MULTILEVEL ANALYSIS IN THE SOCIAL SCIENCES

Daniel COURGEAU* and Brigitte BACCAÏNI*

Like other social scientists, the demographer can study human behaviour at various levels of aggregation. Once a particular level of aggregation has been selected, however, it becomes difficult to change it in the course of the analysis, because the measures, methods and techniques used in one field will not match those used in another.

Demography has for long favoured analysis at aggregated levels. This is possible under the hypothesis – often not stated explicitly – that we are working on sub-populations, each of which is homogeneous with respect to the behaviour being studied. In this case there is little point in considering individual behaviour and characteristics; attention instead focuses on identifying the relations which exist between the classic demographic rates, corresponding to the phenomenon being studied in each sub-population, and the average values of the characteristics also calculated for each sub-population. An analysis of the emigration rates of different regions, for example, would try to link them to the unemployment rates, average incomes, percentage of dependants, etc., found in these regions. This approach is the basis of the usual migration models, which although developed long ago (Young, 1924) are still in use (Puig, 1981; Jacquot, 1994). In the same way it is possible to elaborate regional models of fertility or mortality. The multiregional models of Willekens and Rogers (1978) provide a synthesis of these different approaches under the additional hypothesis of independence between the different demographic phenomena.

Such an analysis can be said to make possible an examination of the effect that the groups being studied have on their own demographic behaviour: the aggregated characteristics that can be measured are then interpreted as being a set of constraints that each sub-population imposes on its members and which will influence their behaviour. So for example such

* Institut national d'études démographiques.

Population: An English Selection, special issue *New Methodological Approaches in the Social Sciences*, 1998, 39-71.

an analysis can reveal a positive association between the rate of unemployment in a region and its emigration rate (Puig, 1981). There is a real danger of concluding from this result that individuals who are unemployed have a higher probability of emigrating from a region, whereas all that is in fact known is that a high rate of unemployment is accompanied by a high rate of emigration, whether the individuals involved be economically active, unemployed or inactive. This type of mistaken inference leads to what is known as the *ecological fallacy*, which occurs when we try to infer individual behaviour from aggregated measures.

It is now more than forty-five years since Robinson (1950) explored these problems using statistical techniques. He showed that the correlations between two characteristics were in general not the same, depending on whether they were measured at the individual level or as proportions in the regions. Thus for the population of the United States in 1930 the correlation between the proportions that were black and illiterate was 0.95 when calculated at the level of nine geographical divisions, whereas for an individual the correlation between the fact of being black and of being illiterate was only 0.20. The conclusion reached by Robinson was unequivocal: an ecological correlation, measured from aggregate data, is not a substitute for an individual correlation.

A number of authors have extended these findings to analyses employing linear and logistic regression techniques (Alker, 1969; Firebaugh, 1978; Piantadosi *et al.*, 1988; Courgeau, 1995; Baccaïni and Courgeau, 1996a). The conclusion is always the same: in the majority of cases, aggregate data analysis is responsible for bias when we wish to make individual-level statements. These biases will be all the greater when the within-area variance for each characteristic is greater than the between-area variance.

This means that analysis must also be conducted at the individual level if we wish to understand human behaviour. It was in response to this need that the event history approach began to be developed more than fifteen years ago. This has led to the elaboration of surveys that collect data about the events that occur in every area of an individual's life, with their precise dates, as well as to the development of new analytical techniques, suitable for linking the different events which can occur in different domains and for measuring the effect of different individual characteristics on these events. Finally, it has been responsible for establishing a new paradigm in demography (Courgeau and Lelièvre, 1997), since attention is no longer directed at homogeneous sub-populations and events that are independent of each other, but instead on the entire individual life history, which is treated as a complex stochastic process. This new paradigm can be expressed by the following hypothesis: in the course of his life, an individual has a complicated trajectory which at any given point in time is dependent on his previous itinerary to date, the information he has been able to accumulate in the past and the conditions prevailing in the society of which he is a member.

The aim in this case is to identify the relationships which exist between a complex individual behaviour that is time-dependent, and various characteristics of these individuals. These characteristics may be fixed once and for all (such as social origin of parents, number of brothers and sisters, place of birth and birth rank) or be time-dependent characteristics which indicate the major stages in the life-course. It is clearly the heterogeneity of populations which is involved here, and the interactions between the different demographic phenomena are central to the analysis. An analysis of the probability of emigrating from different regions, for example, will include the fact that an individual is unemployed or not, his income, number of dependants, and so forth. It is also possible to introduce other, more fixed characteristics such as place of birth, so as to be able to measure the chances of returning for an individual who has left.

This approach successfully combines a number of techniques used in sociology (Tuma and Hannan, 1984), economics (Lancaster, 1990) and demography (Courgeau and Lelièvre, 1989; 1992). Care is needed, however, given that individual behaviour is often considered to be influenced only by the characteristics of the individual or of the next of kin (members of the household or coresident family, for example). The danger here is of committing the *atomic error*, since no attention is paid to the context in which human behaviour occurs. In point of fact, this context certainly does have an influence on individual behaviours and it seems fallacious to consider individuals in isolation from the constraints imposed by the society and milieu in which they live.

This has given rise to the idea of working simultaneously on different levels of aggregation, with the aim of explaining a behaviour which is always treated as individual, rather than aggregated as before. This removes the risk of ecological fallacy, since the aggregated characteristic is used to measure a construction that is different from its equivalent at the individual level. It is introduced not as a substitute but as a characteristic of the sub-population which will influence the behaviour of an individual who belongs to it. And the atomic fallacy is also eliminated once the context in which the individual lives is correctly introduced into the analysis.

This possibility of multilevel or contextual analysis has previously occasioned numerous methodological debates in sociology (Lazarsfeld and Menzel, 1961; Hauser, 1974), but actual applications of these methods have been slow to demonstrate their potential. It is only recently that the use of data files specially designed for such analyses has enabled them to be applied in a variety of human sciences: epidemiology (Van Korff *et al.*, 1992), educational research (Goldstein, 1987, 1995), human geography (Jones, 1993), sociology (Entwistle and Mason, 1985), economics (Geronimus *et al.*, 1996), and demography (Courgeau, 1995; Baccaïni and Courgeau, 1996a). In what follows we shall show in more detail the aims, hypotheses, and methods used when conducting a multilevel analysis and the problems that arise.

I. – Setting up multilevel models

Multilevel analysis aims to study the individual processes which occur in a differentiated space. A range of individual and collective actions are responsible for the creation of spatial structures, such as employment areas, and administrative divisions such as communes and departments, which may change over time. The individuals living in these spatial units behave according to their own characteristics, but also according to the constraints imposed by the living conditions that are particular to each unit: levels of unemployment, average income, population density, presence of a school, and so forth. Thus it can be seen how the individual characteristics and the aggregated characteristics may influence the behaviour of individuals living in each zone in different ways.

It is important to recognize that such an approach is centred on the individual. It is by and through the individual that the various levels of aggregation exist, but this does not prevent the constraints imposed by these levels from causing the individual to adopt a behaviour different from what it would have been without these conditions.

Setting up such an analysis also requires us to distinguish the characteristics to be analyzed according to the level of aggregation being considered.

The characteristic to be analyzed will always be considered here as individual. It may be a binary characteristic: being married or not; a polytomous characteristic: being economically active and in employment, unemployed, or economically inactive; or a characteristic that can be treated as continuous: the individual's height, income, etc.

The explanatory characteristics can be more diverse. We might begin by introducing individual characteristics, like those described above. Next, for a given level of aggregation we could simply aggregate these individual characteristics and estimate the percentages or averages (Loriaux, 1989): the percentage of married individuals, the percentage of economically active in employment and unemployed, the average height of individuals in each spatial unit. More complex analytical procedures can also be applied: as well as average income we could simultaneously introduce the standard deviation of income, or the correlation between income and IQ in each region.

Other characteristics are more global and concern the units in their entirety: population density or number of hospital beds, for example. These do not correspond to any individual characteristic, but they can be aggregated at various levels: thus the number of hospital beds in a region is the sum of the number of beds of each of the departments in the region.

Other collective characteristics are well defined for a given level of aggregation, but cannot be aggregated at larger levels. The political orientation

of a commune, as defined by the party affiliation of its mayor, for example, cannot be aggregated with those of the neighbouring communes which may cover a broad spectrum. This characteristic does not exist at either the individual level, or at the departmental and regional levels.

Such an analysis requires the definition of the different levels at which observation is to take place, and of the ways in which the levels are organized in relation to each other.

The simplest and most widely used structure is hierarchical: individuals live in communes, which are themselves parts of departments, and so on. Each level is formed by the grouping together of all the units of the preceding level. The division employed may be administrative, as in the example above, or of a completely different kind: pupils grouped in classes, which in turn are in schools, which are themselves divided between the state and private sectors, and so on.

The relationship between the levels can be more complicated: individuals grouped in towns of ascending size order, but distinguishing also between administrative, industrial and tourist centres, for example. In this case there is a cross-classification, depending on whether the towns are classified by size or by function. It is of course possible to have relationships that combine hierarchical and cross-classifications. For example, individuals may be classified by type of residential neighbourhood and by the type of place of work (cross-classification), which are themselves placed in a hierarchical classification of departments and regions.

Having set out the aims and hypotheses as well as the types of characteristics to be considered and the various relationships between the possible levels of aggregation, we shall now go on to consider the methods of analysis.

II. – Effect of individual and aggregated characteristics on behaviour without random variation between regions

We begin by considering the simultaneous effect of individual characteristics and different levels of aggregation on a given behaviour, without including the random variations that correspond to these levels. In this case we identify the contextual factors which reflect the conditions in which individuals belong to the various levels of aggregation. The example we have chosen to work on is that of regional migration in Norway.

Analysis of migration flows In an earlier study we used data from the Norwegian Population Register to demonstrate the importance of aggregation effects. We were able to verify the links, initially established in a theoretical way, between the estimations obtained using the various types of model: exponential regression to model

the emigration rates of the regions, logit and event history models to explain the individual risks of migrating in terms of the characteristics of the zones being considered and of the individuals themselves (Baccaïni and Courgeau, 1996a).

The data set

Norway has local population registers in which are recorded the demographic events of the individuals living in the country, and in particular their internal migrations (changes of municipal districts). This register was centralized and computerized in 1964, for all individuals living in Norway at the time of the census of 1 November 1960. The event history data from this register have been combined with data collected in the censuses of 1960, 1970 and 1980.

The file used contains the 54 814 individuals born in 1958, who lived in Norway in 1991 and who had not migrated abroad. For each of these individuals we know the successive changes of region (Norway is divided into 19 regions, see Figure 1). We have considered only the regional emigration flows, observed over a short period of two years, 1980 and 1981, when the individuals were aged 22-23.

A census was conducted in 1980, so we know the various characteristics of the individuals at this date, and we have also been able to establish how long the individual had been living in the region of residence at the start of 1980.

At the individual level eight characteristics have been selected as having a possible effect on the chances of moving out of the region: marital status (married/unmarried), being economically active (active/non-active), type of occupation (farmer/non-farmer), educational level (more than 12 years in full-time education/less than 13 years in full-time education), having children (at least one child/no children), and the level of income (high income; low income; no income).

We were then able to reconstitute the aggregated characteristics for the 19 regions (percentage of individuals having left the region in 1980-1981, percentage of individuals married, percentage of farmers, etc.).

Examining to begin with just the effect of the aggregated characteristics, we showed the similarity of the results obtained using the following three models: an exponential regression to model the rates of emigration, a logit model and an event history model to explain the individual probabilities of migration. The event history model possessed a greater accuracy, however, since it included the length of stay in the region of departure.

The effects of the individual characteristics, on the other hand, are independent of that of the aggregated characteristics. So if the characteristics of the regions and the individual characteristics are introduced simultaneously in a logit or event history model, very different effects are observed

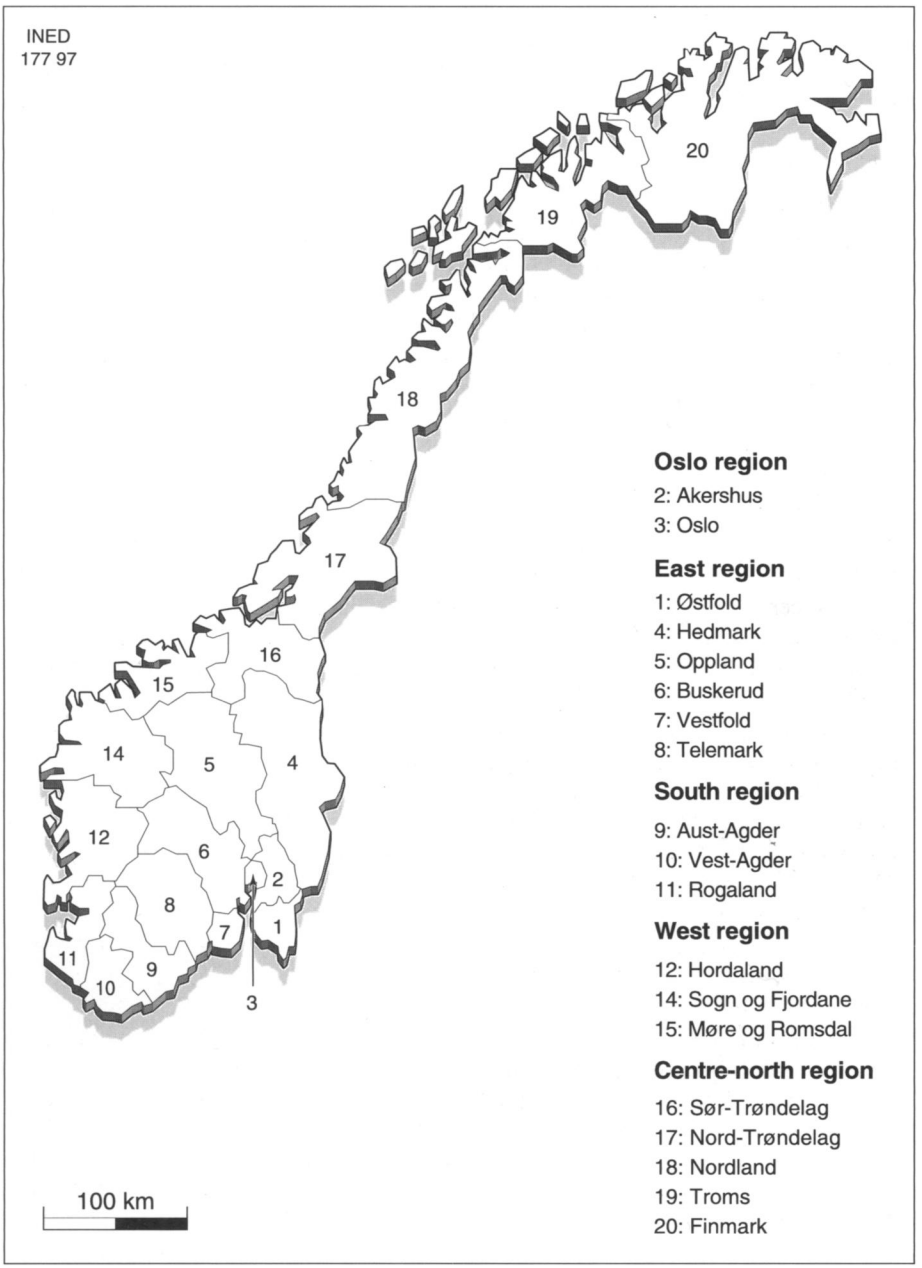


Figure 1. – The regional division of Norway

for the two types of characteristics, sometimes operating in opposite directions (parameters of opposite signs). For example, while the fact of being married increases the probability for a man of leaving his region of residence at age 22, this probability falls when the proportion of married men in the region rises (Table 9).

Conversely, the fact of being a farmer reduces a man's chances of moving out of his region, but this probability rises with the proportion of farmers in the region.

We have shown how this apparent paradox can be explained when the population subjected to the risk is broken down into two distinct groups: the married and unmarried, farmers and non-farmers, and so forth (see III and Figure 6 Schema 1.b).

The independence between the effect of 'macro' characteristics and that of 'micro' characteristics has been confirmed by calculating the coefficients of correlation between the micro and macro parameters: these correlations are in general very weak, in the order of –0.10.

The more complicated case of inter-regional flows

We continued our exploitation of the Norwegian data by examining the flows between regions rather than simply emigration. The aim here is to explain the probability not just of moving out of one's region but also of choosing one destination rather than another.

For the purpose of the analysis the 19 regions have been grouped into 5 large regions (Oslo, East, South, West and Centre-North), in order to limit the number of flows (figure 1).

A number of improvements have been made compared with our previous study. In particular, we have considered the migrations made in the course of the years 1981 and 1982 (rather than 1980-1981). The Census was in fact conducted in November 1980, so in order to know the exact characteristics of the individuals prior to their migration it is preferable to consider the years 1981-82.

New characteristics can be introduced at the individual level. We have considered the length of time in full-time education, distinguishing individuals with less than 10 years in full-time education and those with more than 12 years, compared to those with between 10 and 12 years. Of particular importance is the information that individuals may have about other regions, that is, their links with potential destinations. For this reason we have introduced the fact of having already lived or not in the potential destination regions, the length of previous stays in these regions, and the length of time since the end of the last stay (the age of the individual if there was no stay).

The hypothesis we make here is thus that an individual's chances of migrating to a region are greater if they have already lived there, if this was for a long period, and if it occurred recently.

New variables must also be introduced at the aggregated level to explain the flows between regions.

The three variables defined above can be aggregated, enabling us to use the percentages of individuals in the region who have lived in the other regions (potential destinations), the average length of stay in the various destination regions by the individuals living in the region of origin, the average length of time since the end of the last stay in the different destination regions, for all the individuals of the region of origin.

On the assumption that individuals have strongest relations with individuals who have the same socio-demographic characteristics as themselves, the model must also include the proportions of individuals with the same characteristics in the potential destination regions (percentage of individuals with the same marital status, percentage of individuals with the same profession, percentage of individuals of the same educational level). These variables combine a macro dimension (they are measured at the level of regions) and a micro dimension (they depend on the characteristics of the individuals). Consequently they can only be used in the models that estimate the individual probabilities of migrating (the event history model is used here).

The geographical distance between the regions, whose role as an obstacle to migration has often been demonstrated, can also be included.[1]

We began by considering only the individual characteristics using an event history model, in order in particular to test the effect of previous stays on the choice of a destination. The analysis was thus broken down into five competitive risk models (one model per region of origin, the individuals having a choice of four destinations). Here we present the results for the regions of Oslo and the Centre-North only (Tables 1.a and 1.b).

The binary variable "having made a previous stay in the region" and the discrete variables "length of previous stay" and "time since the end of last stay" are found to be highly redundant. The first, for example, has a very significant impact when introduced on its own but no effect when the other two variables are also included. These three characteristics, which are strongly correlated, must therefore not be introduced simultaneously into the models. The results presented in the tables concern only the effect of a previous stay but the effects of length of stay have also been tested.

The fact of having already lived in a particular region significantly increases the likelihood of returning there when aged 23-24, and the longer the previous stay, the higher is this likelihood. Thus it can be seen that individuals retain special links with the regions in which they lived when younger and it is to these regions that they are the most likely to return.

In many cases, the chances of returning to a region are significantly affected by the length of time elapsed since the end of the last stay there:

---

[1] Our estimate of the distances between regions is based on the distance (in kilometres) between the largest towns of each region.

the further in the past the stay occurred, the lower the chances that an individual will return, as their links with the region are then greatly weakened. This effect is particularly significant in the case of migrations from Oslo to the East or the West, and from the North to Oslo or the West.

TABLE 1a. – EFFECT OF INDIVIDUAL CHARACTERISTICS ON AN INTER-REGIONAL MIGRATION IN 1981-82 (1958 GENERATION, LIVING IN OSLO REGION END 1980)

Charcteristics of individuals	Destination East region		Destination South region		Destination West region		Destination North region	
	Estimated parameter	Standard error	Estimated parameter	Standard error	Estimated parameter	Standard error	Estimated parameter	Standard error
Man	-0.35***	0.09	-0.04	0.20	-0.27	0.18	-0.11	0.14
Economically active	-0.10	0.12	-0.15	0.23	-0.15	0.22	-0.36**	0.16
Married	-0.06	0.12	-0.14	0.24	-0.11	0.22	0.08	0.17
Children	0.11	0.16	-0.71	0.54	-0.04	0.36	-0.44	0.27
Farmer	0.83***	0.25	-0.23	1.01	1.05***	0.43	0.90***	0.35
< 10 years full-time education	0.02	0.12	-0.40	0.38	0.12	0.26	-0.15	0.20
> 12 years full-time education	-0.29**	0.13	-0.30	0.24	0.24	0.21	-0.07	0.17
No income	-0.10	0.25	-0.34	0.63	-0.40	0.63	0.41	0.31
Income < 20 000 krone	0.08	0.14	0.13	0.25	0.68***	0.24	0.15	0.18
Income > 50 000 krone	0.15	0.11	-0.44*	0.24	0.15	0.22	-0.11	0.16
Previous stay in region of destination	1.65***	0.14	3.24***	0.27	2.58***	0.22	2.26***	0.19

*** significant at 1% threshold; ** significant at 5% threshold; * significant at 10% threshold.
Source: Norwegian Population Register, Central Bureau of Statistics, Oslo.

TABLE 1b. – EFFECT OF INDIVIDUAL CHARACTERISTICS ON AN INTER-REGIONAL MIGRATION IN 1981-1982 (1958 GENERATION, LIVING IN CENTRE-NORTH REGION END 1980)

Characteristics of individuals	Destination Oslo region		Destination East region		Destination South region		Destination West region	
	Estimated parameter	Standard error	Estimated parameter	Standard error	Estimated parameter	Standard error	Estimated parameter	Standard error
Man	-0.26**	0.11	-0.13	0.15	0.10	0.21	-0.16	0.16
Economically active	-0.35***	0.12	-0.22	0.16	-0.22	0.23	-0.33*	0.18
Married	-0.25*	0.15	0.07	0.17	0.17	0.25	-0.19	0.21
Children	-1.11***	0.21	-0.41*	0.21	-0.32	0.29	-0.88***	0.27
Farmer	-0.06	0.25	0.27	0.29	-0.19	0.52	0.03	0.35
< 10 years full-time education	-0.43***	0.15	-0.10	0.18	-0.47	0.30	-0.08	0.21
> 12 years full-time education	0.32**	0.15	-0.09	0.23	-0.08	0.34	0.13	0.22
No income	-0.19	0.23	-0.29	0.34	-0.40	0.44	0.10	0.31
Income < 20 000 krone	0.05	0.14	0.13	0.19	-0.01	0.26	0.13	0.21
Income > 50 000 krone	-0.06	0.12	0.21	0.17	-0.21	0.24	-0.14	0.19
Previous stay in region of destination	1.26***	0.18	1.73***	0.21	2.32***	0.28	1.93***	0.23

*** significant at 1% threshold; ** significant at 5% threshold; * significant at 10% threshold.
Source: Norwegian Population Register, Central Bureau of Statistics, Oslo.

When the various individual characteristics do have a significant effect it usually operates in the same direction, regardless of the regions of origin and destination: the probability of changing region is low for men, the economically active, individuals who are married and have children, individuals with a low educational level and those with a high income. These are the same effects already observed for the probability of moving out of one's region (Baccaïni and Courgeau, 1996a and 1996b).

However, a number of effects are seen to operate differently depending on the direction of the migration. Thus the fact of having children, which is in general a brake on migration, increases the probability that individuals in Oslo will migrate to the East region (deconcentration of the Oslo urban region). Likewise, whereas individuals with a low educational level have in general a low probability of leaving their region, the opposite is the case for migrations from the South region to the East region.

The fact of being a farmer also has an effect which differs sharply depending on the region of origin of the individuals. Farmers from the Oslo region have a high probability of leaving it (to go to the East, West or North), whereas in the other regions the fact of being a farmer is, on the contrary, a brake on migration.

Some individual characteristics thus have opposite effects depending on whether we are dealing with migration from region i to region j or from region j to region i. If we consider the two symmetrical migration flows between the Oslo region and the Centre-North region, there are four characteristics with opposite effects depending on the direction of the migration: being married, being a farmer, having had a long full-time education, and having no income.

In the next stage of the analysis, we produced various types of model which included the characteristics of the regions of origin and/or destination.

Introducing these aggregated characteristics necessitates making hypotheses about the processes which can cause an individual to leave a region i and move to a region j. The essential question can be summed up in these terms: does the decision to leave one's region of residence precede or follow the choice of a region of destination? In other words, is the decision to leave region i a consequence of a desire to live in region j, or is the choice of region j a consequence of the desire to leave region i?

The two processes are probably combined, and we consider that it is the comparison of the advantages and disadvantages of the region of origin, relative to the regions of potential destination, which leads an individual to decide for or against making a migration, though the decision depends also on their individual characteristics.

The aim is thus to examine the chances that individuals have of moving to the various different destinations that are available to them, according to, first, their individual characteristics, second, the advantages that a region may offer compared to that in which they live.

To do this we have first considered a model by region of destination, in which the population of the four other regions is that subjected to the risk of moving there. The macro variables to be included in order to demonstrate the effect of the relations with individuals presenting similar socio-demographic characteristics are the ratios between the percentage of individuals with the same characteristic in the region of destination and the percentage in the individual's region of origin.

We give here the results for two regions of destination: the region of Oslo and the Centre-North.

Most of the aggregated characteristics have a very significant effect on the chances of migrating to Oslo when they are introduced alone into an event history model, and slightly less on the chances of migrating to the Centre-North (Table 2).

TABLE 2. – EFFECT OF CHARACTERISTICS OF THE REGIONS ON AN INTER-REGIONAL MIGRATION IN 1981-82, BY REGION OF DESTINATION (1958 GENERATION)

Characteristics of regions of origin and destination	Destination Oslo		Destination Centre-North	
	Estimated parameter	Standard error	Estimated parameter	Standard error
% ratio of individuals with same marital status (a)	2.37***	0.17	1.35***	0.23
% ratio of individuals with same educational level (a)	0.81***	0.07	– 0.63***	0.20
%ratio of individuals with same occupation (a)	0.36***	0.06	0.21***	0.06
Distance between origin and destination	0.001**	0.000	– 0.001*	0.00
% of the population having lived in the region of destination	0.16***	0.02	0.00	0.02

(a): ratio of % in the region of destination and % in the region of origin.
*** significant at 1% threshold; ** significant at 5% threshold; * significant at 10% threshold.
Source: *Norwegian Population Register,* Central Bureau of Statistics, Oslo.

The chances of moving to each of these two regions increase as the proportion of individuals with the same marital status and individuals with the same occupation increase relative to their importance in the individual's region of origin, thus appearing to confirm our hypotheses.

Conversely, the effect of the relative importance of the individuals with the same educational level differs according to the region of destination. Oslo attracts disproportionately more individuals whose educational level is less common in their region of origin than in Oslo, whereas the Centre-North region draws disproportionately more individuals whose level of qualification is more common in their region of origin than in the Centre-North. Individuals with a high educational level do in fact have a greater propensity to migrate than the others, and highly qualified individuals are proportionally more numerous in the Oslo region than in the Centre-North.

Distance is an obstacle for migrants to the Centre-North region, which consequently attracts disproportionately more individuals from nearby regions. This is not the case for migrants to Oslo, who appear to be relatively indifferent to distance.

Conversely, the length of previous stays made by the inhabitants of the region of origin, and how long ago they occurred, (a reflection of the strength of the links between the two regions) have a significant effect only for migrations to the region of Oslo.

The effects of some of these aggregated characteristics change when individual characteristics are also introduced into the model (Table 3). In particular, the effect of the relative importance of individuals with the same educational level in the region of destination and in the region of origin is reversed when the educational level of individuals is taken into account.

TABLE 3. – EFFECT OF CHARACTERISTICS OF INDIVIDUALS AND REGIONS ON AN INTERREGIONAL MIGRATION IN 1981-82, BY REGION OF DESTINATION (1958 GENERATION)

Characteristics of individuals and of regions of origin and destination	Destination Oslo		Destination Centre-North	
	Estimated parameter	Standard error	Estimated parameter	Standard error
Man	– 0.18***	0.05	– 0.04	0.08
Economically active	– 0.33***	0.06	– 0.37***	0.08
Married	– 0.36**	0.16	– 0.24*	0.14
Children	– 1.00***	0.12	– 0.59***	0.15
Farmer	– 0.12	0.15	0.13	0.21
< 10 years full-time education	– 0.57***	0.09	– 0.46***	0.13
> 12 years full-time education	0.78***	0.20	0.59***	0.19
No income	0.00	0.12	0.29*	0.17
Income < 20 000 krone	0.39***	0.06	0.64***	0.10
Income > 50 000 krone	– 0.30***	0.06	– 0.14	0.10
Previous stay in region of destination	0.76***	0.09	2.01***	0.11
% ratio of individuals with same marital status (a)	1.21***	0.35	0.49	0.34
% ratio of individuals with same educational level (a)	– 0.55**	0.25	1.35***	0.48
% ratio of individuals with same occupation (a)	0.36***	0.08	0.07	0.11
Distance between origin and destination	0.001*	0.000	0.000	0.00
% of the population having lived in the region of destination	0.16***	0.02	– 0.01	0.02

(a): ratio of % in the region of destination and % in the region of origin.
*** significant at 1% threshold; ** significant at 5% threshold; * significant at 10% threshold.
Source: *Norwegian Population Register,* Central Bureau of Statistics, Oslo.

In the next stage we changed perspective and worked again with the competitive risk model. The question here is to see how the individuals of a given region may be encouraged to migrate by the relative advantages of the other regions. The behaviour of migrants coming from the different regions can then be compared.

TABLE 4. – EFFECT OF INDIVIDUAL AND AGGREGATED CHARACTERISTICS ON AN INTERREGIONAL MIGRATION IN 1981-82 (1958 GENERATION, LIVING IN OSLO REGION END 1980)

Characteristics of individuals and regions	Destination East region		Destination South region		Destination West region		Destination North region	
	Estimated parameter	Standard error	Estimated parameter	Standard error	Estimated parameter	Standard error	Estimated parameter	Standard error
Man	-0.31***	0.09	0.01	0.20	-0.36*	0.18	-0.12	0.13
Economically active	-0.14	0.12	-0.24	0.25	-0.08	0.22	-0.43***	0.16
Married	-0.17	0.14	-0.24	0.33	-0.15	0.27	0.05	0.20
Children	0.18	0.16	-0.99*	0.54	-0.21	0.36	-0.45	0.27
Farmer	1.22***	0.27	-0.41	1.05	0.91*	0.50	0.39	0.37
< 10 years full-time education	0.01	0.12	-0.61	0.38	0.08	0.26	-0.18	0.20
> 12 years full-time education	0.02	0.15	-0.32	0.34	0.27	0.27	-0.22	0.23
No income	0.00	0.25	-0.48	0.63	-0.53	0.63	0.20	0.49
Income < 20 000 krone	0.08	0.14	0.26	0.25	0.60**	0.24	0.11	0.19
Income > 50 000 krone	0.20*	0.11	-0.54**	0.23	0.13	0.22	-0.14	0.16
Previous stay in region of destination	0.43***	0.09	0.37*	0.21	0.37**	0.18	0.47***	0.14
% ratio of individuals with same marital status (a)	0.40	0.30	1.18**	0.55	0.19	0.54	-0.72	0.48
% ratio of individuals with same educational level (a)	1.54***	0.50	-0.87	0.87	-0.18	0.71	-0.51	0.62
% ratio of individuals with same occupation (a)	-0.32*	0.17	-0.39	0.44	0.08	0.21	0.38***	0.08

(a): ratio of % in the region of destination and % in the region of origin.
*** significant at 1% threshold; ** significant at 5% threshold; * significant at 10% threshold.
Source: *Norwegian Population Register,* Central Bureau of Statistics, Oslo.

TABLE 5. – EFFECT OF INDIVIDUAL AND AGGREGATED CHARACTERISTICS ON AN INTERREGIONAL MIGRATION IN 1981-82 (1958 GENERATION, LIVING IN CENTRE-NORTH REGION END 1980)

Characteristics of individuals and regions	Destination Oslo		Destination East region		Destination South region		Destination West region	
	Estimated parameter	Standard error	Estimated parameter	Standard error	Estimated parameter	Standard error	Estimated parameter	Standard error
Man	-0.40***	0.11	-0.16	0.15	0.11	0.21	-0.11	0.16
Economically active	-0.24**	0.12	-0.09	0.16	-0.26	0.24	-0.30	0.19
Married	-0.28*	0.17	0.19	0.19	-0.16	0.30	0.08	0.22
Children	-1.11***	0.21	-0.37*	0.21	-0.09	0.29	-0.88***	0.27
Farmer	-0.21	0.25	-0.01	0.31	-0.31	0.53	0.11	0.36
< 10 years full-time education	-0.37**	0.15	-0.15	0.19	-0.50	0.30	-0.12	0.21
> 12 years full-time education	-0.48**	0.24	0.30	0.28	-0.25	0.44	0.38	0.28
No income	-0.20	0.23	-0.34	0.34	-0.59	0.45	0.26	0.31
Income < 20 000 krone	0.01	0.14	0.05	0.19	-0.08	0.26	0.14	0.21
Income > 50 000 krone	-0.08	0.12	0.23	0.17	-0.28	0.24	-0.13	0.19
Previous stay in region of destination	-0.88***	0.13	-0.44***	0.15	-0.91***	0.23	-0.44***	0.17
% ratio of individuals with same marital status (a)	0.46	0.47	0.14	0.50	2.58***	0.55	-1.37*	0.74
% ratio of individuals with same educational level (a)	1.21***	0.27	-1.38**	0.62	-0.66	0.88	-0.14	0.43
% ratio of individuals with same occupation (a)	-0.68**	0.34	-2.13***	0.49	-0.18	0.70	0.29	0.49

(a): ratio of % in the region of destination and % in the region of origin.
*** significant at 1% threshold; ** significant at 5% threshold; * significant at 10% threshold.
Source: *Norwegian Population Register,* Central Bureau of Statistics, Oslo.

The results are given for two regions of origin: the region of Oslo and the Centre-North (Tables 4 and 5).

Introducing the aggregated characteristics into these competitive risk models modifies the value of some of the parameters associated with the individual characteristics (compared to what they were in the micro level model seen above). For migrants from Oslo, for example, the fact of being a farmer no longer has a significant effect on their propensity to move to the West or Centre-North regions. For the individuals coming from the Centre-North, introducing the aggregated characteristics causes the effect of a high educational level on the chances of migrating to Oslo to change from positive to strongly negative.

For individuals coming from a given region, the effects of the different aggregated characteristics are very different depending on the region of destination. For example, the attraction of the East region for individuals from Oslo is greatest for those whose educational level is strongly present in the East compared with in Oslo, whereas the opposite is true for the migrations from Oslo to the other regions. The East region, on the other hand, has a disproportionate attraction for individuals from the Oslo region whose occupation is relatively under-represented in the East compared with in the Oslo region, while the opposite is observed for migrations from Oslo to the Centre-North.

The same type of observations can be made about the migrations by individuals from the Centre-North region.

These early analyses, which need to be further developed, show how understanding of migratory processes can be increased by the simultaneous inclusion in the models of the characteristics of the individuals and of the regions of origin and destination. They also illustrate the care needed when interpreting the results.

III. – Analysis with multilevel random variables

The analysis in the preceding pages examined five regions only, which given the size of the sample could have been studied separately. Now let us see what happens when we increase the number of regions in the analysis. In this case the regions studied can be treated as a sample from which information can be drawn about the characteristics of the regions, considered as a population of regions. More specifically, let us consider first the results of a regression analysis.

Regression analysis For this we can use the highly didactic example given by Woodhouse *et al.* (1996). This is a longitudinal study conducted on a cohort of English schoolchildren from the time of their entry to junior classes at age 8 to their entry to secondary

school at age 11. These pupils attended fifty primary (elementary) schools selected at random from among 650 London schools. The aim of the study was to find out if some schools were better than others at promoting the educational progress of their pupils. To do this the authors examined the progress in mathematics as measured by the tests taken at ages 8 and 11, and that they examine at both the individual and school levels.

Let y_{ij} be the 11-year score obtained by the ith pupil of school j, and x_{ij} their 8-year score. We begin by modelling the linear regressions for each school

$$y_{ij} = a_{oj} + a_{1j}\, x_{1ij} + e_{ij} \tag{1}$$

where a_{oj} and a_{1j} are parameters fitted to the jth school, e_{ij} being the random residual of expectation zero and variance σ^2_{ej}.

If there were not many schools it would be possible to estimate as many parameters a_{oj} and a_{1j} as there were schools, by assuming, for example, a variance of e_{ij} independent of the school. These parameters characterizing each school could be compared, but no generalization could be made, since they relate to the schools in the sample and provide no information about the schools as a whole.

If on the other hand we treat the fifty schools as a sample chosen from a population of 650 schools, information of a statistical nature can be inferred about this larger population. Let us see how to go further in formalizing this analysis which introduces two levels of aggregation: the pupil and the school.

It can be seen that one way of introducing the schools into equation (1), is to assume that parameters a_{oj} and a_{1j} are random and that they will therefore vary from one school to another, which is the same as making:

$$\begin{aligned} a_{oj} &= a_o + e_{oj} \\ a_{1j} &= a_1 + e_{1j} \end{aligned} \tag{2}$$

where a_o and a_1 are the mean parameters fitted to all the schools, e_{oj} and e_{1j} the random variables, of expectation zero with the following variances and covariances.

$$\begin{aligned} \text{var}\,(e_{oj}) &= \sigma^2_{eo} \\ \text{var}\,(e_{1j}) &= \sigma^2_{e1} \\ \text{cov}\,(e_{oj}\,,\, e_{1j}) &= \sigma_{eo1} \end{aligned} \tag{3}$$

Formula (1) can then be rewritten as:

$$y_{ij} = a_o + a_1\, x_{1ij} + (e_{oj} + e_{1j}\, x_{1ij} + e_{ij}) \tag{4}$$

which is made up of a fixed part which is independent of the school ($a_o + a_1 x_{1ij}$), and a random part which depends on both the school and the individual.

The various parameters and the variances and covariances can be estimated using generalizations of the least squares method (Goldstein, 1986, 1995). In this way they obtain the estimations given in Table 6, which were computed using the MLn program.

TABLE 6. – PARAMETERS AND STANDARD ERRORS ESTIMATED IN THE MULTILEVEL MODEL RELATING THE 8-YEAR AND 11-YEAR SCORES OF PUPILS

Parameters	Estimation	Standard error
Fixed		
Constant	15.040	1.318
8-year score	0.612	0.043
Random		
School level		
σ^2_{e0} (constant)	44.990	16.360
σ^2_{e01} (covariance)	– 1.231	0.521
σ^2_{e1} (8-year score)	0.034	0.017
Pupil level		
σ^2_e	26.960	1.343

Source: Woodhouse, 1996.

This table shows that all the effects are significant. First of all it can be seen that the higher an individual's 8-year score, the higher his 11-year score, and this regardless of the school he attends. However, depending on the school the variances and covariances of the random variables e_{oj} and e_{1j} will also be significant: the fact that the covariance between e_{oj} and e_{1j} is negative indicates that the higher the average score of the school, the less the 11-year score will depend on the 8-year score. What this means is that some schools do manage to get all the pupils in a given class to a good level in mathematics, regardless of the initial scores of these pupils; conversely, other schools fail to enable pupils whose level in mathematics is already low, to catch up.

These differences between schools can be clearly illustrated by plotting on the same diagram the predicted relationships between the 8-year and 11-year scores in each school, estimated using a multilevel model. This is the same as writing the following relation for each school:

$$\hat{y}_{ij} = a_o + e_{oj} + (a_1 + e_{1j})\, x_{1ij} \qquad (5)$$

where the e_{oj} and e_{1j} are the residuals relative to the model, calculated for each region j. Figure 2 presents these results. The extreme schools are easily identified: in the upper section of the diagram are those in which the 11-year score is weakly related to the initial score, that is the schools which succeed in getting all their pupils to a good level; in the lower section are those in which the 11-year score remains strongly dependent on the initial mark. It is instructive to compare these results with those obtained when

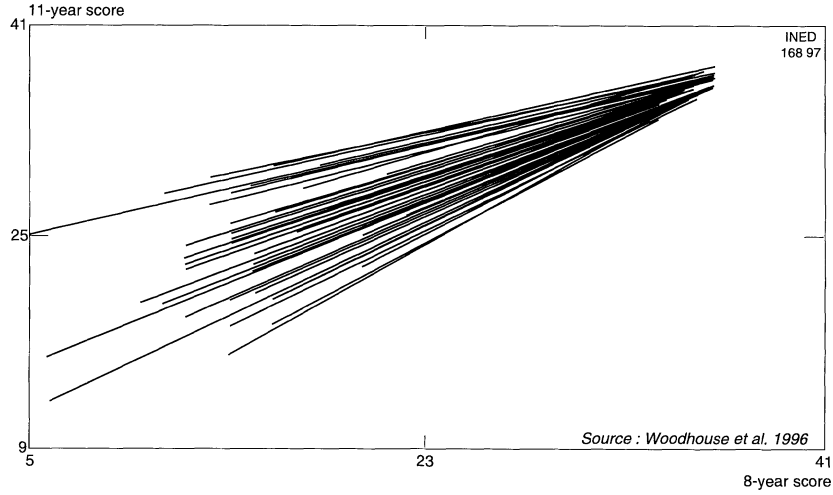


Figure 2. – Plot of predicted 11-year score by 8-year score in each school, using a multilevel model applied to a sample of London schools

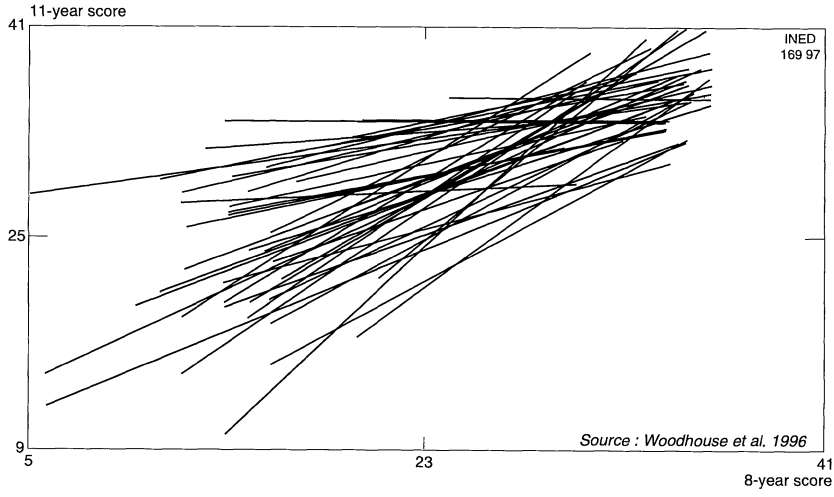


Figure 3. – Plot of predicted 11-year score by 8-year score in each school, using a linear regression applied to each London school of the sample

we apply the regression model (1), to each school, which are shown on Figure 3. This diagram is much less clear than the previous one, because it is based on regressions estimated separately for each school, some of which have a very small number of pupils, causing a poor estimation of the a_{oj} and a_{1j} values.

This analysis can of course be refined by introducing various other characteristics of the individuals or schools (Goldstein, 1995). However, we shall not develop this analysis any further here, but examine in more detail the hypotheses on which these models are based.

Risk of mistaken inference Let it be assumed that a second characteristic, x_{2ij}, which is *independent* of the first, also influences the 11-year score. This might be, for example, a strong parental support in mathematics, as measured by a binary variable, equal to 1 when there is such a support and 0 when there is not. Let it be assumed also that there is a relationship independent of the school attended by the pupil between their 11-year score and the two characteristics (8-year score and parental support). This relationship thus introduces random variables which are also independent of the school the pupil attends. We can then write:

$$y_{ij} = a_o + e_{oij} + (a_1 + e_{1ij})\, x_{1ij} + (a_2 + e_{2ij})\, x_{2ij} + (a_{12} + e_{12ij})\, x_{1ij} \times x_{2ij} + e_{ij} \qquad (6)$$

where e_{oij}, e_{1ij}, e_{2ij}, e_{12ij}, and e_{ij} are random variables of expectation zero, of variance σ^2_{eo}, σ^2_{e1}, σ^2_{e2}, σ^2_{e12}, and σ^2_e, for which all the covariances are nil given that they are independent of zone j and independent of each other. We assume here that the direct effect of parental support is positive, but that the interaction between the 8-year score and parental support is negative: the lower the 8-year score, the greater this effect.

Under these conditions, samples can be simulated which verify them all, for which the size and the number of parents supporting their children are selected at random independently of the school. We have simulated various samples whose parameters are situated in the following intervals:

$$2 \le a_o + e_{oij} \le 7\ ;\ 0{,}25 \le a_1 + e_{1ij} \le 1{,}25\ ;$$

$$21 \le a_2 + e_{2ij} \le 29\ ;\ -0{,}57 \le a_{12} + e_{12ij} \le -0{,}43$$

We give here the results obtained with one of these samples, which are very close to all the others.

If we have no measure of parental support, we can only estimate a model which includes the 8-year score, x_{1ij}. In this case it is easy to confirm that the following relations are verified:

$$a_o + e_{oij} \le a'_o + e'_{oj} \le a_o + a_2 + e_{oij} + e_{2ij}$$

$$a_1 + e_{1ij} \le a'_1 + e'_{1j} \le a_1 + a_{12} + e_{1ij} + e_{12ij}$$

where e'_{oj} and e'_{1j} are random terms which will now depend on the school. The previous model can then be rewritten as:

$$y_{ij} = a'_o + e'_{oj} + (a'_1 + e'_{1j})\, x_{1ij} + e_{ij} \qquad (7)$$

These results appear in Table 7. They show an effect of the school that is highly significant and very close to that obtained in Table 6. Some schools do seem to be successful in getting all their pupils to a good level in mathematics, whatever their initial score, whereas others appear to leave behind the pupils whose initial score is low. As before, Figure 4 shows the estimate of the 11-year scores obtained for each school according to the 8-year score: this is very close to Figure 2 and shows a strong effect of the school, whereas in fact this does not exist. Figure 5 is comparable to Figure 3, and like it shows a less clear cut effect for the school, though this effect is also very similar to that in Figure 3.

TABLE 7. – PARAMETERS AND STANDARD ERRORS ESTIMATED IN THE SIMULATED MULTILEVEL MODEL RELATING THE 8-YEAR AND 11-YEAR SCORES OF PUPILS

Parameters	Estimation	Standard errors
Fixed		
Constant	16.720	1.189
Score at age 8	0.503	0.033
Random		
School level		
σ^2_{e0} (constant)	57.000	14.080
σ^2_{e01} (covariance)	– 1.298	0.373
σ^2_{e1} (score at age 8)	0.030	0.011
Pupil level		
σ^2_e	91.730	2.977

If we now try introducing a supplementary variable, the average 8-year score in each school, into the non-random characteristics, the effect is completely insignificant and produces no change in the model. By contrast, introducing the two individual characteristics and their interaction, while leaving the random terms unchanged, causes all the variances and covariances at the level of the school to become nil as would be expected. The estimated parameters and their standard deviation are shown in Table 8, which becomes completely consistent with the hypothesis of the simulation.

The risk of erroneous inference thus appears to be great when a multilevel model is used, and this even when the omitted characteristic is independent of those introduced in the initial model. Further research is therefore needed in this field in order to give the results of such analysis a greater reliability. An effective precaution involves introducing into the fixed part the largest number of characteristics which have an effect on the phenomenon, so as to minimize the risk of concluding for a random effect which does not in fact exist.

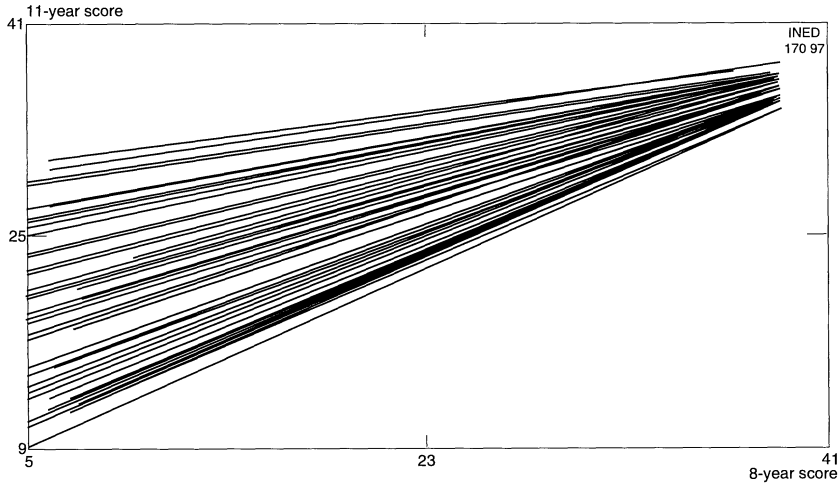


Figure 4. – Plot of predicted 11-year score by 8-year score in each school, using a multilevel model applied to a simulated sample of schools (model 7)

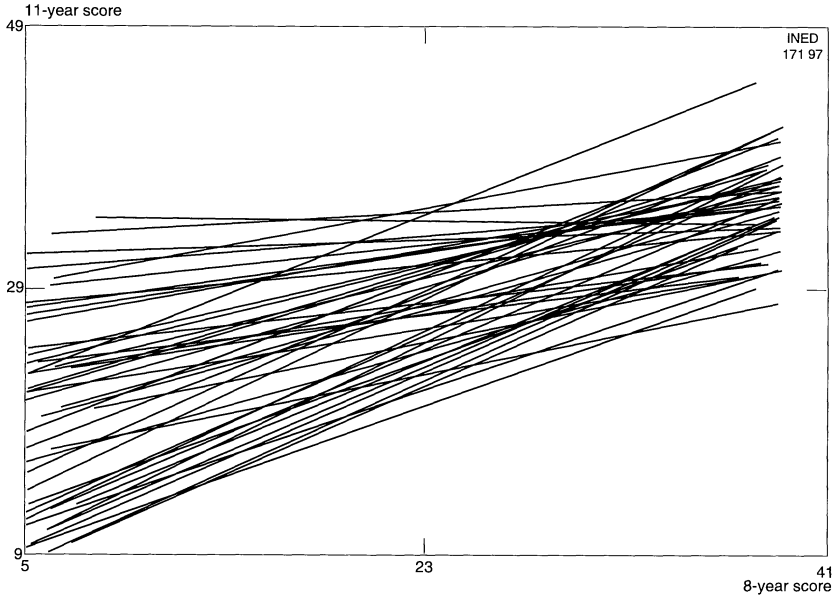


Figure 5. – Plot of predicted 11-year score by 8-year score in each school, using linear regression models applied to each school of the simulated sample (model 7)

TABLE 8. – PARAMETERS AND STANDARD ERRORS IN THE SIMULATED MULTILEVEL MODEL RELATING THE 8-YEAR AND 11-YEAR SCORES OF PUPILS, TO PARENTAL SUPPORT AND TO THE INTERACTION BETWEEN SUPPORT AND THE 8-YEAR SCORE

Parameters	Estimation	Standard errors
Fixed		
Constant	4.410	0.545
8-year score	0.766	0.023
Parental support	25.170	0.783
Interaction	– 0.529	0.033
Random		
Pupil level		
σ_e^2	54.840	1.757

Analysis for discrete response data

Many demographic characteristics are observed in the form of dichotomous or polytomous variables: an individual is married or not, an individual can migrate between n regions, for example. Here we examine in detail the binary case, which with some modifications can be extended to the polytomous case.

Let us assume that we are working with a logit model, where the probability that the characteristic to be estimated, y_{ij}, is 1, is expressed in relation to the explanatory variable, x_{ji}, which is assumed here to be binary:

$$\mathrm{P}\,(y_{ij}=1 \mid x_{ij}) = p_{ij} = \left[1+\exp\left(-\left[a_o + u_{oj} + \left(a_1 + u_{1j}\right) x_{ij}\right]\right)\right]^{-1} \tag{8}$$

It follows that the answers y_{ij}, are distributed according to a binomial distribution of parameters:

$$y_{ij} \sim \mathrm{B}(p_{ij}, 1) \tag{9}$$

In this case we have the following conditional variance:

$$\mathrm{var}\left(y_{ij} \mid p_{ij}\right) = p_{ij}\left(1-p_{ij}\right)$$

The model then becomes a nonlinear model:

$$y_{ij} = p_{ij} + e_{ij} z_{ij}$$

where:

$$z_{ij} = \sqrt{p_{ij}\left(1-p_{ij}\right)}$$

and where:

$$\sigma_e^2 = 1$$

The level 1 variance in this case is equal to unity, and we shall work essentially on the level 2 variances and covariances (Goldstein, 1991).

An application to Norwegian migration

Here we examine the migration flows of the 19 Norwegian regions, for individuals born in 1958 and who migrated in 1980-81 (for more details see: Baccaïni and Courgeau, 1996a). To explain the behaviour observed we have the 8 individual and aggregated characteristics that we defined in part II.

Because individual and aggregate characteristics have a specific effect on the probabilities of emigrating from the regions, we introduce them first for each type of characteristic in a simple logit model and in a multilevel logit model. The results for men are given in Table 9.

The non-random parameters estimated with a multilevel model are in general very close to those we obtain with a simple logit model. This confirms the results (Baccaïni and Courgeau, 1996a) that we mentioned in II. But when the effect of the random terms related to the characteristics is not zero at the regional level, a large increase in the dispersion of these parameters is observed, with an approximate doubling of their standard deviation. Despite this, most of the effects that are significant at the 5% threshold in the simple logit model are also significant in the multilevel model. The only exceptions to this rule are two effects of aggregated characteristics: the positive influence that living in a low-income region had on the chances of migrating becomes non-significant in the multilevel model; on the other hand, regions with a high educational level are found to have a significant effect of reducing the chances of migrating when the multilevel model is used, whereas this was not apparent at all with the simple logit model.

Let us now look in more detail at the combined effect of the fixed parameters and random parameters at the regional level. The logit function for the probability of emigrating from j for the individuals who do not have the characteristic being studied, $\prod_{oj}$, is given by $a_o + u_{oj}$; its between-area variance is equal to $\sigma^2_{\hat{e}o}$. The logit function for the individuals who do have it, $\prod_{1j}$, is the sum $a_o + a_1 + u_{oj} + u_{1j}$; so its between-area variance is equal to: $\sigma^2_{\hat{e}o} + 2\sigma_{eo1} + \sigma^2_{\hat{e}1}$. It will also be valuable to compare these variances depending on whether or not the aggregated characteristic is introduced: so as not to clutter the table (Table 9) the latter estimations are not shown but are quoted directly in the text when this is appropriate.

Let us examine in greater detail the effect of three characteristics that we have represented diagramatically in Figure 6.

Farmers, for example, have a lower probability of migrating than the other occupations. But it is seen that the variances at the regional level of the logits $\prod_{oj}$ and $\prod_{1j}$, are equal, whether or not the aggregated characteristic is introduced. Schema 1.a of Figure 6 shows the values of $\prod_{oj}$ and $\prod_{1j}$ that correspond to farmers and non-farmers respectively, joined for each region by lines, which in this case are parallel to each other. It may be noted that when the percentage of farmers is introduced, the between-region variance

TABLE 9. – ESTIMATION OF THE PARAMETERS AND THEIR STANDARD ERROR (IN BRACKETS) OF THE SIMPLE AND MULTILEVEL LOGIT MODELS WHICH INTRODUCE SIMULTANEOUSLY AN INDIVIDUAL CHARACTERISTIC AND THE CORRESPONDING AGGREGATED CHARACTERISTIC IN 1980 (GENERATION OF MEN BORN IN 1958)

Parameters	Married		Active		Farmer		More than 12 years full-time education	
	Simple logit	Multilevel	Simple logit	Multilevel	Simple logit	Multilevel	Simple logit	Multilevel
Fixed								
Constant	– 1.465 (0.061)	– 1.563 (0.114)	1.586 (0.684)	2.978 (1.625)	– 2.190 (0.043)	– 2.291 (0.149)	– 2.216 (0.076)	– 1.725 (0.217)
Characteristic	0.418 (0.054)	0.393 (0.079)	– 0.540 (0.042)	– 0.588 (0.074)	– 0.401 (0.097)	– 0.406 (0.096)	0.531 (0.058)	0.648 (0.117)
Aggregated characteristic	– 0.057 (0.005)	0.050 (0.008)	– 0.044 (0.009)	– 0.062 (0.021)	0.028 (0.018)	0.028 (0.018)	0.002 (0.008)	– 0.058 (0.024)
Random regional level								
σ^2_{e0} (constant)		0.018 (0.015)		0.045 (0.032)		0.064 (0.029)		0.099 (0.055)
σ_{e01} (covariance)		0.013 (0.012)		– 0.020 (0.027)		0.000		0.107 (0.072)
σ^2_{e1} (characteristic)		0.058 (0.045)		0.060 (0.037)		0.000		0.178 (0.146)
	At least one child		Low income		High income		No income	
	Simple logit	Multilevel	Simple logit	Multilevel	Simple logit	Multilevel	Simple logit	Multilevel
Fixed								
Constant	– 1.307 (0.077)	– 1.373 (0.180)	– 3.053 (0.150)	– 2.590 (0.382)	0.562 (0.306)	– 0.698 (0.670)	– 2.240 (0.083)	– 2.313 (0.290)
Characteristic	– 0.133 (0.079)	– 0.165 (0.098)	0.096 (0.051)	0.125 (2.103)	– 0.195 (0.039)	– 0.256 (0.099)	– 0.065 (0.132)	– 0.074 (0.124)
Aggregated characteristic	– 0.110 (0.080)	– 0.099 (0.026)	0.053 (0.009)	0.025 (0.021)	– 0.004 (0.005)	– 0.022 (0.011)	0.038 (0.029)	0.082 (0.099)
Random regional level								
σ^2_{e0} (constant)		0.033 (0.014)		0.100 (0.035)		0.035 (0.024)		0.067 (0.029)
σ_{e01} (covariance)		0.012 (0.022)		– 0.116 (0.034)		– 0.032 (0.038)		0.000
σ^2_{e1} (characteristic)		0.055 (0.093)		0.156 (0.054)		0.152 (0.068)		0.000

Source: *Norwegian Population Register*, Central Bureau of Statistics, Oslo.

decreases slightly from 0.070 to 0.064, but above all when the percentage of farmers increases, the probability of migrating increases both for farmers and for the other categories alike. We may note, however, that while this effect is not very significant, it does become so when other characteristics are introduced. Schema 1.b presents the average values of $\prod_{oj}$ and $\prod_{1j}$ according to the percentage of farmers in each region: in this case we have two parallel lines, producing a schema similar to that observed for the same category in France (Courgeau, 1995). Farmers have a much lower probability of migrating than the other categories, but the higher the percentage of farmers in a particular region, the higher the probability of migrating for all categories. This result highlights the danger of inferring individual results from results obtained at a more aggregated level: the presence of a large number of farmers in a region results in a higher probability of emigrating for all the categories of population, doubtless due to the greater scarcity of non-agricultural employment in such regions. But this does not mean that farmers have a higher probability of emigrating than the other categories, since it is the exact opposite that is observed.

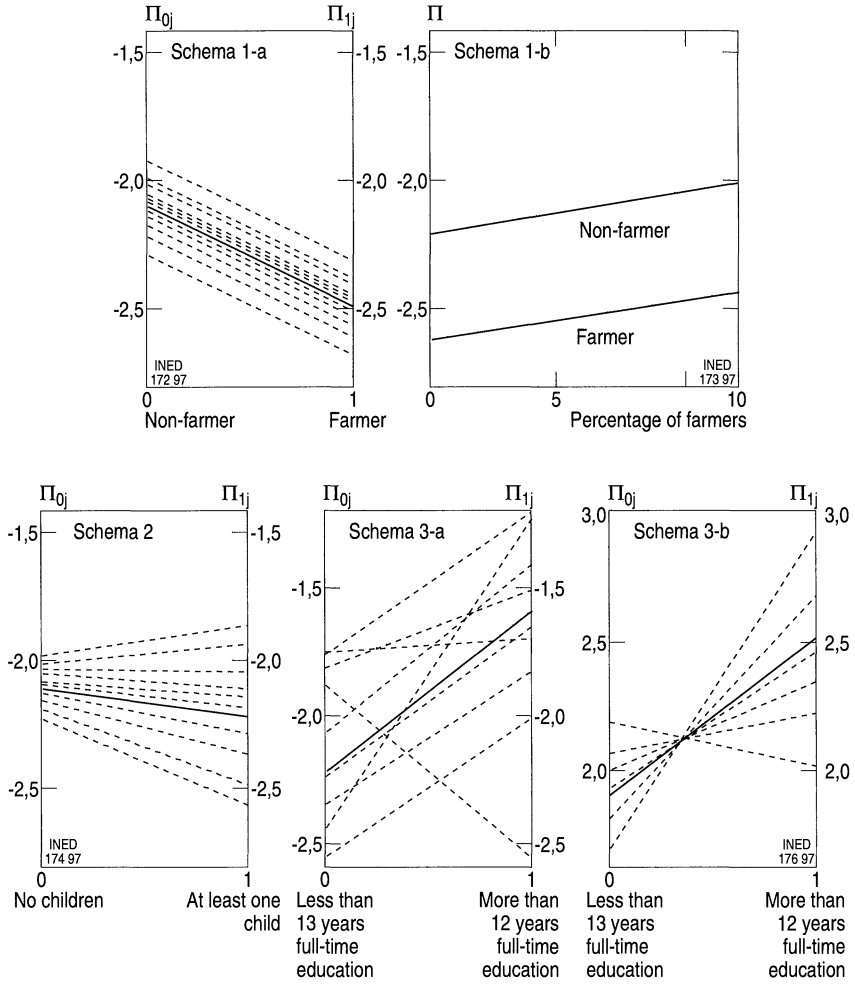
Now let us consider the case of individuals with at least one child. They are found to have a probability of migrating that is always lower than those without children, whether or not we introduce the percentage with at least one child. It is also verified that the variance between regions of the logits of the probability of migrating of those with at least one child (0.174) is three times higher than that of those with no children (0.061), when the percentages with at least one child are not introduced.

The result in this case is Schema 2 in Figure 6, which again shows the values of $\prod_{oj}$ and $\prod_{1j}$: the positive correlation close to unity (0.95) between the random effects characterizes the fan shape of this schema. When this percentage is introduced it has a highly significant effect and, above all, it reduces the between-region variances and covariances by half, so that they lose their significant effect: from 0.06 to 0.03 for the variance of individuals without children, from 0.17 to 0.11 for that of individuals with children and from 0.10 to 0.04 for the covariance. We can therefore say that it does indeed explain a part of these random effects.

To conclude these examples let us examine the case of individuals with more than 12 years of full-time education who have a higher probability of migrating than the others. In this case the correlation between the regional random variables of individuals with more or less than 12 years full-time education is almost nil (0.07): this gives Schema 3a of Figure 6, in which the lines corresponding to the different regions seem to be drawn at random around the average trend joining a_o to $(a_o + a_1)$.

Naturally the other characteristics produce as many different interpretations of the regional situations.

Finally, let us consider a model in which all the characteristics considered are included which have an effect on the regional probability of



Schema 1-a : without including the percentage of farmers
Schema 1-b : including simultaneously the percentage of farmers
Schema 2 : without including the percentage of individuals with at least one child
Schema 3-a : without including the percentage of individuals with more than 12 years full-time education
Schema 3-b : including all the other characteristics

Figure 6. – Effect of three characteristics (being a farmer, having at least one child, and having spent more than 12 years in full-time education) on the logit of the probability of migrating in Norway for the 1958 generation between 1980 and 1981

migrating. Table 10 shows the results of a simple logit model compared with a multilevel model where only the characteristics of educational level are considered to be random between regions. The effects of the non random characteristics are very similar whether the first or second model is used. The case of the farmers now becomes fully significant: the fact of being a farmer always reduces the probability of migrating, whereas a high proportion of farmers in a region has the effect of increasing the probability of migrating for both farmers and non-farmers.

TABLE 10. – ESTIMATION OF THE PARAMETERS AND THEIR STANDARD ERROR (IN BRACKETS) OF THE SIMPLE AND MULTILEVEL LOGIT MODELS WHICH INTRODUCE THE VARIOUS INDIVIDUAL AND AGGREGATED CHARACTERISTICS, WITH A SIGNIFICANT EFFECT ON THE PROBABILITY OF MIGRATING IN 1980-1981 (GENERATION OF NORWEGIAN MEN BORN IN 1958)

Parameters	Simple logit	Multilevel (More than 12 years full-time education)
Fixed		
Constant	2.467 (0.856)	1.711 (1.152)
Married	0.641 (0.061)	0.653 (0.070)
Economically active	– 0.595 (0.046)	– 0.598 (0.085)
Farmer	– 0.226 (0.100)	– 0.208 (0.100)
More than 12 years full-time education	0.520 (0.063)	0.621 (0.082)
At least one child	– 0.467 (0.089)	– 0.467 (0.102)
Low income	– 0.256 (0.063)	– 0.261 (0.067)
High income	– 0.107 (0.051)	– 0.102 (0.084)
No income	– 0.610 (0.140)	– 0.619 (0.133)
Proportion of economically active	– 0.042 (0.011)	– 0.034 (0.014)
Proportion of farmers	0.070 (0.007)	0.074 (0.010)
Proportion with at least one child	– 0.155 (0.012)	– 0.138 (0.010)
Proportion with no income	– 0.078 (0.033)	– 0.100 (0.037)
Random regional level		
σ^2_{e0} (constant)		0.019 (0.009)
σ_{e01} (covariance)		– 0.057 (0.030)
σ^2_{e1} (more than 12 years full-time education)		0.150 (0.108)

Source: *Norwegian Population Register*, Central Bureau of Statistics, Oslo.

By contrast, the fact of having a high income becomes non-significant in a multilevel model. It is the random parameters at the level of the region that are modified the most, in relation to the model in which only the fixed educational level characteristic is included. The between-region variance σ^2_{eo} is reduced to a fifth of what it was (0.018 as against 0.99) by the introduction of the other characteristics. On the other hand, the between-region variance, $\sigma^2_{eo} + 2\sigma_{eo1} + \sigma^2_{e1}$, remains close to what it was (0.055 as against 0.063). Above all, however, the correlation between the regional hazards of individuals with more or less than 12 years of full-time education is –0.99 whereas it was almost nil (0.07) in the earlier model. This

produces Schema 3b of Figure 6 and which is to be compared with Schema 3a: the regions where the probability of migrating for those with less than 13 years full-time education is the lowest, when the effect of all the characteristics is allowed for, will be those where the probability of migrating for those with more than 12 years of full-time education is the highest, and vice-versa. Thus we see emerge a relationship which was obscured by the other characteristics, when these were not included in the model.

In the example used here, it can be concluded that the introduction of a model using the multilevel random variables does not alter the essential of the conclusions obtained with a simple logit model taking into account the characteristics measured at different levels of aggregation. On the other hand, these random variables provide some valuable information about the relationships between probabilities of emigrating from the different regions of individuals who have or do not have a given characteristic. These relations can be visualized by means of diagrams like those in Figure 6. Lastly it can be noted that in our example the effect of these random variables is often statistically not significantly different from zero, which makes the conclusions more tenuous.

IV. – Towards a multilevel event history analysis

The analyses presented so far constitute only a partial application of event history analysis: the individual regressions did include a score for pupils at the beginning and end of a four-year period, but they could not be used to study its evolution over the course of their entire schooling; and while logit analyses and event history models applied to Norwegian migrations considered the events occurring in a two-year period, they did not study a migration of a given rank over the entire stay of the individuals in different regions. So we now have to examine the possibilities and difficulties of conducting a multilevel event history analysis.

Because individuals are going to be followed throughout their stay in a given state, some of their characteristics can change at given times (they marry, change occupation, cease to be economically active, etc.) just as the characteristics of the regions in which they live can be expected to change over time (increase in the percentage of married people or of those with at least one child, changes in the percentage of economically active in the generation, etc.).

We thus need to know these characteristics for the entire life of the individuals; the Norwegian population register provides a record of individuals' marriages and the birth of their children, but it does not monitor their presence in the labour market, their occupation, their income, and so on. This information is only available in a census and we have used the information from that of 1980 to conduct our analyses on the short period

1981-1982. We are therefore unable to conduct this same analysis over the entire length of stay. A multilevel event history analysis thus encounters a problem of data availability. Calculating the regional characteristics on a continuous basis over time with sufficient accuracy requires event history data for much larger numbers than are usually considered. Contextual sample surveys introducing characteristics measured at different aggregation levels, are able to identify the links between individual behaviour and social structures. They appear to be indispensable for carrying out multilevel event history analyses. To do this, we need to "set up systems of observation that are representative of diversified and hierarchical social contexts, by combining in a system of integrated multilevel indicators the contributions of ecological analysis, individual sociological surveys and contextual analysis" (Loriaux, 1987). This is a field which though as yet little explored is essential.

On the other hand analytical techniques do exist for calculating a partial likelihood (Cox, 1972), which is the ratio of the hazard of the individual who experiences the event at a given time, to the sum of the hazard rates of the remaining population exposed to the risk. The product of these likelihoods, calculated for each time an event occurs, can be maximized by introducing several levels of aggregation (Goldstein, 1995). It will be seen that this quickly results in vast data files, which may exceed the size of active memory available, since for each event observed it is necessary to record all the characteristics of each member of the population exposed to the risk, while in addition these characteristics change between one event and the next.

Also, at a given level of aggregation an individual may move to another area during his stay in the population exposed to the risk. This can be shown by considering, for example, the study of fertility in different regions of a country. It is clear that some individuals can be expected to change residence between these regions during their reproductive period. The individuals must therefore be linked to each new region each time they move, and the effect of the aggregated characteristics of these regions will have an effect on their fertility behaviour. Yet this Markov hypothesis (that the behaviour of an individual depends only on the region in which he is at present and that when he arrives in a new region he immediately forgets the constraints of the regions previously inhabited), is scarcely plausible. The conditions need to be made less rigid. A solution is to test the speed of adaptation to conditions in the new region, if this is what is observed, or the conditions for the selection of migrants in the region of origin, if the second hypothesis is confirmed (Courgeau, 1989).

In this way we are led to non-Markov models of demographic behaviour, the complexity of which is to be added to the consideration of multiple levels of aggregation.

The development of authentic event history multilevel models is thus confronted by problems of observation and data availability, as well as by technical and analytical problems which remain largely unresolved.

Conclusions

In the course of this article we have gone from the simplest models, which introduce all the different levels in the form of individual and aggregated characteristics, to more complex models which operate with the random variables specific to each level, and culminating with multilevel event history models, which although they are the most satisfying are also the hardest to apply due to problems of observation and analysis.

These different developments illustrate the rich potential of multilevel models, but also the need to situate them in a coherent theoretical framework.

As a contribution to this, an epistemological reflection is needed in order to determine the significance to be accorded to the different levels of aggregation that can be used. Is it reasonable to interpret the aggregated characteristics as the reflection of the social organization in which we live, and the characteristics specific to each individual as the manifestation of individual liberty (Courgeau, 1996)? In this case what significance is to be attached to the use of a large number of levels of aggregation (from the individual to the household, the district, the commune, the department, the region, and so forth)? Is there not a case for trying to identify the important levels which are not necessarily administrative levels of aggregation (employment areas, for example), which must then be integrated into a more general theory? Finally, there is also a need to articulate these different levels to each other, as they are not independent of each other.

Should we not also try to advance beyond the individual approach adopted here, in which behaviour is explained by characteristics measured at different levels of aggregation? This study may need to be completed by a study of the behaviours that are specific to the different levels and then try to relate them to each other. At the level of a given community, for example, isolated individual actions can bring about awareness of a problem that affects the whole of the community. This could then result in political measures being taken at a more aggregated level. And these measures will in turn of course influence individual behaviour, producing new actions to compensate for their undesirable effects, and so on and so forth.

Lastly, do we not also need to take into account the social structure of the groups being considered? The research presented by Bonvalet, Bry and Lelièvre earlier in this special issue of *Population* illustrates the possibility of doing this when small groups, such as the family or the household, are being examined. Where larger groups are concerned, can they be defined accurately using the average values of individual characteristics or even using variances and covariances? A full treatment of their social structure may also require

taking into account the interactions that exist between the members of the group and the changes over time in their interactions. This is a difficult task, for which new means of observation will have to be developed.

Such an approach thus goes beyond the analytical methods proposed here and looks forward to the development of a theory of human behaviour whose epistemological bases, methods of measurement and analysis remain as yet largely undefined. Future research will confirm or deny the value of such an approach in facilitating the simultaneous study of the different levels of aggregation that are encountered in the social sciences.

Daniel COURGEAU
and Brigitte BACCAÏNI

ACKNOWLEDGEMENTS

Earlier versions of this article were presented and discussed at the international conference on "Analyse en multiples niveaux: problématique générale et méthodologie", 11 October 1996, at Louvain-la-Neuve, in the "Démodynamiques" seminar at INED, 16 January 1997, and in the Franco-Dutch seminar on "Residential mobility and housing choices", 3 and 4 April 1997. We would like to thank Dominique Tabutin for his comments. Finally, our thanks to the Norwegian statistical services who allowed us to use the files produced from the population register and created by Kjetil Sørlie and Øjsten Kravdal.

REFERENCES

ALKER H.-R., (1969), "A typology of ecological fallacies", in Dogan and Rokkan (eds), *Quantitative ecological analysis*, MIT Press, Massachussets, pp. 69-86.

BACCAÏNI B., COURGEAU D., (1996a), "Approche individuelle et approche agrégée : utilisation du Registre de population norvégien pour l'étude des migrations", in J.-P. Bocquet-Appel, D. Courgeau and D. Pumain (eds), *Spatial Analysis of Biodemographic Data*, Congresses & Colloquia N°16, John Libbey/Ined, Paris, pp. 79-104.

BACCAÏNI B., COURGEAU D., (1996b), "The spatial mobility of two generations of young adults in Norway", *International journal of population geography*, vol. 2, n° 4, pp. 333-359.

COURGEAU D., (1989), "Family formation and urbanization", *Population. An English Selection*, n° 1, pp. 123-146.

COURGEAU D., (1995), "From the group to the individual: what can be learned from migratory behaviour", *Population. An English Selection*, n° 7, pp. 145-162.

COURGEAU D., (1996), "Towards a multilevel analysis in social sciences"/"Vers une analyse multi-niveaux en sciences sociales", in J.-P. Bocquet-Appel, D. Courgeau and D. Pumain (eds), *Spatial Analysis of Biodemographic Data*, Congresses & Colloquia N° 16, John Libbey/Ined, Paris, pp. 10-22.

COURGEAU D., LELIÈVRE É., (1989), *Analyse démographique des biographies*, Éditions de l'Ined, Paris, 268 p. English edition (1992), *Event History Analysis in Demography*, Clarendon Press, Oxford.

COURGEAU D., LELIÈVRE É., (1997), "Changing paradigm in demography", *Population. An English Selection*, n° 9, pp. 1-10.

COX D.-R., (1972), "Regression models and life tables (with discussion)", *Journal of the Royal Statistical Society*, B34, pp. 187-220.
ENTWISTLE B., MASON W.-M., (1985), "Multilevel effects of socio-economic development and family planning programs on children ever born", *American Journal of Sociology*, 91, pp. 616-649.
FIREBAUGH G., (1978), "A rule for inferring individual-level relationships from aggregate data", *American Sociological Review*, 43, pp. 557-572.
GERONIMUS A.-T., BOUND J., NEIDERT L.-J., (1996), "On the validity of using census geocode characteristics to proxy individual socio-economic characteristics", *Journal of the American Statistical Association*, 91, pp. 529-537.
GOLDSTEIN H., (1986), "Multilevel mixed linear model analysis using iterative generalized least squares", *Biometrika*, 73, pp. 43-56.
GOLDSTEIN H., (1987), "Multilevel covariance component models", *Biometrika*, 74, pp. 430-431.
GOLDSTEIN H., (1991), "Nonlinear multilevel models, with an application to discrete response data", *Biometrika*, 78, pp. 45-51.
GOLDSTEIN H., (1995), *Multilevel Statistical Models*, Edward Arnold, 178p.+XIV.
HAUSER R.-M., (1974), "Contextual analysis revisited", *Sociological Methods and Research*, vol. 2, n° 3, pp. 365-375.
JACQUOT A., (1994), "1982-1990 : un modèle de déséquilibre pour les marchés régionaux du travail en France", *Revue d'Économie Régionale et Urbaine*, 3.
JONES K., (1993), *Everywhere is nowhere : multilevel perspectives on the importance of place*, The University of Portsmouth Inaugural Lectures, 12 p.
LANCASTER T., (1990), *The Econometric Analysis of Transition Data*, Econometric Society Monographs, Cambridge University Press,
LAZARSFELD P.-F., MENZEL H., (1961), "On the relation between individual and collective properties", in Etzioni (ed.), *Complex Organizations*, Holt, Reinhart and Winston, New York, pp. 422-440.
LORIAUX M., (1989), "L'analyse contextuelle : renouveau théorique ou impasse méthodologique", in G. Wunsch and E. Vilquin (eds), *L'explication en sciences sociales : la recherche des causes en démographie*, Duchêne, , Éditions Ciaco, Louvain-la-Neuve, pp. 333-368.
PIANTADOSI S., BYAR D., GREEN S., (1998), "The ecological fallacy", *American Journal of Epidemiology*, 127, pp. 893-904.
PUIG J.-P., (1981), "La migration régionale de la population active", *Annales de l'INSEE*, n° 44, pp. 41-79.
ROBINSON W.-S., (1950), "Ecological correlations and the behaviour of individuals", *American Sociological Review*, 15, pp. 351-357.
TUMA N.-B., HANNAN M.-T., (1984), *Social Dynamics. Models and Methods*, Academic Press, Orlando, 580p.+XX.
VON KORFF M., KOEPSELL T., CURRY S., DIEHR P., (1992), "Multilevel analysis in epidemiologic research on health behaviors and outcomes", *American Journal of Epidemiology*, 135, pp. 1077-1082.
WILLEKENS F., ROGERS A., (1978), *Spatial Population Analysis : Methods and Computer Programs*, Research Report, IIASA, Laxenburg, Austria, 302 p.
WOODHOUSE G., RASBASH J., GOLDSTEIN H., YANG M., (1996), "Introduction to multilevel modelling", in *Multilevel Modelling Applications*, Woodhouse ed., Institute of Education, London, pp. 9-57.
YOUNG E.-C., (1924), *The Movement of Farm Population*, Cornell Universtiy, Ithaca, New York.

COURGEAU (Daniel), BACCAÏNI (Brigitte). – **Multilevel analysis in the social sciences**

The multilevel approach can be used to study human behaviour taking into account not only individual characteristics but also the fact that these individuals belong to larger geographical units such as communes and regions. This article gives a detailed critical presentation of the aims and formulations of these models. Attention ranges from the most basic models, which introduce the many different levels in the form of individual and aggregated characteristics, to more complex models which operate with the random characteristics specific to each level, and culminates with multilevel event history models. The article concludes with a more general epistemological reflection on the contribution of these models.

COURGEAU (Daniel), BACCAÏNI (Brigitte). – **Analyse multi-niveaux en sciences sociales**

L'approche multi-niveaux permet d'aborder les comportements humains, en tenant compte non seulement des caractéristiques individuelles, mais également du fait que ces individus font partie d'unités géographiques plus larges telles que les communes ou les régions. Une présentation détaillée et critique est faite ici des objectifs et des diverses formulations de ces modèles. Cet article va des modèles les plus simples, qui font intervenir la multiplicité des niveaux sous la forme de caractéristiques individuelles et agrégées, à des modèles plus complexes mettant en oeuvre des aléas propres à chaque niveau, pour aboutir à des modèles biographiques multi-niveaux. Il ouvre à une réflexion épistémologique plus générale sur l'apport de ces modèles.

COURGEAU (Daniel), BACCAÏNI (Brigitte). – **Análisis multi-nivel en ciencias sociales**

El análisis multi-nivel permite abordar los comportamientos humanos teniendo en cuenta no únicamente las características individuales, sino también el hecho de que los individuos forman parte de unidades geográficas tales como los municipios o las regiones. El artículo hace una presentación detallada y crítica de los objetivos y formulaciones diversas de tales modelos. Se va desde los modelos más simples, que hacen intervenir la multiplicidad de niveles bajo forma de características individuales y agregadas, a modelos más complejos que incluyen factores de heterogeneidad en cada nivel, hasta llegar a modelos biográficos multi-nivel. Finalmente, se realiza una reflexión epistemológica general sobre la aportación de estos modelos.

Daniel COURGEAU, Institut national d'études démographiques, 133, bd Davout, 75980 Paris Cedex 20, France, tél: [33] (0)1 56 06 21 07, fax: [33] (0)1 56 06 21 99, e-mail : courgeau@ined.fr