An exploration of childhood antecedents of female adult malaise in two British birth cohorts: Combining Bayesian model averaging and recursive partitioning

John Hobcraft
and
Wendy Sigle-Rushton

Contents

Non-Technical Summary 1
1. Introduction 4
1. Introduction 4
2. Data Description 6
Sample Size 11
3. The Methods 12
3.1. Bayesian Model Averaging (BMA) 12
3.2. Recursive trees 16
4. Initial Explorations 19
4.1 Bayesian Model Averaging (BMA) 21
4.2 Recursive Trees 26
5. Branching Out: Combining Trees and BMA and Beyond 36
5.1 Simple Cross-Pollination 37
5.2 Rooting Around in the Data 38
5.3 A brief look at the tree of life 48
6. Out of sample predictive performance 49
7. Conclusion 54
References 57

CASEpaper 95
March 2005

Centre for Analysis of Social Exclusion
London School of Economics
Houghton Street
London WC2A 2AE
CASE enquiries – tel: 020 7955 6679

Centre for Analysis of Social Exclusion

The ESRC Research Centre for Analysis of Social Exclusion (CASE) was established in October 1997 with funding from the Economic and Social Research Council. It is located within the Suntory and Toyota International Centres for Economics and Related Disciplines (STICERD) at the London School of Economics and Political Science, and benefits from support from STICERD. It is directed by Howard Glennerster, John Hills, Kathleen Kiernan, Julian Le Grand, Anne Power and Carol Propper.

Our Discussion Paper series is available free of charge. We also produce summaries of our research in CASEbriefs, and reports from various conferences and activities in CASEreports. To subscribe to the CASEpaper series, or for further information on the work of the Centre and our seminar series, please contact the Centre Administrator, Jane Dickson, on:

Telephone:	UK+20 7955 6679
Fax:	UK+20 7955 6951
Email:	j.dickson@lse.ac.uk
Web site:	http://sticerd.lse.ac.uk/case

Editorial Note

John Hobcraft is Professor of Social Policy and Demography in the Department of Social Policy and Social Work at the University of York and Wendy Sigle-Rushton is Lecturer in Social Policy in the Department of Social Policy at the London School of Economics and are both Research Associates of the ESRC Research Centre for Analysis of Social Exclusion at the London School of Economics.

Acknowledgements

Our research was supported by the ESRC through CASE. We are grateful to the Data Archive at the University of Essex and to the Centre for Longitudinal Studies at the Institute of Education for making the data for the National Child Development Study and the British Cohort Study 1970 available. The authors would like to thank participants of the ESRC-funded seminar "Approaches to the Analysis of Quantitative Life Course Studies" for helpful comments and suggestions. All remaining errors are our own.

Abstract

We use information from two prospective British birth cohort studies to explore the antecedents of adult malaise, an indicator of incipient depression. These studies include a wealth of information on childhood circumstances, behaviour, test scores and family background, measured several times during childhood. We are concerned both with incorporating model uncertainty and using a person-centred approach. We explore associations in both cohorts using two separate approaches: Bayesian model averaging (BMA) and recursive trees. The first approach permits us to assess model uncertainty, necessary because many childhood antecedents are highly correlated. BMA also aims to produce more robust results for extrapolation to other data sets through averaging over the range of plausible models. The second approach is concerned with partitioning the sample, through a series of binary splits, into groups of people who are as alike as possible. One advantage is that the approach is person-centred in that it retains real groups of respondents. We compare and contrast the insights obtained from the two approaches and use the results from each to inform the other and thus refine our understanding further. Moreover, we explore the claimed added robustness for extrapolation by using a split-sample for the 1970 cohort. The consistency of results across methods and cohorts is discussed throughout.

Keywords: well-being, cohort, Bayesian Model Averaging, recursive trees
JEL classification: I10, C11, C14

Non-Technical Summary

Using data from 1958 The National Child Development Study (NCDS) and the 1970 British Cohort Study (BCS70), we explore the antecedents of adult malaise, an indicator of incipient depression. Our outcome variable, the Malaise Inventory, was designed by Rutter *et al* (1970) and is a 24-item battery of questions designed to identify those individuals at high risk of depression. We restrict attention to female cohort members, and the samples that we analyse comprise all females for whom a malaise indication was available at the relevant adult age. These datasets also include a wealth of information on childhood circumstances, behaviour, test scores and family background, measured several times during childhood. We have constructed a series of childhood measures that, in most cases, summarise information collected in a similar form at each of the three main childhood waves. These include *childhood poverty*; *social class of origin*; *social class of father* (or father figure); *housing tenure*; *family disruption*; *parents' school leaving*; behavioural measures based on several scale items, taken to represent *aggression, anxiety, and restlessness or hyperactivity*, teacher's reports of *mother's and of father's interest in the cohort member's schooling*; *frequent absences from school* (only for NCDS); *contact with the police by age 16* (only for NCDS); and *educational test scores*.

We explore associations in both cohorts using two separate approaches: Bayesian model averaging (BMA) and recursive trees. The first approach permits us to assess model uncertainty, necessary because we have a large number of possible explanatory variables and many of the childhood antecedents are highly correlated. In these circumstances, stepwise methods that choose one "best" model may be misleading. There may in fact be several different models, all of which fit the data similarly well, and all of which are, more or less, defensible. The stronger the evidence for other models, the less certain a researcher can be that any chosen model is the best. This kind of model uncertainty can be problematic, for instance, if different models have different implications for answering research or policy questions. Rather than choose one model, Bayesian model averaging (BMA) identifies a set of plausible models and ranks them from most to least likely.

The second approach, Classification and Regression Trees (CART), is person-centred and involves partitioning the sample, through a series of binary splits, into groups of people who are as alike as possible based on the outcome of interest. Parametric regression techniques, while traditionally the statistical method of choice in the social sciences, often impose strong linearity assumptions and meet with data-based constraints on the extent and order of

interaction terms; when the assumptions underlying parametric regression methods are not met, the resulting model is unlikely to describe the data well. Non-parametric person-centred techniques allow researchers to relax or eliminate many of the restrictive assumptions underlying parametric modelling. The end result can be displayed graphically in tree form and real groups of people in different risk groups can be described.

The BMA results for the NCDS indicate that, using a strict model selection criterion, only five models are retained. Thus despite our a priori concerns, there is not a great deal of model uncertainty. Five childhood antecedents are included in all five models: any childhood poverty, any high anxiety score, any low father's interest in schooling, any frequent school absences, and any lowest quartile test score. In addition, there is positive evidence that female cohort members for whom there were no educational test scores available at any childhood wave, those with some contact with the police, and, more weakly, those with fewer than two test scores in the highest quartile are all more likely to experience malaise at age 33. These eight measures of childhood background correspond to those identified when a stepwise logit model was estimated. The model uncertainty arises from nested subsets of the most likely model involving inclusion of one or two of the three additional but 'uncertain' antecedents.

In contrast, results for the BCS70 sample suggest more model uncertainty, with 10 models being retained. This greater uncertainty may have arisen because, for methodological reasons, we worked with a randomly chosen half of the sample (the 'estimation sample') or from the lack of information on two of the more powerful childhood antecedents that were included in all models for the NCDS. Two factors are in all 10 models: experiencing childhood poverty at any childhood wave and having an educational test score in the lowest quartile in at least one of the three childhood waves. Experience of childhood anxiety is also retained frequently but the higher risk group varies across models (sometimes only those with the most evidence are retained as different from the reference group, other times those with moderate to high evidence are retained). As with the NCDS, there is some, though weaker, indication that having all test scores missing is associated with adult malaise. The only other childhood measure included among the ten possible models for BCS70 reflects family structure whilst growing up. The results for BCS70 are thus a little less consistent and tidy than those for NCDS, and the original stepwise model does not correspond to any of the 10 retained BMA models.

Using the CART method, results for both the NCDS data and the BCS70 show a first split by academic test scores, after which there are substantial differences. Patterns in the NCDS tree suggest that both truancy and test scores could, perhaps, be usefully combined into a summary variable of school performance.

Patterns for the BCS70 sample suggest an interaction between family structure and having experienced *either* low academic test scores *or* having some evidence of childhood poverty. Family structure does not further increase the risk of having a high malaise score for those women who suffered both kinds of disadvantages, but among those who suffered only one, experience of foster care or some forms of single parent family, increases the risk of a high malaise score to levels similar to the doubly disadvantage group.

We then explore the possible gains from bringing together these two analytic approaches and show that the BMA predicted probabilities provide a very good basis for the splits in recursive trees and that the groups from the recursive tree analysis can help to inform regression modelling. We then explore possible interactions or combinations of the childhood predictors in a number of structured, but informal, ways to improve our understanding of the childhood antecedents of female adult malaise. We thus discover important interplays between childhood anxiety and test scores in predicting adult malaise. Furthermore, we used the other random half of the BCS70 sample (the 'validation sample') to explore whether the BMA approach did indeed provide the claimed superiority for 'out of sample' prediction.

Both BMA and CART methods offer new and interesting opportunities for exploring the complex interplays among different elements of childhood disadvantage that affect adult outcomes, such as malaise. We share and have been consistently inspired by Burt Singer's deep concerns about the general linear model as the only tool for data exploration, especially when no serious attempt is made to get beyond main effects. Moreover, Adrian Raftery's plea for better analysis and fitting when using the general linear model also resonate. BMA methods can help identify robust models and predictors but does so with a strong penalty for the inclusion of additional explanatory variables. CART methods can identify ways of interacting and combining the data so that when both are used, perhaps iteratively, more defensible and reliable models can be identified.

1. Introduction

This paper is primarily concerned with applying and combining two fairly distinct methodologies in the context of an exploration of the childhood antecedents of female adult malaise. The problem we face is one shared by those undertaking exploratory analyses of the antecedents of much sociological (or other human) behaviour. One of the difficulties that most analysts face is how to deal with quite large numbers of covariates, in order to 'sort the wheat from the chaff'. As in much social research, many of our covariates are often by nature categorical and, moreover, the problems of dealing with missing information make the use of categorical covariates more desirable.

Several years ago Raftery (1995) provided a thorough account of issues to do with Bayesian model selection in social research. Many social researchers have responded to that *Sociological Methodology* article by adopting the BIC statistic for model testing. But fewer have been enticed by the Bayesian model averaging (BMA) approach, even though that was in many ways the core element of Raftery's approach. The claimed key advantages of BMA were: dealing with model uncertainty; and robustness advantages for out-of-sample prediction (see also Hoeting *et al* 1999). We are convinced that dealing with model uncertainty is desirable and informative and will illustrate some of the benefits through our application to the childhood antecedents of adult female malaise. We also explore the issue of the potential gains for out of sample prediction, using a split-sample approach for one of the two birth cohorts considered here.

The second approach that we explore here is the use of person-centred methods. Readers of *Sociological Methodology* were introduced to an extended example by Singer *et al* (1998). Perhaps that work did not have the influence that the approach merits for two main reasons: the procedures used to group individuals according to their characteristics were very labour intensive; and, in part because the groups were produced without any attention being paid to the outcome being considered, the laboriously obtained groups of persons did not turn out to provide much discrimination or insights about the mental health outcome concerned. However, one of the authors of that study has gone on to extend methods for a more automated approach to person-centred analysis, which is known as recursive partitioning and involves dividing the sample into a series of recursive binary splits, to maximise homogeneity within nodes according to a 'goodness-of-split' criterion, producing a 'tree' that is then 'pruned' to give a more parsimonious grouping (see Zhang and Singer 1999). This approach is related to the 'Automatic Interaction Detection' procedures developed many years ago. Person-centred methods divide into two broad approaches, those which begin with each individual and then group together

those who are most alike according to some criterion, known as cluster analysis, and those that begin with the entire population and successively divide it into groups, often known as Classification and Regression Trees (Breiman *et al* 1984).

One attraction of person-centred approaches to many is that the final groupings correspond to 'real' groups of people. This certainly makes results more accessible to policy makers. Advocates of person-centred approaches often stress this (and other related advantages) as being invaluable through avoiding the complexities of multiple regression models, that attribute probabilistic estimates to individuals, rather than exploring real groups. We regard this distinction as debatable, but nevertheless feel that quantitative approaches to person-centred methods are not explored often enough in social research. We thus explore the insights to be gained from applying the Zhang and Singer (1999) approach to our two birth cohorts.

In exploring these two approaches, it was always our intention to see whether they could cross-inform each other, in the sense that the groups identified through recursive partitioning might either serve to identify interaction terms that had been missed by the regression models or to define a more informative covariate for the BMA models. Conversely, it is also reasonable to ask whether the predicted probabilities derived from the BMA regression modelling do well enough as predictors to dominate the basis for splitting into groups in the recursive trees analysis. A further issue is to examine how well the two approaches do in terms of their ability to 'account for' variation in female adult malaise.

Lastly, we explore how well the two approaches used here do in terms of predictive performance. For the younger birth cohort considered, we randomly split the sample into two halves, an estimation sample and a validation sample, and then explore the out of sample predictive performance of the various models considered.

There is particular value in being able to repeat our analyses for two nationally representative prospective studies of birth cohorts, a group of children born in 1958 (the National Child Development Study, NCDS) and a group born in 1970 (the British Cohort Study 1970, BCS70). Finding, as we do, that the most informative childhood antecedents of adult female malaise are fairly similar across the two cohorts provides some reassurance that our exploratory approach gets close to the underlying mechanisms and that there is continuity across cohorts in the antecedents of female adult malaise that matter.

2. Data Description

The data used are drawn from the National Child Development Study (NCDS) and the British Cohort Study (BCS70), longitudinal studies that have both attempted to follow the lives of around 16,000 people who were born during one week in March 1958 and one week in April 1970 respectively. The NCDS has collected a wide range of information around birth, at ages 7, 11 and 16 during childhood and at ages 23, 33 and 42 in adulthood. BCS70 conducted interviews shortly after birth, at ages 5, 10 and 16 during childhood and at ages 26 (a limited postal enquiry) and 30 in adulthood. In order to make the analyses comparable, we examine malaise as measured at age 33 in NCDS and at age 30 in BCS70. Moreover, we restrict attention to female cohort members since their malaise incidence is higher than for males and the analyses here are intended to illustrate the methods used. The samples that we analyse comprise all females for whom a malaise indication was available at the relevant adult age. In this analysis, the BCS70 sample is randomly split into two samples – the learning and the validation. All models are estimated using the learning sample so that their out of sample performance can be assessed using the validation sample.

Our outcome variable, the Malaise Inventory, was designed by Rutter *et al* (1970) and is a 24-item battery of questions designed to identify those individuals at high risk of depression. The items cover a range of symptoms associated with depression, and, similar to previous work, we classify those individuals answering yes to at least seven of the 24 items as being at high risk of depression (Richman, 1978; Rutter *et al* 1976).

The explanatory variables are, in most cases, constructed as summaries of information collected at more than one childhood wave. However, many cohort members were not interviewed at all ages, and even among those who were interviewed, there is often a good deal of missing information and non-response. Because attrition and non-response appear to be non-random, restricting our sample to those cohort members with complete information could result in serious sample selection issues. We explicitly code missing values for each explanatory variable (for details on NCDS see Hobcraft 1998 and on BCS70 see Sigle-Rushton 2004). This maximizes our sample and allows us to assess whether missing information is likely to be informative. Because most of our explanatory variables summarise information collected at various points in time and because we wanted to exploit as much real information as possible, for each summary variable, only those individuals with no information on the characteristic in all of the childhood waves were classified as missing. Those with at least some information were coded into categories that were constructed with some allowance for missing information.

We have constructed a series of childhood measures that, in most cases, summarise information collected in a similar form at each of the three main childhood waves. This enables capture of any information on disadvantage (even when missing at one or two of the waves) and also permits some measure of repeated incidence of disadvantage, such as depth of childhood poverty. These summary measures of childhood experience are documented and more fully described by Hobcraft (1998 & 2000) for NCDS and by Sigle-Rushton (2004) for BCS70. We have tried to make the handling of these factors comparable over the two cohorts, though this is an ongoing effort, but the same information is not always available in both. Typically, we took information from the three childhood waves, with the item at each age classified into advantaged, intermediate, disadvantaged, and missing categories and summarised information across the combined ages into four categories plus an additional one with all information missing. The usual categories were: a group with clear evidence of disadvantage at two or three waves; a group with one piece of evidence of disadvantage, but possibly some information missing at one or two waves; those with no evidence of disadvantage, but with fewer than two indications of advantage; those with two or three indications of advantage; and all information missing. There are some minor variations in detail, as will be apparent from Table 1.

The summary measures that have been used in this work are shown for the samples used in this analysis in Table 1 and take comparable inputs at each of the three childhood ages unless otherwise indicated, covering:

- *childhood poverty (indicators)*, as measured by: 'experience of financial difficulties' at ages 7, 11, & 16 and by 'receipt of free school meals' at ages 11 & 16 for NCDS; and by receipt of free school meals at age 10, receipt of income support or unemployment benefits at age 10 and self-assessed financial hardship at age 16;
- *childhood poverty (income)* for BCS70 only, using banded family income at ages 10 and 16;
- *social class of origin*, concentrating on three broad groupings of non-manual, skilled manual, and semi- and unskilled manual for the father at the birth of the survey member and the two paternal grandfathers;
- *social class of father* (or father figure), similarly grouped, but for the three childhood ages;
- *housing tenure*, distinguishing renting from local authority (or public housing), owner-occupier, and other;
- *parents' school leaving age* for NCDS only – a combination of whether mother & father left school at the minimum age (measured once for each parent);
- *family disruption*, including having been born outside marriage, experience of care, loss of a parent through death, experience of parental

divorce, and remarriage, with minor variations in the classification, derived from the interview around birth in addition to the three childhood waves;

- behavioural measures based on several scale items, taken to represent
  - *aggression*,
  - *anxiety*, and
  - *restlessness* or *hyperactivity*;

  for BCS70 we have only used ages 5 and 16;
- teacher's reports of *mother's and of father's interest in the cohort member's schooling*, distinguishing very interested and low interest from intermediate groups; for BCS70 this information was only available at age 10;
- *frequent absences from school* (only for NCDS);
- three reports of *contact with the police by age 16*, two from teachers and one from parents (only for NCDS);
- and *educational test scores*, distinguishing lower and upper quartile scores from intermediate ones.

When we apply Bayesian Model Averaging techniques, we treat all of our control variables, other than family type and poverty, as categorical with an explicitly defined missing category. We treat the most advantaged category as the reference category. With the exception of the missing category, all the categorical variables other than family experience have been entered into our models as a series of hierarchically defined dummy variables. Taking the most advantaged category as a reference and creating a dummy for the missing value category, we then created a series of dummy variables. The first was set equal to one for all categories other than the reference and the missing category. This variable identifies those cohort members for whom there is any possibility of disadvantage as we define it for that variable. The next dummy is set equal to one for those members with at least some evidence of disadvantage. A final dummy variable is set equal to one only for those with the clearest evidence of disadvantage.[1]

---

1 The dummy variables for experience of poverty are similarly defined but take into account the larger number of categories for that variable. In other words, there are four, as opposed to three, hierarchal dummies defined for each of the poverty focal variables. Moreover, in NCDS, there is only one dummy for all missing childhood behaviour, rather than three separate all missing dummies for aggression, anxiety and hyperactivity for BCS70.

Table 1: Sample distribution by categories of childhood antecedents for females with adult malaise scores at age 30 in BCS70 and at age 33 in NCDS (per cent)

Childhood Antecedent Categories	BCS70[a]		NCDS Category	
	learning	validation	NCDS	(Where different)
Poverty Indicator				
Not Poor	35	36	40	
Probably Not Poor	30	29	35	
Some Poverty	10	11	11	
Fairly Poor	8	8	8	
Clearly Poor	13	13	3	
All Missing	4	4	3	
Poverty Income				
Not Poor	42	42	N/A	
Probably Not Poor	36	36	N/A	
Some Poverty	8	7	N/A	
Fairly Poor	3	4	N/A	
Clearly Poor	2	1	N/A	
All Missing	9	10	N/A	
Family Type				
Natural throughout	41	43	50	
Natural , partial info	37	38	29	
Ever in care/fostered	4	4	2	
Dissolution, no remarriage	8	6	4	Divorced, no remarriage
Dissolution, remarried	8	7	2	Divorced, remarried
Lone parent at birth, no marriage	1	1	4	Other dissolution, no remarr.
Lone parent at birth, later married	2	1	2	Other dissolution, remarried
All missing	1	1	3	Born out-of-wedlock
			4	All missing
Social Class of Origin				
2-3 nonmanual	16	17	15	
0 IV or V, 0/1 nonmanual	40	37	35	
one IV or V	27	27	32	
2-3 IV or V	10	12	15	
All Missing	7	7	3	
Social Class of Father Figure				
2-3 nonmanual	27	29	25	
0 IV or V, 0/1 nonmanual	43	43	40	
one IV or V	13	14	15	
2-3 IV or V	8	7	14	
All Missing	8	7	5	

Table 1 (continued)

Childhood Antecedent Categories	BCS70[a]		NCDS Category	
	learning	validation	NCDS	(Where different)
Housing Tenure				
2-3 Owner Occ.	8	8	38	
0 Council, 0/1 Owner Occ.	55	56	15	
1 Council	12	12	11	
2-3 Council	23	22	34	
All Missing	3	3	2	
Father's Interest in Education at Age 10				***Father's Interest in Education***
Very Interested	28	28	21	2-3 High Interest
Some Interest	14	15	41	0 Low, 0/1 high
Little Interest	3	3	20	1 Low
No Interest	2	2	7	2-3 Low Interest
Missing	54	52	11	All Missing
Mother's Interest in Education at Age 10				***Mother's Interest in Education***
Very Interested	41	42	32	2-3 High Interest
Some Interest	23	24	41	0 Low, 0/1 high
Little Interest	3	3	18	1 Low
No Interest	1	1	7	2-3 Low Interest
Missing	32	30	3	All Missing
Aggression Scores				
2 Low	22	24	46	2-3 Low
0 High, 0/1 Low	59	58	36	0 High, 0/1 Low
1 High, 1 Low	5	6	12	1 High
2 High or 1 High, 1 Missing	4	4	4	2-3 High
All Missing	10	9	2	All Missing
Anxiety Scores				
2 Low	19	19	27	2-3 Low
0 High, 0/1 Low	60	60	41	0 High, 0/1 Low
1 High, 1 Low	7	8	23	1 High
2 High or 1 High, 1 Missing	5	4	7	2-3 High
All Missing	10	9	2	All Missing
Hyperactivity Scores				
2 Low	20	22	45	2-3 Low
0 High, 0/1 Low	63	62	35	0 High, 0/1 Low
1 High, 1 Low	5	5	13	1 High
2 High or 1 High, 1 Missing	3	3	5	2-3 High
All Missing	10	9	2	All Missing

Table 1 (continued)

Childhood Antecedent Categories	BCS70[a]		NCDS Category
	learning	validation	NCDS (Where different)
Academic Test Scores			
2/3 High Quartile	13	12	18
0 Low, 0/1 High Quartile	49	50	49
1 Low quartile	23	22	16
2/3 Low quartile	9	10	16
All Missing	6	6	1
Contact with Police			
All 3 No	N/A	N/A	30
1-2 No, 0 Yes	N/A	N/A	50
1-2 No, 1 Yes	N/A	N/A	2
Yes>No	N/A	N/A	2
All Missing	N/A	N/A	16
Frequent School Absences			
All 3 No	N/A	N/A	46
1-2 No, 0 Yes	N/A	N/A	33
1+ Yes, 1-2 No	N/A	N/A	17
1+ Yes, 0 No	N/A	N/A	3
All Missing	N/A	N/A	1
Parental School Leaving Ages			
2 Beyond Minimum	N/A	N/A	10
1 Beyond Minimum, Other Missing	N/A	N/A	25
Either or Both at Minimum	N/A	N/A	62
All Missing	N/A	N/A	3
Sample Size	***2823***	***2864***	***5768***
Proportion with High Malaise Score	20	20	12

[a] The BCS70 sample is randomly split into two samples of roughly equal size. This allows us to test the performance of models achieved with the learning sample data using the validation sample data.

The family experience variable is constructed in a similar, but less straightforward, way. The reference category comprises, as mentioned above, those who were living with both natural parents at all three childhood interviews. Those for whom there is no evidence of family disruption, but some doubts because of missing information are identified by a dummy, as are those for whom all information on family structure is missing. An additional dummy picks out those children who have ever been in care. The remaining categories differ slightly for the two cohorts. For NCDS, a further dummy identifies all children born out of wedlock. Four further dummies capture the remaining information: 'disruption' (any divorce or lone parenthood, regardless of

subsequent marital status), 'divorce' (regardless of subsequent marital status), 'remarriage' (regardless of type of disruption), and 'divorced & remarried' to complete the coverage. For the remaining categories in BCS70, we first create a dummy variable that equals one if the cohort member was born to a lone mother or ever experienced a family dissolution. The next variables pick out those among that group who were born to a lone mother, experienced a step-family situation (remarriage following dissolution or marriage to a lone mother), and finally, those who experienced both a dissolution and a step-family arrangement. As constructed, the categorical variables are transformed into 58 hierarchical variables for the NCDS and 53 for the BCS70.

3. The Methods

3.1. Bayesian Model Averaging (BMA)

When faced with a large number of explanatory variables, social scientists often employ model selection techniques like stepwise regression in order to choose one "best" model that is both parsimonious and fits the data well. However, in many instances, there will be several different models, all of which fit the data similarly well, and all of which are, more or less, defensible. The stronger the evidence for other models, the less certain a researcher can be that any chosen model is the best. This kind of model uncertainty can be problematic, for instance, if different models have different implications for answering research or policy questions. Rather than choose just one particular model, Bayesian model averaging (BMA) is an estimation method that explicitly takes model uncertainty into account and identifies a set of plausible models. Researchers can then use the set of defensible models to address the research question at hand and, in many cases, improve out of sample predictive performance (Hoeting *et al*, 1999).

To illustrate Bayesian model averaging methods, we will rely on a specific example that is pertinent to our own application.[2] Suppose our research question is focused on a specific regression parameter, β_1. If we were to choose one best model, it would either include β_1 or it would not. But even if β_1 is not included in the best model, it may, nonetheless, be included in other models that fit the data almost as well. Not taking those other models into account may mean losing some important information about this parameter. Now assume that, instead of choosing just one model (either by choice of variables or selection

---

2 The material in this section is drawn from Raftery (1995) and Hoeting *et al* (1999). We refer interested readers to these publications for a more generalised and rigorous introduction to the method.

techniques) we consider the set of K possible models, which, in most cases, will be rather large. In doing so, we can calculate $\Pr[\beta_1 \neq 0 \mid D]$, the posterior probability that β_1 is in the model, which is

$$\Pr[\beta_1 \neq 0 \mid D] = \sum_{A_1} p(M_k \mid D), \tag{1}$$

where $A_1=\{M_k: k=1,\ldots,K; \beta_1 \neq 0\}$ – that is the set of models that include β_1, and $p(M_k \mid D)$ denotes the posterior model probabilities which gauge the amount of evidence for model M_k. These posterior model probabilities are obtained by Bayes' Theorem as follows:

$$p(M_k \mid D) = \frac{p(D \mid M_k)p(M_k)}{\sum_{\ell=1}^{K} p(D \mid M_\ell)p(M_\ell)}. \tag{2}$$

The more the data support model M_k, the higher the value of $p(M_k \mid D)$, and the higher the posterior odds of model M_k relative to some baseline model. Often the researcher will have no preference for one model over another, so the prior probabilities of each model will all be the same. In other words, $p(M_1) = \ldots = p(M_K) = 1/K$.

Using model averaging techniques, we can also examine the size of β_1 given that it is nonzero. The posterior distribution for β_1 is

$$p(\beta_1 \mid D, \beta_1 \neq 0) = \sum_{A_1} p(\beta_1 \mid D, M_k)p'(M_k \mid D), \tag{3}$$

Where $p'(M_k \mid D) = \frac{p(M_k \mid D)}{\Pr[\beta_1 \neq 0 \mid D]}$.

This posterior distribution can be summarised by its posterior mean and the standard deviation, which provide a point estimator and a Bayesian analogue of the standard error. Raftery (1995) suggests the following approximations

$$E[\beta_1 \mid D, \beta_1 \neq 0] \approx \sum_{A_1} \hat{\beta}_1(k)p'(M_k \mid D) \tag{4}$$

$$SD^2[\beta_1 \mid D, \beta_1 \neq 0] \approx \sum_{A_1} [se_1^2(k) + \hat{\beta}_1(k)^2]p'(M_k \mid D) - E[\beta_1 \mid D, \beta_1 \neq 0]^2$$

where $\hat{\beta}_1(k)$ and $se_1^2(k)$ are respectively, the MLE and standard error of β_1 under the model M_k.

The main obstacle to carrying out this method is the number of models that may have to be considered. With a large number of variables, K may be formidably large. Previous research suggests are two strategies that can be used to reduce the number of models over which we are going to average. First, very unlikely models can be discarded. In addition, models that have more likely sub-models nested within them can also be discarded. When one or both exclusion rules are applied, the remaining models are said to belong to *Occam's window*. When only the first exclusion rule is applied, Occam's window is said to be symmetric, and when both exclusion rules are applied it is said to be strict.

In order to apply either exclusion rule and retain the models that fall within a symmetric or strict Occam's window, we must measure how likely the different models are. When comparing models, it is usually good to establish a baseline model against which all models are compared. This is usually either a null model (with no explanatory variables), M_0, or a saturated model, M_S, in which each data point is fitted exactly. Let's assume our baseline model is M_0 and that we want to compare M_k to it. Evidence for whether or not the data support M_k over M_0 is measured by the posterior odds for M_k against M_0 – that is the ratio of their posterior probabilities. By equation (2), this is

$$\frac{p(M_k \mid D)}{p(M_0 \mid D)} = \left[\frac{p(D \mid M_k)}{p(D \mid M_0)}\right]\left[\frac{p(M_k)}{p(M_0)}\right]. \qquad (5)$$

The first factor on the right hand side of equation (5) is called the Bayes factor for M_k over M_0, denoted by B_{k0}. The second factor on the right hand side of (5) is the prior odds – which, as mentioned above, is often set equal to one when the researcher has no prior preference for either model. Hence, the Bayes factor is equal to the posterior odds when there is no a priori model preference.

When the Bayes factor, B_{k0}, exceeds one, the data support M_k over M_0. Similarly, when B_{k0} falls below one, the data provide more evidence for M_k than for M_0. Raftery (1995) suggests an heuristic interpretation of the Bayes factor – for $B_{k0} \in [1,3]$ there is positive but decidedly weak evidence for M_k over M_0; when $B_{k0} \in (3,20]$ the evidence is positive; when $B_{k0} \in (20,150]$ the evidence is strong; and finally, when $B_{k0} > 150$ the evidence for M_k over M_0 is very strong.

In practice, calculating the Bayes factor often involves high-dimensional and complex integration. The Bayesian Information Criterion (BIC) can provide a good approximation to the Bayes factor, and it is more simple to calculate. When our baseline model is the null model (or any other model nested in M_k) the approximation, denoted BIC'_k, is

$$2\log B_{k0} \approx -\chi^2_{k0} + df_{k0}\log n \qquad (6)$$

Where χ^2_{k0} is the standard likelihood ratio test (LRT) statistic for testing M_0 against M_k, df_{k0} is the number of degrees of freedom associated with the test, and n is, most often, the sample size.

When the baseline model is M_S (or any other model in which M_k is nested), we have $2\log B_{k0} = L^2_k - df_k \log n$, which is denoted as BIC_k. Here, $L^2_k = \chi^2_{Sk}$, the deviance of model M_k and df_k is the corresponding degrees of freedom. In both the BIC'_k and the BIC_k the latter term takes into account both the sample size and the degrees of freedom. Hence, there is a penalty for model complexity. Moreover, the latter term requires more evidence for the inclusion of an additional parameter in large samples. Regardless of the baseline model used, to compare any two models, M_j and M_k, $2\log B_{jk} \approx BIC_k - BIC_j$ and $BIC_k - BIC_j \equiv BIC'_k - BIC'_j$. In other words, any two models (nested or not) can be compared by taking the difference of their BIC (or BIC') values.[3] In either case, the model with the smaller value (more negative) is the preferred model.

To identify those models that fall within a symmetric Occam's window, Madigan and Raftery (1994) suggest that the most likely model be identified and all others with a BIC (or BIC') difference of at least six (very strong evidence against the alternative model) be discarded. This corresponds roughly to odds of at least 20:1 in favour of the best model. To identify those models that fall within a strict Occam's window, all of those models that have more likely sub-models nested within them are removed based on a comparison of the BIC values (which, as mentioned above, penalise model complexity). Typically, the exclusion of models falling outside of Occam's window reduces substantially the set of models taken into consideration.

In our BMA application, we use the bic.logit function in S-Plus. The software uses a leaps and bounds algorithm in order to identify those models that fall within a strict version of Occam's window, and is available on the internet.[4] Because BMA is particularly useful in situations where the sample size is large and the set of potential predictors is large (making the identification of a parsimonious model based on p-values difficult because few variables will be eliminated), it is unfortunate that this program cannot deal effectively with more

---

3 Although the BIC or BIC' must be computed using nested models, two BIC statistics can be compared for any models, nested, or non-nested.

4 This program and other BMA software, all written in S-Plus©, can be downloaded at www.research.att.com/~volinsky/bma.html.

than 30 explanatory variables. When the set of explanatory variables exceeds 30, the program uses stepwise procedures to reduce the set of explanatory variables to 30. We have generally tried to explore reentering those explanatory variables that were backed out by the stepwise procedure, removing those not backed out that did not feature at all in the BMA models and iterating as necessary.

3.2. Recursive trees

Parametric regression techniques, while traditionally the statistical method of choice in the social sciences, often impose strong linearity assumptions and meet with data-based constraints on the extent and order of interaction terms. When the assumptions underlying parametric regression methods are not met, the model is unlikely to describe the data well. In addition, when many and high-order interactions are included in an attempt to allow for non-linear relationships, the model is often difficult to interpret. Non-parametric techniques allow researchers to relax or eliminate many of the restrictive assumptions underlying parametric modelling. Recursive partitioning is one such technique that can aid in the identification of non-linear relationships and in the choice of parsimonious models. The recursive partitioning technique that we use in this application is that of Classification and Regression Trees (CART). This method involves repeatedly partitioning the sample into more homogenous groups based on an outcome variable of interest. The end result can be displayed graphically in tree form and "interpreted as a string of Boolean statements, facilitating conversion of complex output to narrative form." (Zhang and Singer, 1999, p.2). Our summary of the method draws largely from work by Zhang and colleagues and interested readers can refer to these publications for a more comprehensive treatment of both CART and other tree methods (Zhang and Singer, 1999; Zhang and Bracken, 1995).

To illustrate the method we employ in this application, assume we have an outcome variable Y, and a set of p explanatory variables, $x_1,x_2,\ldots,x_p$, where Y is a random variable and the x's are fixed variables. In our example, Y is a dichotomous variable and the x's are ordinal variables with some missing information. We are interested in identifying the relationship between Y and the x's so that we can predict Y based on the value of the x's. Essentially, we want to estimate the probability of the random variable Y conditional on the x's:

$$\mathrm{P}\{Y = y \mid x_1, x_2, \ldots, x_p\} \tag{7}$$

The analysis begins with the unique root node of the tree which is represented by a circle at the top of the tree diagram. This root node contains all observations in the sample from which the tree is derived. At each subsequent layer of the tree, a node can be internal, meaning that additional nodes lie below

it, or, the node can be terminal, meaning no additional nodes lie below it. Internal nodes are represented with circles and terminal nodes with boxes.

Both the root and the internal nodes are partitioned, using the same procedure, into two nodes in the next layer of the tree. These are called left and right daughter nodes, each of which is a sub-set of the internal node above it and, in the diagram, connected to it with straight lines. The partition of the sample becomes progressively more detailed as the layers get deeper, but each subject is eventually assigned to one of the terminal nodes – ideally a node in which all subjects are homogenous with respect to the outcome variable. In practice, the complete homogeneity of terminal nodes is rarely achieved, however.

The partitioning technique is applied layer by layer to each internal node, including the first root node. The goal, for each partition, is to locate a binary split that results in two homogeneous, or pure, daughter nodes. Because no split is likely to achieve total purity, we base our choice on a goodness of a split measure that weighs the impurities of the resulting daughter nodes. Impurity can be measured using any concave function, ϕ that satisfies these three conditions

(i) $\phi \geq 0$;
(ii) for any $p \in (0,1)$, $\phi(p) = \phi(1-p)$, and
(iii) $\phi(0) = \phi(1) < \phi(p)$.

For our estimation we use the entropy impurity function which for node τ_L, is

$$i(\tau_L) = p\ln(p) - (1-p)\ln(1-p). \quad (8)$$

In order to split a given parent node τ into left and right daughter nodes τ_L and τ_R, respectively, the goodness of the split, *s*, is given by:

$$\Delta I(s,\tau) = i(\tau) - P\{\tau_L\}i\{\tau_L\} - P\{\tau_R\}i\{\tau_R\} \quad (9)$$

where $P\{\tau_L\}$ and $P\{\tau_R\}$ are the probabilities that a subject falls within nodes τ_L and τ_R, respectively. Using this measure, the best split is chosen for each predictor variable. If a given ordinal variable, with no missing information, takes on *j* different values among the subjects in a given node, there are *j-1* allowable splits for that variable. For a nominal variable divided into *j* categories, there are $2^j - 1$ allowable splits. In our application, which has a good deal of missing values, we assign subjects with missing information using the "missings together" approach. Essentially, the missings together approach treats those cases with missing information as having either the lowest or the highest value of the variable for which they have missing information. In this way all of the missing cases are assigned to either τ_L or τ_R – either as a category by themselves or with the lower or higher end of each split. For ordinal variables

with *j* values, one of which is missing, there are *2j-3* possible splits. The total number of allowable splits for any given node is, then, the sum of the allowable splits across all the variables. The goodness of fit measures for the best split for each variable are then compared and the one with the largest s is chosen to partition node τ.

To begin building the tree, we first split the root node using the splitting procedure presented above. We continue by attempting to partition its two daughter nodes using the same method, and then their four daughter nodes, on through each layer of the tree. While a given daughter node cannot, by construction, be split in the same way as its parent node, an offspring node may nonetheless split using the same variable. For instance, we might choose to split node τ using ordinal variable x_j so that τ_L contains all those subjects for whom x_j equals –3, -2 or –1 and τ_R contains all those subjects for whom x_j equals 0, 1 or missing. When splitting τ_L, it would not be possible to split x_j at zero because all of those in τ_L take on values below zero. We might find that the best split for τ_L is another, negative, value of x_j, such as –2, however. Notice that in the parent node x_j took on six possible values, so there were nine possible splits, but in splitting τ_L, x_j only took on three possible values, so, while it is still possible to split on x_j, the number of possible splits falls to just three. As we move further down the tree, the number of possible splits for each node, therefore, falls.

We continue splitting each layer of nodes until no offspring nodes can be split any further. When this happens, the tree is said to be saturated. Because the total number of possible splits for a node falls as we more from one layer to the next, the number of permissible splits gradually approaches zero, at which point the tree cannot grow any further. Saturated trees of are limited use, statistically, because the terminal nodes are usually very small, and the trees are often so large that interpreting them can be intractable. In our application, we make the trees more workable by applying a second step called “pruning”. First, we grow the tree until it is saturated. Beginning with the bottom nodes of this tree we begin to prune upwards.

To prune to the tree, we first calculate a statistic for each internal node from the bottom up. This statistic is the Studentized log relative risk of the split to that node, which, for a dichotomous outcome, is the log relative risk that y=1 in the two daughter nodes divided by its standard deviation. In practice, this statistic is biased upward to some extent because the relative risk is calculated as a resubstitution estimate – ie using a similar measure to the one from which the tree was grown. Next, for each internal node, we compare its statistic with those calculated for all of its offspring nodes. If any offspring node has a higher statistic than its ancestor, we replace the statistic of the ancestor with that higher

value. Finally, we examine the size of the statistics for each node. Any node whose statistic falls below some threshold (say 1.96 if we are interested in pruning to a 0.05 significance level) is pruned.

4. Initial Explorations

In this section we explore the childhood antecedents of adult female malaise separately for each of the two approaches considered. In order to introduce the material, we begin with the results of a straightforward stepwise logistic regression, with backwards elimination and possible reentry. This is an approach that is quite often used with large numbers of possible covariates in order to reduce models to a manageable size. For example, we have used stepwise regression procedures extensively in exploring the antecedents of adult social exclusion or multiple disadvantages for these birth cohorts (see Hobcraft 1998, 2000, 2002, 2003, and 2004; Hobcraft and Kiernan 2001; and Sigle-Rushton 2004). Although the limitations of stepwise regression for model selection are well known, we have found reassuring agreement for models derived from both backwards elimination and forwards inclusion.

A key issue with stepwise regression (and other standard regression) procedures is the choice of the p-value for the threshold. The multiple testing involved undoubtedly means that a standard significance level of five per cent or possibly even one per cent is too generous. Both for this reason and in the light of the guidance provided by Raftery (1995, Table 9, p.141) and so strongly endorsed by Hauser (1995, pp.180-181), we have been routinely using a cut-off p-value of either one in a thousand or one in ten thousand in such models as being appropriate with samples of around five thousand.

A further potential problem with standard stepwise regression procedures arises when two variables are close to collinear. Once one enters the model, the other is unlikely to do so.

Table 2 presents results from fitting a backwards elimination logistic model for both birth cohorts, in summary form showing only the significant dummies for the childhood antecedents, eight of 58 for the NCDS and five of 53 for the BCS70. Both models include any incidence of childhood poverty measured by indicators, having any low educational test score, and a measure of childhood anxiety (though with a different cut-point). The model for the NCDS includes two further behavioural measures that are not available for the BCS70, contact with the police and frequent school absences. In addition, the NCDS model includes further cut-points on the educational test scores and an indication of the father's (or father figure's) interest in schooling (measured at more waves in

NCDS than BCS70). The model for BCS70 includes an indication of family structure.

Table 2: Significant childhood antecedents of adult female malaise (backwards stepwise elimination, p=0.001)

Childhood antecedent	NCDS 1958	BCS70
Poverty indicators	Any poverty	Any poverty
Low income	N/A	
Social class of origin		
Social class of father		
Housing tenure		
Parental schooling		
Family structure		(Re)marry
Aggression		
Anxiety	Any high	Not low Missing
Hyperactivity		
Mother's interest		
Father's interest	Any low	
Contact with police	Any	N/A
Frequent school absences	Any	N/A
Educational test scores	Any low 2/3 Low Missing	Any low

Thus we see a fair commonality of models across the two cohorts. In both, any experience of childhood poverty as measured by the indicators is associated with a higher incidence of adult malaise. It is noteworthy that the cut-point in the distribution is the same for both cohorts and that the low income measure available in BCS70 does not appear. Educational test scores also appear in the models for both cohorts, although the NCDS shows more pervasive links of adult malaise to this antecedent. Moreover, both cohorts suggest that having any low test score matters for experience of adult malaise. Childhood poverty and low test scores are amongst the most pervasive antecedents of a wide range of adult disadvantages (see Hobcraft 2000, 2003, and 2004 for NCDS, and Sigle-Rushton 2004 for BCS70).

The other shared component in the stepwise models for the two cohorts is a measure of childhood anxiety, though with a slightly different cut-point for the two cohorts (and noting our omission of an indicator at age 10 for BCS70, which makes the two measures not fully comparable). Thus we see that adult mental health links to childhood behavioural measures. Our broader work shows that links of the three childhood behavioural indicators (anxiety, aggression, and hyperactivity) to adult disadvantages are by no means as pervasive across other adult disadvantage as for childhood poverty or educational test scores (e.g. Hobcraft, 2004).

4.1 Bayesian Model Averaging (BMA)

Table 3A provides summary information on the results for the NCDS 1958 cohort from the BMA models with the strict Occam's window. Only five models were retained; all others were either very unlikely compared with the most likely model (a BIC difference of six or higher), or had a more likely sub-model nested within them (the strict criterion removing a further 15 models). Thus there is not a great deal of model uncertainty. Five childhood antecedents were included in all five models, with posterior effects probabilities of 100 per cent: any childhood poverty, any high anxiety score, any low father's interest in schooling, any frequent school absences, and any lowest quartile test score. These factors also have to be common to all the 20 models included in the symmetric Occam's window (since the additional models can only have further antecedents added to them). In addition, there is positive evidence that female cohort members for whom there were no educational test scores available at any childhood wave, those with some contact with the police, and, more weakly, those with fewer than two test scores in the highest quartile are all more likely to experience malaise at age 33. These eight measures of childhood background correspond to those identified in the stepwise model introduced earlier. The model uncertainty arises from nested subsets of the most likely model involving inclusion of one or two of the three 'uncertain' antecedents in that model. In this situation, we would be fairly comfortable that the most likely model did a pretty good job, containing all the other 'strict' competitor models.

Table 3B provides the analogous results for a strict Occam's window BMA applied to the BCS70 learning sample data. Here there is more model uncertainty, with 10 models being retained. Two factors, experiencing childhood poverty as measured by indicators at any childhood wave and having an educational test score in the lowest quartile in at least one of the three childhood waves, are in all 10 models, providing very strong evidence that they do indeed matter as predictors of female malaise at age 30 for this cohort, as for the 1958 cohort.

Table 3A: Childhood antecedents of adult female malaise in NCDS included in BMA models with strict Occam's window and posterior probabilities

Childhood antecedent	1	2	3	4	5	Effect posterior probability (per cent)
Any poverty (indicator)	*	*	*	*	*	100
Any high anxiety	*	*	*	*	*	100
Any low father's interest	*	*	*	*	*	100
Any frequent school absences	*	*	*	*	*	100
Any low test score	*	*	*	*	*	100
Missing test scores	*	*	*		*	90.8
Any contact with police	*	*		*		78.4
Less than two high test scores	*		*			60.9
Model posterior probability (%)	46.2	23.0	14.7	9.2	6.9	100

Table 3B: Childhood antecedents of adult female malaise in BCS70 included in BMA models with strict Occam's window and posterior probabilities

Childhood antecedent	1	2	3	4	5	6	7	8	9	10	Effect posterior probability (per cent)
Any poverty (indicator)	*	*	*	*	*	*	*	*	*	*	100
Any low test score	*	*	*	*	*	*	*	*	*	*	100
Probably anxious	*	*	*			*					68.8
(Re-)married lone parent	*		*	*			*		*		63.4
Missing anxiety	*	*									55.2
Missing tests			*	*	*	*					25.9
Any high anxiety				*	*		*	*			22.8
Either anxiety (prob. or any)	*	*	*	*	*	*	*	*			91.6
Model posterior probability (%)	38.7	16.4	8.1	6.7	5.6	5.5	5.5	5.0	4.4	3.9	100

Experience of childhood anxiety appears in two guises, with posterior effect probabilities of 68.8 per cent for 'probably anxious' and of 22.8 per cent for any high anxiety. Since these two alternative indicators of anxiety never appear in the same model, the overall posterior probability of one or other measure of childhood anxiety appearing is 91.6 per cent, which constitutes fairly strong evidence. Why is there more uncertainty for BCS70 on anxiety than for NCDS? There are several possibilities: anxiety is only measured in useable form at two of the three childhood waves for BCS70, giving a less secure summary measure; there is genuine uncertainty as to where in the anxiety distribution the cut should fall and the near collinearity shows up in the model uncertainty; the smaller sample used with BCS70, because of the split into two halves affects uncertainty (though the BIC criterion should avoid some of the problems of more conventional frequentist testing); and the other important measures in NCDS (frequent school absences and contact with the police) may somehow bring the anxiety measure into sharper focus.

As with the NCDS, there is some, though weaker, indication that having all test scores missing is associated with adult malaise; unlike the 1958 cohort, however, there is nothing to suggest that a further split on the test scores matters at all for the 1970 cohort,. The only other childhood measure included among the ten possible models for BCS70 reflects family structure whilst growing up and indicates either having been born to a lone parent who subsequently married or having a parental marriage dissolution followed by remarriage and has a posterior effect probability of 63.4 per cent.

The results for BCS70 are thus a little less consistent and tidy than those for NCDS, with greater model uncertainty, a lack of nesting of all 10 models within the most likely one, and the original stepwise model not corresponding to any of the 10 retained BMA models (though containing only antecedents that were also identified by the BMA procedure).

Since the symmetric BMA Occam's window produced 20 models for NCDS and 64 for BCS70, it is not feasible to examine the terms included in each model here. Rather, we use the summary information on posterior effects probabilities shown in Table 4. We shall not dwell on the elements already considered for the strict windows. Suffice it to observe that the posterior effect probabilities for these antecedents are (unsurprisingly) very similar for both the strict and the symmetric windows. However, it is worth examining the additional childhood antecedents that appear infrequently in the models and always make the models less likely though more complex according to the BIC. In general, the posterior effects probabilities for these additional childhood factors are low and usually low enough to suggest evidence for the null hypothesis that they do not matter in determining adult malaise.

Table 4: Effect posterior probabilities for NCDS and BCS70 in BMA models with both strict and symmetric Occam's windows

Childhood antecedent	NCDS		Childhood antecedent	BCS70	
	Strict	Symmetric		Strict	Symmetric
Number of models	6	20	*Number of models*	10	64
Any poverty	**100**	**100**	**Any poverty**	**100**	**100**
Any low test	**100**	**100**	**Any low test**	**100**	**100**
Any truancy	100	100			
Any high anxiety	**100**	**100**	***Prob low/any anxiety***	***91.6***	***92.3***
Any low father interest	100	100			
Missing tests	90.8	92.4	**<2 Low anxiety**	**68.8**	**78.2**
Any police contact	78.4	82.3	**Missing anxiety**	**55.2**	**67.2**
<2 High tests	60.9	62.7	(Re)marry LP	63.4	58.7
			Missing tests	25.9	24.6
			Any high anxiety	**22.8**	**22.5**
Ever in care		16.7	2/3 high aggression		19.7
<2 Owner-occupier		4.7	Dissolution& remarry		14.1
2/3 high truancy		4.2	<2 Owner-occupier		10.5
Any hyperactivity		4.0	Any public housing		6.7
2/3 high aggression		2.7	Any hyperactivity		5.1
Any low mother interest		1.8	<2 non-manual father		2.2
Missing police		1.6	Ever in care		1.6
Missing behaviour		1.5	2/3 manual origin		1.4
			<2 non-manual origin		1.1
			Missing aggression		1.0
			Missing hyperactivity		0.9

But there are some grounds for not dismissing this information out of hand. For example, having ever been in care for the 1958 cohort has a posterior effect probability of 16.7 per cent. In earlier work, using a conventional five per cent cut-off for the stepwise modelling, having been in care was powerfully associated with a wide range of disadvantaged adult outcomes (Hobcraft 1998). However, this is a rather rare group, covering about two per cent of the population. Intuitively, when using categorical variables, it would make sense to use a less stringent testing criterion for rare groups and this need has also been discussed in the context of the BIC, which uses the total sample size and not the group size (see Weakliem 1999 and discussion). In this instance, we thus believe that BMA points us towards a relationship that probably does matter.

The remaining, very weak posterior effects probabilities for the NCDS probably do not add much insight. However, they may give hints about measurement issues: mother's and father's interest in schooling are fairly highly correlated, yet the father's interest dominates quite clearly; weak indications of the other dimensions of childhood behaviour (aggression and hyperactivity) also appear; as does a further indicator of frequent school absences; and missing values on two important factors, anxiety and contact with the police, just get noticed.

For BCS70 there is a slightly stronger indication than for NCDS that childhood aggression may matter for adult malaise. The substantial subset (85%) of the (re)marry group, with dissolution followed by remarriage appears as an alterative in some models (PEP=14.1 per cent). There is also some weak indication that parental housing tenure during childhood is linked to malaise at age 30 for this cohort (PEPs of 10.5 and 6.7, totalling 17.2 per cent), as it was even more weakly for NCDS (PEP=4.7 per cent). For BCS70, but not for NCDS, there is a scatter of weak indications that social class matters a little, whether of the father or of origin.

Thus BMA provides many useful insights that are not easily obtained with other approaches. Where there is powerful collinearity (not seen here) the models can alternate between those that contain one or the other of two closely related factors (we discovered this bonus by accident). The recognition and quantification of model uncertainty is valuable and can prevent over interpretation of results from simpler models. Some looser similar insights can be gained from a careful examination of the variables that are just included or just not included in stepwise models or by fitting a fuller model and examining significance levels in terms of p-values rather than arbitrary cut-offs.

There are a few difficulties that deserve recall. Firstly the algorithm used, bic.logit, can only search among up to 30 covariates, because the number of possible models is vast (2^{30} or about 10^9). Conventional backwards stepwise elimination is used to reduce the initial list of covariates to 30 (once the algorithm is patched). Thus care is needed if there are covariates that are close competitors: one is likely to be removed by the stepwise procedure. We have verified model selection by retaining only those childhood measures among the 30 used for the BMA search that were included in the symmetric window and then supplementing these with those that were eliminated by the stepwise procedure. We advise such (tedious) caution.

A second major problem comes in the presentation of results. We all know the difficulties of presenting any complex model in an accessible way. When the real result is to say that we need to average over several (often very many) 'acceptable' models this becomes near impossible. We have shown that several

key insights can be obtained from the posterior effects probabilities, but presenting the parameter estimates for each model can only be achieved in graphical form when there are many models (see, for example, figure 4, p.28 of Hoeting *et al* 1999). The difficulty is that (at least in our models) the parameter estimates for each covariate have a small range, excepting the probability mass at zero when not included in the models. We thus find it unhelpful to present an 'average' parameter estimate weighted by the model posterior probabilities, even though this is available within the algorithm used.[5] It is the models that are averaged. This presents little difficulty for the derivation of predicted probabilities, which simply involves the laborious fitting of each of the BMA models in turn, derivation of the fitted values, and averaging these using the model posterior probabilities as weights; all of this is fairly easily accomplished by editing the output from bic.logit into a Stata do-file. But we do not really have a clue as to how to present BMA detailed results to policy makers in any accessible form beyond the posterior effects probabilities and summary charts on the parameter estimates.

4.2 Recursive Trees

4.2.1 Recursive tree results using the BCS70 learning sample

We use recursive trees to explore the relations between the childhood background factors listed in Table 1 and scoring high on the malaise inventory at age 30. Both entropy impurity measures and relative risks are used to describe the risks of having a high malaise score in any two sub-populations of the learning sample. The entropy impurity of any split of node τ into left and right daughter nodes τ_L and τ_R, respectively, is given by $P\{\tau_L\}i\{\tau_L\} + P\{\tau_R\}i\{\tau_R\}$ where $P\{\tau_L\}$ and $P\{\tau_R\}$ are the probabilities that a subject falls within nodes τ_L and τ_R, respectively, and $i(\tau_L)$ and $i(\tau_R)$ are given in equation (8) above.[6] For descriptive purposes, we also present the 95% confidence interval for the relative risk, but, because the relative risks are biased upward, these are inappropriate for significance testing.

Using data from a randomly selected sample of women (n=2813) from the British Cohort Study, the tree-based method, pruned to a significance level of 0.005, produced Figure 1. Here, we see that the childhood test score summary was used to split the entire learning sample (the root node). Two subpopulations result from this split, and comparing nodes 2 and 3, it emerges that women who

---

5 Although as equation (3) makes clear, the value we want is $p(\beta_1 \mid D, \beta_1 \neq 0)$, the algorithm we use averages over all the models, including those for which $\beta_1 \neq 0$.

6 Because in assessing goodness of split (equation 8 above), $i(\tau)$ is the same for all splits of node τ, a ranking of impurity from smallest to highest will also provide a ranking of goodness of fit.

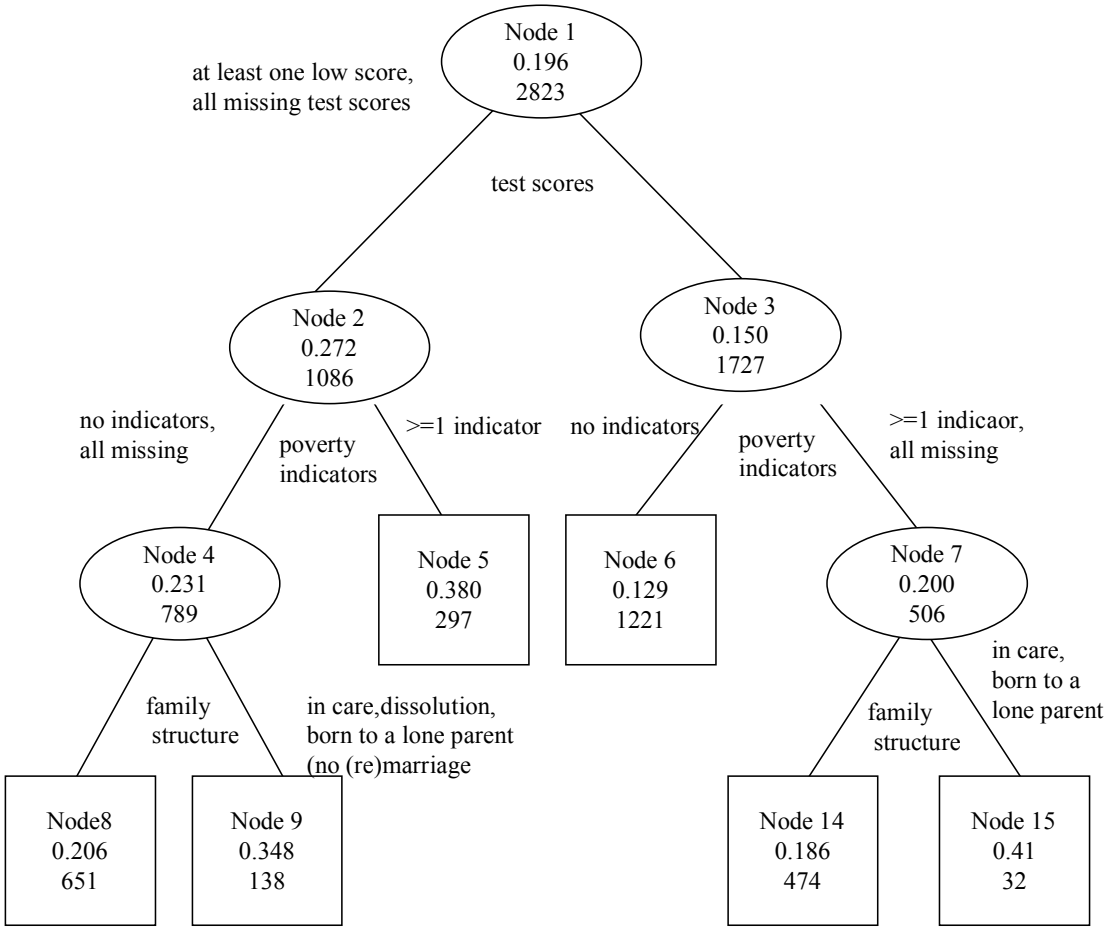
had at least one bottom quartile set of test scores during childhood are more likely than women with better test scores to have a high malaise score at age 30. The missings together approach places those women with no information on their academic test scores at either of the three childhood waves into the higher risk group (relative risk (RR) = 1.81, 95 percent confidence interval (CI) 1.5-2.2).

Nodes 2 and 3 were each split using the measure of childhood poverty that combines three indicator measures (free school meals, receipt of universal benefits, and reported financial difficulties), but the sub-populations were split somewhat differently. Node 2 was split so that those women with at least one indicator of child poverty form one, higher risk daughter node, and those without any indicators that suggest poverty, as well as those with missing information for all three indicators, form the other, lower risk node (RR=1.65, 95 percent CI 1.2-22). Those women located in node 3, who performed better on their childhood academic tests, are split into one subpopulation that has no evidence of childhood poverty and no more than one missing value on the three indicators and another subpopulation that has more evidence of childhood poverty, including those with missing information (RR=1.54, 95 percent CI 1.2-2.0). Although nodes 2 and 3 are not split in exactly the same way, the relative risks of childhood poverty are surprisingly similar once we have conditioned on academic test scores during childhood.

In the next layer of the tree, nodes 4 (containing those women with low test scores and low evidence of childhood poverty) and 7 (containing those women with higher test score and higher evidence of childhood poverty) are both split via childhood family structure. Once again, the split is not identical, but in no way dissimilar. Node 4 is split into a higher risk group containing those women who had ever been in foster care, experienced a family dissolution, or were born to a lone mother who never (re)married. All other measures of family experience, including those who have no information on childhood family structure are placed in the lower risk subpopulation (RR=1.69, 95 percent CI 1.1-2.5). Node 7 is split into a higher risk group of women who had ever been in foster care or who were born to a lone parent (regardless of whether or not the parent subsequently married). All other family structure categories are placed in the lower risk node (RR=2.19, 95 percent CI 1.0-4.6). This pattern suggests an interaction between family structure and having experienced *either* low academic test scores *or* having some evidence of childhood poverty. Family structure does not further increase the risk of having a high malaise score for those women who suffered both kinds of disadvantages, but among those who suffered only one, experience of foster care or some forms of single parent family, increases the risk of a high malaise score to levels similar to the doubly disadvantage group. About 38% of those women with low test scores and

relatively high evidence of childhood poverty have a high malaise inventory score at age 30. Among those with only one of these disadvantages, the percentages are about 35% for poor academic performers with little evidence of poverty, and 41% for better academic performers with higher evidence of poverty.

Figure 1: Tree drawn using the BCS70 learning sample, pruned to significance level 0.005, showing Node, proportion with high malaise, and number of cases



4.2.2 Recursive trees for the NCDS sample

Figure 2 presents a tree grown using data from the NCDS sample of women (n=5768). Although the tree grown in Figure 1 was pruned to a significance level of 0.005, the tree presented in Figure 2 is pruned to a more stringent significance level of 0.001. We make this decision both to maintain the simplicity of the tree and because the NCDS sample is more than twice the size of the BCS70 learning sample. Following the advice of Raftery (1995), we apply more stringent significance levels to the large sample, because the use of typical levels of significance is likely to be too inclusive.

Similar to what we found for the BCS70 data, Figure 2 shows that the childhood test score summary was used to split the entire population (the root node). Two subpopulations result from this split, and comparing nodes 2 and 3, it emerges that women who had at least one bottom quartile set of test scores during childhood are more than twice as likely as women with better test scores to have a high malaise score at age 30. Once again, the missings together approach places those women with no information on their academic test scores at either of the three childhood waves into the higher risk group (RR = 2.18, 95 percent CI 1.9-2.6).

Nodes 2 and 3 were each split using a summary of truancy – a measure that was not available in the BCS70 data. Both sub-populations were split so that those with any evidence of truancy form the higher risk group. Additionally, those with missing information on all measures of truancy are placed with the lower risk group in both cases. Although nodes 2 and 3 were split in the same way, truancy has a slightly stronger effect among those individuals with better test scores. When node 3 is split via the truancy summary, those cohort members with some evidence of truancy are more than twice as likely to have a high malaise inventory score at age 30 (RR=2.2, 95 percent CI 1.7-2.8). In contrast, among those with at least one low test score observation, the split via truancy is less dramatic. Compared to other women in node 2, those with some evidence of truancy are 1.67 times as likely to have a high malaise inventory score (95 percent CI 1.3-2.1). Interestingly, when nodes 2 and 3 are split, those groups of women with either one bottom quartile set of test scores and no evidence of truancy (node 4) or no bottom quartile set of test scores and some evidence of truancy (node7) both have similar risks of high malaise (15.6% and 16.5% chance of high malaise in nodes 4 and 7 respectively). This pattern suggests that truancy and test scores could, perhaps, be usefully combined into a summary variable of school performance.

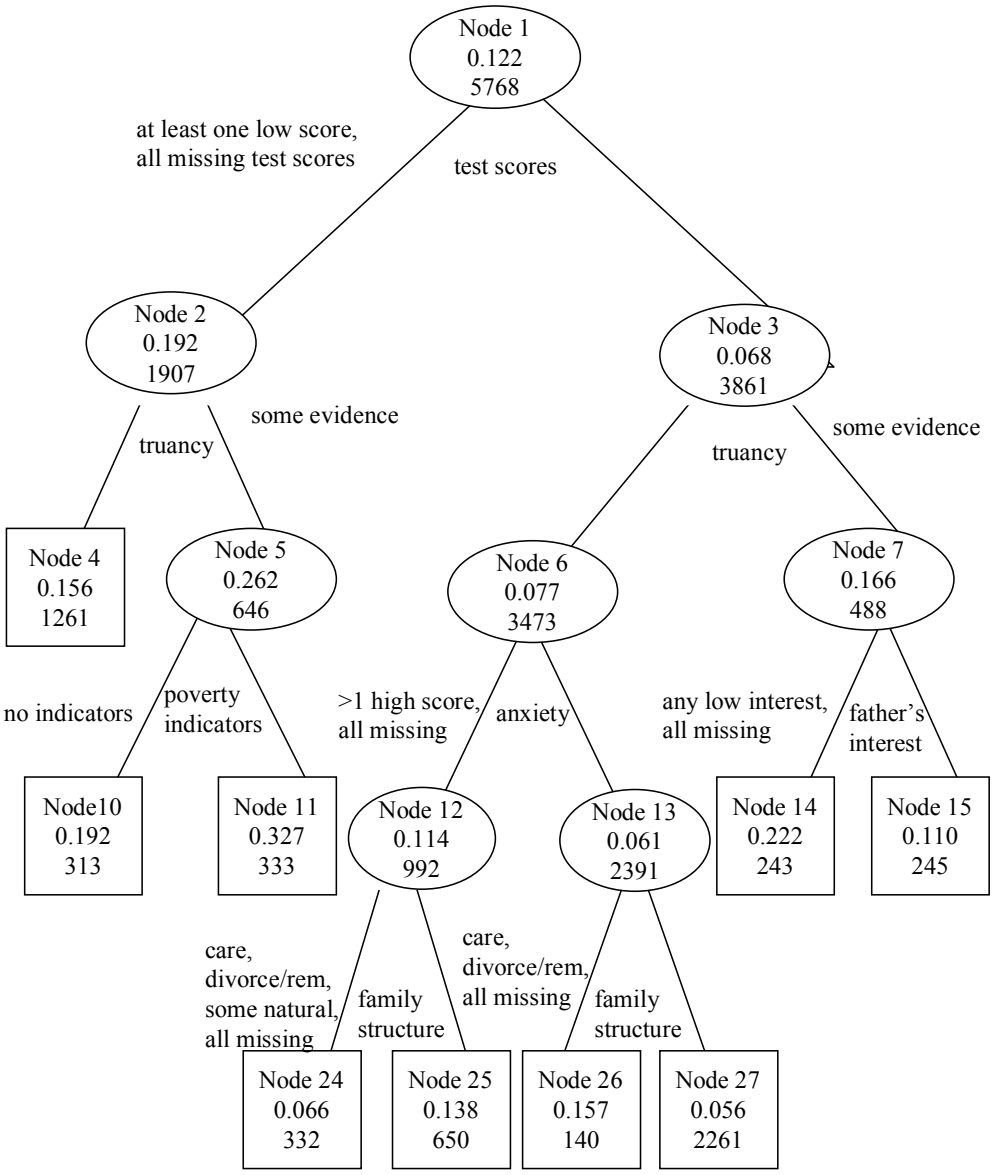
In the next layer of the tree, nodes 5, 6, and 7, are each split using different variables. Node 5 is split using the poverty indicator summary so that those women with low test scores and some evidence of truancy are even more likely to have a high age 30 malaise score if there is any evidence of poverty during childhood (RR=1.71, 95 percent CI 1.2-2.5). This higher risk group also includes those women with missing information on all childhood poverty indicators. Among the most advantaged group (no low test scores and no evidence of truancy) in node 6, having any high anxiety scores during childhood (or missing information at all ages) increases the risk of having a high age 30 malaise score (RR=1.86, 95 percent CI 1.4-2.4). Finally, among those women with no low test scores and some evidence of truancy, it is the summary of father's interest in their education that best differentiates that group according to adult mental health. For those women located in node 7 who also have at least

one childhood report of low paternal interest in education, the risk of having a high malaise score at age 30 is about twice as high as women with more interested fathers (RR=2.02, 95 percent CI 1.2-3.3). This high risk group also contains women with no information on father's interest in their education at any of the childhood waves.

In this layer, we find similar risks of having a high malaise score among that group of women who have no low test scores, some evidence of truancy, but no evidence of low paternal interest in education (node 12) and another group of women who have no low test scores, no (or missing) evidence of truancy, but at least one high anxiety score (node 15). The risks of high malaise for these groups are 0.114 and 0.110, respectively.

In the fourth, layer of the tree, only nodes 12 and 13 are split, but both are split via the family type variable. Among those women with no evidence of low test scores, no evidence of truancy, but at least one high anxiety score, there is a *lower* proportion with high malaise inventory scores at age 30 among those who also had ever been in care (n=12), ever lived in a stepfamily following parental divorce (n=18), appear to have lived with both natural parents throughout childhood but have some missing information (n=257), or have all information missing (n=45). The higher risk group, which is dominated by those who have clear evidence of living with both natural parents at all childhood waves (n=546), are more than twice as likely to have a high malaise score (RR=2.09, 95 percent CI 1.3-3.4). This split is not easily interpreted and not very intuitive. On the other hand, the partition of node 13 via the family type variable shows that among women with no low test scores, no evidence of truancy, and no high anxiety scores there are *more* women with a high malaise score among those who had ever been in care (n=21), ever lived with a stepfamily following a parental divorce (n=31), or had all family structure information missing (n=88; RR=2.83, 95 percent CI 1.7-4.6). Thus, both splits by family structure pick out groups that are dominated by those with missing or partial information. This apparent crossover split is not at all easy to interpret and a researcher confronted with such information may choose either to prune this split from the tree or to see whether slightly different, but more sensible splits have similar levels of impurity (Zhang and Singer, 1999).

Figure 2: Tree drawn using the NCDS sample, pruned to significance level 0.001, showing Node, proportion with high malaise, and number of cases



4.2.3 Hidden Branches

While an examination of the factors selected to construct the tree in Figures 1 and 2 is instructive, it is also useful to examine the nearly selected factors (Zhang and Singer, 1999). Following Zhang and Bracken (1995), we present these "hidden" relations in tabular form: Table 5 for the tree built using the BCS70 learning sample and Table 6 for the tree built using the female NCDS cohort members. The first column of these tables indicates the number of the node along Figures 1 and 2 – for example, the number 1 refers to the root node. The second column presents the name of the variable that provides a competitive split of that node, and column 3 identifies the higher risk

subpopulation based on that split. Column 4 presents the impurity of the split based on the entropy impurity function – the measure used to identify the best split for each node. Finally, the last two columns report the relative risk and its 95 percent confidence interval – information used in the pruning process. For the purposes of these tables, we ordered competitive splits by the entropy measure used to determine the split and only show the top three competitors, although more are shown where the entropy measure to 3 decimal places is the same. In addition, we have only included those where the 95 per cent confidence interval around the relative risk does not include 1.0. To illustrate, the third row of Table 5 tells us that a competitive split based on childhood housing tenure for node 1 in Figure 1 is whether the woman was observed living in public housing at one of the childhood waves. The impurity of the split is 0.49, and there are more women with a high malaise score among those women observed living in public housing at least once (RR=1.5, 95 percent CI 1.3-1.9).

Table 5: Competitive splits for the nonterminal nodes for high malaise score (Figure 1), BCS70

Node	Variable	Higher Risk Subpopulation	Impurity	Relative Risk	95% CI
1	tests	At least one low set of scores*	0.485	1.8	1.5-2.2
1	povind	At least one indicator	0.488	1.7	1.4-2.1
1	tenure	At least one public house obs.	0.490	1.5	1.3-1.9
1	famtype	Care, diss/rem, lone at birth/no rem	0.491	1.7	1.3-2.1
2	povind	At least one indicator	0.574	1.6	1.2-2.4
2	famtype	Care, dissolution, lone at birth/no rem	0.576	1.6	1.2-2.1
2	fathclass	<2 nonmanual obs*	0.580	1.7	1.1-2.8
2	aggress	2 high scores	0.580	1.7	1.0-2.8
3	povind	No ind + 2 missing or >0 indicators*	0.418	1.5	1.2-2.0
3	tenure	At least one public house obs.*	0.419	1.5	1.2-2.0
3	anxiety	<2 low scores*	0.420	1.7	1.2-2.4
3	famtype	Both natural parents, dissolution	0.420	1.9	1.1-3.1
4	famtype	Care,dissolution, lone at birth/no rem	0.532	1.7	1.1-2.5
7	famtype	Care, lone at birth	0.492	2.2	1.0-4.6
7	tenure	At least one public house obs.	0.493	1.6	1.0-2.5

Not surprisingly, given the subsequent best splits of nodes 2 and 3, Table 5 shows childhood poverty provides the second best, competitive split of the root node. Both the impurity measure and relative risk measures are extremely close. Tenure provides the next best competitive split, and it provides competitive splits in the next layer of the tree, as well as for node 7. Consequently, housing tenure, although not selected as one of the best splits in Figure 1, may, nonetheless, be a useful factor to consider.

Moving on to examine the competitive splits for node 2 (conditional on having made the "best" split of the root node), we find that both the social class of the father figure and childhood aggression provide competitive splits. Because neither of these factors provides competitive splits of node 3, this pattern suggests that there may be an interaction between low test scores and having a lower social class (measured by the occupation of the father). Additionally, there may be an interaction between low test scores and high aggression scores. Among the subpopulation with higher test scores in node 3, high childhood anxiety provides a competitive split, also suggesting a possible interaction.

No factors other than childhood family structure provide competitive splits for node 4 (those women with at least one bottom quartile test score but no evidence of childhood poverty). For node 7, only tenure provides an additional competitive split.

Table 6 provides information on competitive splits for the NCDS data. The second best split of the root node is the summary of paternal interest in education – a factor that contributes a best split much later in the tree. The truancy summary, which provides the best split of nodes 2 and 3, provides the third best competitive split for the root node. Among all three candidate variables, both the impurity measure and relative risk measures are extremely close. Maternal interest in education and the poverty summary provide the next two best competitive splits, and both provide competitive splits in the next layer of the tree. Although no node is ever split via maternal interest in education it is worth noting that this summary variable provides a competitive split for nodes 5 and 7 as well. Note in Table 5, this variable provided none of the best competitive splits. This may be due to the fact that the NCDS measure summarises three measures while the BCS70 measure only relies on information collected at age 10. Whatever the reason, maternal interest in education, although not selected as one of the best splits in Figure 2, may be a useful factor to consider. Alternatively, given the high correlation between maternal and paternal interest in education, it may be worth considering whether the two measures could be combined to form a stronger indicator.

Table 6: Competitive splits for the non-terminal nodes for high malaise score (Figure 2), NCDS

Node	Variable	Higher Risk Subpopulation	Impurity	R R	95% CI
1	tests	at least one low test score*	0.361	2.18	1.9-2.6
1	finted	little interest at least once*	0.362	2.12	1.8-2.5
1	truant	some evidence of truancy	0.362	2.24	1.9-2.7
1	minted	little interest at least once	0.363	2.10	1.8-2.5
2	truant	some evidence of truancy	0.481	1.67	1.3-2.1
2	povind	at least one indicator*	0.481	1.67	1.3-2.1
2	minted	little interest at least once	0.483	1.58	1.3-2.0
2	finted	little interest at least once*	0.483	1.62	1.3-2.1
3	truant	some evidence of truancy	0.293	2.16	1.7-2.8
3	finted	little interest at least once	0.294	1.95	1.5-2.5
3	povind	at least one indicator	0.295	1.88	1.5-2.4
3	anxiety	at least one high score*	0.295	1.72	1.4-2.2
3	minted	little interest at least once	0.295	1.83	1.4-2.4
5	povind	at least one indicator*	0.563	1.71	1.2-2.5
5	famtype	all/some nat, divorce, other one parent	0.566	1.70	1.1-2.6
5	minted	little interest at least once*	0.567	1.62	1.1-2.4
6	anxiety	at least one high score*	0.267	1.86	1.4-2.4
6	tests	<2 high scores	0.269	1.71	1.2-2.3
6	povind	yes>no	0.269	2.11	1.4-3.2
6	famtype	all/some natural, one parent, other remarry	0.269	1.67	1.2-2.3
6	finted	little interest at least once	0.269	1.64	1.2-2.2
6	police	some evidence of contact	0.269	2.62	1.4-4.8
7	finted	little interest at least once*	0.438	2.02	1.2-3.3
7	aggress	at least one high score	0.441	1.91	1.1-3.2
7	minted	<2 very interested	0.442	2.09	1.1-4.0
7	povind	all not poor	0.442	1.87	1.1-3.3
12	famtype	all natural, lone parent at birth, divorce (no remarry), other one parent	0.349	2.09	1.3-3.4

13	famtype	ever in care,widowed parent, stepfamily	0.227	2.83	1.7-4.6
13	fathclass	<2 non-manual	0.228	1.96	1.3-2.9
13	tenure	<2 owner occupier	0.228	1.81	1.3-2.6
13	tests	< 2 high scores	0.229	1.94	1.3-3.0
13	povind	all not poor*	0.229	2.41	1.4-4.1
13	finted	<2 very interested	0.229	1.64	1.1-2.3
13	origclass	<2 non-manual obs.*	0.229	1.84	1.1-3.0

Given that the root nodes in Figures 1 and 2 are split in the same way, we can compare the competitive splits for nodes 2 and 3 across samples in order to identify a set of interactions with test scores that are likely to have common effects on adult malaise in both samples. We compare the best competitors in Table 6 to Table 5. The second best split of node 2 in Table 6 is via the childhood poverty indicator. This variable provides the best split of node 2 in Table 5. Mothers' and fathers' interest in education contribute the third and fourth competitive splits of node 2 in Table 6, but neither of these provide a competitive split in node 2 of Table 5.

Moving on to node 3, we find that the impurity measure of the first five competitive splits ranges from 0.293-0.295, so the differences between their goodness of split is rather small. After truancy, which provides the best split of node 3 in the NCDS sample, we find that paternal interest in education offers the next best competitive split. The third best split of node 3 is via the poverty indicator summary. Despite differences in the construction of the two measures, the poverty summary variable provided the best competitive split of node 3 in the BCS70 sample. The next best split of node three is via the childhood anxiety summary variable, and closely behind is the split of maternal interest in education. Of these last two competitive splits, the anxiety summary provides a competitive split of node 3 in Table 5.

Among women with any low test scores and any evidence of truancy (node 5 in Table 6 and figure 2), only three factors provide competitive splits. After the poverty indicator, the second best split is via the family type variable but, as with the splits via family type in nodes 12 and 13, this is not easily interpreted. The only other competitive split of node 5 is provided by the summary of maternal interest in education.

For node 6, comprising the most advantaged women with no low test scores and no evidence of truancy, academic test scores, childhood poverty, family type, paternal and maternal interest in education, and contact with the police all offer

competitive splits. The range of the impurity measures is narrow, suggesting that all factors are close competitors.

For node 7, comprising women with no low test scores but some evidence of truancy, aggression, mother's interest in education, and the child poverty indicator provide competitive splits. Looking across nodes 5, 6, and 7, we find that the poverty indicator variable offers a competitive split of all three. In contrast, anxiety is not a competitive factor for either node 5 or node 7. Finally, father's interest in the cohort members' education provides a competitive split of both node 6 and node 7, suggesting an interaction between having no low academic test scores and paternal interest in education.

There are no competitive splits of node 12. For node 13 the two closest to the best split, are the summary of the father's social class, and housing tenure. In addition, academic tests, childhood poverty, paternal interest in education, social class of origin, aggression, mother's interest in education, contact with police and parent's school leaving ages all provide competitive splits of node 13. Given the fact that the population of the node 13 is still rather large (n=2391), it is not surprising that there are still competitors to the split via family type. Nonetheless, after splitting by family type, both daughter nodes become terminal.

In the preceding discussion, we have assumed, as we moved down the tree, that the previous splits have followed those in Figure 1 or Figure 2. We could, alternatively, glean additional information by looking for patterns along alternative tree structures. This may provide additional information about the underlying structure of the data, but is beyond the scope of this current application.

5. Branching Out: Combining Trees and BMA and Beyond

In this section we provide several illustrations of further ways in which insights can be gleaned from bringing the two approaches together and from using the results to extend the analysis in other ways. We have chosen to use the NCDS sample for these purposes, partly because the subsequent out-of-sample validation analysis uses BCS70, but also because there was less apparent model uncertainty and more covariates are available.

We begin by summarising comparable information on goodness of fit for the NCDS using the two approaches. We do this using three indicators: the BIC, as used to judge goodness of fit for the BMA procedure; the likelihood ratio deviance; and the ROC (receiver operating characteristic). Unlike BIC, neither

the deviance nor the ROC penalise for model complexity. Table 7 provides these summary indications and show some of the inherent problems of judging the fit of models. According to the ROC statistic, the BMA models all 'fit' better than the model derived from the recursive trees, though the difference is small at about 0.01. The deviance measure, on the other hand, suggests that the regression model based on the categorical nodes fits better. Lastly, our preferred BIC measure provides strong evidence that the BMA-derived models are all to be preferred to that derived from the nodes of the tree, with the BIC-difference being over 10 compared with the 'best' BMA model.

Table 7: Goodness of fit for several logistic models for malaise, NCDS

Model	ROC	BIC	Deviance	Degrees of freedom
Initial BMA strict window	0.6871	-209.8 to –206.1	248	N/A
Initial BMA symmetric	0.6908	-209.8 to –203.9	252	N/A
Initial BMA 'best' model	0.6867	-209.8	248	8
Nodes from Figure 2	0.6761	-198.9	268	8
BMA with counts and interactions – 'best' model – section 5.2.1	0.6912	-227.6	297	8

5.1 Simple Cross-Pollination

Since recursive tree methods essentially generate information about interactions, one post-initial strategy is to add the predicted probabilities from the BMA results into the variables considered for the tree. If the BMA approach, using averages of general linear models without interactions, actually accounts well for the variation in incidence of malaise in the sample we would expect the predicted probabilities to dominate at least the early splits in the trees. Splits on other antecedents could then be indicative of risk factors that applied over part of the range of underlying propensities to experience malaise. When we introduce the BMA predicted probabilities into the recursive trees alongside the full range of the child antecedents we do indeed find that the initial splits are dominated by the BMA predicted probabilities. Indeed, if the tree was pruned to a significance level of 0.0005 (a less stringent cut-off than we have used in some earlier work using stepwise regression, see Hobcraft 2000), only splits on the BMA predicted probability are included, with four resultant terminal nodes identifying cut-points at 0.065, 0.102, and 0.213 and including 1,079, 2,363, 1,740 and 585 women respectively. The only further splits to occur at the 0.001 level concern two groups, the smallest group with the highest level of risk, average malaise of 30.6 per cent and the largest group at

the second level of risk. For the high risk group, the tree identifies childhood family disruption as a major source of variation, with the 109 women who had been in care (n=47), had no coresident father at birth (n=39), or had all missing information on family (n=23) having a malaise incidence of 44 per cent, compared with 27.5 per cent for the remainder. For the largest group, the next split identifies those with two or more high test scores (n=315) as being at a lower risk of malaise (4.7 per cent, compared with 9.2 per cent for the remainder). Subsequent splits among the remainder group are uninformative and not considered here. A further example of this approach is given for the validation sample from BCS70 later.

The alternative cross-pollination approach is to introduce information about the nodes from the recursive trees into regression-type analysis. We had hoped that recursive trees would help to identify clear indications of interaction terms to introduce into regression analysis, but the recursive binary splitting often leads to quite complex 'twigs' emerging. We have considered two approaches to exploring the benefits of recursive trees.

The first approach is to include the resultant terminal nodes in BMA analysis, but exploration using dummy variables to identify the nodes has proven unsatisfactory. An alternative is to introduce the observed (same as predicted) probabilities for members of each node into a BMA analysis and explore what is identified as improving the fit. This is illustrated for the BCS70 validation sample later.

The second approach is to use the results from the recursive trees to give insights into meaningful and missed interactions that can be included in regression analysis. Again, as discussed above, there are problems in getting clear indications of such interactions, especially from the more complex trees for the NCDS. Inevitably many branches lead to intermediate or final nodes with fairly similar incidence of malaise. In the earlier discussion of Figure 2, the similar incidence of malaise for those with low test scores and no frequent school absences and those with no low test scores and some frequent school absences were noted, suggesting a possible combination of two disadvantages into three categories, neither, either, or both might be helpful. However, as we move further down the tree things get much untidier, with different factors coming into play. Moreover, the results are sometimes counterintuitive or very difficult to interpret, even where the same factors are involved in the split as instanced for nodes 24 to 27 in Figure 2.

5.2 Rooting Around in the Data

The lack of real success in identifying interactions from the recursive trees led us to explore two further, more data intensive approaches. These retain some of

the spirit of the BMA and trees combination and draw considerable inspiration from the underlying philosophy of Singer *et al* (1998). The first additional exploration uses regression techniques to illuminate possible interactions and suggests the need for an interaction between educational test scores and the anxiety measures. It also combines some of the results from BMA with a modification of the Boolean OR approach advocated by Singer *et al* (1998). The second additional approach is an attempt to retain the spirit of the very painstaking examination of the data advocated by Singer *et al* (1998), but economising on effort to some extent by using the BMA results to suggest where to look for combinations into real groups.

5.2.1 Some Mechanical Digging

Both the BMA and recursive trees analyses suggested that educational test scores were among the most powerful antecedents of adult malaise (this power of educational test scores has been found very consistently in analyses of a whole range of adult disadvantages using information from the NCDS – see Hobcraft 1998, 2000, 2003, and 2004). Table 8 shows the results from fitting stepwise regressions separately within each of the test score classifications used in the best fitting child model as identified by BMA. Particularly striking are the odds ratios for childhood anxiety, which suggest a clear interaction of anxiety with test scores: those with any low test scores need two or more low anxiety scores to have lower adult malaise; those with two or three high test scores but also two or three high anxiety scores are also at excess risk of adult malaise; and the largest, intermediate group on test scores (no low, but <2 high) are at greater risk of adult malaise if they had any high anxiety score as children. This result is fairly robust since locking in all of the non-test dummies from the best child model and exploring the need for additional significant childhood antecedents only picks out the same anxiety interactions. The proportions experiencing high malaise as adults by the identified groupings are shown in Table 9. The most striking incidence of malaise is for the small group of women who performed very well on educational tests but were very anxious, among whom 23 per cent (or 12 women) experienced high malaise at age 33.

The other set of results that can be recalled from the BMA analysis for NCDS, shown in Tables 3A and 4, indicate the cluster of antecedents other than test scores and anxiety that were included in the best fitting child model nearly all had posterior probabilities of 100 per cent, with any police contact being the only exception at around 80 per cent. Although the BMA effect sizes were not shown, they are all of fairly similar magnitude. This suggests that a simple count of these 'strong' disadvantages might be worth considering as a more parsimonious description of childhood disadvantage. In essence, a sum of indicator variables serves to identify the first level of a set of Boolean OR statements (Singer *et al* 1998) covering all of the four items of any occurrence

of poverty, low father's interest, frequent school absence, or contact with the police. The intrinsic advantage of a sum of these indicators is that it can also give us some information regarding multiple childhood disadvantages too. Since the symmetric BMA identified several childhood antecedents with low posterior effect probabilities it also seemed worth posing the question as to whether a count across these 'weak' disadvantages (ever in care, two or three high aggression scores, any high hyperactivity score, any low mother's interest in schooling, and strong indications of frequent school absences) might add further explanatory power to our models or trees. The incidence of malaise and the numbers of women involved for these groups are shown in Table 10.

Table 8: Odds ratios from stepwise regression models within test score groups (p=0.003)

Antecedent	BMA best model	Any low tests	No low tests & <2 high tests	Two or three high tests
Number of women	5768	1822	2821	1040
Any poverty	1.53	1.55	---	---
Fairly poor	---	---	1.72	---
Any low father's interest	1.49	---	1.54	3.06
Any low mother's interest	---	1.42	---	---
Any contact with police	1.71	1.79	---	6.45
Any frequent school absence	1.60	1.45	1.84	---
Not 2/3 low anxiety	---	1.63	---	---
Any high anxiety	---	---	1.57	---
2/3 High anxiety	---	---	---	6.89
Any low tests	1.54	---	---	---
No low tests & <2 high tests	1.58	---	---	---
All tests missing	4.00	---	---	---

Table 9: Incidence of high malaise by test scores and anxiety

Anxiety	Test scores		
	Any low	No low <2 high	2-3 high
2-3 high	0.209 (1366)	0.132 (831)	0.231 (52)
1 high			
No high <2 low		0.086 (831)	0.048 (988)
2-3 low	0.140 (456)		

Table 10: Incidence of high malaise by counts of strong and weak childhood disadvantages

Count	**0**	**1**	**2**	**3**	**4**
Strong disadvantages	0.076	0.127	0.202	0.305	0.396
N	3080	1541	766	328	53
Weak disadvantages	0.087	0.148	0.241	0.412	0.25
N	3532	1675	456	97	8

Note: Strong disadvantage includes a count of any: poverty, contact with police, frequent school absences and low father's interest in schooling; weak disadvantages are: ever in care, 2-3 high aggression, any hyperactivity, any low mother's interest in schooling and strong frequency of school absence.

Results from incorporating dummy variables for these test-anxiety interactions and counts of strong and weak indicators of disadvantage are shown in Table 11. We note that the dummy variables for the counts of disadvantage were again coded hierarchically (1+, 2+, 3+, 4). We see that a possible penalty of introducing these new items into the BMA is an increase in model uncertainty, especially for the strict criterion. This is also reflected in the lack of effect posterior probabilities of 100 per cent, except for the indicator of the experience of any one or more of the strong childhood disadvantages. However, what we do observe is that the new measures do indeed dominate the plausible models, with the convenient indication here being taken as an effect posterior probability of above two-thirds. This subset also comprises the best BMA model and the odds ratios from that model are shown in the final column of Table 11. In interpreting these odds ratios it is important to recall the

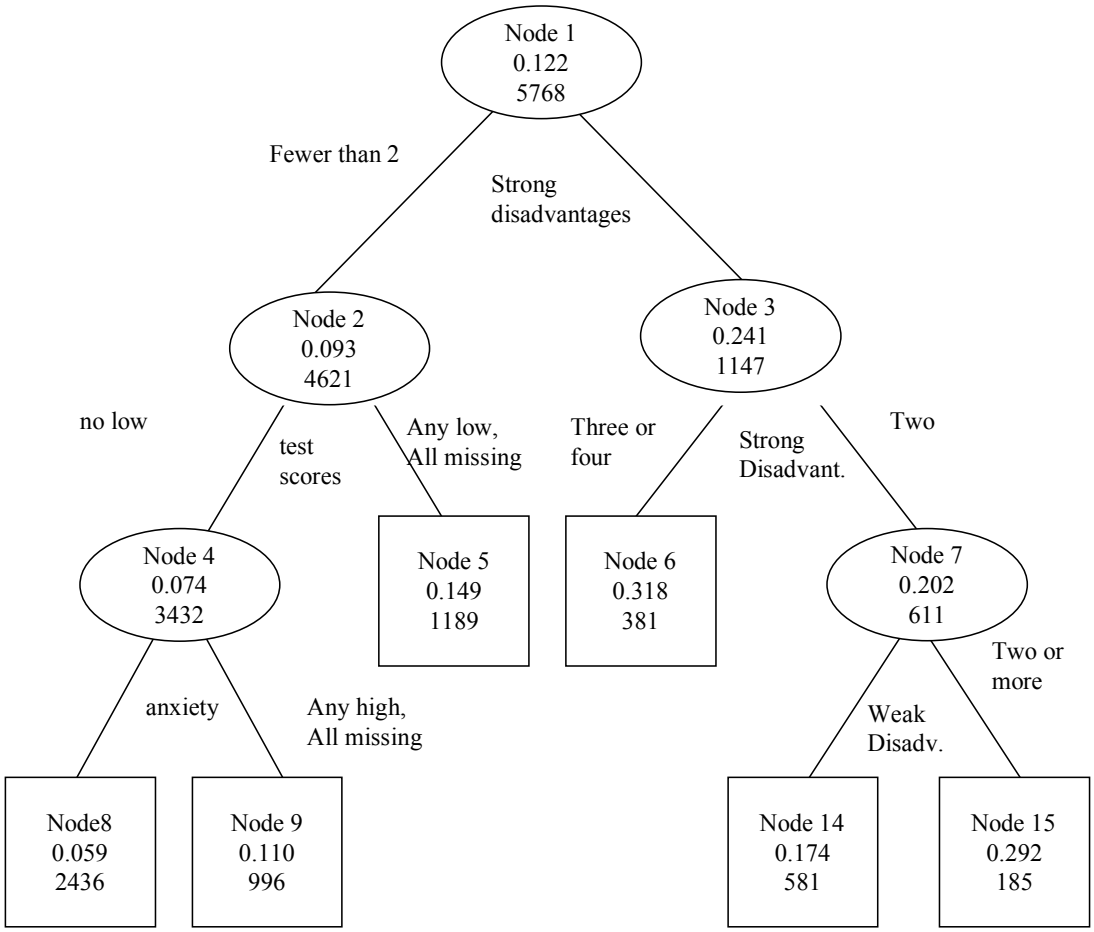
hierarchical coding of the counts, so that the odds ratio for the three or more strong disadvantage group is 3.81 (=1.57*1.53*1.58) compared with the no strong disadvantage group. No component of either the strong or weak disadvantage counts appear at all on its own and no other dummy appears at all outside the counts and the test scores and anxiety measures already implicit in the interactions. Thus the new model seems to be doing well in capturing variation. This proves to be the case if we return to Table 7, which summarises the various goodness-of-fit measures for this model (with counts and interactions). This new 'best' model has a better ROC score, a higher deviance, and a substantially higher BIC score than any of the earlier ones.

Table 11: Effect posterior probabilities for BMA models with counts and interactions for female adult malaise, NCDS

Childhood antecedent	**Strict**	**Symmetric**	**Odds ratios in 'best'**
Number of models	*33*	*45*	
1+ Strong disadvantages	100	100	1.57
2+ Strong disadvantages	94.0	95.0	1.53
2-3 High tests & 2-3 High anxiety	91.0	92.4	3.79
No low tests &<2 High tests & any high anxiety	86.1	86.0	1.60
All tests missing	82.7	85.4	2.88
Any low tests & not 2-3 low anxiety	76.9	80.5	1.98
2+ Weak disadvantages	68.9	69.8	1.50
3+ Strong disadvantages	68.6	67.4	1.58
3+ Weak disadvantages	25.3	27.9	---
No low tests&<2 High tests	23.1	30.8	---
Any low test	19.0	23.7	---
Any high anxiety	8.1	6.8	---

Figure 3 shows the tree grown with the addition of these counts and interactions, pruned at a significance level of 0.0005. The story is very similar, with tests, anxiety and the two counts dominating splits, but some potentially interesting insights that we do not pursue further here, suggesting that the counts discriminate better within the right hand fork that has the higher incidence of malaise, identified by two or more strong disadvantages, and that tests and anxiety play a greater role for the relatively advantaged, lower malaise incidence left fork.

Figure 3: Tree drawn using the NCDS sample, pruned to significance level 0.0005, showing proportion with high malaise, and number of cases



5.2.2 Grubbing around using tools rather than bare hands

Singer *et al* (1998) used an exceedingly labour intensive approach to obtaining groups of people for their person-centred analysis, which involved examining constructed detailed biographies of individuals and manually searching for patterns of characteristics. We wished to retain some of the spirit of their approach whilst finding a less labour-intensive way of proceeding. Our initial hope was that recursive trees would prove convincing in this respect, but the automation is too great and less sophisticated tools are required.

After some experimentation, we adopted the following approach. We began with the results of the BMA and in particular the elements of the best model (or any of the models under the strict criterion). We then used the collapse facility in Stata to obtain a file with counts of individuals in cells corresponding to the full 2^5* 5 cross-classification; of the 160 possible cells there were only observations in 103. The file also contained the mean value on adult malaise and all of the other childhood characteristics for each combination of test scores

with indicators of any incidence of childhood poverty, anxiety, low father's interest in schooling, contact with the police, and frequent school absences. The resulting file can usefully be sorted by size of the cell or by the summary proportion experiencing malaise and visually inspected using the Stata browser or data editor facility. Moreover attention can be restricted to larger cells, sorted by malaise scores (or any other criterion). Thus there is much scope for semi-automated exploration of the patterns. By far the largest group (n=1225) were those with no low but fewer than two high test scores and no other 'best-model' childhood disadvantage, with 6.3 per cent experiencing malaise at age 33. The second largest group (n=614) comprises those with two or three high test scores and no other 'best-model' disadvantage, with only 3.6 per cent experiencing adult malaise. The association between having no low test scores and not being disadvantaged in other respects is clear.

The patterns on the childhood covariates for the cells containing more than 50 women are shown in Table 12. It is striking how test score groups broadly order the combinations ranked by malaise incidence. Moreover, there is a clear tendency for sub-ordering within test score groups by the count of the number of best-model disadvantages. Although contact with the police is fairly clearly associated with increased risk of malaise it is rare for women (only four per cent in this sample) and thus does not fall into the largest groups. Unlike the previous analysis, there is no special indication that anxiety interacts with test scores more than other childhood disadvantage in its association with malaise incidence. The patterns shown here are indicative of a much more thorough examination of smaller cells too. What emerges from this first pass, in our view, is the possibility of adopting a summary interaction measure combining the test scores with the count of best-model disadvantages. The groupings adopted are shown in Table 13.

Table 12: Cells from collapsed NCDS data with more than 50 observations, ordered by grouped malaise incidence

Malaise incidence (per cent)	Sample size	Poverty	Anxiety	Father interest	Police	Frequent school absence	Count	Tests
Very low								
0.036	614	0	0	0	0	0	0	2-3H
0.038	53	1	0	0	0	0	1	2-3H
0.055	181	0	1	0	0	0	1	2-3H
Low								
0.063	1225	0	0	0	0	0	0	0-1H
Next								
0.082	219	0	0	1	0	0	1	0-1H
0.087	173	1	0	0	0	0	1	0-1H
0.098	102	0	0	0	0	1	1	0-1H
0.104	347	0	0	0	0	0	0	Low
0.114	88	0	1	1	0	0	2	0-1H
0.115	462	0	1	0	0	0	1	0-1H
0.125	96	1	1	0	0	0	2	0-1H
Next								
0.143	112	1	0	1	0	0	2	Low
0.147	184	0	1	0	0	0	1	Low
0.150	60	0	0	0	0	0	0	Miss
0.161	62	1	0	1	0	0	2	0-1H
0.163	80	0	0	0	0	1	1	Low
0.169	65	0	0	1	0	1	2	0-1H
0.172	87	0	1	1	0	0	2	Low
0.173	110	0	0	1	0	1	2	Low
High								
0.181	199	0	0	1	0	0	1	Low
0.191	110	1	0	0	0	0	1	Low
Very high								
0.235	51	0	1	1	0	1	3	Low
0.268	56	1	1	1	0	0	3	Low
0.308	120	1	0	1	0	1	3	Low
0.342	73	1	1	1	0	1	4	Low
Count		9	9	12	0	7		

Table 13: Suggested interaction pattern for test scores and counts of best-model childhood disadvantage

<table>
<tr><td rowspan="2">Count of best-model childhood disadvantages</td><td colspan="4">Test scores</td></tr>
<tr><td>Any low</td><td>No Low
<2 high</td><td>2-3 High</td><td>All missing</td></tr>
<tr><td>0</td><td>0.104
(347)</td><td>0.063
(1225)</td><td rowspan="2">0.039
(929)</td><td rowspan="4">0.188
(85)</td></tr>
<tr><td>1</td><td rowspan="2">0.172
(1039)</td><td>0.101
(981)</td></tr>
<tr><td>2</td><td>0.147
(435)</td><td rowspan="2">0.207
(111)</td></tr>
<tr><td>3+</td><td>0.310
(436)</td><td>0.228
(180)</td></tr>
</table>

The results of a stepwise regression with forward selection (backwards selection gives very similar results, but a slightly worse fit, removing a different interaction dummy) and p=0.003 using this new interaction term are shown in Table 14, as are the goodness of fit statistics. The only terms retained are all but one of the dummies for this test-disadvantage interaction; no other childhood factor enters the model at all (test scores were of course omitted). Judged by the BIC and the deviance, but not by the ROC, this model provides a better fit than any of the original BMA results or the initial model with the nodes. However the improvement is not as great as for the model considered in section 5.2.1.

Table 14 also recalls the odds ratios from the best BMA model for the analysis of section 5.2.1 with test-anxiety interactions and the counts of strong and weak disadvantages. Finally we explore whether combining the insights gained from these two analyses and further introducing a count of the five disadvantages included in determining the test-disadvantage count of this section (any indications of poverty, father's interest, police, and school absences, corresponding to the strong disadvantages of the previous section plus any low anxiety). This is also coded hierarchically. The resultant stepwise model is shown in Table 14 and provides the best fit yet on all of the measures of goodness of fit, with only seven degrees of freedom.

We have thus achieved quite substantial gains in model fit from the two exploratory approaches in section 5. No doubt further progress could be made with more digging. But there is, of course, a danger of overfitting with such detailed exploration.

Table 14: Odds ratios from stepwise logistic regression (p=0.003) results for models with test-disadvantage interactions from section 5.2.2 and for models combining the counts and interactions from sections 5.2.1 and 5.2.2

Count or interaction terms	Test-disadvantage count interactions plus child antecedents	Test-anxiety interactions and counts of disadvantages and childhood (BMA best)	Both sets plus count of disadvantages From 5.2.2
Test Disadvantage Count Interactions			
Low & 0		xxx	
Low & 1 or 2	2.30	xxx	
Low & 3+	4.95	xxx	
<2 high & 0	Ref.	xxx	Ref
<2 high & 1		xxx	
<2 high & 2	1.90	xxx	
<2 high & 3+	3.26	xxx	
2-3 high & 0 or 1	0.45	xxx	0.38
2-3 high & 2+	2.89	xxx	
All missing & any	2.56	2.88	
Test-Anxiety Interactions			
Low & <2 low	xxx	1.98	1.56
Not low, < 2 high & any low	xxx	1.60	
2-3 high & 2-3 high	xxx	3.79	4.27
Count of strong disadvantages			
1+	xxx	1.57	
2+	xxx	1.53	1.51
3+	xxx	1.58	1.60
Count of weak disadvantages			
2+	xxx	1.50	1.55
Count of five disadvantages Section 5.2.2			
1+	xxx	xxx	1.56
Any other childhood antecedent	**None**	**None**	**None**
Deviance	276	297	313
Degrees of freedom	7	8	7
BIC	-215.0	-227.6	-252.1
ROC	0.6792	0.6912	0.6940

Note: xxx means not considered, blank means not included in model**.**

5.3 A brief look at the tree of life

Since we also have results from the malaise inventory at ages 23 and 42 for the NCDS sample, one way of assessing the value of the insights gained from the detailed analysis of the results at age 33 is to see how well they suffice to 'explain' malaise at the other ages. We have explored the antecedents of continuity and change in malaise scores at ages 23 and 33 in more detail elsewhere (Hobcraft 2003 and 2004), but without the possible benefits from the detailed explorations undertaken here. The malaise inventory is meant to capture a fairly stable underlying propensity towards depression, so there should be quite a lot of continuity across ages, although we have also shown that experience of unemployment or of a divorce between ages 23 and 33 is associated with increased risk of malaise (Hobcraft 2003 and 2004). We ignore the inconvenient element involved in moving backwards through the life-course when considering malaise at age 23.

Table 15 shows a simple illustration of the predictive value of the symmetric BMA predicted probabilities. The first panel shows the odds ratios for high malaise scores at age 23 and only four further childhood antecedents enter the model, suggesting that the BMA at age 33 captures many of the important antecedents of malaise at age 23. Similarly, only three additional covariates appear in the model for high malaise scores at age 42, again showing that the BMA at 33 is of considerable predictive value at other ages, despite other life course changes that may have occurred.

Table 15: Stepwise logistic regression results using BMA predicted probabilities for age 33 malaise and other childhood antecedents in models of malaise at ages 23 and 42 in NCDS

Malaise at age 23	**Odds ratio**
BMA at 33 symmetric	276.8
Fewer than 2 Low anxiety	1.53
Fewer than 2 Owner-Occupier	1.39
All behaviour missing	3.50
Fewer than 2 high tests	1.53
Malaise at age 42	
BMA at 33 symmetric	77.21
Fewer than 2 father non-manual	1.39
Any high hyperactivity	1.38
All father's interest missing	1.45

6. Out of sample predictive performance

In this section we examine the predictive performance of both Bayesian Model Averaging and Recursive Tree methods, comparing both to the more typical backward stepwise selection approach. Using the selected models estimated with data from our learning sample, we explore how well these models perform when we apply them to a different, validation sample. We use Receiver Operating Characteristic (ROC) curves in order to summarise out of sample predictive performance of these very different techniques (Hanley, 1989). The area under the ROC curve can reach a maximum of 1 and provides information on the predictive performance of a model. A random prediction model should have a ROC area equal to 0.5, so any defensible model should have an area in the range of 0.5-1.

Table 16 shows the area of the ROC curves when the various models chosen using our learning sample are applied to data from our validation sample. With a ROC-curve area of 0.561, the backward stepwise model chosen with a significance level of 0.003 performs slightly less well than either the BMA or the tree method. The BMA models, both with and without a strict Occam's window perform best out of sample, and the recursive tree (pruned at a 0.005 significance level) has a ROC curve area that is similar to that of the backward stepwise method. While the BMA methods do perform marginally better, it is evident from these measures that a great deal of variation in our outcome variable is not explained by any of these methods.

Table 16: Out of sample predictive performance, BCS70 Validation sample

Method	ROC
Stepwise (0.003)	0.561
	(0.01)
BMA Strict	0.566
	(0.01)
BMA Symmetric	0.575
	(0.01)
Recursive Tree	0.562
	(0.01)

To explore further how well these modelling techniques perform out of sample, we first calculate, for the validation sample, three estimates of the predicted probability of having a high malaise score. The first estimate PPBMA1 averages

over the set of models that fell within a strict Occam's window, and the second estimate, PPBMA2 averages over the set of models that obtained when Occam's window was symmetric. Finally, the predicted probabilities for the recursive tree model, PPTREE, are assigned based on the frequency of high malaise within each terminal node. For example, a woman with one set of bottom quartile test scores and some evidence of childhood poverty would be assigned a predicted probability of high malaise of 0.380. This is because her characteristics would place her in node 5 of the estimated tree, which among the learning sample contained 297 women, 113 of whom had a high malaise score. Next, we apply BMA and recursive tree methods using the validation sample, but in addition to our initial set of explanatory variables, we add one of the predicted probability measures. In what follows, we present a subset of these models in order to explore the extent to which the different model selection techniques summarise our data and its relationship with adult malaise scores. Although we do not provide the results from all nine runs (three predicted probabilities times three estimation methods), what we do present is largely representative of our findings.

Table 17 presents the set of models chosen when we apply BMA techniques with a strict Occam's window to our validation sample and include PPBMA1 as an additional explanatory variable. Here we see that the set of defensible models is small, but that model uncertainty is high. The model that contains just PPBMA1 is retained, but it is the least likely of the chosen models. Nonetheless, PPBMA1 is retained as significant in all of the models, so there is very strong evidence that PPBMA1 belongs in the model. Childhood hyperactivity is also retained as significant in the two best models in Table 17, but the evidence for the inclusion of this measure in the model is weak. Neither of the other two factors that are retained in at least one of the models has positive evidence of a significant effect.

Hence, while the BMA with a strict window does not explain much of the variation in our outcome measure, it does summarise the information in the data set and its association with the outcome variable rather well.

Table 17: BMA, strict, predicted probability entered as an explanatory variable, estimated using BMA strict on validation sample, BCS70 data.

	1	2	3	4	5	Effect posterior probability (%)
PPBMA1	*	*	*	*	*	100
Any hyperactivity	*	*				76.5
Missing interest	*			*		47.7
High aggression			*			11.4
Posterior Model Probability	41.0	35.5	11.4	6.7	11.4	100

Table 18 presents results similar to those in Table 17, except, PPTREE is entered as an additional explanatory variable in place of PPBMA1. The set of models that obtain from this application of BMA is also small, and the set of variables that is retained is very similar to those presented in Table 17. The variable PPTREE is retained in three of the five defensible models and appears in the two most likely ones. Nonetheless, there is positive, but not strong, evidence for its effect. Similar to what we found in Table 17, childhood hyperactivity also appears, once again with weak (but nearly positive) evidence. Childhood hyperactivity provided competitive splits (although not one of the top three presented in Table 5) to nodes 1 and 3 in the learning sample, and using a symmetric Occam's window (but not a strict) was retained, but with a PEP of just 5.1%. Why it should emerge with such higher evidence in the test sample is not clear. No other retained variables have positive evidence, but interestingly, having two high aggression scores during childhood is also associated with having a high malaise score in the second "best" model. This is also similar to the competitive splits that emerged via the aggression variable in nodes 1 and 3 of the learning sample, and using a symmetric Occam's window (but not a strict) was retained with a PEP of 19.7% – perhaps because the predicted probability reduces the number of variables in each model, it is possible for this variable to be retained within a strict Occam's window because model complexity is reduced. Once again, why anxiety does not emerge as well, is not clear. Finally, housing tenure is retained in one of the models. Recall that housing tenure frequently offered a competitive split all along the construction of the tree in Figure 1 – usually split into those with at least one observation of public housing. In addition, like hyperactivity and aggression, the variable was retained in some BMA models using a symmetric Occam's window with a PEP of 6.7%. Although there is not positive evidence for an effect of either of these variables, the fact that they were seen as providing competitive splits in the original construction of the tree may merit further attention.

Table 18: Recursive trees predicted probability entered as an explanatory variable, variable, estimated using BMA strict using validation sample, BCS70 data

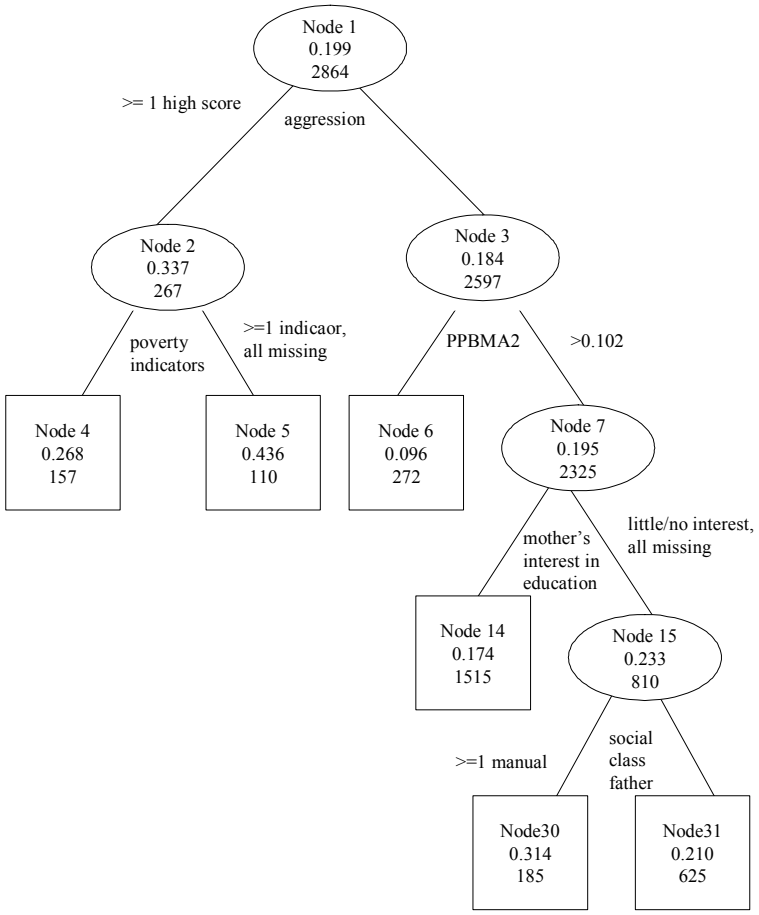
	1	2	3	4	5	Effect posterior probability (%)
PPTREE	*	*			*	88.6
Any hyperactivity	*		*	*		73.7
Some council housing			*	*		11.4
missing mother's interest			*			6.0
High aggression		*				21.5
	62.3	21.5	6.0	5.5	4.8	100

Figure 4 presents a tree built using the validation sample with PPBMA2 entered as an additional explanatory factor. The root node is not split via PPBMA2, but rather by the aggression variable – women with at least one high childhood aggression score form the higher risk subpopulation. An examination of the competitive splits for this node reveals that the PPBMA2 measure provides the second best split (PPBMA2>0.25 identifies the higher risk sample) and the impurities of both splits are extremely close. When root node is split via the aggression variable, the impurity measure is 0.493 and when it is split using PPBMA2, the impurity measure is 0.494. The difference in the relative risks is somewhat larger, however. When the root node is split via the aggression variable, the relative risk is 1.8 compared to 1.6 for the split via PPBMA2. Similarly, while node 2 is split via the poverty indicator summary (RR=1.6, 95 percent CI 1.0-2.7), once again PPBMA2 provides the next best competitive split (RR=1.6, 95 percent CI 1.0-2.7). Node 3 is split via PPBMA2, and the same variable once again provides a competitive split for node 7. Although PPBMA2 does not completely dominate the growth of the recursive tree in Figure 4, it does emerge as a consistently competitive factor. Nonetheless, why it is that aggression emerges as such a strong predictor in the test sample and not the learning sample (where aggression provides only the eighth best split) must simply result from sampling variability.

In this section, we examined the out of sample predictive performance of the Bayesian Model Averaging and Recursive Tree methods. Although out of sample predictive performance based on the ROC curve was disappointing, both methods appear to summarise the data sets fairly well. By introducing predicted probabilities as explanatory variables, we find that in a validation sample, our models do summarise the data fairly well. This is especially true when the predicted probabilities calculated from BMA methods are used as explanatory

variables in another BMA run. The BMA predicted probabilities do not dominate trees grown using the validation sample, but they do emerge as consistently competitive factors throughout. These findings suggest that our data do not explain the variation in having a high malaise score very well, but that the model selection techniques exploit the information that is there in a reasonably satisfactory way.

Figure 4: Tree drawn using the BCS70 test sample, PPBMA2 entered as an explanatory variable, pruned to significance level 0.005, showing Node, proportion with high malaise, and number of cases



7. Conclusion

It is easy to get lost in the thickets of exploratory data analysis, but we have illustrated many of the returns that can be gleaned through a variety of quite intensive approaches. Because our backgrounds are quantitative and rooted in regression analyses we may have failed to glean as much from the recursive trees as might someone from a more qualitative background.

However, we confess to some disappointment with the benefits of the automated recursive tree approach. There are two areas (at least) where we have concerns. Firstly, the binary nature of the splitting leads to quite complex combinations very rapidly, especially with sizeable data sets as here. Second, the impurity measure seems to favour splits that retain groups of fairly equal size and thus do not clearly identify some of the interactions that we discovered through more laborious processes. On the other hand, the option, however clumsy and limited, to control the splitting process is useful and can provide interesting insights into measures that are close competitors as in section 4.2.3. The software used, RTREE, does not permit examination of alternative splits within the same variable though.

There is considerable advantage though to the variables being considered for splitting one at a time. We have made use of the 'missings together' approach, suggested by Zhang and Singer (1999), which is very useful indeed when dealing with the situation we usually face, where the categories for childhood disadvantage on most antecedents are strictly ordered with the exception of those for whom all information is missing; the option to group these at either end of the hierarchy is thus useful. However, with a more complex antecedent, like our family structure, simply treating partially structured or partially ordered groups as nominal does not suffice. However, though not yet operationalised, it would be possible to include several further versions of our family structure variable that identified the partial hierarchies implicit in our coding of the dummy variables. Moreover, the option exists to include all sorts of collinear versions of the covariates in a recursive tree analysis. For example, our summary measures of childhood experience regarding each antecedent typically take observations at each of three childhood waves and combine them. There is nothing to prevent us including the component measures at each of the three childhood ages as well as the combination, in order to explore whether the combination actually does perform better or whether experience at a particular age matters more. The possibilities are limitless.

We have become enthusiasts for learning about model uncertainty and have found the posterior effect probabilities from BMA consistently illuminating. It

is helpful to have the results both for the strict and the symmetric Occam's window models and we have demonstrated ways of taking advantage of this above. The main difficulty with BMA, beyond the technical discussions of Raftery (1995) and Weakliem (1999) and their commentators, is the presentation of the results. In several cases above, we have resorted to reporting the odds ratios or parameter estimates from the best fitting model according to the BIC. The other indicative alternative is to present the model posterior probability weighted average of the non-zero parameter estimates, adjusted for the posterior effect probability. Neither is satisfactory for non-linear models, where only the model posterior probability weighted average of the predicted probabilities seems really defensible. But although this is perhaps a useful tool for out of sample prediction (see section 6 and section 5.3), it is not what can be usefully presented to policy makers.

We hope that this paper has also shown some of the rewards that be obtained from a lot of data exploration that is illuminated but not completely driven by insights from BMA and recursive trees. We have emphasised the exploratory nature of most of our analysis, although cautiously drawing attention to continuities across cohorts and across the life-course of one of our cohorts as helping to give confidence that we are discovering something real and valuable. We are unashamed empiricists in this respect and regard much so-called theory as being closer to adopting a disciplinary set of blinkers. It is no accident at all that we have made great effort to consider a wide range of childhood antecedents of malaise. This broad philosophy is better illustrated by our concerns with social exclusion or multiple disadvantages across a wide range of domains, including economic, social, welfare, demographic, and health outcomes (Hobcraft 1998 and Sigle-Rushton 2004) and to late adolescent and early adult experiences as intermediate elements in pathways to adult disadvantages (Hobcraft & Kiernan 2001, Hobcraft 2000, 2002, 2003, and 2004). Of course, we have no antipathy to genuine theoretical insights, nor to drawing on cumulative knowledge. But most theory in the social sciences perhaps needs disciplining as a strong Bayesian prior, so that the constraints imposed are made explicit.

Discovering complex interplays among different elements of childhood disadvantage that affect adult outcomes, such as malaise, is a difficult task and all too rarely pursued with vigour. We share and have been consistently inspired by Burt Singer's deep concerns about the general linear model as the only tool for data exploration, especially when no serious attempt is made to get beyond main effects. Moreover, Adrian Raftery's plea for better analysis and fitting when using the general linear model also resonate. In this paper we have tried to indicate some time-intensive, but ultimately rewarding, approaches that began by using BMA and recursive trees directly, but then moved beyond these simple

mechanical approaches to try to combine some of the philosophy underlying Burt Singer's approaches with varying degrees of automation. We trust the we have indicated too that the general linear model is not yet dead, but that main effects only models should be interred.

References

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks/Cole.

Hanley, J.A. (1989) 'Receiver operating characteristics (ROC) methodology: The state of the art'. *Critical Reviews in Diagnostic Imaging* 29: 307-335.

Hobcraft, J. (1998) *Intergenerational and Life-Course Transmission of Social Exclusion: Influences of Childhood Poverty, Family Disruption, and Contact with the Police*. CASEpaper 15, Centre for Analysis of Social Exclusion, London School of Economics

Hobcraft, J. (2000) *The Roles of Schooling and Educational Qualifications in the Emergence of Adult Social Exclusion*. CASEpaper 43, Centre for Analysis of Social Exclusion, London School of Economics

Hobcraft, J. (2002) 'Social Exclusion and the Generations'. In J. Hills, J. LeGgrand and D. Piachaud (Editors) *Understanding Social Exclusion.* Oxford: Oxford University Press

Hobcraft, J. (2003) *Continuity and Change in Pathways to Young Adult Disadvantage: Results from a British Birth Cohort*. CASEpaper 66, Centre for Analysis of Social Exclusion, London School of Economics.

Hobcraft, J. (2004) 'Parental, Childhood and Early Adult Legacies in the Emergence of Adult Social Exclusion: Evidence on What Matters from a British Cohort'. In P. Lindsay Chase-Lansdale, K. Kiernan and R. Friedman (Eds.) *Human Development across Lives and Generations: The Potential for Change*. Cambridge: Cambridge University Press.

Hobcraft, J. and K. Kiernan (2001) 'Childhood Poverty, Early Motherhood and Adult Social Exclusion'. *British Journal of Sociology* 52: 495-517.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky (1999) 'Bayesian Model Averaging: A Tutorial'. *Statistical Science* 14(4).

Madigan, D.M. and A.E. Raftery (1994) 'Model selection and accounting for model uncertainty in graphical models using Occam's Window'. *Journal of the American Statistical Association* 89: 1335-1346.

Raftery, A.E. (1995) 'Bayesian Model Selection in Social Research'. In P.V. Marsden (Ed.) *Sociological Methodology* 1996:111-195. (with Comments by: A. Gelman and D.B. Rubin; R.M. Hauser; and Rejoinder by A.E. Raftery) Oxford: Blackwell.

Richman, N. (1978) 'Depression in Mothers of Young Children'. *Journal of the Royal Society of Medicine* 71: 489-493.

Rutter, M., J. Tizard and P. Graham (1976) 'Isle of Wight Studies: 1964-1974'. *Psychological Medicine* 16: 689-700.

Rutter, M., J. Tizard and K. Whitmore (1970) *Education, Health and Behaviour*. London: Longman.

Sigle-Rushton, W. (2004) *Intergenerational and Life-Course Transmission of Social Exclusion in the 1970 British Cohort Study*. CASEpaper78, Centre for Analysis of Social Exclusion, London School of Economics.

Singer, B., C.D. Ryff, D. Carr and W. Magee (1998) 'Life Histories and Mental Health: A Person-Centered Strategy'. In A.E. Raftery (Ed.) *Sociological Methodology* 1998: 1-51.

Weakliem, D.L. (1999) 'A Critique of the Bayesian Information Criterion for Model Selection'. *Sociological Methods and Research* 27(3): 359-443 (with Comments by D. Firth and J. Kuha; A. Gelman and D.B. Rubin; A.E. Raftery; Y. Xie; and Reply by D. Weakliem).

Zhang, H.P. and M. Bracken (1995) 'Tree-based risk factor analysis of preterm delivery and small-for-gestational-age birth'. *American Journal of Epidemiology* 141: 70-78.

Zhang, H. and B. Singer (1999) *Recursive Partitioning in the Health Sciences*. New York: Springer-Verlag.