ISBA 2000, Proceedings, pp. 000–000
ISBA and Eurostat, 2001

ANALYSIS OF SURVEY DATA WITH BAYESIAN NETWORKS

MARCO RAMONI and PAOLA SEBASTIANI
Children's Hospital Informatics Program, Harvard Medical School, Boston, MA and
Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA

SUMMARY

This paper uses Bayesian modeling techniques to analyze a data set extracted from the British General Household survey. The models used are Bayesian networks, which provide a compact and easy to interpret knowledge representation formalism. An issue considered is the need for automated Bayesian modeling.

Keywords: AUTOMATED MODELING; BAYESIAN NETWORKS.

1. INTRODUCTION

The General Household Survey is a yearly survey, based on a sample of the general population resident in private households in Great Britain. The goal of this survey is to provide continuous information about the major social fields of population, housing, education, employment, health and income. Since the survey covers all these topics, it provides users with the opportunity to examine not only each topic separately, but also their mutual interplay. Summary of the statistical findings are published by the British Office of National Statistics, and are typically presented via contingency tables relating two or three variables at a time (Thomas *et al.*, 1998). We believe that this communication style fails one of the primary objective of the survey, which is to offer, to a non-technical audience, an up-to-date picture of living in Great Britain.

To avoid the fragmentation of the overall information, one should try to build a model that associates a large number of variables. To be a communication tool, however, such a model needs to be easily understandable, and easy to use. Understandability and usability being the requirements, we focus on Bayesian networks (Bns), which are known for providing a compact and easy-to-use representation of probabilistic information (Lauritzen, 1996). A Bn has two components: a directed acyclic graph and a probability distribution. Nodes in the directed acyclic graph represent stochastic variables and arcs represent directed stochastic dependencies among these variables. Thus, the graph provides a simple summary of the dependency structure relating the variables. The probability distribution for the network variables decomposes according to conditional independencies represented by the directed acyclic graph, and

each component — a conditional probability table — quantifies the remaining directed dependencies. The graph is an effective way to describe the overall dependency structure of a large number of variables, thus removing the limitation of examining the pair-wise associations of variables. Furthermore, one can easily investigate undirected relationships between the variables, as well as making prediction and explanation, by *querying* the network. This last task consists of computing the conditional probability distribution of one variable, given that values of some variables in the network are observed. Nowadays there are several efficient algorithms for probabilistic reasoning, which take advantage of the network decomposability (Castillo *et al.*, 1997), and commercial programs such as `Bayesware Discoverer` (available at `http://www.bayesware.com`) or `Hugin` (available at `http://www.hugin.com`) implement these algorithms.

The problem to be addressed, and we believe is one of the reasons for the slow gain of popularity these models deserve in the statistical community, is how to practically build a Bn from a large data set using Bayesian methods. This is considered in the next section. In Section 3 we analyze a data set extracted from the 1996 General Household Survey. The model selected is a network that displays a global picture of living in Britain and discovers interesting associations among variables describing the household wealth, the socio-economic status and the ethnic group of the head of the household.

2. OVERVIEW OF AUTOMATED LEARNING

A Bn is a directed acyclic graph (DAG) and a probability distribution. Nodes in the DAG represent stochastic variables $X = \{X_1, \ldots, X_v\}$ and directed arcs from *parent* nodes to a *child* node represent conditional dependencies. Each conditional dependence is quantified by the set of conditional distributions of the child variable given the configurations of the parent variables. Marginal and conditional independencies encoded by the DAG (Lauritzen, 1996), provides the factorization of the joint probability distribution

$$p(x_{1k}, ..., x_{vk}) = \prod_{i=1}^{v} p(x_{ik}|\pi_{ij}). \tag{1}$$

Here, $x_{1k}, ..., x_{vk}$ is a combination of values of the variables in $\mathcal{X}$. The variable Π_i denotes the parents of X_i so that x_{ik} and π_{ij} denote the events $X_i = x_{ik}$ and $\Pi_i = \pi_{ij}$, and π_{ij} is the combination of values of the parent variables Π_i in the event x_k.

The problem we consider next is learning a Bn from data. We describe this as an hypothesis testing problem. So, we suppose to have a set $\mathcal{M} = \{M_0, M_1, ..., M_g\}$ of Bns for the discrete random variables $X_1, ..., X_v$, and each Bn represents an hypothesis on the dependency structure relating the variables. We wish to choose one Bn after observing a sample of data $\mathcal{D} = \{x_{1k}, ..., x_{vk}\}$, for $k = 1, \ldots, n$. If $p(M_h)$ is the prior probability of M_h, one Bayesian solution to the model selection problem consists of choosing the network with maximum posterior probability $p(M_h|\mathcal{D}) \propto p(M_h)p(\mathcal{D}|M_h)$. The quantity $p(\mathcal{D}|M_h)$

is the *marginal likelihood* and it is computed by averaging out θ^h from the likelihood function $p(\mathcal{D}|\theta^h)$, where θ^h is the vector parameterizing the distribution of $X_1, ..., X_v$, conditional on M_h. Hence $p(\mathcal{D}|M_h) = \int p(\mathcal{D}|\theta^h)p(\theta^h)d\theta^h$, and $p(\theta^h)$ is the prior density of θ^h, conditional on M_h. The computation of the marginal likelihood requires the specification of a parameterization of each model M_h, and the elicitation of a prior distribution for Θ^h. The use of Hyper-Dirichlet priors for θ^h (Cowell *et al.*, 1999) with hyper-parameters α_{ijk} provides the following solution for $p(\mathcal{D}|M_h)$

$$p(\mathcal{D}|M_h) = \prod_{ijk} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \tag{2}$$

where n_{ijk} is the sample frequency of (x_{ik}, π_{ij}), and $n_{ij} = \sum_k n_{ijk}$ is the marginal frequency of π_{ij}. Note that we are making the assumption that there are no missing data in the sample and that sample cases are exchangeable.

In principle, given a set of Bns with prior probabilities and a complete data set, one can compute their posterior probability distribution and select the network with maximum posterior probability. However, as the number of variables in the data set increases, the size of the search space makes the task infeasible. Thus some heuristic is required to reduce the dimension of the search space. Fortunately, under some particular model priors, the posterior probability of each model M_h factorizes, thus allowing for local computations. This property can be fully exploited by imposing an "order" over the variables which transforms model selection into a sequence of locally exhaustive searches. We will describe a greedy search algorithm to reduce the complexity of each locally exhaustive search when the model space is still too large.

The marginal likelihood $p(\mathcal{D}|M_h)$ in (2) has a multiplicative form and this fact, together with the assumption that the network prior probabilities are *decomposable* (Heckerman *et al.*, 1995), provides a factorization of each model posterior probability. A prior for M_h is termed decomposable if it admits the factorization $p(M_h) = \prod_i p(M_h^i)$. Thus, decomposable priors are elicited by exploiting the modularity of a Bn, and are based on the assumption that the prior probability of a local structure M_h^i of a Bn is independent of the other parts M_h^j. This prior factorization, together with the factorization of the marginal likelihood, ensures that the posterior probability of the Bn M_h can be written as $p(M_h|\mathcal{D}) \propto \prod_i p(M_h^i|\mathcal{D})$ for $p(M_h^i|\mathcal{D}) \propto p(M_h^i)p(\mathcal{D}|M_h^i)$. Thus, the network posterior probabilities are decomposable and, in the comparison of models which differ only for the parent structure of a variable X_i, only the quantity $p(M_h^i|\mathcal{D})$ matters. Thus, for fixed i, the comparison of two local network structures M_i^j and M_i^l specifying different parents of X_i can be done by simply evaluating the product of the local Bayes factor $BF_{j,l} = p(D|M_i^j)/p(D|M_i^l)$, and the ratio $p(M_i^j)/p(M_i^l)$. This comparison is independent of any other associations among the variables $\{X_1, \cdots X_v\} \setminus X_i$.

Now, the problem is how to exploit this posterior probability decomposability. One approach, proposed by Cooper and Herskovitz (1992), is to restrict the model search to a subset of all possible networks which are consistent with an

order relation $\succ$ on the variables in $\mathcal{X} = \{X_1, ..., X_v\}$. The order relation $\succ$ is defined by $X_j \succ X_i$, if X_i cannot be parent of X_j in any network in $\mathcal{M}$. In other words, rather than exploring networks with arcs having all possible directions, this order limits the search to a subset of networks in which there are interesting directed associations. At first glance, the requirement for an order among the variables could appear to be a serious restriction on the applicability of this search strategy. However, we have used this approach in real applications, and one example is in Sebastiani *et al.* (2000). From a modeling point of view, specifying this order is equivalent to specifying the hypotheses to be tested and some careful screening of the variables in the data set may avoid the surprise of selecting a not very sensible model or explore uninteresting associations. In the next section, we will consider the problem of selecting an order among the variables in a real application.

This order imposed on the variables, induces a set of q_i possible parents for each variable X_i, say $P_i = \{X_{i1}, \ldots, X_{iq_i}\}$, and one way to proceed, which produces the sequence of locally exhaustive searches, is to implement an independent model selection for each variable X_i as follows. For each variable X_i, we define $\mathcal{M}^i$ to be the set of networks given by the possible combinations of parents. The set of networks can be displayed on a lattice with q_i levels, each level having models in which the associated DAG specifies k parents for X_i. The first level of the lattice contains the model M_0^i in which X_i has no parents. The second level contains the q_i models M_i^j in which X_{ij} alone is parent of X_i, and so on. For each variable X_i, the exhaustive search consists of evaluating the posterior probability of each model in the lattice so that the model with maximum posterior probability in the lattice can be identified. The global model is then found by linking together the local models for each variable X_i.

Although the order among the variables greatly reduces the dimension of the search space, this locally exhaustive search should explore a lattice of 2^{q_i} models for each variable X_i and, for large q_i, this may be infeasible. A further reduction is obtained via a *greedy search strategy*, also known as K2 algorithm, (Cooper and Herskovitz, 1992). The K2 algorithm is a bottom-up strategy, so that simpler models are evaluated first. For each variable X_i, rather than computing the posterior probability of all networks in the set $\mathcal{M}^i$, the search moves up in the lattice as long as in the level just explored there is at least one network with posterior probability higher than posterior probabilities of the networks in the precedent level. The search starts by evaluating the marginal likelihood $p(\mathcal{D}|M_0^i)$ of the Bn M_0^i, that specifies independence of X_i on the variables in P_i. The next step is the evaluation of the marginal likelihood $p(\mathcal{D}|M_j^i)$ of the k_i Bns M_j^i in which only one of the parents X_{ij} is selected from P_i. If the maximal marginal likelihood $p(\mathcal{D}|M_J^i)$, for some J, is greater than $p(\mathcal{D}|M_0^i)$, the parent X_{iJ} is accepted and the search proceeds in the same manner by trying to add one of the parents from $P_i \backslash X_{iJ}$ to the Bn selected. If none of the k_i Bns has a marginal likelihood greater than $p(\mathcal{D}|M_0^i)$, the model M_0^i is accepted and the search moves to some other variable. Clearly, this heuristic search has the limitation that it can end up in a local maximum, and one should

Variable	Description	State description
Region	Region of birth of Hoh	England, Scotland and Wales
Ad_fems	No of adult females	0, 1, ≥ 2
Ad_males	No of adult males	0, 1, ≥ 2
Children	No of children	0, 1, 2, 3, ≥ 4
Hoh_age	Age of Hoh	17-36; 36-50; 50-66; 66-98 (years)
Hoh_gend	Gender of Hoh	M, F
Accomod	Type of accommodation	Room, Flat, House, Other
Bedrms	No of bedrooms	1, 2, 3, ≥ 4
Ncars	No of cars	1, 2, 3, ≥ 4
Tenure	House status	Rent, Owned, Soc-Sector
Hoh_reslen	Length of residence	0-3; 3-9; 9-19; ≥ 19 (months)
Hoh_origin	Ethnic group of Hoh	Cauc., Black, Chin., Indian, Other
Hoh_status	Status of Hoh	Active, Inactive, Retired

Table 1. *Description of the variables. Hoh denotes the Head of the Household. Numbers of adult males, females and children refer to the household.*

be aware of this risk, when interpreting the model eventually selected. Other search strategies have been proposed to address this problem, see Cowell *et al.* (1999) and references thereafter.

3. ANALYSIS

The data set consists of 9033 British households. Since 1981, the household is defined as a single person or a group of people who have the address as their only or main residence and who share either one meal a day or the living accommodation. Data were selected from the British General Household Survey[1], which was conducted between April 1996 and March 1997 by the British Office of National Statistics in Great Britain.

In order to show the potential usefulness of our methodology, we selected 13 variables describing British households in terms of composition (Ad_fems, Ad_males, Children, Hoh_age, Hoh_gend), regions of United Kingdom (Region), one ethnicity indicator (Hoh_origin), one mobility indicator (Hoh_reslen) and economic indicators of the household (Accom, Bedrms, Ncars, Hoh_status, Tenure). Variables and their states are summarized in Table 1. This group of variables is fully observed in the data set extracted from the survey.

The modeling of the data was carried out with the program `Bayesware Discoverer` which implements the model search approach described before. The approach described in the previous section requires the variables to be discrete. Therefore, the first step of the analysis was to discretize continuous variables into 4 bins of approximately equal proportions. Before this step, variables having a skewed distribution were transformed in a logarithmic scale.

[1] Crown Copyright 1996. Used by permission of the Office for National Statistics.

Many integer-valued variables — as those indicating the number adult males or females in the household — were appropriately recoded and states observed with a low frequency were grouped into a unique state. We then choose an order among the variables to limit the space of models to be explored. After careful considerations, the order imposed on the variables was:

$$\mathsf{Region} \succ \mathsf{Hoh\_origin} \succ \mathsf{Hoh\_gend} \succ \mathsf{Ad\_fems} \succ \mathsf{Ad\_mal} \succ \mathsf{Hoh\_age} \succ \mathsf{Hoh\_status} \succ \mathsf{Children} \succ \mathsf{Tenure} \succ \mathsf{Hoh\_reslen} \succ \mathsf{Accomod} \succ \mathsf{Bedrms} \succ \mathsf{Ncars}.$$

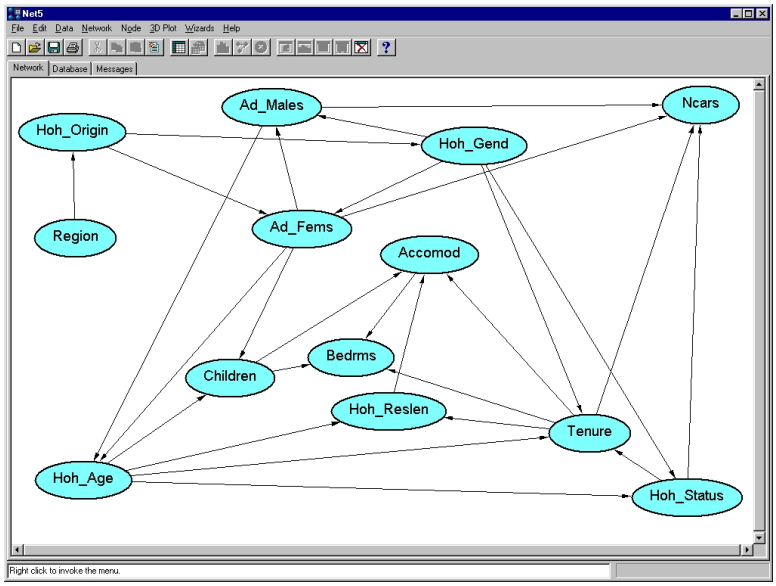
The order was chosen on the basis of the following considerations. Recall that the relation $X_j \succ X_i$ implies that X_i cannot be parent of X_j and, therefore, excludes all those models in which there is a directed arc from X_i to X_j. It is typically of interest to look at the distribution of household-describing variables conditional on the region where the household is (Thomas *et al.*, 1998). Therefore, we left Region as first node in the order, so that it was considered as parent of every other node. For the same reason, we left the ethnic group of the head of the household (Hoh_origin) as second variable. The gender of the head of the household (Hoh_gend) was chosen as third node in the order, as it is likely to affect the remaining ones. There is no particular preference between the order on the number of adult females and males in the household. Therefore, we choose the order Ad_fems $\succ$ Ad_males for pure convenience. The idea that the age of the head of the household (Hoh_age) affects the household wealth (which is described by the variables Hoh_status, Children, Tenure, Hoh_reslen, Accomod, Bedrms, Ncars) suggested to put the variable Hoh_age next in the order. The remaining order was chosen in a similar way, on the basis of possible cause-effect relationships between the remaining variables.

We used this order to build 4 models, using the K2 algorithm, uniform prior probabilities on the possible networks, and Hyper-Dirichlet model-parameter priors. We chose hyper-parameters $\alpha_{ijk} = \alpha/c_{ij}$, where c_{ij} is the number of cells in the conditional probability table quantifying the dependency of X_i on the parents Π_i. We chose four values for the *global precision* $\alpha = 1, 5, 10, 20$ to evaluate the effect of changing prior information on the model selected by looking at the differences between the networks topologies, and their different predictive capabilities. This last aspect was evaluated by computing the classification accuracy of the four networks. Full details of the analysis are in Sebastiani and Ramoni (2000) and led to select the network learned with $\alpha = 5$, which is depicted in Figure 1 and described in the next section.

5. RESULTS AND DISCUSSION

The network in Figure 1 shows important, directed dependencies and conditional independencies. The dependency of the ethnic group of heads of the households on the variable Region reveals a more cosmopolitan society in England than Wales and Scotland, with a larger proportion of Blacks and Indians as head of households. The ethnic group is parent of the gender of the head of the household and the number of adult females in the household, and separates both variables from Region, thus showing that differences in the household are

Figure 1. *The Bn induced for* $\alpha = 5$.



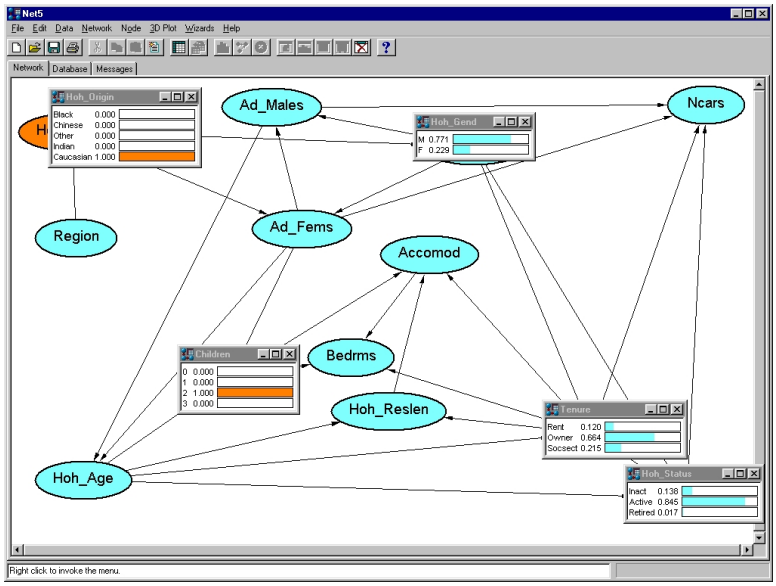
to be attributed to the ethnic group of the head of the household rather than differences in England, Wales or Scotland.

The working status of the head of the household (Hoh_status) is independent of the ethnic group given gender and age, and shows that young female heads of household are much more likely to be inactive than male heads of household (40% compared to 6% when the age group is 17–36). This difference is attenuated as the age of the head of the household increases. The dependency of the gender of the head of the household on the ethnic group implies that Blacks have the smallest probability of having a male head of the household (64%) while Indians have the largest probability (89%).

The age of the head of the household depends directly on the number of adult males and females, and shows that households with no females and two or more males are more likely to be headed by a young male while, on the other hand, households with no males and two or more females are headed by a mid age female. There appear to be more single households headed by an elder female than an elder male. Also the composition of the household changes in the ethnic groups: the most interesting fact is that Indians have the smallest probability of living in a household with no adult males (10%), while Blacks have the largest probability (32%).

The tenure status of the accommodation depends directly on the age, gender and status of the household head. On average, the largest proportion of British households are in owned accommodations (75%), when the age of the head of the household is between 36 and 66 years. Younger heads of household have

Figure 2. *An example of query with the Bn induced for* $\alpha = 5$.



a higher chance of living in rented accommodations (20%), while senior heads of household have a larger chance of living in accommodations provided by the social service (32%). These figures however change dramatically when the gender of the head of the household is taken into account. Young female heads of households have only 27% of being in an owned accommodation compared to 65% for males. This probability raises up to 52% when the household head is an elder female compared to 69% for elder males. Households are more likely to be in an accommodation provided by the social service when the head is an inactive female rather than an inactive male.

The number of bedrooms is directly affected by the number of children in the household, the type of accommodation and its tenure status. Households with two or more children are more likely to be in three bedroom flats or houses, although accommodation provided by the social service is slightly smaller than those rented or owned by the head of the household. Houses are more likely to have a larger number of bedrooms than flats: the most likely number of bedrooms of an owned house is three, compared to one in a flat. Interestingly, flats provided by the social sector are more likely to be one-bed flats, while rented and owned flats are most likely to be two-beds flats. The length of residence is directly dependent on the age of the head of the household and the tenure status of the accommodation and shows that the length of residence in rented accommodations or those provided by the social service is shorter than that in owned accommodations.

By querying the network, one may investigate other undirected associations

and discover that, for example, the typical Caucasian mid family with two children has 77% chances of being headed by a male who, with probability .57, is aged between 36 and 50 years. The probability that the head of the household is active is .84, and the probability that the household is in an owned house is .66. Results of these queries are displayed in Figure 2. These figures are slightly different if the head of the household is, for example, Black and the probability that the head of the household is male (given that there are two children in the household) is only .62 and the probability that he is active is .79. If the head of the household is Indian, then the probability that he is male is .90 and the probability that he is active is .88. On average, the ethnic group changes slightly the probability of the household being in an accommodation provided by the social service (26% for Blacks, 23% for Chinese, 20% Indians and 24% Caucasians). Similarly, black heads of household are more likely to be inactive than heads of household from different ethnic groups (16% Blacks, 10% Indians, 14% Caucasians and Chinese) and to be living in a less wealthy household, as shown by the larger probability of living in accommodations with a smaller number of bedrooms and of having a smaller number of cars. The overall picture is that of households headed by a Black to be less affluent than others, and this would be the conclusion one reaches if the gender of the head of the household is not taken into account. However, the dependency structure discovered shows that the gender of the head of the household and the number of adult females make all the other variables independent of the ethnic group. Thus, the model extracted suggests that differences in the household wealth are more likely caused by the different household composition, and in particular by the gender of the head of the household, rather than racial issues.

6. CONCLUSIONS

In this analysis, we focused on networks learned by using uniform model priors and sets of independent, symmetric Dirichlet distributions as prior distribution for each model parameters. The advantage of using these prior distributions is that they can be elicited by assigning the global prior precision α and this produces consistent model comparisons. However, symmetric Dirichlet distributions are known to be too invariant (Forster and Smith, 1998), so that they model in the same way different dependency structures. Although invariance seems to be a desirable property to produce objective results, on the other hand, one may wish to use a class of model parameter priors which encode different prior information. An interesting question is to devise a class of prior distributions which maintains the consistency of model comparisons, feasibility of computations, and provides the user with more modeling freedom. Furthermore, the analysis was carried out by discretizing continuous variables, thus raising the issue of the effect of the discretization on the analysis. We are currently working on the implementation of a more general learning algorithm which can work with both continuous and discrete variables.

One further issue is related to the publications of the results found with the method described here. A Bn is not just the DAG displaying the dependency

structure selected, conditional on the data. It is also a probability distribution, and as such, the best way to publish the results is by giving the whole Bn, thus allowing users to make their own queries. Given the increasing importance that the World Wide Web is assuming in nowadays communication system, publication of the network over the WWW offers a way to display results without giving direct access to the original data, so that data confidentiality is preserved.

7. ACKNOWLEDGEMENTS

This research was supported by Eurostat under contract EP29105. Material from the General Household Survey 1996 is Crown Copyright; has been made available by the Office for National Statistics through The Data Archive and has been used by permission. Neither the ONS nor The Data Archive bear any responsibility for the analysis or interpretation of the data reported here.

REFERENCES

Castillo, E., Gutierrez, J. M., and Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models*. Springer, New York, NY.

Cooper, G. F., and Herskovitz, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York, NY.

Forster, J. J., and Smith, P. W. F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response (with discussion). *Journal of the Royal Statistical Society, B*, **60**, 57–70.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combinations of knowledge and statistical data. *Machine Learning*, **20**, 197–243.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford, UK.

Sebastiani, P., and Ramoni, M. (2000). Analysis of survey data via Bayesian networks. Technical report. Available from the authors

Sebastiani, P., Ramoni, M., and Crea, A. (2000). Profiling customers from in-house data. *ACM SIGKDD Explorations*, **1**, 91–96.

Thomas, M., Walker, A., Wilmot, A., and Bennet, N. (1998). *Living in Britain: Results from the 1996 General Household Survey*. The Stationary Office, London, UK.