1 Introduction

Multilevel analysis is a methodology for the analysis of data with complex patterns of variability, with a focus on nested source of such variability, e.g. students in classes, employees in firms, suspects tried by judges in courts, animals in litters, longitudinal measurements of subjects. When analyzing nested data, it is key to take into account that each level of nesting is associated with variability that has a specific interpretation. For example, there is variability between students and there is variability between schools. One could reach misleading conclusions if one does not recognizes both sources of variability and does not interpret them accordingly.

Multilevel analysis is a collection of statistical methods for the analysis of nested data. The term 'multilevel analysis' is widespread in the social sciences and biomedical sciences. The terms 'hierarchical linear models', 'mixed models' and 'random coefficient models' refer to techniques used in multilevel analysis. There are two bodies of literature that have contributed to what is nowadays known as multilevel analysis: contextual analysis and mixed effects models. Contextual analysis is a development in the social sciences focusing on the effects of social context on individual behavior. Mixed effects models are statistical models in the analysis of variance (ANOVA) and in regression analysis where it is assumed that some of the coefficients are fixed and others are random. Multilevel analysis was formed based on the realization that, in contextual modeling, the individual and the context are distinct sources of variability and should both be modeled as random influences. The main statistical model of multilevel analysis is the hierarchical linear model, an extension of multiple linear regression to a model that includes nested random coefficients.

1.1 Multilevel data: when the multilevel structure is not of primary interest

A common assumption in statistical modeling is the assumption of independence of samples. Under simple random sampling with replacement, the result of one selection is independent of the result of any other selection and each individual in the population has the same chances of being sampled. However, other sampling schemes are more cost-effective. For example, it is cheaper to visit 100 schools and interview 10 students per school than to travel to 1000 schools and interview one student per school. Multistage sampling is an example of a more judicious sampling design: the population of interest consists of subpopulations (clusters) and selection takes place via those subpopulations. If there is only one subpopulation level, the design is a two-stage sample. Students are grouped in schools, hence the populations of students consists of subpopulations of schools that contain students. In a random two-stage sampling, a random sample from the primary units (schools) is taken at the first stage, and then the secondary units (students) are sampled at random from the selected primary unit (school). A graphical representation of two-stage sampling design is presented in Figure 1. Relevant terminology for two-stage sampling is presented in Table 1, while other examples of primary and secondary units are presented in Table 2.

It is of key importance to properly account for the fact that the secondary units (students) were not sampled independently of each other: after selecting a primary unit (school), the chances of the secondary units (students) in that primary unit increase. If a primary unit is not selected, the secondary units that belong to that primary unit cannot be selected. The implication is that multistage sampling induces *dependence* of observations. For example, the dependence between students in the same school may be due to:

1. students within a school sharing the same school environment
2. students within a school sharing the same teachers
3. students within a school affecting one another by direct communication
4. students within a school affecting one another by developing shared group norms
5. students within a school coming from the same neighborhood

Strongly correlated achievement levels of students within particular schools as compared to students from other schools are indicative that probable causes of achievement are related to the organizational unit (the school). If so, we may wish to make inferences on the schools (which schools are better for students?) as well as on the students (which students perform best). In this case variables are defined at the primary unit level (schools) and other variables are defined at the secondary unit level (students).

Ignoring the dependence of samples in a multistage design can lead to error especially by underestimating *uncertainty* (e.g., standard errors, p-values). The reason is simple: the data contain less information if the samples are dependent as compared to the case when the samples are independent. However, the estimates obtained by ignoring dependence are (nearly) unbiased and consistent (i.e., the estimates converge to the true value as the number of observations in each micro-level unit becomes large). Therefore the main error induced by ignoring dependence is in the statement of uncertainty about the estimates rather than in the estimates themselves.

A special type of nesting is defined by repeated measurements/longitudinal data. The macro-level units are the subjects, while the micro-level units are measurement occasions. Since the same quantity is measured for each subject at several time points, the dependence of the different measurements for a given subject is of primary importance in longitudinal data. Growth curve data is similar to repeated measurements data, but in this case a pattern of expected over time. For

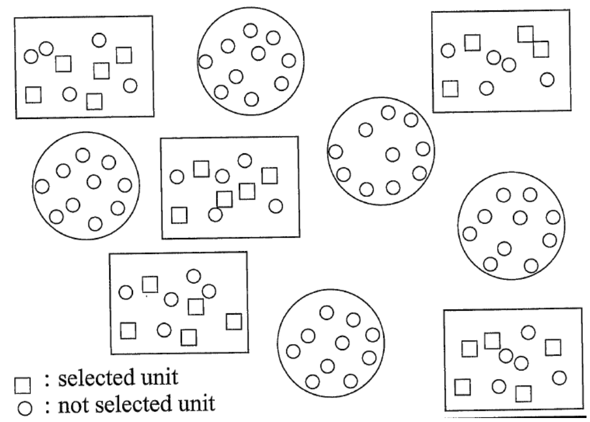


Figure 1: Example of two-stage sampling. A population consists of 10 subpopulations, each with 10 micro-units. A sample of 25% is taken by randomly selecting 5 out of 10 subpopulations and within these randomly sampling 5 out of 10 micro-units.

Example:	Schools	Students
	macro-level units	micro-level units
	macro-units	micro-units
	primary units	secondary units
	clusters	elementary units
	level-two units	level-one units

Table 1: Equivalent terminology in two-stage sampling design

Macro-level	Micro-level
schools	teachers
classes	students
neighborhoods	families
families	children
companies	employees
jawbones	teeth
litters	animals
doctors	patients
interviewers	respondents
judges	suspects

Table 2: Examples of units at the macro and micro level

example, recording the heights of individuals (increasing than stationary over time) or looking at criminal offending by age (increasing than decreasing over time). Another key type of nesting is time series cross-sectional data — see Figure 2. On each macro-level unit, measurements are taken sequentially in time, often at equally spaced intervals (e.g., every year).

1.2 Multilevel data: when the multilevel structure is of primary interest

This is the case when we want to make inferences about primary (macro-level) units as well as secondary (micro-level) units. Examples:

- Is there a relation between student test scores and their teachers' experience?
- Do employees in multinational companies earn more than employees in other firms?
- Is the sentence differential between black and white suspects different between judges? Is so, can we find the characteristics of judges to which this sentencing differential is related?

Multilevel propositions can be represented as in Figure 3. Types of propositions include:

1. about micro-units
    - e.g. ”boys lag behind girls in reading comprehension”
2. about macro-units

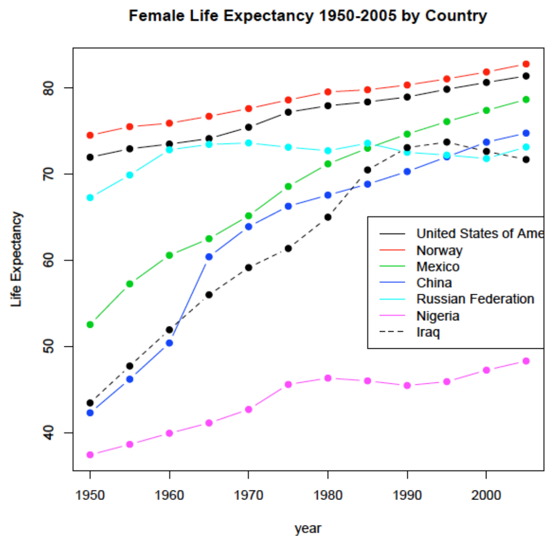


Figure 2: Example of time series cross-sectional data: life expectancy over time. Here countries are macro-level units and time periods are micro-level units.

- e.g. "crime rates are higher in neighborhoods with bad housing conditions"

3. about macro-micro relations

    - e.g. "a child living in a high crime neighborhood will tend to have lower test scores"
    - e.g. "a child living in bad housing conditions will affect classroom climate"

One option to proceed with the statistical analysis of multi-level data is aggregation. That is, the micro-level data are aggregated to the macro level. In the simplest form, one can work with the averages for each macro-level unit. Such an approach is acceptable if interest lies in macro-level proportions. But aggregation can lead to errors for micro and macro-micro proportions:

1. Shift of meaning: aggregated variables refer to macro-level variables, not micro units. E.g., the average test score in a school refers to the school, not directly to the students.

2. Ecological fallacy:

    - A correlation between macro-level variables cannot be used to make assertions about micro-level relations. E.g., the number of immigrants in a neighborhood could be related to the proportion of people with extreme right-wing views. But that does not mean that immigrants tend to have extreme right-wing views.

3. Neglect the original data structure. Please see the code below.

```
# Dataset with 25 individuals in 5 groups (g) and one predictor variable (x)
g <- rep (1:5, times=rep(5,5))  # Grouping variable
x <- c(1:5,2:6,3:7,4:8,5:9)
y <- x-2*g+rnorm(25,0,.3) # Simulate outcome variable y
par (mfrow=c(1,2))
plot (x,y,type="n")
for (i in 1:5) points (x[(5*(i-1)+1):(5*i)],y[(5*(i-1)+1):(5*i)],pch=i)

# So within each group, y is *postively* related to x:
#  y = b0 + b1 x + eps,
#   where b1=1.

# Form group means
ymean <- rep(NA,5)
xmean <- rep(NA,5)
for (i in 1:5) {
  ymean[i] <- mean (y[g==i])
  xmean[i] <- mean (x[g==i])
}
points (xmean,ymean,pch=19)

summary (lm (ymean ~ xmean))

# So if we regressed the group means, we get a *negative* relationship,
#  y = b0 + b1 x + eps
#   with b1 = -1, and R2=0.9988 and t=-50 !
# This is the exact reverse of the correct answer.

summary (lm (y~x))

# So if we regress y on x without controlling for the grouping variable,
#  we get a very small and non-significant coefficient.

# Question: What are the group differences after controlling for x?
# Micro-level analysis: Regression of y on x
res1 <- residuals (lm(y~x))
res1means <- tapply (res1,g,mean)
plot (1:5,res1means,col="blue",pch=19, xlab="Group",ylab="Adjusted mean of y")
# So when we do the analysis on the micro scale, we see that there's
# a big difference between groups after controlling for x.
```

```

# Macro-level analysis (aggregate first):
# Regress the y group mean on the x group mean
res2 <- residuals (lm(ymean~xmean))
points (1:5, res2, col="red",pch=19)
# So when we aggregate first, we get the impression that there are no
#  group differences after controlling for x.
# So aggregating is very misleading in this case.
```

Listing 1: Code illustrating the dangers of neglecting the original data structure. Courtesy of Adrian Raftery.

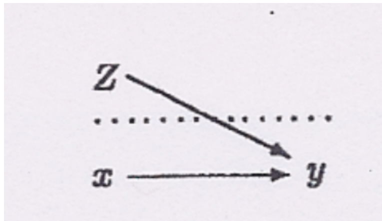
At the opposite end of aggregation, one can choose to ignore the measures of macro-level variables and work only with micro-level data. This is called disaggregation: we treat multilevel data as micro-level data. However, statistical analyses based on disaggregation can lead to overestimating the effective amount of information.

- Example: Do older judges give shorter sentences?
- Two-stage sample: 10 judges sampled, and then 10 trials sampled per judge (100 trials sampled in all)
- One could do regression using trials as units of analysis (disaggregation).
- But this would be pretending that there are 100 independent observations, whereas in fact there are only 10 (the judges).

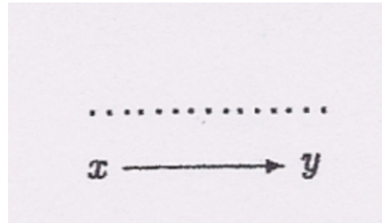
If data are analyzed at the micro-level, correlation has to be taken into account. This is the design effect of two-stage sampling.

- Example: Do black suspects get longer sentences?
- We could address this from the 100 trials from the previous example.
- But if there are systematic differences between judges, we would overestimate the significance of any differences.
- The design effect implies that the ”effective” sample size is less than 100.
- It involves the intraclass correlation coefficient.

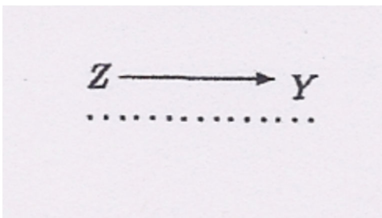
Multilevel statistical models are always called for if we are interested in propositions that connect variables defined at different levels, the micro and the macro, and also if a multistage sample design has been employed.



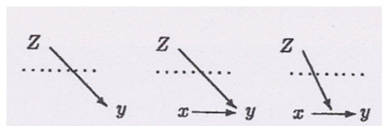
(a) The structure of a multi-level proposition



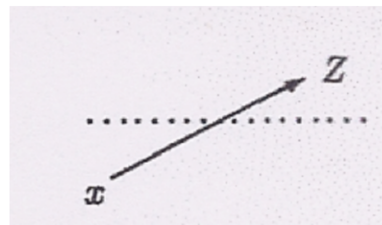
(b) The structure of a micro-level proposition



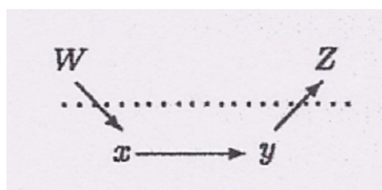
(c) The structure of a macro-level proposition



(d) The structure of macro-micro propositions



(e) The structure of a micro-macro proposition



(f) A causal macro-micro-micro-macro chain

Figure 3: Various types of multilevel propositions.

2 Multilevel structures

In this section we present graphical displays for data that come from populations with a complex structure. We start by identifying the atomic units which are the units of the lowest level of the system. Often, but not always, the atomic units are individuals, e.g. students. The atomic units are subsequently grouped into higher-level units, e.g. schools. We follow the convention that students are at level 1 and schools are at level 2 in the resulting structure. Hierarchical structures arise when the lower-level unit nests in one and only one higher-level unit. They can accommodate a wide range of study designs and research questions.

2.1 Two-level hierarchical structures

Figure 4 is a *unit diagram* which aims to show the underlying structure of a research problem in terms of individual units; the nodes on the diagram are specific population units. In this case the units are students and schools which form two levels (or classifications). The lower units form the student classification (St1, St2 etc.) and the higher units form the school classification (Sc1,..., Sc4). This unit diagram is just a schema to convey the essential structure of students nested within schools. In a real data set we would have many more than four schools and 12 students. The hierarchical structure means that a student only attends one school and has not moved about. Such a structure may arise when we are interested in school performance and we make repeated measurements of this by assessing student performance for multiple students from each school. This structure is likely to give rise to dependent data, in the sense that students in the same school will often have a tendency to be similar on such variables as exam performance. Figure 5 gives another representation of the nested structure from Figure 4. This is a *classification diagram* with one node per level. Nodes joined by a single arrow indicate a nested (strict hierarchical) relationship between the classifications. Classification diagrams are more abstract than unit diagrams and are particularly useful when the population being studied has a complex structure with many classifications.

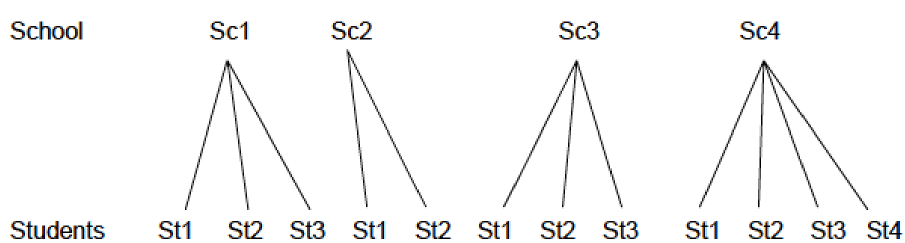


Figure 4: Unit diagram of a two-level nested structure: students in schools

Sample Size. A question that often comes up is how many units are needed at each level. It is difficult to give specific advice but there are some general principles that are worth stating now. The key one is the target of inference: in other words, are the units in your dataset special ones that you are interested in in their own right, or are you regarding them as representatives of a larger population which you wish to use them to draw conclusions about? If the target of inference in an educational study is a particular school then you would need a lot of students in that school to get a precise effect. If the target of inference is between-school differences in general, then you would need a lot of schools to get a reliable estimate. That is, you could not sensibly use a multilevel model with only two schools even if you had a sample of 1000 students in each of them. In the educational literature it has been suggested that, given the size of effects that are commonly found for between-school differences, a minimum of 25 schools is needed to provide a precise estimate of between-school variance, with a preference for 100 or more schools. You would not normally omit any school from the analysis merely because it has few students, but at the same time you will not be able to distinguish between-school and between-student variation if there is only one student in each and every school. Note that schools with only one student still add information to the estimates of the effects of the explanatory variables on the mean. There are, of course, some contexts where some or all of the higher-level units will have only a few lower-level units. An extreme and common case is when individuals are at level 1 and households are at level 2, because then the sample size within a level 2 unit is typically less than five people. This need not be a problem if the target of inference is households in general because the quality of estimates in this case is based on the total number of households in the sample and it should be possible to sample a large number of these. If the target of inference is a specific household, however, parameters will be poorly estimated because a single household has very few members.

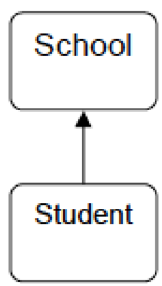


Figure 5: Classification diagram of a two-level nested structure: students in schools

A lot of confusion is related to the following question: When is a variable to be treated as a classification or level as opposed to an explanatory variable? For example, school type (e.g., state vs. private) is a classification of schools so why not redraw Figures 4 and 4 as a three-level multilevel model – see Figure 6. School type is certainly a way of classifying schools and as such it is a classification. However, we can divide classifications into two types which are treated in different ways when modeling: (i) random classifications and (ii) fixed classifications. A classification is a *random classification* if its units can be regarded as a random sample from a wider population of units. For example the students and schools in our example are a random sample from a wider population of students and schools. However, school type or indeed student gender has a small fixed number of categories. There is no wider population of school types or genders to sample from. State and private are not two types sampled from a large number of school types, and male and female are not just two of a possibly large number of genders. Students and schools, however, can be treated as a sample of students and schools to which we want to generalize. The distinction between fixed and random classifications has important implications for how we handle the classifying variable in a statistical analysis. Strictly speaking, for a classification to correspond to a level in a multilevel model, it must be a random classification. It turns out that school type is not a random classification; it is better to conceive of this as a fixed classification, and that is why we would treat school type as a variable not as a level.

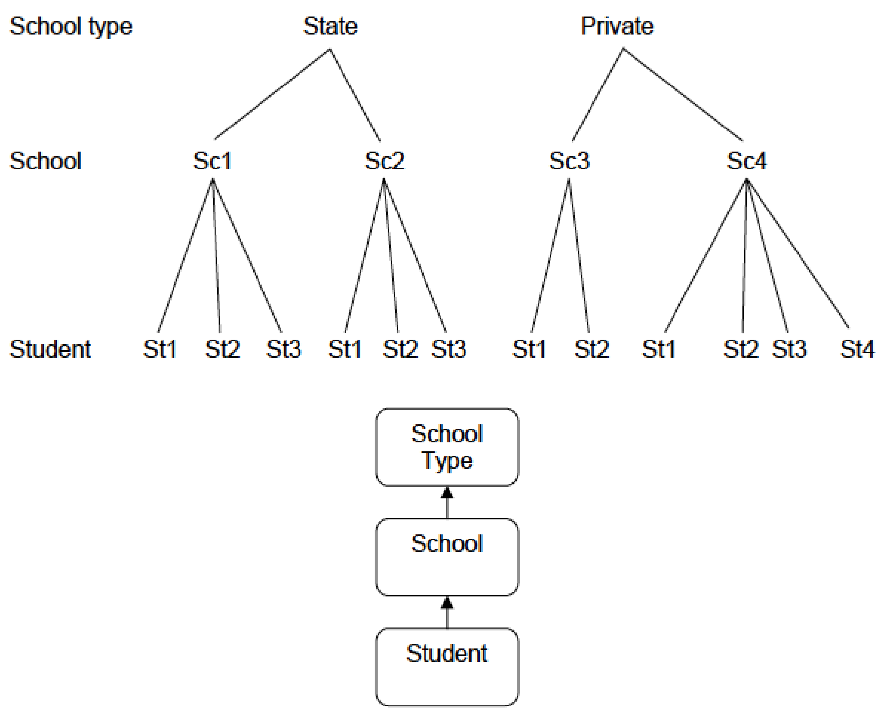


Figure 6: Unit and classification diagrams for a three-level nested structure: students in schools in school types

2.2 Other examples of a two-level structure

Other common examples of two-level nested structures are people within households, patients within hospitals, and people within neighbourhoods. All these examples and the students-in-schools example arise when the real world has a multilayered structure, that is, the levels exist in the population. However, the multilevel structure might also be imposed through the study design and data collection. We now cover other interesting examples of two-level nested structures that arise from common types of research design that can be found in social science research:

- Repeated measures, panel data
- Multivariate designs
- Multistage survey designs

2.2.1 Repeated measurements within individuals, panel data

We assume we have recorded measures on an individual on two time occasions, e.g. a prior test score and a current test score for students. We can analyze change (that is, progress) in this situation by specifying the current attainment as the response and including prior score as an explanatory variable. However, when there are measurements on more than two occasions there are advantages in treating occasion as a level nested within individuals. Such a two level strict hierarchical structure is known as a repeated measurement or panel design — see Figure 7. This occurs when we have repeated measurements (level 1) over time on a number of people (level 2). We think of measurement occasion as being nested within individuals and this is an example of when the atomic units are not individuals. These structures apply when we are analyzing the extent and nature of variation between individuals in their patterns of growth. With repeated measurements we can often expect quite strong correlations across time within individuals. For example, a tall person is likely to continue to be tall across time and their value this year is likely to be related to the value last year.

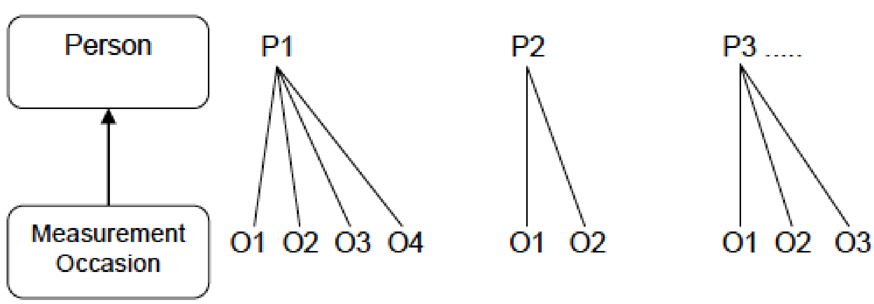


Figure 7: Classification and unit diagram for a two-level repeated measures design

2.2.2 Multivariate responses within individuals

Sometimes we may wish to model more than one response. For example, we may wish to consider jointly English and mathematics exam scores for students because the two responses are likely to be related. We can regard this as a multilevel structure with subjects (English and maths) nested within students – see Figure 8.

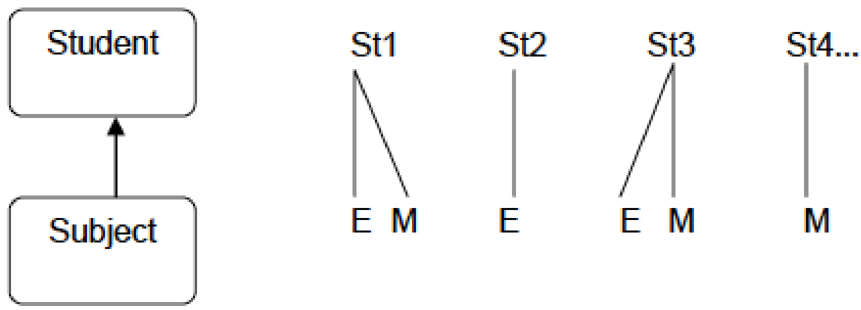


Figure 8: Classification and unit diagrams for a multivariate response model

2.2.3 Two-stage sample survey design

In many large scale face-to-face surveys, the data collectors may adopt a two-stage design to minimize interviewer costs. In a study of voting, for example, the researchers may first select constituencies, the so-called primary sampling unit (PSU), and then select individuals within these areas. This selection procedure leads to a sample that is geographically clustered — see Figure 9. There would of course normally be many more than the four clusters shown here.

2.3 Three-level hierarchical structures

As a first example we will consider an analysis of students nested within classes in schools. The structure is a strict hierarchy if each student belongs to one and only one classroom and each classroom group is found in one and only one school, as is often the case. The unit and classification diagrams are given in Figure 10.

Other three-level nested structures include:

- Repeated measures within students within schools. This allows us to look at how learning trajectories vary across students and schools.

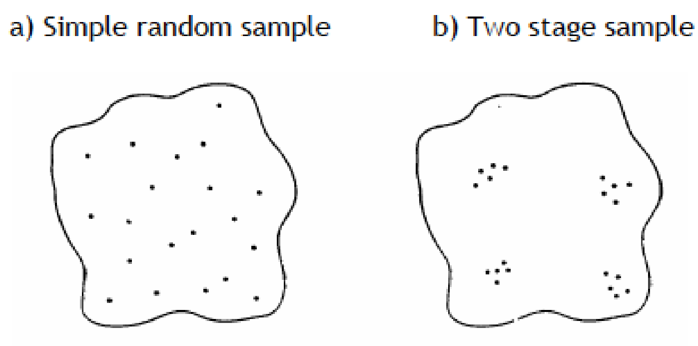


Figure 9: A schematic map of respondents obtained from a simple random sample contrasted with that from a two-stage sample design

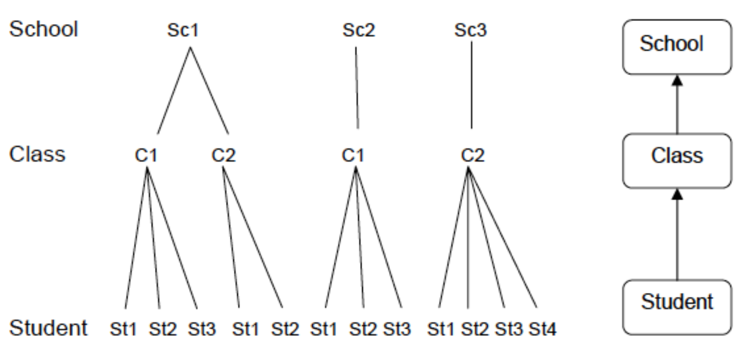


Figure 10: Unit diagram and classification diagram for the three-level structure of students within classes within schools

- Multivariate responses on four health behaviours (drinking, smoking, exercise and diet) on individuals within communities. Such a design allows the assessment of the correlation between these behaviors at the individual level and at the community level, and to do so taking account of other characteristics at both the individual and community level.

Figure 11 shows the unit and classification diagrams where we have exam scores for groups of students who entered school in 1990 and a further group who entered in 1991. The model can be extended to handle an arbitrary number of cohorts. This is an example of a repeated cross-sectional design: students within cohorts within schools.

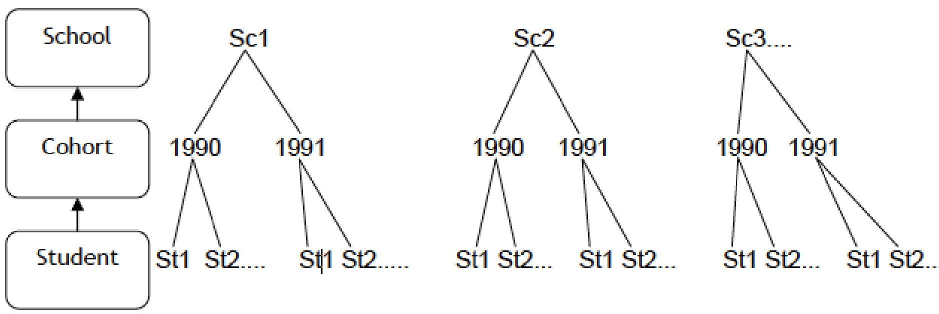


Figure 11: Classification and unit diagrams for students within cohorts within schools

2.4 Four-level hierarchical structures

Here are a few examples:

- Student within classes within schools within local education authorities
- Multivariate responses within repeated measures within students within schools
- Repeated measures within patients within doctors within hospitals
- People within households within postcode sectors within regions

The last example of a hierarchical structure is a doubly nested repeated measures structure. Suppose we have repeated measures within students within cohorts within schools – see Figure 12. Cohorts are now repeated measures on schools and tell us about stability of school effects over time. Measurement occasions are repeated measures on students and can tell us about studentsearning trajectories. The measurement occasions refer to the age of the student rather than the time, so that measurement occasion 1 for all students is when they are in the first year of school, for example, rather than measurement occasion 1 being 1991, say, for all students. Note that there is no requirement for this to be the case in general: in a different example, measurement occasions could be calendar years, for example, with different individuals having different ages on the same measurement occasion.

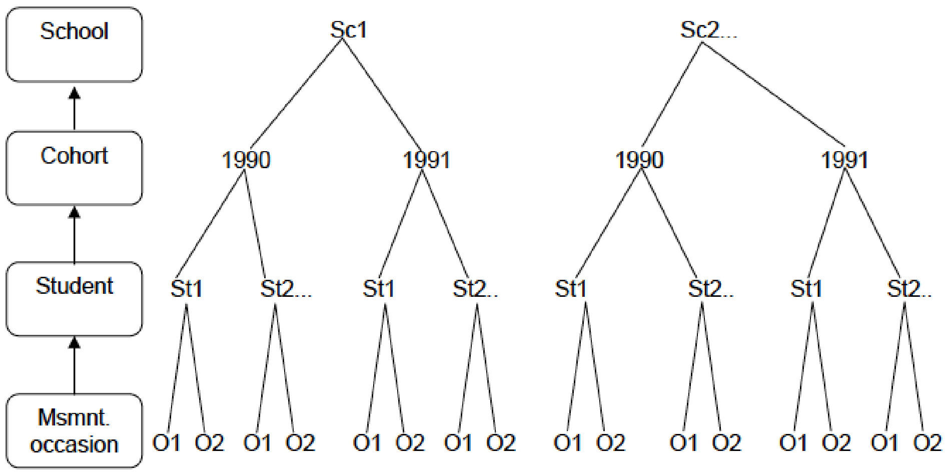


Figure 12: Unit and classification diagrams for a doubly nested repeated measures structure

2.5 Non-hierarchical structures

So far all our examples have been of an exact nesting of a lower level unit in one and only one higher-level unit. That is, we have been dealing with strict hierarchies. But social reality can be more complicated than that! In fact it turns out that we need two non-hierarchical structures which, in combination with strict hierarchies, will be able to deal with all the different types of designs, realities and research questions that we will meet:

- Cross-classified structures
- Multiple membership structures

The existence of non-hierarchical structures is the reason why the term ”multilevel models” is typically preferred to ”hierarchical models.”

2.5.1 Cross-classifications: students cross-classified by school and neighborhood

Consider the unit diagram in Figure 13 which links students to their area of residence and to the school they attend. You can see that there is a non-nested structure, for example:

- Students 1 and 2 attend School 1 but come from different areas
- Students 6 and 10 come from the same area but attend different schools

Area is not nested within school and school is not nested within area. Students lie within a cross-classification of school by area. Students are nested within schools and students are nested within areas. However, schools and areas are cross-classified, so the nodes for school and area are not connected on the classification diagram of Figure 14.

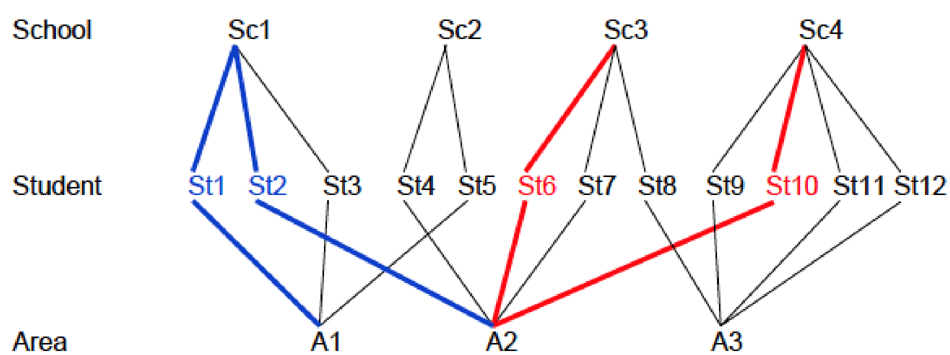


Figure 13: Unit diagram for a two-way cross-classified structure

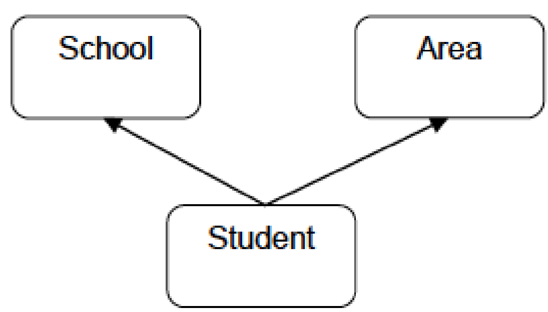


Figure 14: Classification diagram for a two-way cross-classified structure

2.5.2 Repeated measures within a cross classification of patients by clinician

We consider a health setting with repeated measures on patients but patients being assessed by different clinicians at different times. This is depicted in Figure 15 where, for example, patient 1 is seen by clinician 1 at occasions 1 and 2 then by clinician 2 at occasion 3. This is another example of a cross-classification.

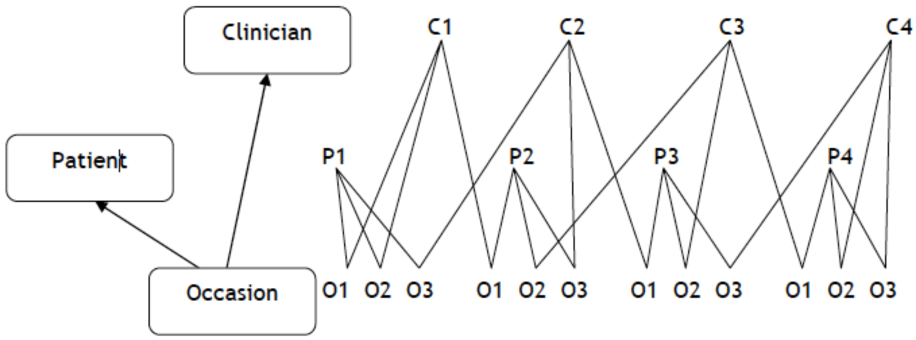


Figure 15: Classification and unit diagrams for cross-classification of measurement occasion by patient and clinician

Other examples of two-way cross-classifications include:

- Exam scores within a cross-classification of student and examiner, where a student paper is graded by more than one examiner to get an indication of examiner reliability.
- Students within a cross-classification of primary school by secondary school. We may have studentsxam scores at age 16 and wish to assess the relative effects of primary and secondary schools on attainment at age 16.
- Patients within a cross-classification of general practitioner (GP) practice and hospital.

In many of these examples, individuals are seen as occupying more than one set of contexts, for example students may be influenced by residential setting and school setting.

2.6 Combining structures: hierarchies, cross-classifications and multiple membership relationships

It is possible to combine the three types of structure (strict hierarchy, cross-classification and multiple membership) to reflect social reality and the research designs employed to study it. Consider the structure depicted in Figure 16. Students are nested within a cross-classification of school by area. However

- Student 1 moves in the course of the study from residential area 1 to 2 and from school 1 to 2
- Student 8 has moved schools but still lives in the same area

- Student 7 has moved areas but still attends the same school

Now in addition to schools being crossed with residential areas, students are multiple members of both areas and schools. In the classification diagram, students are connected to schools and areas by double arrows to represent two sets of multiple membership relationship. School and Area are not connected by arrows indicating that the lower level units, students, lie within a cross-classification of school by area.

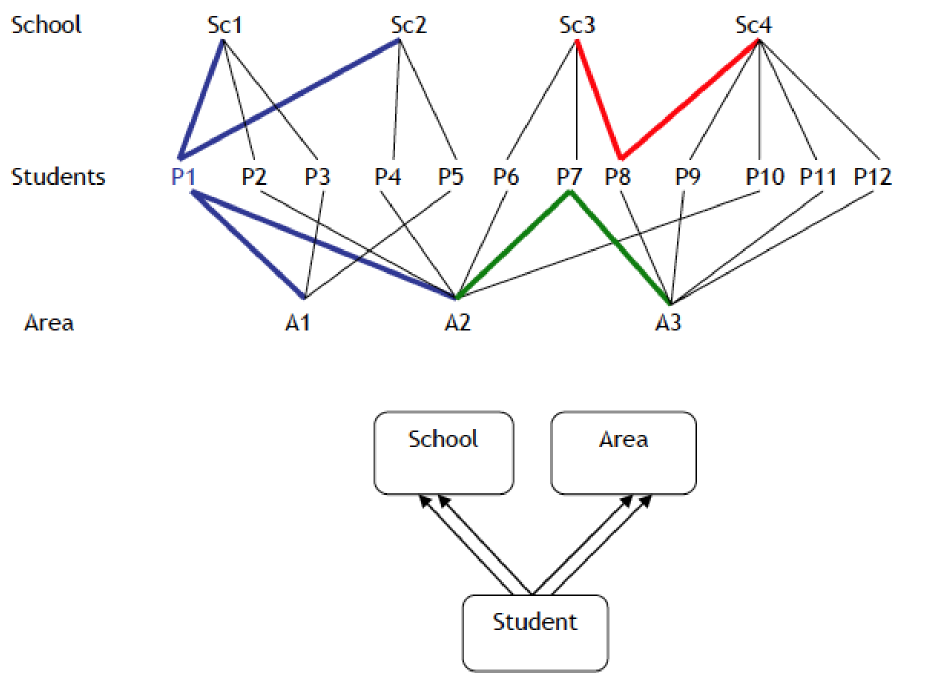


Figure 16: Unit and classification diagrams for cross classifications and multiple memberships: students, areas and schools

Another example of a non-hierarchical structure relating patients to nurses, hospitals and GPs is shown in Figure 17. In summary:

- Patients are multiple members of nurses
- Nurses are nested within hospitals
- Patients are nested within GPs
- Nurses are crossed with GP
- Hospitals are crossed with GPs

This relatively complicated structure is more succinctly expressed in the classification diagram from Figure 18.

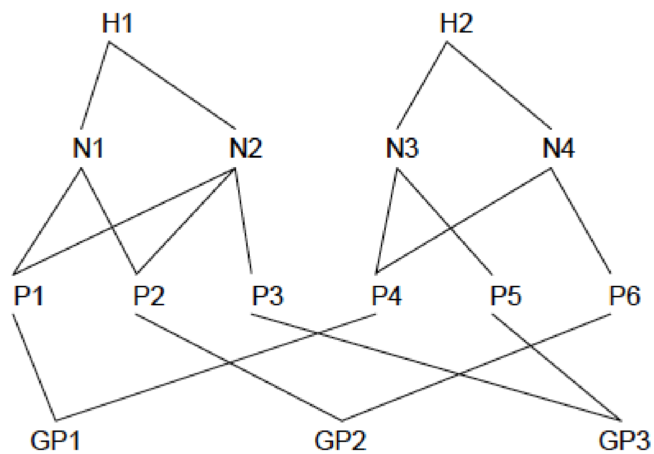


Figure 17: Unit diagram showing a complex mixture of multiple membership, hierarchical and crossed relationships

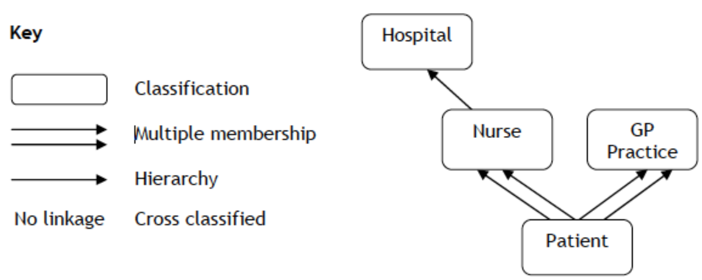


Figure 18: Classification diagram showing a complex mixture of multiple membership, hierarchical and crossed relationships