PTRC 1998

PROTOTYPICAL SAMPLE ENUMERATION AS A BASIS FOR FORECASTING WITH DISAGGREGATE MODELS

Andrew Daly
Hague Consulting Group

The objective of disaggregate modelling as applied to travel demand forecasting is to explain the choices made by individual travellers. This approach has proved very successful as a basis for the development of models and through the technique of sample enumeration disaggregate models have also been used successfully for short-term forecasting. However, longer-term forecasting is not possible with the straightforward applications of sample enumeration. To fill this gap, the technique of *prototypical* sample enumeration has been developed.

The function of prototypical sample enumeration is to provide a basis for the application of disaggregate travel demand models. Additionally, it allows the changing nature of the population to be taken into account in forecasting. Given the substantial changes that are currently taking place in the structure of Western European populations, such as reductions in household size, increasing employment of women and a general 'greying' of the population, the resulting changes in travel demand can also be quite large. Vrolijk *et al.* (1987) predict a 19% growth in car kilometres over a 27-year period simply from changes in the population (including 6% overall growth), quite separately from increases in car ownership and licence holding. Thus predictions of changed population structure contribute to the forecasting of traffic changes in themselves, in addition to providing a basis for the application of other models.

The first section of this paper introduces the background of disaggregate modelling and sample enumeration in which prototypical sampling is applied. The second section states simply the background and main objectives of the method. Section 3 explains the details of the two principal methods that are used in practice to achieve those objectives: Iterative Proportional Fitting and Quadratic Optimisation. Some examples from practical studies are given in Section 4 and the paper concludes with a discussion of the main advantages and disadvantages of the method.

1. DISAGGREGATE MODELS AND SAMPLE ENUMERATION

Disaggregate modelling is an approach to predicting travel demand which focusses on the choices made by individual travellers. Travel demand, whether expressed as the number of cars on a road in the peak hour, revenues for a proposed public transport system or total car sales, is thus seen as the aggregation of a large number of individual decisions. The success of the approach has most often been seen in its ability to obtain good estimates of the factors influencing travellers' choices, i.e. in model development.

A key characteristic of disaggregate modelling is the statistical approach that it inherently takes to the analysis of data. This approach recognises that it is not possible to predict correctly how each individual in a population will be have, but this does not prevent information being obtained on the variables that *influence* – rather than *determine* – behaviour. The model for each individual is then formulated as

$$\Pr\{\ c_i = k \mid K_i, S_i\ \} \ = \ p_k(K_i, S_i) \qquad (1)$$

giving the *probability* that the choice c_i of individual i, whose characteristics are S_i, will be alternative k from the choice set K_i (which has availability and characteristics specific to individual i and his or her journey). It is a primary objective of the modelling then to specify how the alternatives in K are described and which characteristics S are relevant. A further important task in the modelling is to determine the form of p and estimate the values of unknown parameters that appear in it.

Modelling and in particular the estimation of unknown parameters requires data. In this context the need is for information about a sample of travellers, described by their characteristics S, each of whom has been observed to make a choice from a set of alternatives K whose characteristics are known. Then by confronting a model of the form of equation (1) with the data, the form of the model and/or the values of the unknown parameters can be adjusted until the fit of the model to the data is optimised according to some principle, often that of maximum likelihood.

The model in the form (1) is often of great value in itself. For example, estimated time and cost parameters can be examined, their ratio giving a 'value of time' that can be used for the evaluation of transport improvements. Hypotheses can be tested, for example that values of time increase with income, or that improvements in accessibility lead to an increase in the total amount of travel. But the greatest value of such a model is its applicability for forecasting.

In order to make useful forecasts a means must be found to *aggregate*, to derive from a model predicting the behaviour of individuals a forecast of the behaviour of an entire population. An important point is that it is not correct simply to set K and S to the average population values and apply equation (1) as if the entire population behaved like a mass of identical average individuals: *this overstates the response to changes*, an effect known as aggregation bias which has long been recognised (e.g. Daly, 1976, Gunn, 1984). Similarly, the model (1) cannot be used directly to calculate elasticities, again this leads to an overstatement of responsiveness.

A technique which does not have this disadvantage is that of sample enumeration. Essentially, sample enumeration simply applies the model (1) to each member of a sample in turn. Then, *if the sample is representative*, the sum of the forecasts for each individual is the unbiased forecast for the whole population. Formally, the expected demand Q_k for an alternative k is given by

$$Q_k \ = \ \Sigma_i\ w_i \ .\ p_k(K_i, S_i) \qquad (2)$$

where w_i is the expansion factor or weight attached to individual i in the sample in order to make its sum representative of the population. Very often, the sample used for forecasting is the same sample used for model estimation, while the weights w are determined by the sampling process used: often they are the inverse sampling probabilities.

The advantages of sample enumeration using the basic equation (2) are its simplicity and convenience. The estimated model (1) can be used directly, the sample is often available directly from the estimation process, with base values for K and S already specified for the estimation process. The computer run time for the forecast is usually small and can be controlled by taking a sub-sample if that is required. The forecasts are unbiased.

It is important to note that the procedure of sample enumeration is entirely independent of the form of the model that is used for forecasting: logit, linear, whatever model is used can be applied in this way.

The primary disadvantage of sample enumeration is that a representative sample may not be available, perhaps because the model is being transferred in time or space. In particular this will always be true when a forecast is required over any considerable period, so that a base-year sample can no longer be considered representative. The problem of the sample will also occur when a forecast is required of details of behaviour over a wide area, e.g. when traffic flows are to be assigned to a network.

The name 'sample enumeration' appears first to have been used by Ben-Akiva and Atherton (1977) although the procedure was also in use, without explicit naming, by others at the same time (Daly and Zachary, 1977, is probably not the only case). The limitations of sample enumeration for long-term forecasting (note "short-range" in the title of the Ben-Akiva/Atherton paper) and for area-wide work were also known at an early stage and alternative aggregate forecasting procedures were available to apply disaggregate models in those cases (Ben-Akiva *et al.*, 1978).

However, aggregate forecasting methods bring a number of disadvantages with them and in many cases it is clear that the sample enumeration method would be preferable. The contrast between the two basic methods of aggregation, segmentation and sampling, was set out by Daly (1982). The segmentation approach calculates demand as the total of demand from each of a large number of segments. Within the general concept of segmentation may also be included geographical segmentation, i.e. zoning. However, because of the limitations of computer power and data availability, the number of socio-economic segments must be limited when the number of zones is at all large. The question arises of the 'competition' between detailed socio-economic representation and detailed locational representation, i.e. detailed zoning. In contrast the sampling approach does not attempt complete coverage in any dimension but focusses instead on representativeness and maximising the accuracy that can be achieved from a sample of a given size. Mathematically, the two approaches are alternative methods of *integration* of the demand function. They can be compared on the basis of which gives the greater accuracy for a given expenditure of computer time.

The conclusion is that the advantages of sample enumeration are substantial in some circumstances and therefore that it would be advantageous to be able to apply the technique more widely. A means was therefore required for generating representative samples for circumstances different in space or time from those for which real samples are available.

2. PROTOTYPICAL SAMPLING

The most obvious way to produce samples representative of future conditions is to generate an artificial population which has, as far as is known, the characteristics of the future population. However, the forecasts that are generally available – e.g. from planning authorities – typically refer to aggregate statistics such as age-sex population distribution, rather than the composition of individual households. A method is therefore required for generating a sample of households that is internally consistent, i.e. that it 'looks like' a typical population, while also achieving consistency with such aggregate statistics as are available.

A systematic approach was taken to the development of such samples as bases for the application of the Netherlands National Model, work published in a series of papers in the mid 1980's (Gunn *et al.*, 1983, Gunn, 1984 and Daly and Gunn, 1985). In the first of these papers, a series of detailed tests is presented which illustrate the possibility of substituting 'artificial' samples for real ones as a basis for model applications without losing significant accuracy in the model forecasts. Essentially a kind of independence or orthogonality could be established between socio-economic and locational characteristics. The second and third papers elaborate the detailed methods used, also summarised and developed further in the following section of this paper. The novel development of this work was to use an adjustment of the expansion weights, the w's of equation (2), to adapt a sample of 'foreign' origin to be representative of a series of target areas. These target areas were then chosen to be the zones for which the forecasting model was to be run.

For The Netherlands, convenient samples could be drawn from the National Travel Survey (OVG). It was found that these samples into could be re-weighted to fit all of the zones of the National Model, provided account was taken of two special area types: city central areas, which were found to contain an exceptional number of older single people, mainly women; and polder areas, which were found to contain an exceptional number of families with children.

Subsequently, it has been found possible to apply similar procedures in other areas. Minor developments have been made to the procedure as required by the availability of data in these other areas, these are explained in Section 4 below.

The objective of the method is thus to use an existing household sample to produce a sample that is or will be representative of one or more target areas. The key method used for adjusting the samples is the adjustment of the expansion weights present on the survey records. The following section discusses the possible ways in which these weights can be adjusted.

3. OPTIMISATION

Two major groups of procedures which have been used in practice to set up prototypical samples are Iterative Proportional Fitting (IPF) and Quadratic Optimisation. Both these methods rely on the availability of a detailed sample of households which is not directly representative of a specific target area or year. The detailed sample may refer to another area (larger, smaller or elsewhere), another year or both. The objective of the procedure is to create samples that are representative of target areas, given data for those target areas that is much less detailed in character.

3.1 Iterative Proportional Fitting

IPF is a procedure of repeated factoring of a multi-dimensional matrix to meet marginal totals in each dimension. It is often applied in transportation planning for the 'balancing' or origin-destination matrices, commonly under the name of the Fratar method.

Beckman *et al.* (1996) give a detailed exposition of a prototypical sampling procedure based on IPF and adapted to exploit specific tables and slightly censored samples ("PUMS") available to the public from the US census. The way in which the details of the method are derived from the data that is available – the table numbers in the census publications are given in the paper – indicates the degree to which prototypical sampling can be adapted to varying circumstances, although of course it also means that the specific method used in the paper cannot be applied outside the US. The recent date of the paper and the careful review the authors give of work by others in this field gives confidence that they are able to give us a proper reflection of the best procedures for iterative proportional fitting.

The procedure adopted by Beckman *et al.* is to use the published tables to give totals of the numbers of households falling in each of a series of categories for each of a series of variables. Each table gives the distribution of a single variable in a small area (a census tract), to which some aggregation of categories and other manipulations of the categorisation are applied. The sample data, which relates to a larger area, is used to construct a cross-classification scheme, termed by the authors a "multiway" table, implicitly assuming that the correlations between the categorical variables in the larger area also apply in the smaller area. The IPF method is employed at this point, repeatedly correcting marginal totals to achieve a perfect match to the categorical totals of each of the variables. Thus for each of the small areas a multiway table is constructed that matches the census marginal distributions exactly while also matching – as far as possible – the correlations given in the larger area.

Once these multiway tables have been constructed, households are sampled from the large-area sample in proportion to the probabilities indicated by those tables. The procedure set out by Beckman *et al.* samples whole numbers of households which of course involves some minor deviations from the target probabilities due to rounding. The output of the procedure is thus a sample which contains the correct number of households for the target area (i.e. no weighting is needed).

For a specific study area (in New Mexico), Beckman *et al.* show that their procedure could not be improved by either of two plausible heuristic amendments and that the overall distributions of some variables not used in constructing the sample are reasonable. However, the relative frequencies of households in the multiway tables will not be the same as those in the base sample, because of the repeated factoring necessary to match the marginal totals. The extent of this deviation is not reported and could presumably be quite large because it is not controlled in any way.

While the Beckman *et al.* application was applied and reported for a specific data source, it is clear that the IPF method could be applied to data of different structure, such as would be available in other countries. Substantial amendments would of course be required to the details of the procedure. Similarly, the method could be used to generate samples that could be considered representative of future populations, providing the marginal distributions of key variables could be forecast.

3.2 Quadratic Optimisation

The construction of prototypical samples by the quadratic optimisation method ('QUAD') rests on the recognition that the data for the target area and the base sample may be inconsistent. That is, the method **balances** the need to meet the target area marginal totals against the wish to retain the detailed relationships between the frequencies of different household types indicated by the base sample. Weights can be given to the relative divergences: in this sense QUAD is a generalisation of the IPF method, which gives exact matches to the marginal totals but sacrifices faithfulness to the original detailed sample.

A further difference between most applications of QUAD and the IPF method explained by Beckman *et al.* is that QUAD constructs its detailed samples by weighting or re-weighting the records of the base sample, rather than by drawing from the base sample with fixed probabilities. This difference has the minor advantage that the rounding errors found in IPF are eliminated, but its more important advantage is that it avoids the additional step of drawing the sample. The output is thus a sample whose size is predetermined and independent of the target area; the fit to the target area is achieved by the weighting.

Re-weighting is applied to all of the households in each of a series of categories, pre-defined to cover the main dimensions of interest for the prediction of travel behaviour. In the applications of the procedure that have been made to date, 40-50 categories have typically been found to be sufficient, defined with respect to such variables as household size, numbers of adults and workers, the age of the household head and (unless explicitly modelled) the licence holding of the household.

QUAD is called quadratic optimisation because it can be specified in the form of optimisation with respect to the new frequencies ϕ_c of households of each category c of a quadratic function for each target area, i.e. (following Daly and Gunn, 1985),

$$\phi = \text{argmin}(Q), \qquad Q = \Sigma_t\, w_t \,.\, (z_t - \Sigma_c\, \phi_c \,.\, x_{tc})^2 + \Sigma_c\, (\phi_c - f_c)^2 \tag{3}$$

and

w_t	is the weight attached to the importance of meeting target t;
z_t	is the value per household of target statistic t in the current area;
x_{tc}	is the average amount of target variable t for a household in category c; hence $(\Sigma_c\, \phi_c.x_{tc})$ is the predicted total of statistic t;
f_c	is the frequency of household category c in the base sample.

The first term in Q clearly represents the error in not meeting the target marginal totals for each variable z, while the second term represents the divergence from the current distribution of households over the categories. The weights w are introduced so that differential importance can be given to meeting each of the different targets or that the balance between consistency with targets and consistency with base population can be adjusted. In fact, in most applications it has been found satisfactory to set all the w's to 1. Setting large values of w would cause QUAD to find a distribution of households that matched the target totals very well at the expense of substantial departures from the original distribution, i.e. a solution like that given by IPF.

Note that all terms of Q are on a per-household basis.

The simple form of Q makes it in principle easy to optimise. Given any starting value of ϕ, the global minimum of Q is always at the value ϕ^* given by

$$\phi^* = \phi - Q'(\phi) \,.\, Q''(\phi)^{-1} \tag{4}$$

where Q' and Q'' are the first and second derivatives respectively of Q with respect to ϕ, i.e. Newton's calculation, which converges directly for a function which is exactly quadratic such as Q. The calculation is particularly easy if the starting value is taken at $\phi = 0$.

However, reality requires that constraints be imposed on the values of ϕ, e.g. that $\phi \geq 0$, and there is no guarantee that Newton's calculation will give such a result. The procedure that can be used in this case is then an iterative calculation, in four steps as follows.

1. Specify minimum values ϕ_{min} for ϕ and set $\phi_0 = \phi_{min}$ and $i = 0$.

2. Perform Newton's calculation as in equation (4) above deriving $\phi_{i+1} = \phi_i - Q'(\phi_i).Q''(\phi_i)^{-1}$.

3. Check whether all free values of $\phi_i \geq \phi_{min}$ and that Q' ≥ 0 for all constrained values of ϕ; if so, terminate.

4. Otherwise adjust any ϕ values that are less than ϕ_{min} to ϕ_{min}; free any ϕ values which are constrained and for which Q' < 0; set i = i+1 and repeat from Step 2.

This algorithm can be proved to converge to the overall optimum in a finite number of steps, because the set of constraints $\phi \geq \phi_{min}$ form a convex set while the function Q is concave. Each iteration of the algorithm gives a reduction in the value of Q. This theoretical result is however of limited value, because the number of steps might be quite large. If the number of categories is of the order of 50, as is commonly the case, the maximum number of steps could theoretically be 2^{50}, approximately 10^{15}. In practice, the number of steps turns out to be very limited: typically convergence is achieved in 5 or 6 iterations.

The values ϕ_{min} can in principle be chosen to be any (non-negative) limits that seem sensible, such as 10% of the frequency of each household category in the base sample. Their function is to prevent unusual, perhaps erroneous, target data from generating an impossible – or nearly impossible – future population distribution.

3.3 Discussion

It is clear from the presentations above that both the IPF and QUAD methods are somewhat arbitrary in the criteria that they adopt for optimality. No basis is put forward for the choice of criterion, although it could be expected that a statistical analysis of the likely errors in the base sample and the target data could lead to a more well-founded choice of criterion and could also probably illuminate the basis of specification of the weights in the QUAD method. This remains an area for future research.

Nevertheless, it is also clear that both methods are robust and unlikely to lead to totally unrealistic results. This finding is based not only on the simple forms of the methods but also on the practical experience that has been built up in their use.

In choosing between the methods, preference is given to QUAD because of its ability to deal with inconsistent inputs, the fact that it achieves a balance between base data and targets and its flexibility in adjusting the weights to meet different circumstances. For example, for short-term forecasting one would wish to remain close to the base household distribution, i.e. keep relatively low values of w; for the longer term it would be possible for the household distribution to change more and more attention should therefore be paid to the target forecasts, i.e. the values of w should be increased.

The programming implementation of QUAD is not complicated and runs very quickly on modern computers. The calculation of Q''^{-1} requires the inversion of a matrix of size 30-50, but this can be performed by an efficient algorithm in a fraction of a second.

4. EXAMPLES FROM PRACTICAL STUDIES

To illustrate the range of applicability of the QUAD method, examples are given from three practical studies.

4.1 Netherlands National Model

The first application of the QUAD method as described above was made for the Netherlands National Model (Daly and Gunn, 1985). For this application, approximately 1000 households were drawn from the National Travel Survey (OVG), about 0.02% of the total population. A categorisation into 120 groups was set up, based on household size, numbers of employed household members, the presence or not of a female worker and the age of the household head; however, it was found that 96% of the households fell into just 44 of these categories and ϕ values for these 44 were calculated for each of the 345 zones to adjust the initial weights which had been attached to the OVG records (and adjusted for the further sub-sampling to 1000 households). The ϕ_{min} constraints were set to zero. The target forecasts available gave 11 statistics for each zone: 8 age-sex population groups, employed persons by sex and the total number of households.

In practice, it was found necessary to divide the country into three areas (Amsterdam, the polders and the rest), which had very different population structures and which could not be accurately described by a single base distribution (the f values in equation 3). Further adjustments were made when inconsistencies were identified in the target totals (which were derived from various sources). In operation, the method proved robust and has been applied for a very wide range of planning scenarios and for years ranging from the very short term to the very long term.

4.2 Norwegian 'Climate' Model

For the Norwegian 'Climate' study, a model was developed of private travel for all modes in the entire country (TØI and HCG, 1990). As a base for the application of this model a prototypical sample was set up, containing 2942 households from the National Travel Survey of 1984/5, about 0.2% of all the households in Norway. A total of 17 'target' data items were forecast for each of the 454 *kommuner*: 10 age-sex population groups, total population, population in built-up areas, employed persons by sex, single-person and total households and the total disposable income. Categories were defined to divide the population into 30 groups on the basis of household size, the age of the household head and the employment status of the household head and (if present) partner. Because no detailed expansion data was available, 'naïve' initial expansion factors were defined simply to reproduce the total number of households in Norway. The ϕ_{min} limits were set to 20% of these naïve values, limits which proved rarely to operate in practice.

In operation, the QUAD method gave ϕ values yielding acceptable results for all the targets except income. In fact, it could not be expected that the substantial increases in income that were projected (more than 100% growth over the 40-year forecasting period) could be realised simply by shifts within the existing distribution of household incomes. There must also be an overall growth of income for **all** household types. For this reason, income was dropped as a target variable and the forecast income growth was implemented as an overall increase in the income of each household, once the other changes in population structure had been forecast. Essentially, the assumption was that the **shape** of the income distribution would remain as in the base year for each household type.

The Norwegian application thus followed explicitly the general lines set out in the Netherlands National Model, adapting as required to the local data structures.

4.3 Stockholm Integrated Model System

For this model system, known as SIMS, a prototypical sample of 3533 households was taken from the 1986/7 home interview survey, roughly 0.5% of the total population of the study area. These were divided into 51 categories on the basis of employment, household size, the presence of children and the licence holding of the heads of household. This last distinction was added because SIMS does not predict licence holding from a model, unlike the Netherlands and Norwegian models, but takes it as an exogenously determined input. It was again found to be adequate to set $\phi_{min}=0$. Three area types were defined in the area, for each of which a specific original distribution over household categories (f) was defined.

An important issue in this work was the form in which the target data was available. Excellent data was available, but for the fully detailed level of 850 zones only the numbers in each of 10 age-sex population groups could be provided. Information about the distribution of household size (three statistics) and employed persons by sex could be provided only at a more aggregate level of 40 zones; the total population was included at each level to maintain consistency. Because of this data structure, a two-stage procedure was adopted.

First, at the level of 40 zones, using the full set of 16 targets, the QUAD procedure was used to forecast a changed distribution of the households. Then, these 40 sets of household distributions were used as 'initial distributions' for a further optimisation at 850-zone level, using just the 11 targets that were available at that level. In this way account could be taken of the full content of the data available.

The results gave excellent agreement to the targets, particularly once inconsistencies between the target data and the household survey had been explained and partially corrected. These results were then used as the basis for the applications of the model, both for base-year and for future-year work.

5. SUMMARY and CONCLUSIONS

The objectives and methods of prototypical sampling have been set out and illustrated with some information from practical applications.

Prototypical sampling is a method for generating samples that are as far as possible representative for 'target' areas or years for which no 'native' sample is available. It relies on the existence of a sample for another area or year and some information about aggregate expectations for the target areas.

Two methods are commonly used for calculating the appropriate frequencies with which households should appear in the prototypical sample. The IPF method exploits iterative factoring to match the aggregate totals exactly, while the QUAD method uses quadratic minimisation to balance the divergence from those totals against the changes from the original distribution over household types.

While neither method offers a rigorous statistical basis for its calculations, the QUAD method is generally to be preferred because of its greater flexibility and ability to accept inconsistent inputs. Either method can be considered to be reliable in practice and both have been applied several times.

The adaptability of the QUAD method was illustrated by reference to three major modelling studies, with differing data availability requiring minor or major adaptations of the basic method. In each case the adapted method was able to reproduce the target aggregate statistics to an acceptable level of accuracy without diverging too far from the base distribution of the household types.

Prototypical sampling offers a good basis for creating samples to which disaggregate models can be applied, meaning that the efficient sample enumeration procedure can be used much more widely. It further contributes to forecasting by allowing account to be taken of shifts in the distribution of the population, which themselves can have significant impact on forecasts.

References

S. Algers, A. Daly, P. Kjellman and S. Widlert, "Stockholm Model System (SIMS): Application", presented to Seventh World Conference on Transport Research, Sydney, 1995.

R. J. Beckman, K. A. Baggerly and M. D. McKay, "Creating Synthetic Baseline Populations", *Transp. Res. A*, **30**, pp. 415-429, 1996.

M. Ben-Akiva and T. J. Atherton, "Methodology for Short-Range Travel Demand Predictions", *J. Transp. Econ. and Policy*, **XI**, 3, Sep. 1977.

M. Ben-Akiva, B. C. Kullman, L. Sherman and A. J. Daly, "Aggregate Forecasting with a System of Disaggregate Travel Demand Models", PTRC Summer Annual Meeting, 1978.

A. J. Daly, "Model Split Models and Aggregation", Local Government OR Unit *Transportation Working Note*, 1976.

A. J. Daly, "Policy Analysis using Sample Enumeration – an application to car ownership forecasting", PTRC Summer Annual Meeting, 1982.

A. J. Daly and H. F. Gunn, "Cost-Effective Methods for National-Level Demand Forecasting", IATBR Conference, Noordwijk, 1985.

A. J. Daly and S. Zachary, "The effect of free public transport on the journey to work", TRRL Report SR 338, 1977.

H. F. Gunn, "Artificial Sample Applications for Spatial Interaction Models", Colloquium Vervoersplanologisch Speurwerk, The Hague, 1984.

H. Gunn, P. Fisher, A. Daly and H. Pol, "Synthetic Samples as a Basis for Enumerating Disaggregate Models", PTRC Summer Annual Meeting, 1983.

TØI and HCG, "A Model System to Predict Fuel Use and Emissions from Private Travel in Norway from 1985 to 2025", Institute of Transport Economics and Hague Consulting Group, Report to Ministry of Transport, 1990.

P. Vrolijk, H. Gunn and T. van der Hoorn, "Waar komt de groei vandaan?" (What causes the growth?), presented to Colloquium Vervoersplanologisch Speurwerk, The Hague, 1987.