which do this. However, it doesn’t necessarily make sense to use geographic space as districts close geographically may vary greatly in socio-economic and demographic characteristics. A more sophisticated approach is needed where a metric of ‘distance’ between districts which is not based on geographic distance but on demographic and economic characteristics. In effect one would be re-drawing the district map of England and Wales so that districts with similar demographic and economic features become neighbours. In that situation a simple spatial smoothing function could be applied.

3.3) Incorporating the Two-Stage Design

This section describes some very initial thoughts on how one can account for the two stage design.

The model in section 3.1 is postcode based and ignores the fact that postcodes are only selected from selected enumeration districts. The model can be considered in terms of enumeration districts where Y_{aidjc} , the response, is an enumeration district count. This means the Y_{ai} ’s in the model need to be estimated from the sampled postcodes in enumeration district i . Ericksen (1973, 1974) actually considers this problem as his basic unit is the PSU and the value for this is estimated. This is effectively a measurement error model where the true value of the response (the true enumeration district total) is never observed but an unbiased estimate is. The effect is to introduce an extra term of variation into the model which represents this sampling error within the basic model unit. Under the assumption that the measurement error is uncorrelated with the response Ericksen fits the model ignoring the measurement error and estimates the measurement error from the survey estimation of the PSU totals.

3.4) Discussion

This paper has initially described the basic sampling strategy which is proposed for the CCS to obtain county level age-sex specific estimates of underenumeration. The regression approach described in this paper is a standard method and could be used to allocate the underenumeration to a lower level of aggregation, perhaps district. However, research is needed to assess whether it is efficient for very small areas of aggregation. The teams current position is that the strategy described in CCS Working Paper 2 is preferred for sub county

estimates not least because it can make estimates to individual level, the ultimate goal of the ONC project.

An implicit assumption behind all the proposed models is that the CCS obtains a 100% count in the sampled postcodes. This is not a very likely assumption and for this reason it is currently proposed that, preferably two, administrative records will be used in a capture recapture analysis to make final county level estimates. This strategy will be the subject of research in the next few months.

4) Case Study of the Design for Hampshire

Hampshire was chosen purely for convenience to examine the feasibility of the design proposed in Sections 2.3 to 2.7. However, Hampshire was considered a reasonable choice as an ‘average’ county. It has some deprived enumeration districts but is not one of the counties dominated by them. The final result gives a stratification strategy and sample allocations for a designed RSE of 0.5 percent.

4.1) The Method Used

4.1.1) *Selecting EDs*

SASPAC was used to get information on the age-sex counts for all Enumeration Districts (EDs) at the last Census (1991). The age groups used for males and females were 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-79, 80-84, and 85+. These were chosen as analysis at the last Census suggested undercount to be reasonably homogeneous within these groups with the 45-79 group suffering negligible undercount, even at a district level. The age-sex counts were used to select those EDs with zero population at the last Census. Of the 113,062 EDs in England and Wales 3,489 had zero population at the last Census. These were excluded from the rest of the analysis and will need to be treated separately within each county.

4.1.2) *Hard to count Index*

As the proposed method of estimation uses the ratio estimator it relies on a linear relationship between the observed 2001 Census count and the CCS count. This will only be true after some control for the socio-economic make-up of EDs. This was done using a hard to count index which was calculated for all EDs with non-zero population in 1991. The variables used were:

$$\text{Percent Unemployed} = \frac{\text{Unemployed Aged 16+}}{\text{Economically Active Aged 16+}} \times 100$$

$$\begin{aligned} \text{Percent Shared or No Indoor Toilet} \\ = \frac{\text{Total Households With No or Shared Indoor Toilet}}{\text{Total Households}} \times 100 \end{aligned}$$

$$\text{Percent No Car} = \frac{\text{Households With No Car}}{\text{Total Households}} \times 100$$

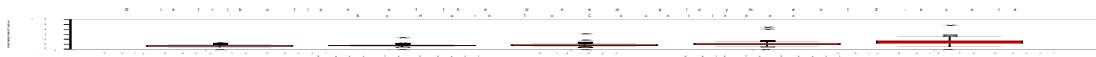
$$\begin{aligned} \text{Percent Permanent Rented} \\ = \frac{\text{Renting Households in Permanent Accommodation}}{\text{Total Households}} \times 100 \end{aligned}$$

$$\text{Percent Temporary} = \frac{\text{Households In Temporary Accommodation}}{\text{Total Households}} \times 100$$

$$\text{Percent PPR} > 1.5 = \frac{\text{Households With More Than 1.5 Persons Per Room}}{\text{Total Households}} \times 100$$

$$\text{Percent Lone Parents} = \frac{\text{Households With Lone Parent}}{\text{Total Households}} \times 100$$

For each variable a z-score was calculated based on the national distribution of each variable. The variables were all designed such that an increasing z-score implied increasing hard to count in the ED. An overall hard to count score was given to each ED by summing the individual scores. To turn this into an index the quintiles of the overall score were used. This produced a hard to count index with five levels, level one being the EDs with the lowest hard to count score and five being those with the highest. Below in Figure 4.1 is a box plot of the distribution of one of the component variables, the percent of economically active people who are unemployed, by the hard to count index. It should be noted that this is a simple hard to count index which has been calculated for this expository analysis.



The figure shows that z-scores for the percentage unemployed are increasing as the index increases as you would expect. Appendix III contains the plots for the other six component variables. The general pattern is similar across all the plots but stronger in some than others.

4.1.3) Principal Component Analysis and Cluster Analysis

As suggested in Section 2.5 Principal Component Analysis was used to reduce the number of size variables. It was found that across Hampshire the first two principal components of a correlation matrix analysis accounted for 86 percent of variability in the 24 original age-sex variables with remaining components accounting for very little. The first component was just a weighted average of all the 24 original variables while the second was a contrast between the old and young populations of the enumeration districts. This can be seen in Table 4.1 below.

Table 4.1 - The First Three Principal Components For Hampshire

Principal Component	Proportion of Variability	Cumulative Proportion	Description
1	0.776	0.776	Weighted Average
2	0.086	0.862	Contrast of 45+ males and 40+ females against the rest.
3	0.036	0.898	

The first two principal components were used in a Cluster Analysis. A positive constant was added to principal component two to make all the scores greater than or equal to zero (this does not change the variance of the variable but it makes sense for a variable which represents a size measure which is intrinsically positive). Centroid and Average Linkage were tried in the clustering algorithm but the final choice was Ward Linkage. Not only did this perform 'best' it also made sense, as for Ward linkage the clustering criterion minimises the sums of squares within clusters and therefore variability. The clustering was run to produce a specified number of clusters to be used as size strata within the hard to count groups. Clusters of two or less were rejected and any outliers were treated as special cases to be added to the sample at the end.

Two sets of plots were examined to see if the clustering was producing sensible answers. The first was a set of scatter plots, one for each group of the hard to count index. These scatter plots showed whether the clustering was really finding individual clusters. Figure 4.2 is the scatterplot for the very hard to count enumeration districts with fifteen clusters, and four outliers, which was the final clustering for that group.

Ward Linkage Cluster Analysis (15 Clusters)

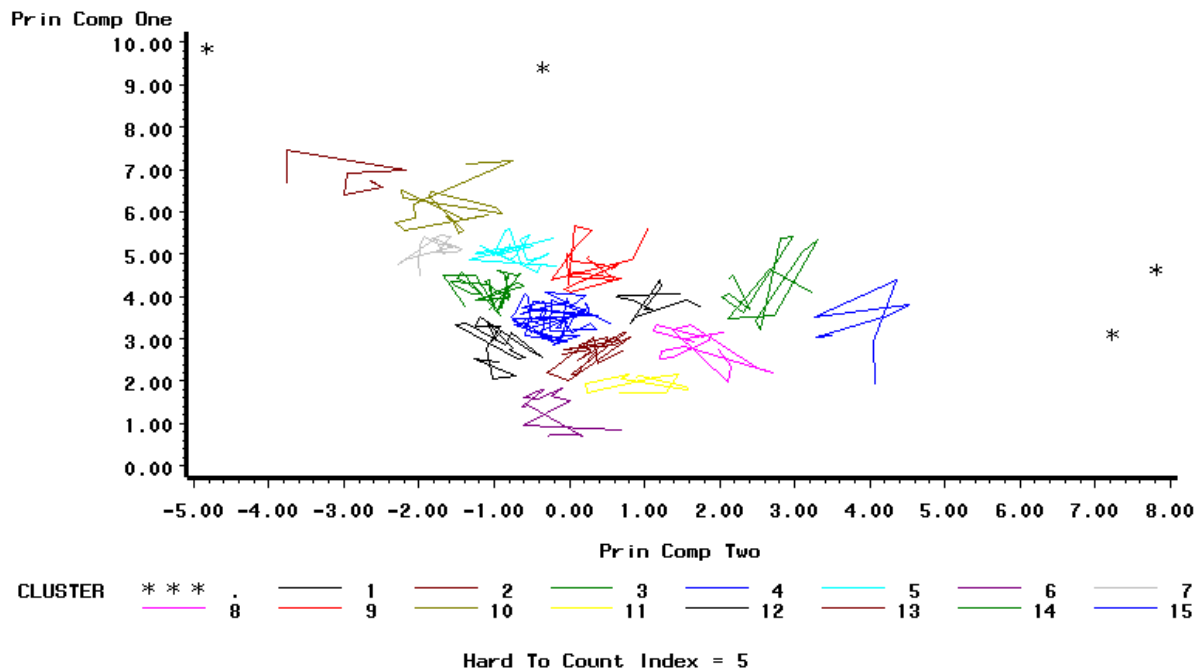


Figure 4.2 shows that the clusters are all separate with the four outliers clearly being different from the main cloud of points. However, it also shows that there is no evidence to say there are really fifteen separate clusters. In this situation this is not so important as any stratification method, such as Dalenius Hodges, will sometimes split units which may more naturally go together. The other four scatter plots can be seen in Appendix IV. Each of these has twenty-five clusters.

The second set of plots were box plots of the cluster means for the two clustering variables and the design variable showing their distributions within each level of the hard to count index. As for the final set of size strata each stratum sample was small (most equal one) these plots give an idea of the sample distribution in each hard to count group. Figure 4.3 demonstrates this using the plot for the first principal component, a weighted average of the 24 original age-sex variables.

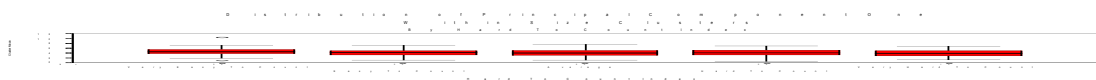


Figure 4.3 shows that the distribution across the groups is pretty much the same. This makes sense as the grouping and clustering variables are not calculated using the same information. The plots for principal component two and the design variable show a very similar pattern which suggests that the sampling distribution in each group will be very similar. These plots can be seen in Appendix V.

4.1.4) Sample Size Calculations

For a given set of strata a total sample size and a set of stratum sample sizes were calculated using the formulae set-out in Section 2.6 and Appendix I. The resulting sample sizes were used to decide on whether to increase the number of clusters created in the cluster analysis. The number of size strata created by the clustering were increased until there were no longer significant gains from increasing the number of size strata in a hard to count group. The final choice was not considered 'optimal' but efficient and robust. The final choice had very fine strata which spreads the sample of enumeration districts through-out each of the hard to count groups.

4.2) Sample Size Results of the Case Study

Several different numbers of size clusters were tried. Initially ten clusters in each hard to count group were tried. Table 4.2 gives the final sample allocations for this clustering.

Table 4.2 - Sample Sizes for Ten Clusters Per Level of Hard to Count Index

Level of Index	Sample Size ¹	Outliers	Total Number of Enumeration Districts ²	Total Sample Fraction ²
1	74	0	892	0.08
2	61	0	717	0.09
3	58	0	678	0.09
4	51	0	601	0.08
5	28	2	341	0.09
TOTAL	272	2	3229	0.08

1. Based on an RSE of 0.5 percent

2. Outliers Included

From Table 4.2 one can see that the sampling fraction of enumeration districts is quite high. This does not take into account second stage sampling of postcodes within enumeration districts but not all postcodes in the sampled enumeration districts will be counted. The two outliers are added to the sample making a total of 274 enumeration districts. Clearly the sample sizes in each hard to count group are still large and there are more gains to be had from increasing the number of clusters. Tables 4.3 and 4.4 are for fifteen and twenty clusters per hard to count group respectively.

Table 4.3 - Sample Sizes for Fifteen Clusters Per Level of Hard to Count Index

Level of Index	Sample Size ¹	Outliers	Total Number of Enumeration Districts ²	Total Sample Fraction ²
1	58	0	892	0.07
2	49	0	717	0.07
3	43	0	678	0.06
4	36	0	601	0.06
5	22	4	341	0.08
TOTAL	208	4	3229	0.07

1. Based on an RSE of 0.5 percent
2. Outliers Included

Table 4.4 - Sample Sizes for Twenty Clusters Per Level of Hard to Count Index

Level of Index	Sample Size ¹	Outliers	Total Number of Enumeration Districts ²	Total Sample Fraction ²
1	49	0	892	0.05
2	37	2	717	0.05
3	33	0	678	0.05
4	31	0	601	0.05
5	21	6	341	0.06
TOTAL	171	8	3229	0.06

1. Based on an RSE of 0.5 percent

2. Outliers Included

Comparing Tables 4.3 and 4.4 shows that there are no gains in group five from increasing the number of clusters. This is especially true as the number of outliers increases and the suggestion is that outliers are treated as separate strata of size one. It is clear though that for the other four groups there are still potential gains to be made. Table 4.5 has a variable number of clusters per group.

Table 4.5 - Sample Sizes for a Mixed Number of Clusters Per Level of Hard to Count Index

Level of Index	Number of Clusters	Sample Size ¹	Outliers	Total Number of Enumeration Districts ²	Total Sample Fraction ²
1	25	42	0	892	0.05
2	25	34	2	717	0.05
3	25	32	0	678	0.05
4	25	29	0	601	0.05
5	15	20	4	341	0.07
TOTAL	115	157	6	3229	0.05

1. Based on an RSE of 0.5 percent
2. Outliers Included

Table 4.5 is not ‘optimal’ but it does represent a good design. For most groups the sample size is about the same as the number of clusters suggesting there are no real gains to be made from more clusters. This represents a five percent sample of enumeration districts within Hampshire to get an RSE of 0.5 percent at the county level. Again it should be remembered that this does not include second stage sampling of the postcodes within enumeration districts. Appendix III gives the complete breakdown of this allocation across all the strata and it shows that most stratum samples are one.

4.3) Conclusion

The case study has shown that the proposed design is feasible. The stratification has led to significant gains in the required sample size. The next stage of the design as this will determine the final sample size in terms of postcodes. It will also, of course, determine cost. The clustered nature of the design, selecting postcodes within enumeration districts, should help keep costs down.

5) Conclusions

This paper describes the basis for a complete strategy from the design of the 2001 CCS to estimation of true populations from the CCS and Census for district level populations. The design depends on a large amount of information being available from the 1991 Census to get an efficient design. The direct estimation for age-sex groups within hard to count by county strata from the CCS proposes using a Ratio Estimator with the raw Census counts. The design does not depend on using this estimator which should lead to gains in efficiency to balance out designing on old information.

The strategy proposed to make small area adjustments uses a standard small area regression approach which should be acceptable to the user community. The success of this strategy depends on there being a good correlation between the Census and CCS counts for the postcodes. This is why a separate model is needed for each age-sex group as it is well known that this has a strong effect on undercount. However, it may be possible to look at some groups together and therefore to reduce the number of groups.

The next stages of the research will be:

- i) extend the work in this paper to make national sample size and cost calculations.
- ii) compare the effectiveness of the simple strategy to make small area adjustments with the more sophisticated approach described in CCS Working Paper 2, paying particular note to the power of the simple estimates at very low levels of aggregation.

The Steering Committee is asked to comment on:

- i) the overall design strategy.
- ii) the potential for giving extra weight to the strata which are expected to be harder to count.

References

- Brewer, K. R. W. (1963).
Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process.
Australian Journal of Statistics Vol. 5 pp 93-105.
- Ericksen, E. P. (1973).
A Method for Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas.
Demography Vol. 10 pp 137-160.
- Ericksen, E. P. (1974).
A Regression Method for Estimating Population Changes of Local Areas.
JASA Vol. 69 pp 867-875.
- Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953).
Sample Survey Methods and Theory, Volume 1 pp 483-486.
New York by John Wiley and Sons, Inc., 1953.
- Royall, R. M. (1970).
On finite population sampling under certain linear regression models.
Biometrika Vol. 57 pp 377-387.

Appendix I - Sample Size Calculations

For the super-population model ξ given in section 2.4 the BLUP for the total of an age-sex group is the stratum by stratum expansion estimator given by:

$$T_{\xi} = \sum_{hd} N_{hd} y_{Shd} \quad (1)$$

where y_{Shd} is the sample mean for the CCS enumeration district count. For this estimator and model the variance of the estimation error is given by:

$$\text{var}_{\xi}(T_{\xi} - T) = \sum_{hd} (N_{hd}^2/n_{hd})(1 - n_{hd}/N_{hd})\sigma_{hd}^2 \quad (2)$$

For a given total sample size of n enumeration districts optimal allocation is used to get the individual stratum population sizes such that:

$$n_{hd} = n \cdot \frac{N_{hd}\sigma_{hd}}{\sum_g N_g\sigma_g} \quad (3)$$

For a given population quantity such as the total T with estimator T_{ξ} you can measure how accurate your estimator is using the relative standard error (RSE) defined as:

$$\text{RSE}(T_{\xi}) = \frac{\{\text{var}_{\xi}(T_{\xi} - T)\}^{1/2} \cdot 100}{T} \quad (4)$$

We want to design for an RSE of α percent. Considering the variance formula you can write it as:

$$\text{var}_{\xi}(T_{\xi} - T) = \sum_{hd} (N_{hd}^2\sigma_{hd}^2/n_{hd} - N_{hd}\sigma_{hd}^2) \quad (5)$$

Now only the first term depends on the sample sizes. Substituting for n_{hd} in terms of n using (3) in the first term of (5) gives:

$$\begin{aligned} \text{var}_{\xi}(T_{\xi} - T) &\leq \sum_{hd} N_{hd}^2\sigma_{hd}^2 \cdot \{\sum_g N_g\sigma_g / (nN_{hd}\sigma_{hd})\} \\ &= \{\sum_{hd} N_{hd}\sigma_{hd}\}^2 / n \end{aligned} \quad (6)$$

Using (6) as the variance in the RSE formula gives (with extra built protection) the approximate sample size required for an RSE of α percent as:

$$n = \frac{10^4 \{\sum_{hd} N_{hd}\sigma_{hd}\}^2}{\alpha^2 T^2} \quad (7)$$

For the actual calculations a design variable is used in place of the Y_i 's as these are obviously unknown and the required RSE is 0.5 percent.

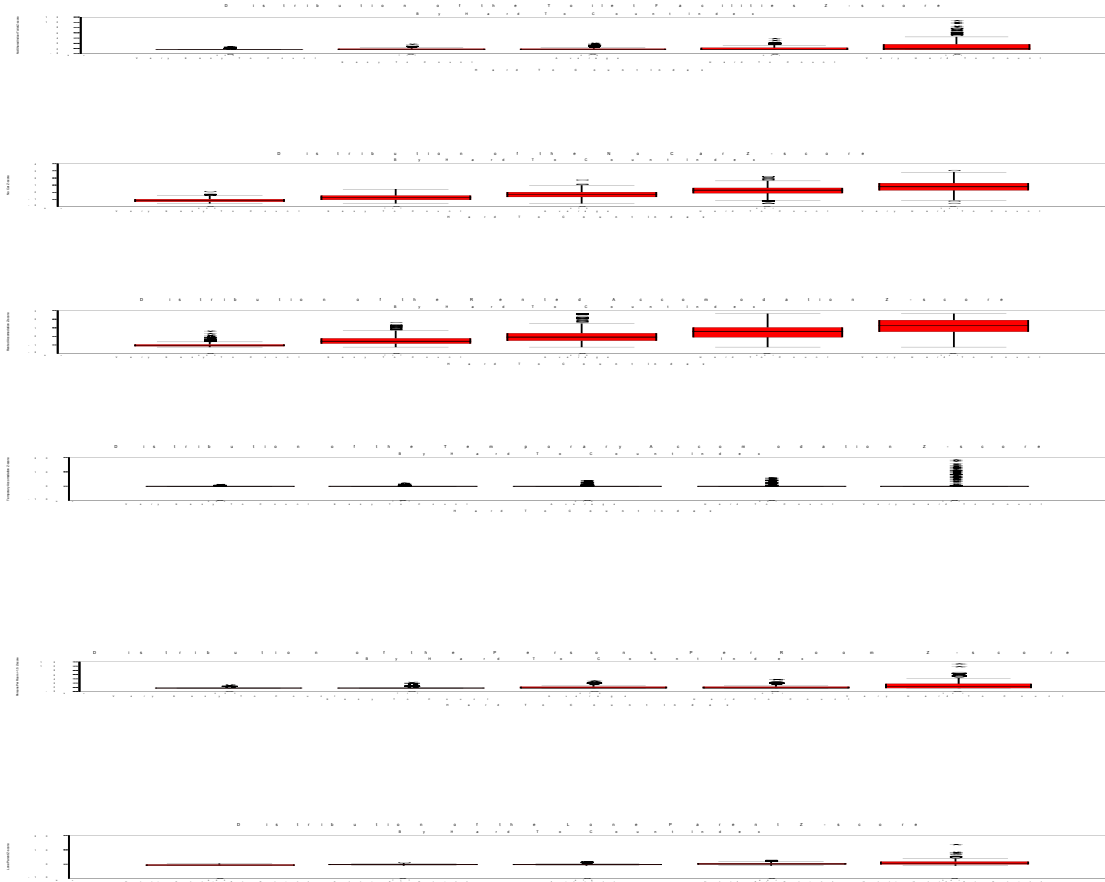
Appendix II - Sample Allocation to Hard to count Groups and Clusters

Hard to Count Index	Cluster	Population (ED's)	Sample Size (ED's)
1	1	40	1
1	2	102	3
1	3	45	2
1	4	47	2
1	5	72	3
1	6	47	1
1	7	46	2
1	8	45	1
1	9	43	3
1	10	69	2
1	11	27	2
1	12	31	1
1	13	32	2
1	14	19	2
1	15	26	1
1	16	30	2
1	17	18	1
1	18	56	3
1	19	38	2
1	20	25	1
1	21	7	1
1	22	10	1
1	23	9	1
1	24	5	1
1	25	3	1

Hard to Count Index	Cluster	Population (ED's)	Sample Size (ED's)
2	1	43	1
2	2	40	2
2	3	80	2
2	4	33	2
2	5	38	1
2	6	42	2
2	7	43	2
2	8	18	1
2	9	40	2
2	10	54	2
2	11	28	2
2	12	15	1
2	13	30	1
2	14	22	1
2	15	47	2
2	16	16	1
2	17	12	1
2	18	19	1
2	19	20	1
2	20	26	1
2	21	13	1
2	22	13	1
2	23	10	1
2	24	9	1
2	25	4	1
3	1	55	1
3	2	22	1
3	3	44	1
3	4	46	2
3	5	46	2
3	6	32	2
3	7	59	2
3	8	18	1
3	9	48	2
3	10	28	1
3	11	45	2
3	12	40	2
3	13	28	1
3	14	22	1
3	15	9	1
3	16	15	1
3	17	37	1
3	18	23	1
3	19	14	1
3	20	12	1
3	21	13	1
3	22	9	1
3	23	7	1
3	24	3	1
3	25	3	1

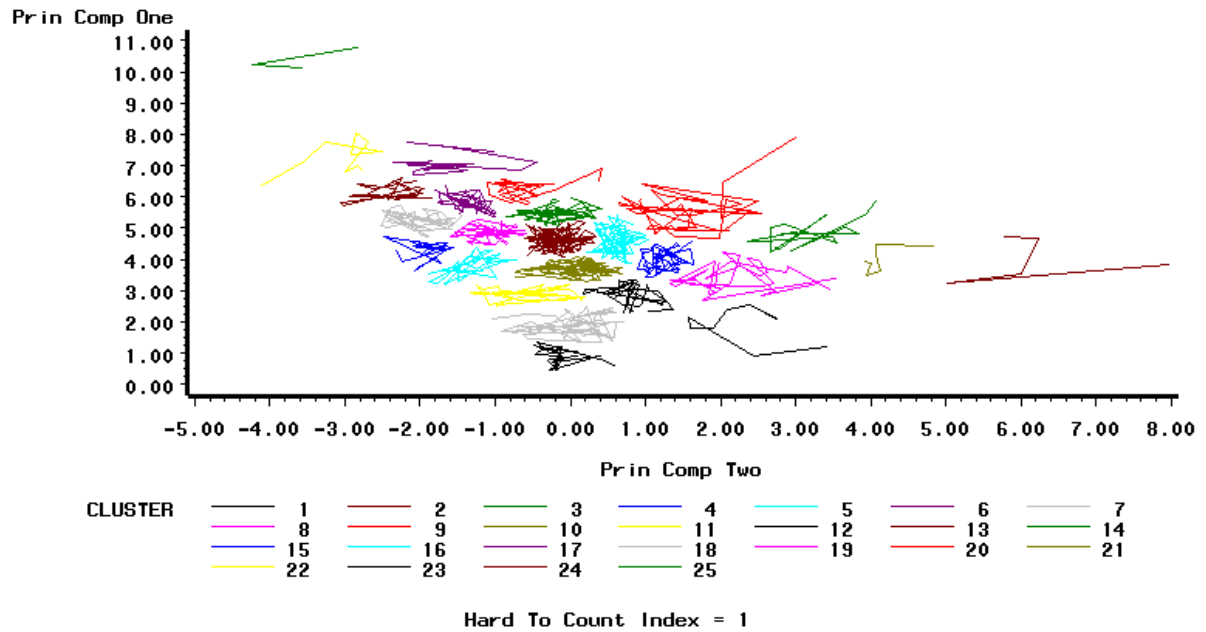
Hard to Count Index	Cluster	Population (ED's)	Sample Size (ED's)
4	1	34	1
4	2	58	2
4	3	49	2
4	4	76	2
4	5	33	1
4	6	23	1
4	7	48	2
4	8	27	1
4	9	30	1
4	10	24	1
4	11	29	1
4	12	27	1
4	13	18	1
4	14	16	1
4	15	6	1
4	16	10	1
4	17	16	1
4	18	13	1
4	19	10	1
4	20	8	1
4	21	11	1
4	22	10	1
4	23	16	1
4	24	6	1
4	25	3	1
5	1	24	1
5	2	29	2
5	3	40	2
5	4	64	3
5	5	28	1
5	6	15	1
5	7	17	1
5	8	20	1
5	9	21	1
5	10	20	1
5	11	12	1
5	12	14	1
5	13	8	1
5	14	17	2
5	15	8	1

Appendix III - Distributions of the Z-scores for the Component Variables in the Hard To Count Index

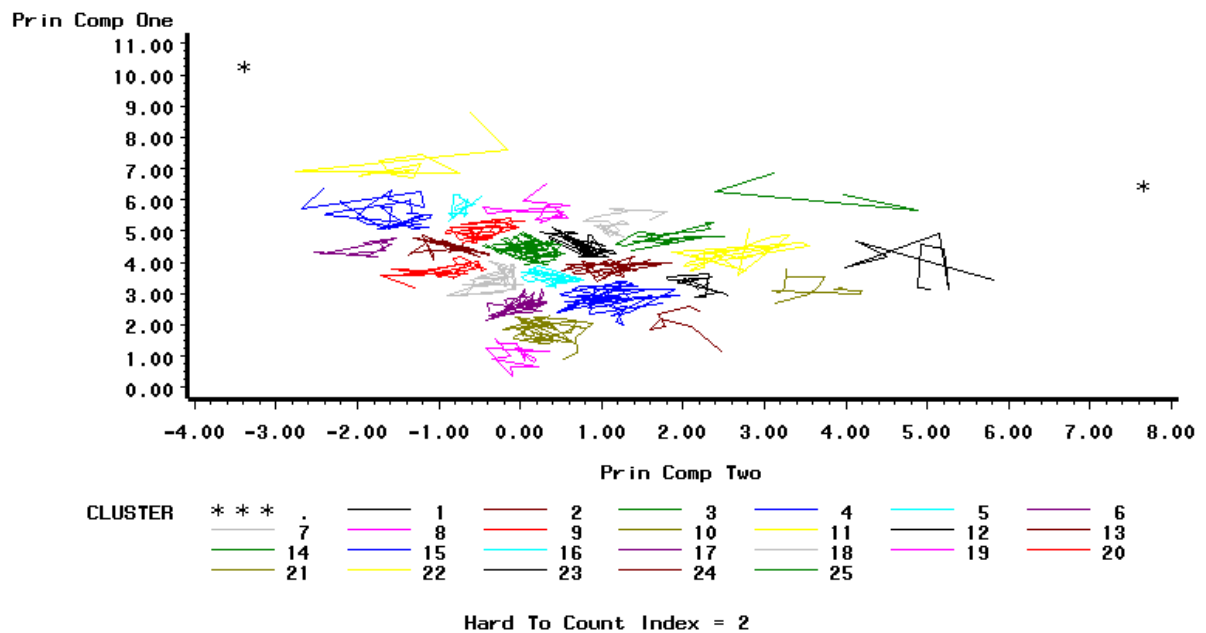


Appendix IV - Scatter Plots of the Final Cluster Analysis

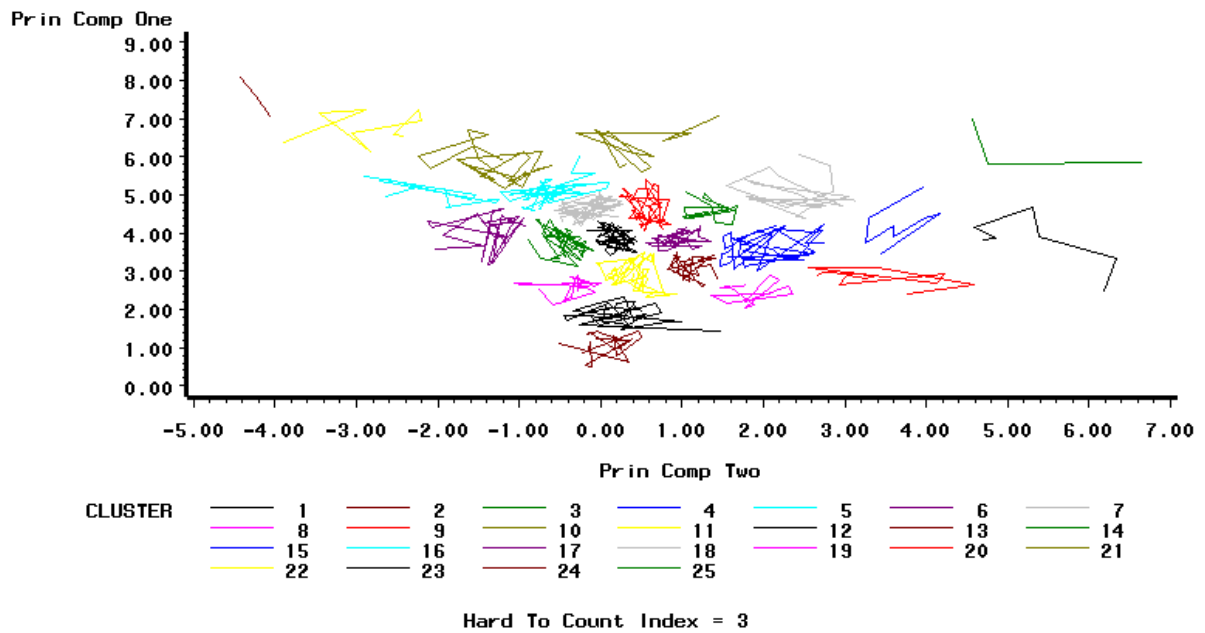
Ward Linkage Cluster Analysis (25 Clusters)



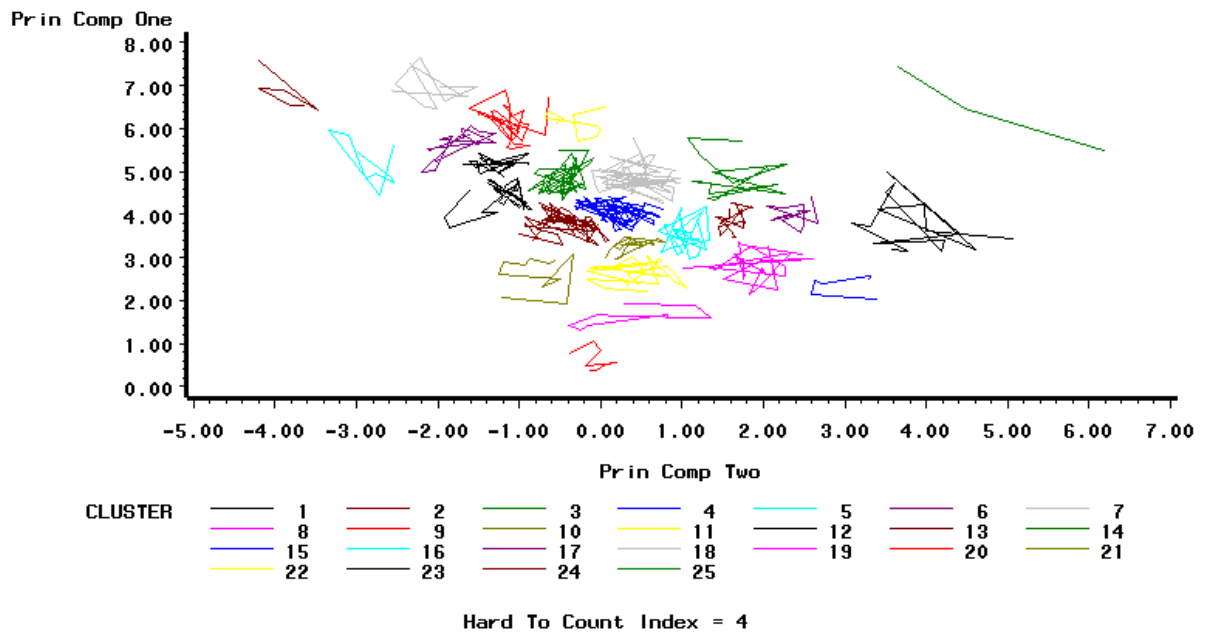
Ward Linkage Cluster Analysis (25 Clusters)



Ward Linkage Cluster Analysis (25 Clusters)



Ward Linkage Cluster Analysis (25 Clusters)



Appendix V - Distributions of the Cluster Means for Principal Component Two and the Design Variable

