Multilevel models in the study of dynamic household structures[1]

by

Harvey Goldstein, Jon Rasbash, William Browne, Geoffrey Woodhouse

Institute of Education, University of London

and

Michel Poulain

FNRS – GéDAP, Université catholique de Louvain

Summary

A modelling procedure is proposed for complex, dynamic household data structures where households change composition over time. Multilevel multiple membership models are presented for such data and their application is discussed with an example.

Acknowledgements

This work was partly supported by the Economic and Social research Council (UK) under its programme for the analysis of large and complex datasets and other research grants. Our thanks are due to Daniel Courgeau, Jenny Jenkins and Tom O'Connor for helpful discussions.

Keywords

Complex data structures cross classification, event duration models, extended multiple membership model, fuzzy set, household structures, multilevel model, multiple membership model, repeated measures data.

[1] A version of this paper was presented at the European Population Conference; The Hague, Sept. 1999)

1. Introduction

Murphy (1996) discusses a key issue in the analysis of population household data. He points out that households are dynamic, changeable units whose definition over time is problematical. In longitudinal or panel studies of household composition and its influence on individuals, these dynamic structures raise difficult issues of interpretation: the number of possible household structures over time is extremely large and difficult to summarise.

He suggests that a fruitful way of looking at this is from the point of view of the individual, where an individual, such as a student, may 'belong' to several different household units their characteristics shared among these units. Murphy makes an analogy here with ‘fuzzy set’ theory and this is one theme we shall develop below.

Our main objective is to extend Murphy’s discussion by positing explicit statistical models for studying dynamic household structures which allow many of the difficulties to be resolved within a formal framework whose complexity of structure matches that of the system being studied. The class of models we use are known as ‘multilevel’ or ‘random coefficient’ models which have been developed since the mid 1980s so that they are now capable of dealing with a wide variety of data structures including those generated by the dynamics of household composition. These models allow for both continuous and discrete responses as well as for hierarchies and crossings of units at any level of a data hierarchy such as the individual or the family. They can handle data as repeated measurements or as durations as in event history models.

We begin by describing the basic multilevel model and follow this with various extensions of increasing complexity showing how these can be used to describe household dynamics.

2. Population structures

Human populations are structured in complex ways, but particularly exhibit hierarchical groupings, whereby, for example, individuals are grouped within households. A

household is defined as a group of individuals with or without family links living together in the same dwelling. When data from populations are modelled it is important to take account of such structures if the data values are related to them. Thus, for example, individuals within a household tend to be more alike in terms of attitudes and behaviours than individuals from different households. Failure to take account of such structures can lead to incorrect inferences. In addition the properties of such structures and their influences on responses are important to understand and hence to build into statistical models. Multilevel models attempt to do this by explicitly incorporating information about population structures into the model and estimating associated parameters.

In addition to such hierarchies population units can be cross-classified, for example a child will generally belong to both a particular school and a geographical neighbourhood. Likewise, individuals may belong to a household and to one geographical neighbourhood where they live and another where they work. A more complex structure arises where individuals may belong to more than one unit of the same type, for example individuals may work in more than one location. An important case occurs in longitudinal studies where individuals may pass from one unit to another, such as children who change schools or individuals who move between households. We will develop these two examples later. Since the mid 1980s the methodology for model specification and fitting has developed steadily and it is now possible to fit all these kinds of model. In the next section we will briefly review the basic theory. Following this we will show how more complex models can be specified and then discuss some applications.

3. The basic multilevel model.

For simplicity consider a simple data structure where a response is measured on individuals in a number of areas, together with one or more covariates. Instead of areas we could think of households, schools, etc. We wish to model a relationship between the individual response and the explanatory variables, taking into account the possibility that this relationship may vary across areas. The response might be a continuous variable such as income or survival time, or a discrete variable such as a voting preference or death. We shall assume in what follows that we are dealing with a continuously distributed response, and for simplicity that this has a Normal distribution. Extensions to other kinds

of responses follow similar lines and these are discussed by Goldstein (1995) and we briefly refer to one such extension, an event duration model, in our discussion of the example. We shall refer to the areas as higher level units and individuals as lower level units. In the present case we just have two levels with areas as level 2 units and individuals as level 1 units. A simple such model can be written as follows

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{ij} \\ \mathrm{var}(e_{ij}) &= \sigma^2_{e0} \\ \mathrm{var}(u_j) &= \sigma^2_{u0} \end{aligned} \qquad (1)$$

where y_{ij} is the response and x_{ij} the value of a single explanatory variable (covariate) for the *i*-th individual in the *j*-th area. The slope coefficient β_1 is for the present assumed to be the same for all the areas while the random variable u_{0j} represents the departure of the *j*-th area's intercept from the overall population intercept term β_0. The first two terms on the right hand side of (1) constitute the fixed part of the model and the last two terms describe the random variation. As mentioned we shall develop the model initially assuming that the random variables have a (multivariate) Normal distribution. This model could be viewed as a standard analysis of covariance if we treated each u_{0j} as a fixed parameter to be estimated. Such a model however will often be inappropriate, for the following reasons.

First, we may have a very large number of areas, leading to a very large number of separate parameters to estimate. Secondly, some of the areas may have very few individuals, so that their individual departures will be poorly estimated. Most importantly, we may be interested in treating the areas as a sample from a *population* of areas and wish to make general inferences about the likely behaviour of other areas in this population rather than, or in addition to, providing separate estimates for each area in the sample. For all these reasons it will usually be more appropriate to regard u_{0j} as random and to write

$$u_{0j} \sim N(0, \sigma^2_{u0}) \qquad e_{0ij} \sim N(0, \sigma^2_{e0})$$

Note, however, that we are also at liberty to 'fix' one or more of the u_{0j} using an associated dummy (0,1) variable as an explanatory variable, for example if we knew that it was special and should not be considered as a member of the same population as the remainder. This is often useful for exploring 'outliers' (Langford and Lewis, 1998). We can elaborate (1) by allowing the coefficient β_1 to vary across areas and rewrite the model in the more compact form

$$
\begin{aligned}
&y_{ij} = \beta_{0ij} x_0 + \beta_{1j} x_{1ij} \\
&\beta_{0ij} = \beta_0 + u_{0j} + e_{ij} \\
&\beta_{1j} = \beta_1 + u_{1j} \\
&U = \{u_{0j} u_{1j}\}^T \quad E(U) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{cov}(U) = \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}, \quad \text{var}(e_{ij}) = \sigma^2_e
\end{aligned}
\tag{2}
$$

This model is often referred to as a 'random coefficient model' by virtue of the fact that the coefficients β_{0ij} and β_{1j} in the first equation of (2) are random quantities, each having a variance with a covariance between them. As more explanatory variables are introduced into the model, so we can choose to make their coefficients random at the area level thereby introducing further variances and covariances, and this will lead to models with complex covariance structures. One of the aims of multilevel modelling is to explore such potential structures and also to attempt to explain them in terms of further variables.

Having fitted such a model we can obtain estimates for the individual 'residuals' $(u_{0j}, u_{1j}, e_{0ij})$ at either level by estimating their expected values (or other functions of their distributions), given the data and model estimates. Thus, for example, we can estimate $E(u_{0j}, u_{1j} | Y, \beta, \theta)$ where

$$
\beta^T = \{\beta_1, \beta_2\} \quad \theta = \{\sigma^2_{\text{u0}}, \sigma_{u01} \sigma^2_{u1}, \sigma^2_{e0}\} \tag{3}
$$

and substituting model estimates for the unknown parameters. The multilevel model is here described in non-Bayesian terms. For a full Bayesian specification of this model we would need to add prior distribution assumptions for the parameters in (3). The interested reader is referred, for example, to Rasbash et al. (2000) for details with worked examples.

These procedures are all implemented in the software package *MLwiN* (Rasbash et al., 2000).

4. Cross classification of units

Across a wide range of disciplines it is commonly the case that data have a structure which is not purely hierarchical. Subjects may be clustered not only into hierarchically ordered units (e.g., students nested within classes, within schools), but may also belong to more than one type of unit at a given level of a hierarchy. In this exposition we use educational data to illustrate our ideas and we shall then apply the models to population structures. Thus, a student might be classified as belonging sequentially to a particular combination of primary school and secondary school, in which case the student will be identified by a *cross classification* of primary schools and secondary schools. Alternatively, a particular student may spend a proportion of time in one school and the remaining proportion in another school. In this case, the student has *multiple membership* of units at a given level of clustering.

Raudenbush (1993) and Rasbash and Goldstein (1994) present the general structure of a model for handling complex hierarchical structuring with random cross classifications. For example, assuming that we wish to model the achievement of students taking into account both the primary and the secondary school attended by each student, then we have a cross classified structure, which can be modelled as follows:

$$\begin{aligned} &y_{i(j_1 j_2)} = (X\beta)_{i(j_1 j_2)} + u_{j_1} + u_{j_2} + e_{i(j_1 j_2)}, \\ &j_1 = 1,\ldots J_1, \quad j_2 = 1,\ldots J_2, \quad i = 1,\ldots N \end{aligned} \tag{4}$$

in which the score of student *i*, belonging to the combination of primary school j_1 and secondary school j_2, is predicted by a set of fixed coefficients $(X\beta)_{i(j_1,j_2)}$. The random part of the model is given by two level 2 residual terms, one for the primary school attended by the student (u_{j_1}) and one for the secondary school attended (u_{j_2}), together with the usual level 1 residual term for each student. We note that the latter may be further modelled to produce complex level 1 variation (Goldstein, 1995, Chapter 3).

5. The multiple membership model

Considering now just the secondary schools, suppose that we know, for each individual, the weight w_{ij_2}, associated with the j_2-th secondary school attended by student *i* with $\sum_{j_2=1}^{J_2} w_{ij_2} = 1$. These weights, for example, may be proportional to the length of time a student is in a particular school during the course of a longitudinal study. Note that we allow the possibility that for some (perhaps most) students only one school is involved so that one of these probabilities is one and the remainder are zero. Note that when all level 1 units have a single non-zero weight of 1 we obtain the usual purely hierarchical model. We can write for the case of membership of just two schools {1,2}:

$$\begin{aligned} &y_{i(1,2)} = (X\beta)_{i(1,2)} + w_{i1}u_1 + w_{i2}u_2 + e_{i(1,2)} \\ &w_{i1} + w_{i2} = 1 \end{aligned} \qquad \text{(5a)}$$

and more generally:

$$\begin{aligned} &y_{i\{j\}} = (X\beta)_{i\{j\}} + \sum_{h\in\{j\}} w_{ih}u_h + e_{i\{j\}} \\ &\sum_h w_{ih} = 1, \quad var(u_h) = \sigma_u^2 \\ &var(\sum_h w_{ih}u_h) = \sigma_u^2 \sum_h w_{ih}^2, \end{aligned} \qquad \text{(5b)}$$

Thus, in the particular case of membership of just two schools with equal weights we have

$$w_{i1} = w_{i2} = 0.5, \quad \text{var}(\sum_h w_{ih}u_h) = \sigma_u^2 / 2$$

Where a student does not belong to a school the corresponding weight is zero. Thus (5) is a 2-level model where the level 2 variation among secondary schools is modelled using the set of weights for each student across all schools as explanatory variables. A similar formulation can be used to model the case where, for some students, there is no identification of the school(s) to which they belong. If we are able to assign a set of probabilities of membership among a subset of schools, however, then utilising the

(square root) of these probabilities as weights (standardised to sum to 1) we can still carry out a valid analysis (Hill and Goldstein, 1998).

An extension of (5) is also possible and has important applications, for example in modelling spatial data. In this case we can write

$$
\begin{aligned}
&y_{i\{j_1\}\{j_2\}} = (X\beta)_{i\{j\}} + \sum_{h\in\{j_1\}} w_{1ih}u_{1h} + \sum_{h\in\{j_2\}} w_{2ih}u_{2h} + e_{i\{j\}} \\
&\sum_h w_{1ih} = W_1, \ \sum_h w_{2ih} = W_2, \quad var(u_{1h}) = \sigma^2_{u1} \quad var(u_{2h}) = \sigma^2_{u2} \\
&cov(u_{1h}, u_{2h}) = \sigma_{u12}, \qquad j = \{j_1, j_2\}
\end{aligned}
\tag{6}
$$

There are now two sets of higher level units which influence the response. In spatial models one of these sets is commonly taken to be the area where an individual (level 1) unit occurs and the other set consists of the neighbouring units which have an effect. The total weights for each set will need to be carefully chosen; in spatial models the W_1, W_2 are typically chosen each to equal 1 (see Langford et al, 1999 for an example). Another application of such a model for household data is where households share facilities, for example an address. In this case the household that an individual resides in will belong to one set and the other households at the address will belong to the other set. We can readily extend (6) to the case of multiple sets - which we refer to as the 'extended multiple membership model' – and this will allow us additionally to incorporate multiple spatial structures into household models.

In terms of households we can have two kinds of multiple membership. The most frequent is the case where, as in the case of students changing schools, an individual sequentially moves from one household unit to another. But we can also observe individuals who alternate between households and may be considered as simultaneously belonging to more than one[2]. . Both these cases can be dealt with, and combined together, by using weights which reflect time spent within each household.

Multiple membership models bear a close relationship to fuzzy sets (see for example, Manton et al, 1994, for an introduction, and Haberman, 1995 for a critique) where

[2] This is the case for example for individuals aged 60 and over who are living alone part of the week but cohabitate with a partner during the second part of the week. The same may be observed for young people leaving home progressively.

individual units also can belong to several groups at a time, with 'membership coefficients' being equivalent to our weights. There appears to be no explicit application of fuzzy set theory, however, to general hierarchical structures.

We now look at ways in which we can utilise the models introduced so far, in a systematic way to describe population structures of some complexity.

6. Dynamic household composition

Many different kinds of complex population structures exist. We shall describe in detail the specification and analysis of one of these which is of particular interest and look at other examples in a final discussion section.

In studies which follow households over time we shall consider that a household is defined at a given time solely by its composition, that is all individuals living together in the same dwelling Due to individuals leaving existing households to enter others or to form new households, household composition is changing all the time. To simplify our approach we shall consider that all changes of composition of an existing household result in a new household composition and thus in a new household. According to this definition a birth or a death will produce a new household. The same is true for all migrations, except if the whole household is moving together at the same time. If the existence of a household per se is assumed to influence the individual measurements of interest, e.g. attitudes or behaviours, then a particular individual will be expected to acquire influences from all the households that they 'belong' to during the course of a study. These households will be differentiated by their type (e.g. in terms of the numbers and ages of children within them, income or the number and types of adults) which will also be expected in general to influence measurements. The following model illustrates a simple case where individuals are measured regularly with information available on how long they spent in each household.

In this formulation the total sample of households is defined as the ‘superset’ consisting of the union of the samples which exist over time. Thus, for example, a young person may leave a household to start another household with other young people and this new household will be added to the total sample of households. In general we would wish to

measure all the people in this new household, because this will then enable us properly to characterise that household and model its influence. Let us assume that the practical problems associated with an increased burden of measurement can be overcome. To see how we might model this structure consider the following study.

We have a sample of households at two occasions. At the first occasion all the household members are measured in the set of households H_1. Some of these households ($H_{1,1}$) remain intact, while at the second occasion, among the total second set of households (H_2), some of the occasion 1 households ($H_{1,2}$) have amalgamated, or lost or acquired one or more members. Some of these new households will include people who were not present in the sample at the first occasion or were present but the identification of their household is unknown ($H_{2,1}$). Others of these new households will include people who have been present in different but known households at the first occasion ($H_{2,2}$). There may also be households at occasion 2 which contain no members who were in occasion 1 households ($H_{2,3}$).

Assume that the response variable (y) of interest is continuous, e.g. individual income, and we wish to model y at the second occasion as a function of its value at the first occasion, together with person characteristics. Household characteristics are specific to each occasion and we assume that such measurements are available – we shall discuss later a special case where they are not. A simple model, using (5a) and generalising the notation can be written as

$$\begin{aligned} & y^{(2)}_{i(j_1,j_2)} = \alpha y^{(1)}_{i(j_1,j_2)} + (X\beta)_{i(j_1,j_2)} + w_{ij_1}u_{j_1} + w_{ij_2}u_{j_2} + e_{i(j_1,j_2)} \\ & w_{ij_1} + w_{ij_2} = 1 \\ & \text{var}(u_{j_1}) = \text{var}(u_{j_2}) = \sigma^2_u, \quad \text{var}(e_{i(j_1,j_2)}) = \sigma^2_e \end{aligned} \tag{7}$$

where j_1, j_2 index the households at occasions 1 and 2 and the superscript also refers to the occasion at which the measurement is made and we have a common between-household variance. This model assumes that a person belongs to at most 2 households, but can be extended to the multiple household case using (5b). For the set $H_{1,1}$ one of the weights is zero since only one household is involved. For the set $H_{2,2}$ each person will

have two weights. Since the analysis is conditional on the first measurement, a reasonable choice is to make them proportional to the time spent in each household between occasions 1 and 2. Other choices are possible, for example giving relatively more weight to the most recent household. In other kinds of model, such as a repeated measures model we might choose weights proportional to time measured from an origin prior to the first measurement and care will be needed in making a choice.

The major problem is set $H_{2,1}$ for which we will have generally no data at occasion 1 and no identification of the household either; similar issues arise for the case of completely new households at occasion 2. These individuals will provide relevant information when the response variable measured on the individual present at time 1 is influenced by the characteristics of the other household members in set $H_{2,1}$. This will be particularly important when these characteristics are time dependent, for example, in the case when, say, a change in income between time periods occurs. In some cases the missing data could, in principle, be handled by a suitable imputation procedure (see for example Goldstein and Woodhouse, 1996) and in this situation the assumption of completely missing at random will generally be reasonable.

If the identification is unknown but we do know that the households actually belong to H_1, then we can assign a *probability* of belonging to one of the first occasion households and use the procedure described by Hill and Goldstein (1998) for estimation. For those who entered the sample for the first time at occasion 2, this will generally not be possible and the following procedure can be used.

We assume that for each of these individuals, we know or can estimate the weights w_{i1}, w_{i2}, possibly using the mean weights derived from those sample members with known weights. We also assume that they come from distinct households, although if there is information that some come from the same household this can be incorporated. We now write (6) as

$$y^{(2)}_{i(j_1,j_2)} = \alpha y^{(1)}_{i(j_1,j_2)} + (X\beta)_{i(j_1,j_2)} + (1-\delta_i)(w_{ij_1}u_{j_1} + w_{ij_2}u_{j_2}) + \delta_i(w_{ij_1}u^*_{j_1} + w_{ij_2}u_{j_2}) + e_{i(j_1,j_2)}$$
$$w_{ij_1} + w_{ij_2} = 1 \qquad (8)$$

where δ_i is 1 if a person belongs to $H_{2,1}$ and zero if not. The random effect $u^*_{j_1}$ is specific to the set $H_{2,1}$ and if we assume that it comes from a population with the same characteristics as H_1 will have a variance constrained to be equal to σ^2_u. To fit this model we use the same device as in the general multiple membership model, for $u^*_{j_1}$, defining a set of dummy variables with coefficients random at the highest level whose variances are constrained to be equal.

As an alternative to imputation for the missing occasion 1 variable we can write a modified version of (8) for the members of $H_{2,1}$ as

$$y^{(2)}_{i(j_1,j_2)} = (X\beta)^*_{i(j_1,j_2)} + \delta_i(w_{ij_1}u^*_{j_1} + w_{ij_2}u^*_{j_2}) + e^*_{i(j_1,j_2)}$$
$$w_{ij_1} + w_{ij_2} = 1 \qquad (8a)$$

Where there are households at occasion 2 with members of both $H_{2,1}$ and $H_{2,2}$ we have u_{j_2}, $u^*_{j_2}$ from the same household and this therefore allows us to estimate the correlation between these two random terms and thus provides an efficient modelling procedure.

In practice, and especially if the numbers are small, the set $H_{2,1}$ can be omitted from the analysis. While this will reduce statistical efficiency it will not lead to biases if we can assume that the joint distribution of the characteristics of this set is the same as the remainder of the sample. In practice, however, this may not be reasonable since one might expect the less stable households to have different characteristics. For example, suppose an explanatory variable is measured at the household level and is an aggregation of individual characteristics. If we use the average income at the first occasion of all the second occasion household members as a predictor and if we exclude all the households that contain members of $H_{2,1}$ we may introduce biases, since these will tend to be the more volatile households. Instead we can use the aggregate measure based just upon those individuals for whom it is available; this will then constitute an explanatory variable measured with error, where the error variance is known or can be estimated. Woodhouse et al (1996) discuss how to handle such models.

The extension to the case where individuals may belong to more than 2 households is relatively straightforward, so long as the multiple weights are available, although complexities will be introduced where multiple households are involved and some identifications are missing. We now turn to the interpretation of the results of these kinds of analyses.

7. Further elaborations

The results of an analysis such as that of model (7) will yield the following parameter estimates: a between-household variance, a between-person variance and a set of fixed coefficients representing the regression of the response on its previous value and covariates measured at either occasion. Such a model, however will generally not capture the full complexity. Thus, for example, the variation among some kinds of household members may differ, perhaps according to age. Likewise, the variation among households may differ according to household or individual characteristics. Age again may be important, or educational and socio-economic level. We can introduce random coefficients to accommodate such structures. Thus, we may find that the between-household variance is an increasing function of age – that the average income, say, of younger persons within a household varies less than the average income of older persons. Or we may find that the variation in attitudes between persons within a household increases with educational level. Such findings and the possibility of explaining such variation by incorporating further covariates, will often be as important as the values of the fixed (regression) coefficients. In longitudinal repeated measures studies which study trends in, say, opinions with respect to time, the comparative stability of such trends according to individual or household characteristics can readily be studied by these complex models.

In some situations we may wish to define certain patterns of multiple membership as household types in their own right. For example there may be pairs of households which exchange members on a regular basis, say a child spending part of their time with a mother and part with a father living apart. A pair of such households may be more usefully viewed as a single household in terms of its members rather than as two separate households with regularly changing membership.

In some cases we may also wish to relax our strict definition of a household unit as being composed of a given set of individuals. Thus, for certain kinds of event, we may choose not to define a new household, but simply to record a changing characteristic which will appear in our model. We might wish to do this for, say, the death of a particular kind of person or the birth of a child. In such cases, however, no other household is created or destroyed, and we cannot adopt such a procedure within our modelling framework when this occurs, for example if we try to define households solely in terms of their 'head'

Finally we can introduce multiple, correlated, responses giving multivariate models, These can be incorporated within the same framework as above using the general procedures described by Goldstein (1995, Chapter 4).

8. Example

The data are taken from the population of Charleroi, Belgium (Population Register) where, between the 1st January 1995 and the 1st January 2000, 65,000 individuals who lived in a set of selected addresses within the town were followed. For each individual, over a 5 years period, there is information on their household membership. For present purposes the total duration is divided into 10 semesters and the household membership recorded at 11 occasions, every six months. In addition there is information about marital status, household position, gender and nationality. The household address is also recorded and in some cases there are several households living at one single address. In principle, sharing an address may be an influential factor and address could be incorporated within an extended multiple membership model as described above. For simplicity, we shall ignore this possibility in the following analysis.

Interest centres on the length of time a household survives intact. Once one or more individuals leave a household it ceases to exist and new households are formed. The length of time that is spent in a specific household composition can be expected to be a function of individual and household characteristics, including factors such as age and sex of individuals, size and age distribution within the household. Each individual, during the course of the study, will be part of one or more households for varying lengths of time so that the basic data structure is that of repeated (within) individual measures, with individuals belonging to one or more households, that is, a multiple membership model of

the type described above. If we simply measure, for each occasion, the duration length for an individual in a specific household, then we could use this as the response in our model. A more efficient alternative is to use an event duration model where the probability of belonging to a new household composition is modelled as a function of time spent in the household (see discussion). In both cases, for our data, there will be a slight underestimation of duration length since, for those individuals present during the first period it is not known how long they have been at that address. A serious difficulty for present purposes, is that there is confounding between individual and household. Thus, during the time an individual belongs to a particular household composition, all individuals in that household by definition will have the same value for the response variable. If one person leaves the household the composition will change for all members of the original household and the length of time spent by all members in that household is strictly the same. This implies that once household variation is incorporated into the model there will be no between-individual variation, and this is in fact what happens when the model is fitted. This will not be the case if we choose other responses like voting preferences, attitudes, individual income… for which all members of a given household will not have the same response.

In the present case therefore, to illustrate our models, we have chosen as the response variable the average duration of stay for an individual in all households up to and including the current one. Thus, assume an individual belongs to household A at time 1 (T1) and time 2 (T2) and to household B at T3, T4 and T5, and we assume her to have spent 0.75 years in the first household (A) and 1.25 years in the second (B). In this case, the first response value is 0.75 and the second is 1,0. During the first occasion, ending at (T2) they will be designated as belonging to household A only with a weight of 1.0 and at the second occasion, ending at (T5) will be designated as belonging to households A and B, each with weight 0.5, etc. Other weighting systems are possible, for example giving less weight to previous households. We note that this is a somewhat extreme example in that, while there is no longer complete confounding between individual and household, there remains a strong association since the average durations for all individuals within a household includes a component which is the same *current* household duration length.

For this reason we will expect the between-individual variation to be smaller than that between households.

The model is as follows

$$y_{ij\{k\}} = (X\beta)_{ij\{k\}} + u_j + \sum_{h \in \{k\}} w_{ijh} v_h + e_{ij\{k\}}$$
$$\sum_h w_{ijh} = 1, \quad var(u_j) = \sigma_u^2, \; var(v_h) = \sigma_v^2 \qquad (9)$$

where *i, j, k* refer to the levels occasion, individual and household. Since there is no repetition within the cells of the classification of individuals by households, level 1 variation represents that which is not accounted for by the multiple membership structure.

The distribution of average duration is skewed and we might wish to transform the data, for example using Normal scores, but for presentational purposes we remain with the original scale; in fact a Normal score analysis produces very similar general inferences.

Table 1 shows the results from three separate analyses using different predictor variables.

Table 1. Multiple membership model for average (cumulative) duration of stay			
	Model A	**Model B**	**Model C**
Fixed	**Estimate (s.e.)**	**Estimate (s.e.)**	**Estimate (s.e.)**
Intercept	1.438 (0.0078)	1.626 (0.0086)	1.494 (0.0116)
Gender		0.0058 (0.0046)	-0.0072 (0.0048)
Size of household		-0.0814 (0.0012)	-0.0774 (0.0012)
Age – 30 years		0.0034 (0.00014)	0.0060 (0.00020)
Spouse – head of house			0.0536 (0.0042)
Child – head of house			0.0822 (0.0052)
Is married			0.0684 (0.0054)
Is Belgian nationality			0.0246 (0.0074)
Random			
Between household variance	1.46 (0.0132)	1.390 (0.0128)	1.364 (0.0124)
Between individual variance	0.140 (0.0016)	0.128 (0.0012)	0.128 (0.0012)
Residual (level 1) variance	0.0108 (0.00012)	0.0100 (0.00012)	0.0100 (0.00012)
-2 * Log-Likelihood	84660.7	79742.5	79159.8

We see that, as expected, the between–individual variance is much smaller than that between households, about 10%. If an individual in fact moves between 10 different households over the whole period then the household contribution to the variance should be equivalent to the between – individual variance as is the case, for the household contribution is $\sum_{h=1}^{10} w_h^2 \sigma_v^2 = \sigma_v^2 / 10 \cong \sigma_u^2$.

We also see that age is associated with a longer duration and household size with a shorter, with little gender difference. Heads of households tend to have shorter durations than spouses and children have the longest. Married people have longer durations and being of Belgian nationality is associated with longer durations. We have not studied any interactions among these variables.

9. Conclusions

Other kinds of complex structures exist in real populations and in principle the methods we have described can be extended to these situations. One example is the study of the way in which extended family structures influence individual characteristics, such as physical or attitudinal ones, and here our models overlap and extend some of those used in genetics (see for example, Sham, 1998). A separate paper using these models is in preparation.

We have shown how the use of multiple membership models provides a powerful procedure for describing complex population structures, including those of a dynamic kind where these change over time. The use of these models provides greater efficiency in analysis and also allows structures to be defined and explored which are difficult or impossible to handle with conventional techniques. It needs to be stressed, however, that in practice we may expect to encounter difficulties when fitting complex models, both in terms of obtaining satisfactory numerical convergence and interpreting results. Likewise, the weights are predetermined and the choice of weights is clearly important and in practice it will often be useful to try different weighting systems and observe their effects.

Murphy (1996) points to the need to be able to model the interrelationships among individuals in a household and suggests that network models may be appropriate. In the models discussed in the present paper, such relationships among individuals within households are often implicit and modelled via the correlation structures within and between households. For example if we are studying the relationship between income and a political attitude variable then we could choose these both as response variables. Fitting a bivariate response at both individual and household level would allow us to study the correlation between these measures at individual level *within* households and to see how far this might be explained by further covariates such as age. We could further allow for different associations among different kinds of household members, as in the examples we have discussed.

While our exposition has been in terms of Normally distributed responses, binary and other discrete outcomes can be handled. The models can also be extended in straightforward ways to accommodate multivariate responses including cases of mixtures of continuous and discrete data (Goldstein, 1995) and to weights applied to units at different levels of a hierarchy.

An important application of multilevel models is to event history data (Goldstein,1995, Chapter 9). Thus, we may wish to model the time that each household member is in employment and to see whether there are distinguishable household effects. In our example we have seen that care needs to be exercised since separating effects may not be possible when the durations for the different kinds of units are partially or totally confounded.

If the models we have discussed are to be applied there are important implications for the type and extent of data needed. For studying changing household compositions, longitudinal or panel data on individuals and households are required with detailed information on transitions as in our example. Such data are difficult to gather and the use of administrative population registers is one possibility.

We believe that the approach in this paper resolves, at least in principle, the long standing issue about how dynamically changing households are to be defined as units. By generalising the definition of a household unit we are able fully to model the complexity arising from changing structures.

References

Goldstein, H. (1995). *Multilevel Statistical Models*. London, Edward Arnold: New York, Wiley.

Goldstein, H. and Woodhouse, G. (1996). *Multilevel models with missing data.* Eleventh International workshop on statistical modelling., Orvieto, Italy.

Haberman, S (1995), Review of *Statistical applications using fuzzy sets*, J. American statist. Assn., 90, 1131-1133.

Hill, P. W. and Goldstein, H. (1998). Multilevel modelling of educational data with cross classification and missing identification of units. *Journal of Educational and Behavioural statistics* **23**: 117-128.

Langford, I. and Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society, A.* **161**: 121-160.

Langford, I., Leyland, A., Rasbash, J. and Goldstein, H. (1999). Multilevel modelling of the geographical distribution of diseases. *Journal of the Royal Statistical Society, C.* **48**: 253-268.

Manton, R. G., Woodbury, M. A., and Tolley, H. D. (1994). *Statistical applications using fuzzy sets.* New York, Wiley.

Murphy, M. (1996). The dynamic household as a logical concept and its use in demography. *European Journal of Population,* 12, 363-81.

Rasbash, J. and Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model. *Journal of Educational and Behavioural statistics* **19**: 337-50.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., et al. (2000). *A user's guide to MlwiN (Second Edition).* London, Institute of Education.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross sectional and longitudinal research. *Journal of Educational Statistics* **18**: 321-349.

Sham, P. (1998). *Statistics in human genetics*. London, Arnold:

Woodhouse, G., Yang, M., Goldstein, H. and Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society, A.* **159**: 201-12.