Modelling with Chain Event Graphs

Katarzyna Kobalczyk

Department of Statistics, University of Warwick, United Kingdom
katarzyna.kobalczyk@warwick.ac.uk

Abstract. Chain Event Graphs (CEGs) are a family of graphical statistical models derived from well-known probability trees. They form a generalisation of Bayesian Networks, providing an explicit representation of context-specific conditional dependencies within their topology. This report demonstrates on a real cohort study how CEGs enable us to depict various hypotheses about the data generation mechanisms. We argue that CEGs, in contrast to the standard framework of generalised linear models, can exhibit dependencies between multiple variables in a much more intuitive way. We also present how CEGs can be used for statistical inference with incomplete data set, identifying if the data are missing at random and extracting further conclusions from the patterns of missingness. We additionally discuss the problem of data discretisation and propose a method for supervised discretisation without leaving the framework of tree-based models.

Keywords: Chain Event Graphs, Missing data, Missing not at random, Data discretisation, Tree-based models

1 Introduction

There is a lot of potential in statistical models which are based on graphical representations of a problem providing an excellent way of explaining dependencies and interactions between multiple variables. They may be a useful approach to translating questions such as 'Do responses by novice drivers to surveys about accidents depend on age and skill?' or 'Is the missingness of responses among selected groups of respondents caused purely by chance?'

In particular, the Bayesian network was developed to give a way of representing complex relationships in real problems informed by data and expert judgements. However, the Bayesian network has its limitations. It may not necessarily be an easy, natural or even feasible method to use in specific problem descriptions. Tree diagrams are widely used in many disciplines, especially in computer science, probability and decision theory, to depict the structure of a process. They offer a practical tool with profound representational power and can be used as a framework for transforming an explanation of a process into a well-defined probability model. By adding colours to the edges and vertices of the event trees and transforming their graphs, we can derive a new, richer class of models called Chain Event Graphs (CEGs). This class of models can be used for analysing a wide range of problems in a coherent, comprehensive and transparent manner with the advantage that the conclusions can be easily read back to the decision-maker.

This report presents Chain Event Graphs in different modelling scenarios on the example of a particular study of learners and new drivers carried out by the British Department for Transport. The report begins with a description of the cohort study forming the basis of this analysis. Section 3 provides an introduction to graphical models, specifically the Bayesian networks. In section 4, we introduce Chain Event Graphs and simultaneously apply this framework in the context of modelling the accident liability of new drivers who took part in the study. In section 5 we show how CEGs can become useful in the case of incomplete data set containing missing values. We demonstrate how CEGs can help us identify different types of missingness (MCAR, MAR, MNAR) and how we can deduce directly from the graph that the MAR and MNAR assumptions are unlikely to hold. We also give an overview of how various hypotheses about the mechanisms behind missing responses in the study can be represented explicitly with CEGs. We further analyse the best fitting CEG and transform it to a more compact form allowing for a more transparent representation. Section 6 contains a discussion on transforming a continuously distributed random variable into its discrete analogue and the data pre-processing steps involved. We present how staged trees can be directly used to define alternative partitions of the sample space without losing on the parsimony of the models. We demonstrate how the proposed method can generate CEG models which score significantly better compared to CEG models based on data split into bins of equal frequencies. Finally, in section 7, we discuss the issues encountered during the research, conclusions from the project and possible improvements to the methods presented in the report.

2 Cohort II: A Study of Learner and New Drivers

'Cohort II' was a major six-year study funded by the British Department for Transport. It provides a picture of how 'cohorts' of learner drivers in Great Britain undertake driver training and testing, and of their subsequent experiences as new drivers. It followed the first large-scale investigation of new drivers, the Cohort I study in 1988–89. The aims of the study were:

- to investigate how people learn to drive, including the number of hours of tuition and practice, and to compare this to outcomes from the theory and practical driving tests;
- to assess the impact of changes to the testing regime, specifically the hazard perception test which was introduced during the period of study;
- to explore new drivers' experiences and attitudes to driving; and
- to identify their level of accident involvement over time.

Every three months, from November 2001 to August 2005, a cohort of 8,000 practical driving test candidates was sent postal questionnaires. Each person who passed the practical test and responded to the original survey on learning to drive (LTDQ) was subsequently followed with further postal questionnaires at 6, 12, 24 and 36 months after completing the original survey (DEQ1-4). Cohorts A to H received all four questionnaires. Subsequent cohorts received just the first three or the first two questionnaires because of the overall project duration. The sample initially comprised 42,851 learner drivers, however not all of these passed their practical tests to be involved in the subsequent Driver Experience Questionnaires. The sample of new drivers in Cohort II varied from over 10,000 at six months after the practical test to just fewer than 3,000 at three years after taking the test. Tables 1 and 2 show the number of samples and respondents to each survey by cohort and sex. The proportion of female and male respondents to the surveys was imbalanced with 26776 (62.5%) female and 16078 (37.5%) male respondents to the initial Learning to Drive Questionnaire (LTDQ). The proportion of men among the respondents was steadily decreasing in each subse-

Cohort	LTDQ	DEQ sample (LTDQ pass respondents)	DEQ1	DEQ2	DEQ3	DEQ4
A	3001	1247	696	506	342	323
B	3118	1445	787	576	447	399
C	3082	1423	723	568	399	389
D	3086	1491	781	586	439	370
E	2804	1230	636	444	324	293
F	2956	1173	630	441	303	277
G	2792	1316	582	453	291	295
H	2883	1385	712	553	397	412
I	2439	1221	642	469	339	
J	2776	1289	654	474	333	
K	2408	1180	552	428	281	
L	2448	1192	579	434	291	
M	2179	1038	506	328		
N	2246	1082	549	373		
O	2320	1124	501	405		
P	2316	1096	520	400		
	42854	19932	10050	7438	4186	2758

Table 1: Samples and number of respondents to LTDQ, DEQ1, DEQ2, DEQ3 and DEQ4 by cohort.

	Women		Men	
Questionnaire	n	%	n	%
LTDQ	26776.0	62.5	16078.0	37.5
DEQ1	6410.0	63.8	3640.0	36.0
DEQ2	4825.0	64.9	2613.0	34.8
DEQ3	2792.0	66.7	1394.0	32.8
DEQ4	1878.0	68.1	880.0	31.1

Table 2: Number of respondents and response rates to LTDQ, DEQ1, DEQ2, DEQ3 and DEQ4 by sex.

quent waves of the Driving Experience Questionnaire. Section 5.2 presents how CEGs can become useful in expressing possible hypothesis about the missingness of the responses.

One of the objective of the 'Cohort II' study was to determine the influence of a range of variables on young drivers' accident liability. The original report from the study [11] concludes with a multivariate regression model describing the relationship between key driver's characteristics and the number of reported accidents. It introduces a 'base model' on four variables: sex, age, experience and exposure. Later additional variables are added to the 'base model' and by the means of statistical significance testing the factors which influence accident liability are identified. In this modelling context, age is taken as the age at which the respondent passed the practical driving test. The experience measure is the number of years the respondent has been driving. Exposure is a composite measure that includes the annualised mileage driven within the reporting period plus 10 times the annualised number of days on which the driver has driven; derived artificially to optimise the fit of the models. In section 4.1 we present how Chain Event Graphs, in particular Ordinal Chain Event Graphs, can be used as an alternative to classical Generalised Linear Modelling, using the Cohort II study as an example. We argue that CEGs can represent dependencies between multiple variables in an easier to interpret way than what can usually be achieved through a careful examination of the interaction coefficients of a GLM model. Additionally, standard GLM approaches require the data to contain only complete observations. In section 5 we discuss how CEGs can provide a framework for including observations with missing values, identifying if the data are missing at random and extracting further conclusions from the patterns of missingness. Due to the decreasing number of respondents in every iteration of the survey, in section 4 we only consider the first wave of the Driver Experience Questionnaire - up to 6 months after passing the driving test (DEQ1). For modelling changes in probability distributions over time dynamic extensions of Chain Event Graphs have recently been developed [1,4], although their application is beyond the scope of this report.

3 Introduction to graphical models

A graphical model is a probabilistic model for which a graph expresses the conditional dependence structure between random variables. Probabilistic graphical models use a graph-based representation as the foundation for encoding a distribution over a multi-dimensional space and set of independence assumptions that hold within a given distribution. These representations sit at the intersection of statistics and computer science, relying on concepts from probability theory and graph algorithms. The graphical structure of these models make them a particularly useful tool for dealing with complex dependencies between many variables; therefore, they have been employed in many real-world applications, including recent advancements in machine learning for medical diagnosis or speech recognition.

3.1 Bayesian networks

Perhaps the most widely known probabilistic graphical model is a Bayesian network. The class of discrete Bayesian networks has a very close relationship to Chain Event Graphs; it forms a subclass of these models. Because many of the developments enabling inference of BNs motivated the development of CEGs, we briefly introduce them before discussing CEGs.

The core of the Bayesian network representation is a directed acyclic graph (DAG) $\mathcal{G}$, whose nodes are the random variables in our domain and whose edges intuitively correspond to the direct influence of one node on another. Several equivalent definitions of a Bayesian network have been offered.

Definition 1 (Beyesian netwrok). *Let $\boldsymbol{X} = \{X_1, X_2, \dots, X_m\}$ be a vector of random variables. Suppose its joint probability density function factorises as $\prod_{i=1}^{n} p(x_i|\boldsymbol{x}_{Q_i})$, where $Q_i \subseteq \{X_1, \dots, X_{i-1}\}$ (with the exception of $Q_1 = \emptyset$). A Bayesian network (BN) on $\boldsymbol{X}$ is a set of the $m-1$ conditional independence statements together with a DAG (Directed Acyclic Graph) $\mathcal{G} = (V, E)$ such that:*

1. *The set of vertices V of the DAG is given by $\{X_1, X_2, \dots, X_m\}$*
2. *A directed edge from X_i into X_s is in the edge set E of $\mathcal{G}$ if and only if X_i is a component of the vector $\boldsymbol{X}_{Q_j}$ where Q_j is the parent set of X_s for all $1 \leq i, j \leq m$.*

We make the above definition concrete with an example from the 'Cohort II' study. Hence, let X_a describe the age of the individual passing the driving test, X_s their sex, X_f their frequency of driving during the first 6 months after passing the driving test. We also let Y be a binary variable describing whether an individual reported at least one non-low speed accident during the first 6 months after passing the test. For simplicity, we define X_a to take values in three age intervals: 16-18, 19-21, 21-80. X_f can take one of the three values with the following encoding: 1 = Everyday, 2 = 4-6 days per week, 3 = Between once a week to to once a fortnight. Due to a small number of observations, in all models from this section, we exclude drivers who reported a frequency of driving below once a fortnight. Fig 1 presents three possible BNs estimated from the 'Cohort II' data set using the `bnlearn` R package [10]. The graphical representation allows us to directly read off the conditional independence assumptions from the DAGs in Fig 1. For example, in a) we have that $Y \perp\!\!\!\perp X_f, X_s \mid X_a$ and that the factorisation of the joint probability mass function of (X_a, X_s, X_f, Y) is given by:

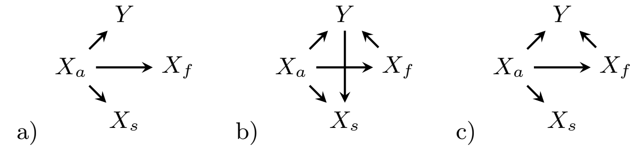


Fig. 1: Three plausible Bayesian Networks for the 'Cohort II' data set.

$$p(x_a, x_s, x_f, y) = p(x_a)p(x_t|x_a)p(x_f|x_a)p(y|x_a)$$

In b) we have that $X_s \perp\!\!\!\perp X_f \mid X_a, Y$ and the mass function with ordering (X_a, X_f, Y, X_s) factorises as:

$$p(x_a, x_s, x_f, y) = p(x_a)p(x_f|x_a)p(y|x_a, x_f)p(x_s|x_a, y)$$

Observe that the ordering (X_a, X_f, Y, X_s) in a) would give the same graph and the same factorisation of $p(x_a, x_s, x_f, y)$, but the ordering (X_a, X_s, X_f, Y) is not compatible with the BN from b). We cannot necessarily read a unique ordering from a DAG of a BN but all such orderings must correspond to the same factorisation. Finally in c) we have that $Y \perp\!\!\!\perp X_s \mid X_a$ but contrary to the DAG in a) $Y \not\perp\!\!\!\perp X_f \mid X_a$. A possible ordering of the variables in c) is (X_a, X_s, X_f, Y) with the joint probability mass function factorising as:

$$p(x_a, x_s, x_f, y) = p(x_a)p(x_s|x_a)p(x_f|x_a)p(y|x_a, x_f)$$

It is now clear how a Bayesian network representation can easily encode some conditional independence statements of the model. More conditional independence statements can be inferred from the graph using a d-separation property - the formal definition of which can be found in [5] - and through a more detailed examination of the conditional probability tables. However, in certain cases, the BN does not provide a rich enough structure to incorporate all information obtainable from the data set. This is the case, for example, when the conditional independence statements of the problem are asymmetric. In our example, we might want to encode the following assumptions that $Y \perp\!\!\!\perp X_s \mid X_f = \text{Everyday}$ but $Y \not\perp\!\!\!\perp X_s \mid X_f = \text{4-6 days per week}$. Asymmetric assumption of this kind cannot be represented simply by the directed edges between variables in the BN. This motivates the introduction of a new class of graphical models.

4 Chain Event Graphs

Chain Event Graphs are a family of graphical statistical models derived from well-known probability trees. Their interpretation is different from that of a BNs introduced in the previous section: in particular, in the event-tree framework, vertices always correspond to events in an underlying probability space and not to random variables. Every edge $e = (v, v')$ depicts the possibility of transitioning from the donating vertex v to the receiving vertex v' and can be labelled by a corresponding transition probability $\theta(e)$. This is in contrast to the Bayesian network where edges represent certain dependencies between the donating and receiving variables.

By introducing the concepts of *stages* and *positions* which group the vertices of a tree according to their associated conditional probabilities, an event tree becomes a *staged tree*. The tree structure is then tightened by transforming it into a new, more compact object called a *Chain Event Graph (CEG)*. In this section, we define these concepts and support them with illustrative examples based on the 'Cohort II' Study from Section 2. We adopt the notation as introduced in [5].

4.1 Tree based models on the example of 'Cohort II' study

A finite graph denoted as a pair $\mathcal{T} = (V, E)$ with vertex set V and edge set $E \subseteq V \times V$ is called a *tree*, if it is connected and has no cycles. We call the set of vertices $\text{pa}(v) = \{v' \mid \text{there is } (v', v) \in E\}$ the *parents* of $v \in V$ and $\text{ch}(v) = \{v' \in V \mid \text{there is } (v, v') \in E\}$ the set of *children* of $v \in V$. A vertex without parents is called a *root* of the tree, usually denoted $v_0 \in V$, and vertices without children are called *leaves*. Non-leaves, or inner vertices, are also called *situations*. We call a pair $\mathcal{F}(v) = (v, E(v))$ of a vertex $v \in V$ together with its emanating edges $E(v) = \{(v, v') \in E \mid v' \in \text{ch}(v)\}$ a *floret*.

Definition 2 (Probability tree). *Let $\mathcal{T} = (V, E)$ be an event tree with parameters $\theta(e) = \theta(v, v')$ associated to all edges $e = (v, v') \in E$. We call the vector $\boldsymbol{\theta}_v = (\theta(e)|e \in E(v))$ of all parameters associated to the same floret a vector of floret parameters. If all labels are strictly positive probabilities and the components of all floret parameter vectors sum to unity, so $\theta(e) \in (0, 1)$ and $\sum_{e \in E(v)} \theta(e) = 1$ for all $e \in E$ and all non-leaves $v \in V$, then the pair $(\mathcal{T}, \boldsymbol{\theta}_{\mathcal{T}})$ of graph $\mathcal{T}$ and vector of all labels $\boldsymbol{\theta}_{\mathcal{T}} = (\theta(e)|e \in E))$ is called a probability tree.*

Tree-based models are especially useful where there exists a natural ordering of the variables of interest. In our example we want to gain knowledge about the influence of X_a, X_s, X_f on the target variable Y. For this reason, we would want Y to be the last variable in the tree. The order of the other three variables is less obvious. We decide on the ordering $\boldsymbol{X} = (X_a, X_s, X_f, Y)$ which is supported by both BNs a) and c) in Fig 1. Fig 2 illustrates an event tree defined on $\boldsymbol{X}$. To every node v of this tree, we can attach a floret parameter vector $\boldsymbol{\theta}_v$ thought of as a vector of conditional transition probabilities emanating from that situation. For instance, $\boldsymbol{\theta}_{v_{25}} = (0.935, 0.065)$ means that male drivers who passed their test after the age of 21 and reported driving their car daily are estimated to have a 6.5% chance of being involved in at least one accident during their first 6 months after passing the test.

If two or more vectors of floret parameters take the same values, fewer parameters are needed to describe the full probability distribution over $\mathcal{T}$. This allows us to merge situations sharing the same floret parameter vectors into *stages* and obtain a new tree-based model - a *staged tree*.

Definition 3 (Staged Tree). *Let $(\mathcal{T}, \boldsymbol{\theta}_{\mathcal{T}})$ with graph $\mathcal{T} = (V, E)$ and labels $\boldsymbol{\theta}_{\mathcal{T}} = (\boldsymbol{\theta}_v|v \in V)$ be a probability tree. Whenever two floret parameter vectors are equal $\boldsymbol{\theta}_v = \boldsymbol{\theta}_w$ up to a permutation of their components we say that their vertices v and w are in the same stage, for $v, w \in V$. To every stage we assign one unique colour. A probability tree together with the partition of its vertices into stages is called a staged tree.*

In all of the graphs presented in this report colours are only applied to stages including more than one node. If nodes are not coloured, they are assumed to be in separate stages. Fig 3 shows the fitted staged tree obtained from the probability tree in Fig 2. Stages were found using the AHC algorithm, the details

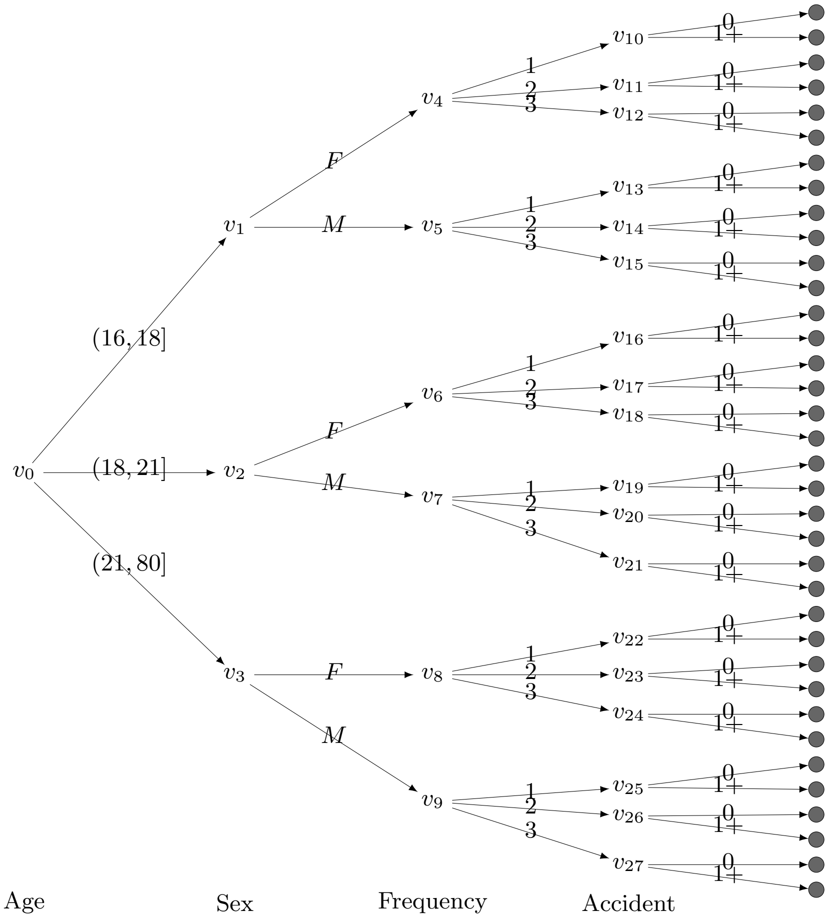


Fig. 2: Example of an event tree, $\mathcal{T}$, on four variables: age, sex, frequency of driving and accident involvement.

of which can be found in [7] and which is further discussed in section 6. In this example, colouring of the tree gives us the following set of stages:

$$u_0 = \{v_0\}, u_1 = \{v_1\}, u_2 = \{v_2\}, u_3 = \{v_3\},$$
$$u_4 = \{v_4, v_5\}, u_4 = \{v_6, v_7, v_8, v_9\}, u_6 = \{v_{10}, v_{14}, v_{19}\},$$
$$u_7 = \{v_{11}, v_{15}, v_{16}, v_{17}, v_{20}, v_{25}, v_{27}\}, u_8 = \{v_{12}, v_{22}, v_{26}\},$$
$$u_9 = \{v_{18}, v_{21}, v_{23}, v_{24}\}, u_{10} = \{v_{13}\}$$

Stage u_8 (coloured red) includes women aged 16-18 driving between once a fortnight to once a week, women above the age of 21 driving daily and males above the age of 21 driving 1 to 3 days per week. All these groups of respondents share the same floret paramter vector $\boldsymbol{\theta}_{u_8} = (0.981, 0.042)$ corresponding to an estimated probability of being involved in an accident of 4.2%.

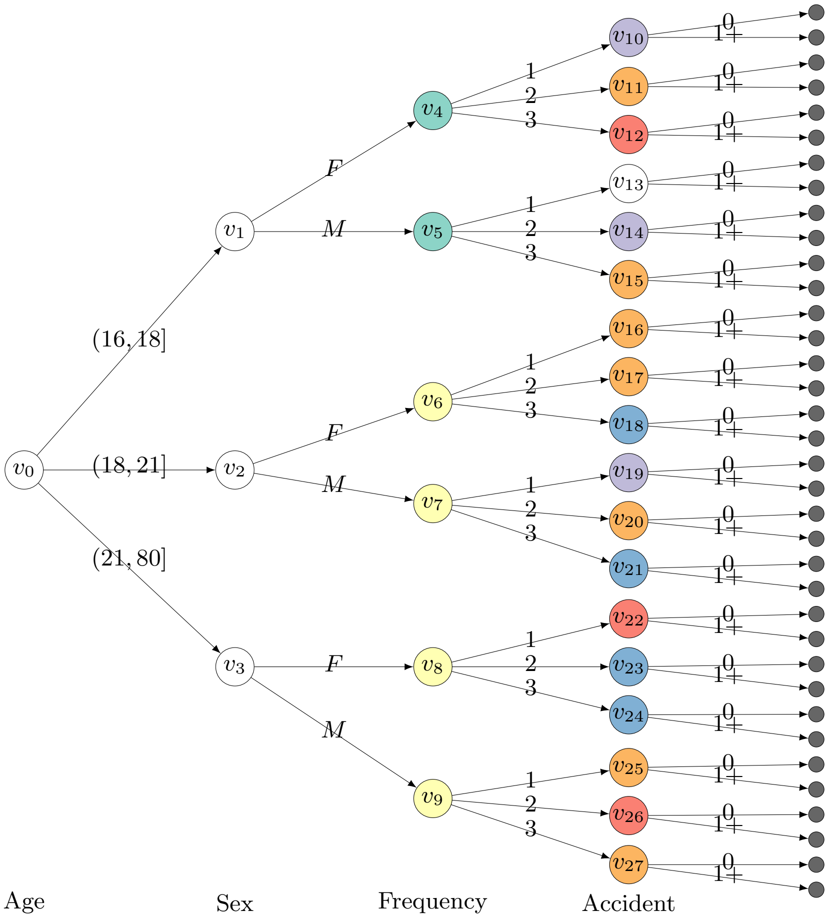


Fig. 3: Example of a staged event tree, $\mathcal{T}$, on four variables: age, sex, frequency of driving and accident involvement.

A coarser partition $W_{\mathcal{T}}$ of the vertices of a staged tree is given by the definition of *positions*. We say two vertices v, v' are in the same position, w, if their subtrees $(\mathcal{T}(v), \boldsymbol{\theta}_{\mathcal{T}(v)})$ and $(\mathcal{T}(v'), \boldsymbol{\theta}_{\mathcal{T}(v')})$ have the same topology such that there exeists a bijection between the edges of the two subtrees and the conditional probabilities associated with the corresponding edges are the same.
In the above example, v_4 and v_5 are in the same stage, but not in the same position. While the topologies of the subtrees $\mathcal{T}(v_4)$ and $\mathcal{T}(v_5)$ are in a one-to-one correspondence, the conditional probabilities downstream of v_4 and v_5 are not. However, all the vertices of u_6 are both in the same stage and position. In fact the same trivially holds for u_7, u_8, u_9 and u_{10}. In this example, there are no non-trivial positions. If, however, node v_{16} would be coloured in purple (i.e. belong to stage u_6), then nodes v_6 and v_7 could be merged into the same position; v_6 and v_7 are in the same stage, and the subtrees $\mathcal{T}(v_6)$ and $\mathcal{T}(v_7)$ together with the conditional probabilities associated with their edges would be in a one-to-one

correspondence. All leaves are always in the same position, which we denote as $w_\infty \in W_\mathcal{T}$ and call a *sink node*. The root is always the only element in position $w_0 \in W_\mathcal{T}$, also called the root. We can now construct the following graph:

Definition 4 (Chain Event Graph). *Let $(\mathcal{T}, \boldsymbol{\theta}_\mathcal{T})$ be a staged tree with graph $\mathcal{T} = (V, E)$. Denote the set of positions of this tree by $W_\mathcal{T}$. We build a new labelled graph $(C(\mathcal{T}), \boldsymbol{\theta}_\mathcal{T})$ as follows: $C(\mathcal{T}) = (W, F)$ is a graph with vertex set $W = W_\mathcal{T}$ given by the set of positions in the underlying staged tree. Every position inherits its colour from the staged tree. F is a set of possibly multiple edges between these vertices with the following properties. If there exist edges $e = (v, v')$, $e' = (w, w') \in E$ and v, w are in the same position, then there exist corresponding edges $f, f' \in F$. If also v', w' are in the same position, then $f = f'$. The labels $\theta(f)$ of edges $f \in F$ in the new graph are inherited from the corresponding edges $e \in E$ in the staged tree $(\mathcal{T}, \boldsymbol{\theta}_\mathcal{T})$. We call the labelled graph $(C(\mathcal{T}), \boldsymbol{\theta}_\mathcal{T})$ the Chain Event Graph (CEG) with underlying staged tree $(\mathcal{T}, \boldsymbol{\theta}_\mathcal{T})$.*

Following the above definition, we transform the staged tree in Fig 3 to obtain a CEG structure given in Fig 4. The graph from Fig 4 is a special type of a CEG - an *ordinal* CEG - as first defined in [2] - which is restricted to problems with a binary outcome variable occurring last in the tree. It provides a more easily interpretable graphical representation of the standard CEG by imposing a particular ordering on the positions of the graph. An ordinal CEG with respect to Y is a CEG where positions in each vertex subset associated with a variable X_i, are vertically aligned in descending order with respect to the predictive probability $\mathbb{P}(Y = 1|(C(\mathcal{T}), \boldsymbol{\theta}_\mathcal{T}))$.

This compact representation enables us to provide a plausible story of how different driver characteristics influence their risk of involvement in an accident, where the conclusions can be easily read off directly from the CEG. The predictive probabilities of an accident associated with the positions $w_{13}, w_{19}, w_{27}, w_{22}, w_{23}$ in the final layer of the CEG are: 13.6%, 9.6%, 6.5%, 4.2%, 1.9% respectively. The ordinal CEG allows us to immediately identify the groups of respondents with the highest accident liability: young men aged 16-18 driving every day. Similarly, the group with the second highest accident liability are men aged 16-18 driving 4-6 days per week, men aged 18-21 driving every day and women aged 16-18 driving every day. Another advantage of the ordinal CEG is that we can immediately observe that overall male respondents are associated with a higher risk of non-low speed accidents than female respondents and that the younger groups of respondents are associated with higher accident risk than the older groups. Similar and arguably more precise conclusions could be made by examining the coefficients of a generalised linear model, yet the CEG representation provides us with a framework for presenting these conclusions without having to introduce the quantitative measures.

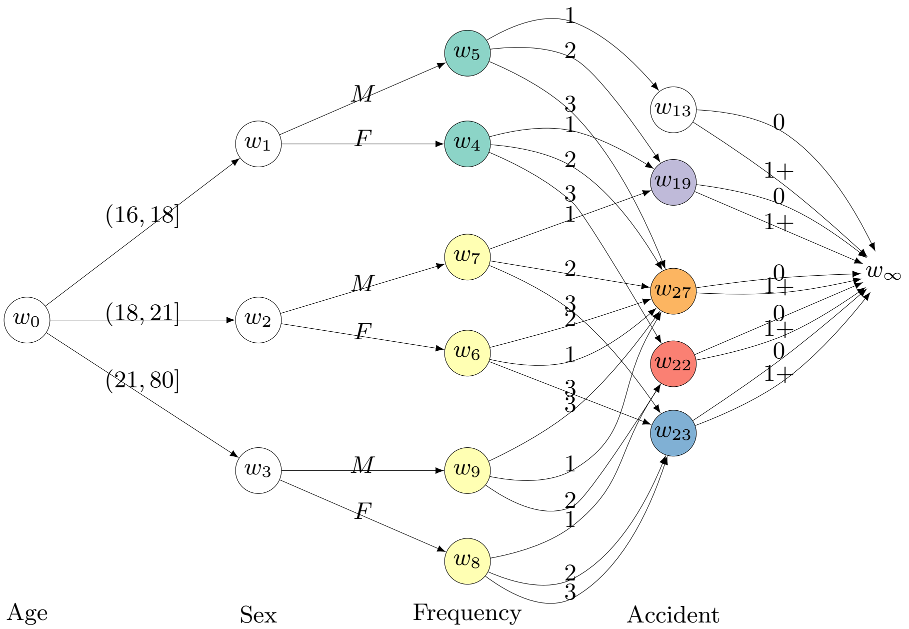


Fig. 4: CEG derived from the staged tree in Fig 3.

5 Chain Event Graphs and missing values

5.1 Missingness of the mileage

In the previous section, we constructed a CEG model which allows us to categorise the drivers into groups of increasing accident liability. We used three explanatory variables: age, sex and frequency. Instead of the frequency as a measure of exposure to road accidents, we may instead choose to rely on the reported mileage. However, out of the 9491 observations considered for this model 757 (8%) have a missing response to the question about the mileage. Those observations could simply be discarded, yet CEGs allow for not only including the incomplete observations into the model but also to reason about the patterns of the missingness. For a more detailed discussion on the CEG models of missingness refer to [2].

As previously, let X_a describe the age, X_s sex, and Y the accident involvement during the first 6 months after passing the test. Let also X_m describe the annualised mileage taking one of the three values coded as: 1 = above 12000, 2 = 4000 - 12000 miles, 3 = below 4000 miles. We further introduce a new variable R_m indicating whether X_m is missing or not. Fig 5 presents the most probable CEG on $\boldsymbol{X} = (X_a, X_s, R_m, X_m, Y)$ found by the AHC algorithm. The positions $w_{27}, w_{47}, w_{48}, w_{19}$ in the final layer of the CEG are ordered with respect to the predictive probability of accident involvement: 14.1%, 8.6%, 4.9%, 2.7%, respectively. When considering the potential impact of the missing data on the final findings, it is important to consider the underlying reasons why the data

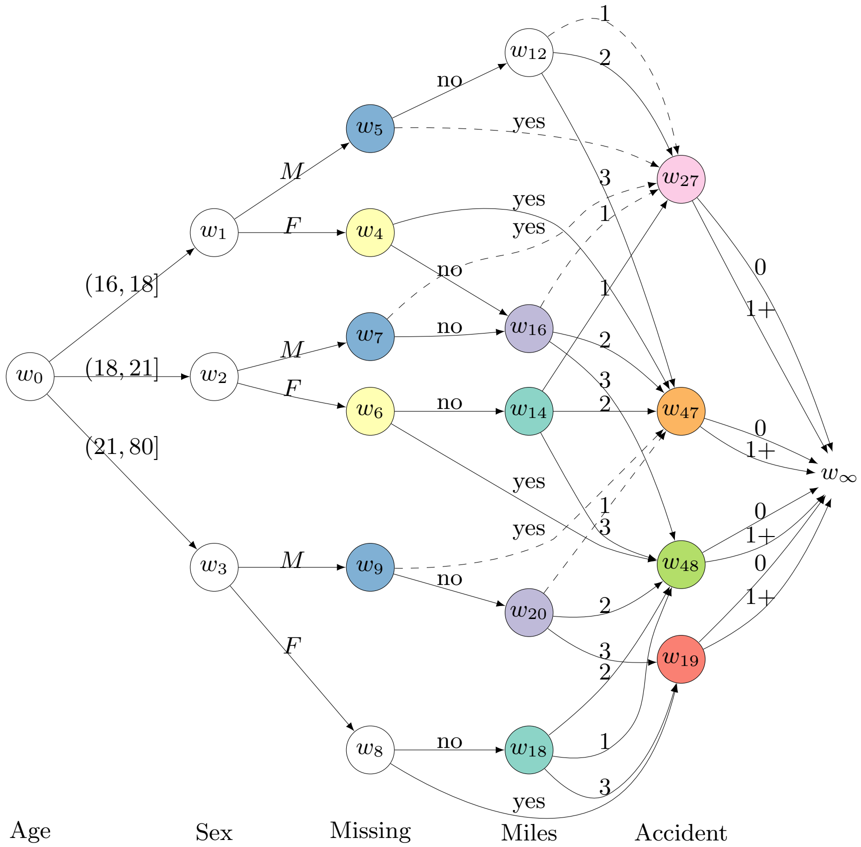


Fig. 5: CEG on five variables $\boldsymbol{X} = (X_a, X_s, R_m, X_m, Y)$. Dashed edges highlighting the coincidence of missing response for men with the highest mileage.

are missing. Missing data are typically grouped into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR); as proposed in [3].

If the probability of X_m missing is the same for all age-sex groups, then the data could be hypothesised to be MCAR. This would require that R_m is independent of X_s and X_a and we would expect all positions on the level of the R_m variable to be in the same stage. This is not the case as in the CEGs from Fig 5 we can distinguish 3 different stages for R_m. These correspond to the following groups of respondents: all men (blue stage), women aged 16-21 (yellow stage), and women above the age of 21 (white stage). Hence, the assumption that the data are MCAR is unlikely to hold. Among those three groups, it is the women aged 16-21 who are the most likely not to provide an estimate for the miles driven (11%), followed by women above the age of 21 (9%), while the non-response to the question on miles among men is estimated to be about the level of 3%. However, the data can be missing at random only conditionally on certain values of another variable. From the CEG in Fig 5 we can observe that all men end up in the same stage on the R_m level (blue). We can therefore deduce that:

$$R_m \perp\!\!\!\perp X_a \mid X_s = M$$

That is, conditionally on the respondent being a male, the missingness of the mileage is independent of their age. If we were to hypothesise that among certain age groups, say the youngest drivers aged 16-18, the missingness of the mileage is independent of sex - $R_m \perp\!\!\!\perp X_s \mid X_a = (16, 18]$, then the two edges emanating from the node w_1 which corresponds to the 16-18 age group should lead to the same stage. Again, this is not the case as the edges lead to two different stages (blue and yellow); the hypothesis that the missingness of the mileage is independent of sex is, therefore, unlikely to be true; even conditionally on a particular age group.

When data are missing at random (MAR) the missingness process is independent of the missing values given the observed values, so that

$$\mathbb{P}(R_m \mid X_s, X_a, X_m) = \mathbb{P}(R_m \mid X_s, X_a)$$

Under the assumption of MAR, the edges labelled "yes" emanating from the nodes corresponding to R_m should lead to the stages whose predictive probability of accident involvement is a weighted average of the predictive probabilities of accident involvement for all levels of mileage (1, 2, 3) in a given age-sex group. This again is not the case. Interestingly, we observe that for all men, the missing mileage coincides with the positions of high mileage (above 12000 mi). Yet for all women, the missing mileage leads to the same positions as very low (below 4000 mi) or average (4000 - 12000 mi) mileage. To determine whether this means that data are unlikely to be MAR, it is necessary to additionally calculate the weighted average of the probability of an accident and compare this with the true probability of an accident for the missing category given an age-sex group. On the example of male drivers aged 19-21, under the CEG model considered, the predictive probability of an accident is 14.1%, 8.6% and 4.9% given high, average and low mileage respectively. Hence, if the data are MAR a young man aged 19-21 for whom the mileage is missing should have an accident probability of $14.1 \times \frac{93}{865} + 8.6 \times \frac{349}{865} + 4.9 \times \frac{423}{865} = 7.4\%$ where among men aged 19-21 93 reported high mileage, 349 reported average mileage and 423 reported low mileage. However, we see that the edge describing the missingness of the mileage for men aged 19-21 leads to the same position as the edge for high mileage with a noticeably higher predictive probability of (14.1%).

5.2 Modelling study drop-outs with CEGs

As was already noted in section 2 one major obstacle to treat the 'Cohort II' as a longitudinal study is the dropping response rate within subsequent waves of the Driving Experience Questionnaire. In this section, we apply Chain Event Graphs to represent various hypotheses about the patterns of non-response to the subsequent waves of the survey. Hence we let X_s be the variable describing the gender of a respondent and X_1, X_2, X_3, X_4 be four binary variables indicating the response to surveys DEQ1, DEQ2, DEQ3 and DEQ4 respectively. Here, the only natural ordering of the variables is: $\boldsymbol{X} = (X_s, X_1, X_2, X_3, X_4)$. The corresponding event tree is a binary tree with five levels.

Representing possible models of non-reponse with CEGs

We now consider different assumptions about the patterns of non-response to the survey and express these hypothesis with adequate CEGs.

Example 1 (Independence). The simplest model one can think of is to assume that all the variables are independent of each other, i.e $\perp\!\!\!\perp_{i=0}^{4} X_i$. In terms of an $\boldsymbol{X}$-compatible staged tree representation of the model, we would draw an event tree with binary florets where all florets which lie along the same level are also in the same stage. The corresponding CEG in Figure 6 is an $\boldsymbol{X}$-compatible representation of the binary independence model. This scenario corresponds to the situation where the response data is missing completely at random (MCAR).

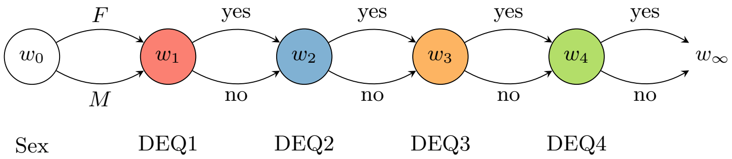


Fig. 6: $\boldsymbol{X}$-compatible CEG representation for the binary independence model from Example 1 (MCAR).

Example 2 (Biased Coin). Another simple hypothesis might be to model the responses to the questionnaires as a series of a possibly biased coin tosses, where the bias depends only on the sex of a respondent. That is, all women have the same probability of responding to every questionnaire and all men have the same probability of responding to every questionnaire. Because the 'coin tosses' are assumed to be independent, all florets corresponding to the sub-tree of the event tree rooted at position corresponding to a single sex are in the same stage. The corresponding CEG in Figure 7 is an $\boldsymbol{X}$-compatible representation of the biased coin flip model. This scenario corresponds to the situation where the response data is missing at random (MAR) but not completely at random.

Example 3 (Counting Responses). In this example we draw our attention to the number of currently missing responses to the survey. The hypothesis of this example is as follows: Given the wave of the survey, the response to the next questionnaire depends only on the number of already recorded responses. The missigness of responses is additionally independent of sex. If we label the edges of the $\boldsymbol{X}$-compatible staged tree with 1 and 0 to indicate the presence of response in every survey, then the vertices of the staged tree are in the same stage, if and only if they are on the same level of the tree and the sum from the edges of their root-to-vertex paths is equal. Fig 8 represents the CEG for this hypothesis. For instance, in this scenario, responding to the first survey and not responding to

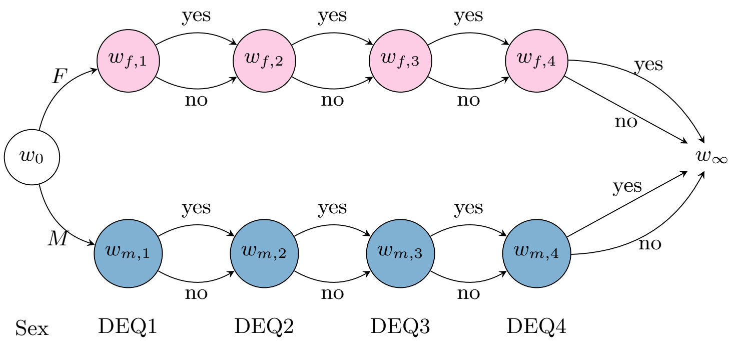


Fig. 7: $\boldsymbol{X}$-compatible CEG representation for the biased coin-flip model from Example 2 (MAR).

the second - path {1,0} - leads to the same position $w_{2,1}$ as not responding to the first survey, but responding to the second one - path {0, 1}. Similarly, all three paths {1, 1, 0}, {1, 0, 1} and {0, 1, 1} lead to the same position $w_{3,2}$ interpreted as: out of the first three surveys, response to exactly two was recorded.

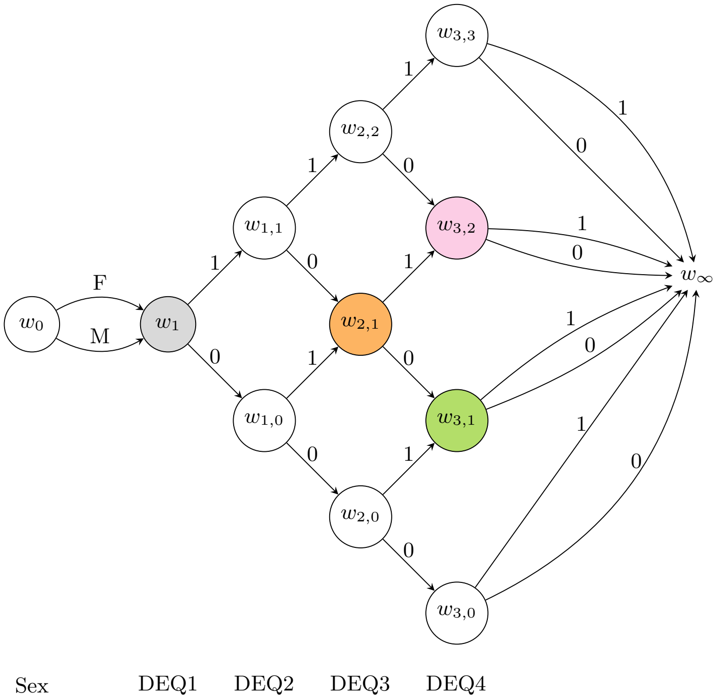


Fig. 8: $\boldsymbol{X}$-compatible CEG representation for the model from Example 3

Best-scoring CEG of the non-reponse

In the previous subsection, we presented several examples of how a modeller might approach the problem of missing responses and how these hypotheses might be expressed with a CEG. If these were the only competing models we could use a scoring criterion such as the AIC or the Bayes Factor as described in [5] for a Bayesian approach in model selection, to choose the best model out of the three. However, exploratory data analysis suggest that we should not expect any of the three examples presented to be an adequate representation of the problem. We, therefore, introduce another model which was found using the AHC algorithm. The corresponding CEG is presented in Fig 9. Clearly, the structure of this graph is less easy to interpret at first sight. However, the graph illustrates several asymmetries which allow us to draw a number of conclusions that are not obtainable from a regular BN.

1. The probability of response to the first survey is dependent on sex. The probability of response to the second survey is dependent on both the sex and the response to the previous survey.
2. Given that the pattern of responses to the first and second survey is $\{0, 1\}$, i.e. $X_1 = 0$ and $X_2 = 1$, the probability of responding to DEQ3 and DEQ4 is independent of sex. (The paths following edges $\{F, 0, 1\}$, $\{M, 0, 1\}$ are incident on position w_{12}).
3. Given that the pattern of responses to the first and second survey is $\{1, 0\}$, i.e. $X_1 = 1$ and $X_2 = 0$, the probability of responding to DEQ3 is the same for men and women (positions w_{13} and w_9 are in the same stage), but the probability of responding to DEQ4 differs for men and women (w_{13} and w_9 are in the same stage, but are not merged into one node).

By following the root to sink paths we can observe how the probability of responding to the next survey increases when the response to the previous survey was also recorded. However, due to the complexity of this graph, it is not as easy to examine, when presented in its full form. We, therefore, suggest a more compact representation by introducing a new variable $X_{1,2}$. taking values $\{00, 01, 10, 11\}$ which represents the responses to the first two waves of the survey. The new CEG on $\boldsymbol{X} = (X_s, X_{1,2}, X_3, X_4)$ is presented on Fig 10. With this graph, observations 2. and 3. are easier to spot. The union of blue and orange stages represents the event of answering one survey out of the first two. Given that this event has occurred, the response to the third survey is independent of sex. Comparing the probabilities associated with blue and orange stages we can observe that it is the response to the second survey, rather than the first, which is associated with higher chances of responding to the third survey (31% vs. 16-20%). Positions w_6 and w_{10} correspond to the event of answering both DEQ1 and DEQ2, yet the probability of this event happening is higher for female than for male respondents (33% vs. 24%). Given that response to the first two surveys was recorded, it is again the female respondents who are more likely to answer the third survey (60% vs. 55%). Paths $\{F,\ 11,\ 1\}$ and $\{M,\ 11,\ 1\}$ lead to position w_{18} which has the highest estimated probability of answering the final survey DEQ4 out of all positions in the final layer of the graph. This is perhaps not very surprising. An analogous observation can be made for the paths $\{F,$ 00, 0$\}$ and $\{M,\ 00,\ 0\}$ which lead to position w_{19} having the lowest probability of answering to DEQ4.

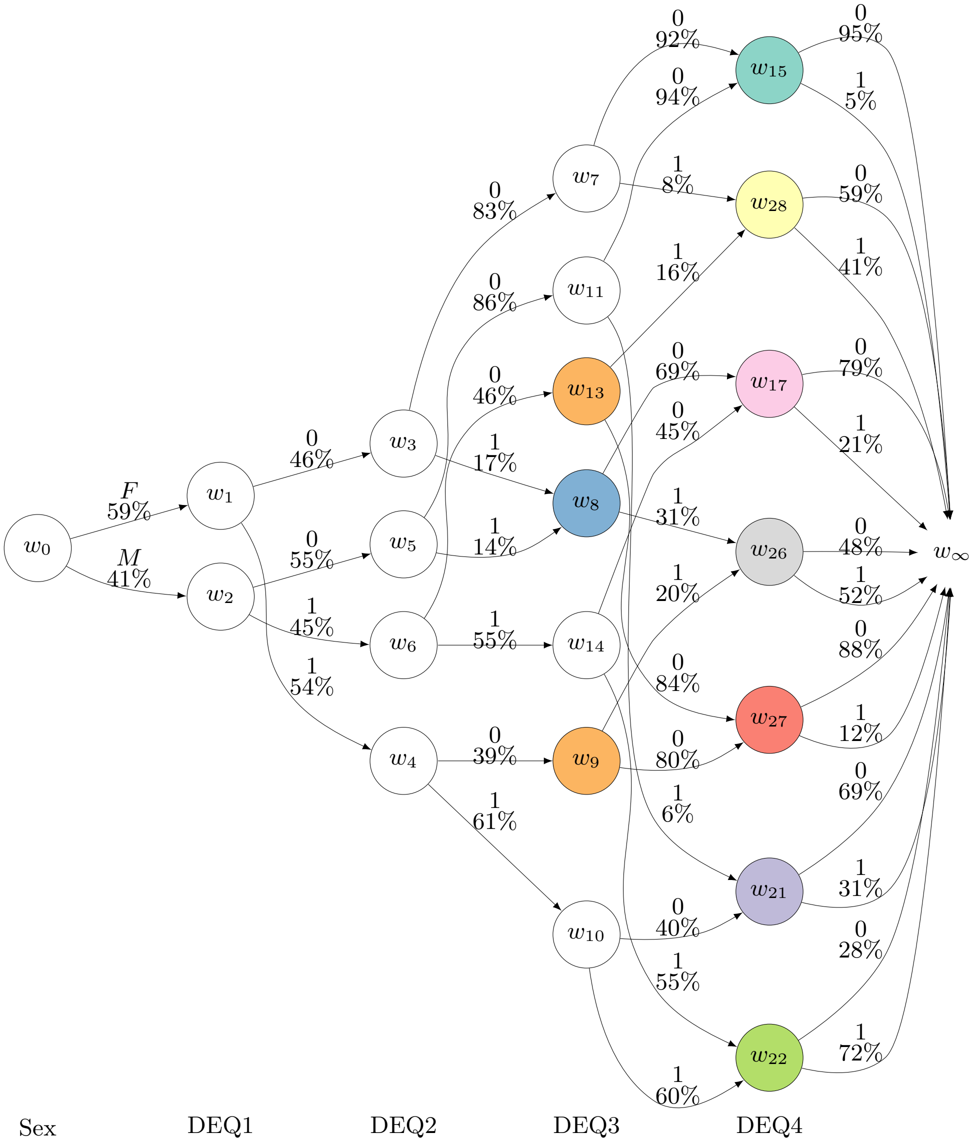


Fig. 9: The highest scoring CEG found by the AHC algorithm together with the transition probabilities.

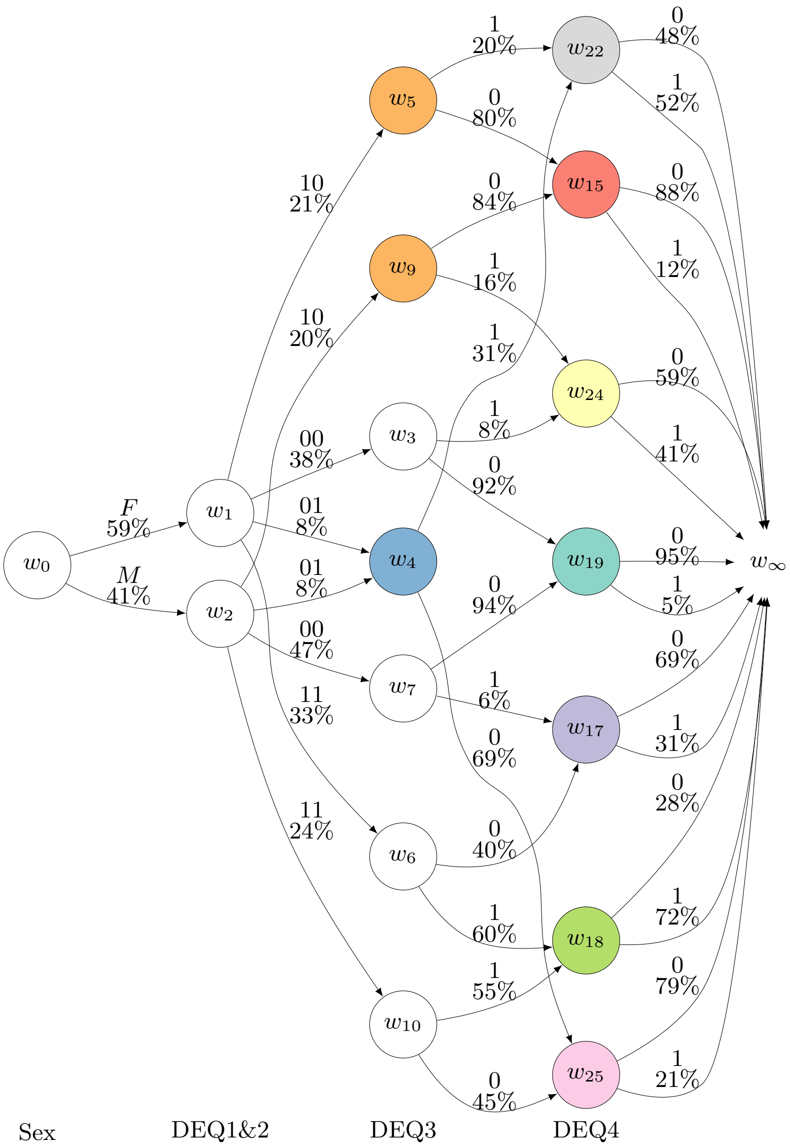


Fig. 10: Compact representation of the survey response model.

6 Discretising continuous variables for CEG models

6.1 Introduction to the problem

Chain Event Graphs are a class of graphical models suited for modelling discrete events. Hence, in all of the previous models, variables such as age or mileage - originally taking values on the continuous scale - had to be discretised into a fixed number of intervals. A natural question may arise about the optimality of this process. Whenever continuous data are discretised, there is always some amount of discretisation error associated with a certain loss of information. Typically data are discretised into partitions of K equal lengths/width (equal intervals) or $K\%$ of the total data (equal frequencies). All previous models adopted the latter approach of equal frequencies. In this section we discuss whether we can improve on our models by defining different discretisation rules, what is the optimal number of the discrete outcomes, what thresholds between the intervals on the continuous scale should be chosen, and what exactly is considered under the term *optimal*.

The equal-width and equal-frequencies approaches fall into the category of *unsupervised* discretisation methods, which are based on the intrinsic data distribution of an individual variable. As an alternative, we may consider *supervised* methods - an informative way of discretisation that takes into account the state of the target variable to inform and optimise the discretisation of each individual explanatory variable. There is a large number of supervised discretisation algorithms available, including the Fayyad & Irani's MDL method [6], CAIM algorithm [9], and many others that are based on entropy minimisation. A drawback of supervised discretisation algorithms is that the thresholds they produce are often physically meaningless. In addition, supervised algorithms may produce potentially spurious discretisation thresholds that fit to noise in the data rather than thresholds that increase the predictive power of a model.

In the context of the 'Cohort II' study, we want to discretise the two explanatory variables: age and mileage, with accident involvement being our target variable. Initial experiments in discretising these features involved metrics that are used in decision tree learning: Gini impurity and Information gain. Binning the variables according to these metrics often led to an increase in the final score of the CEG models; yet as already alluded, they provided thresholds that would not be considered as meaningful for a real-world application. An age threshold of 18.441 in determining the accident risk category is not necessarily the most sensible or practical.

6.2 Direct application of staged trees for the discretisation process

Motivated by the above discussion we suggest a method for discretising continuous variables that is based on the same mechanisms used for searching over the CEG model space.

To this end, let us assume that $\boldsymbol{X} = (X_1, \ldots, X_t, Y)$ and that X_t is a variable which needs to be discretised according to the target variable Y. Instead of treating X_t as a continuous variable, in the first pre-procssing step, X_t is quantised into a large number of possible outcomes $\Omega = \{\omega_1, \dots, \omega_N\}$. The granularity level of the initial quantisation will typically depend on the application context. Suppose we are given $K \in \mathbb{N}$ with $K \ll N$. The goal is to find a partition of Ω into K classes $C_1, \dots, C_K$ and define a new, 'coarser' random variable $\tilde{X}_t$ with the probability mass function:

$$\mathbb{P}(\tilde{X}_t = C_n) = \sum_{\omega_i \in C_n} \mathbb{P}(X_t = \omega_i).$$

In the case when the outcomes $\omega_1, \dots, \omega_N$ admit a natural ordering $\omega_1 \prec \dots \prec \omega_N$, we further require that every class consists of a series of consecutive outcomes. That is if $\omega_i, \omega_j \in C_n$ with $i < j$, then $\omega_k \in C_n$ for all $k = i+1, \dots, j-1$. We wish to find the best value of K and the corresponding partition of Ω into K classes such that the CEG defined on $\tilde{\boldsymbol{X}} = (X_1, \dots, \tilde{X}_t, Y)$ is the best scoring CEG according to some optimality criterion (e.g. BIC).

In our example, the first quantisation step is to treat age as a random variable with integer-valued categories: $16, 17, 18, \dots, 80$. This is the finest levels of granularity that we wish to accept in the final discretised form of age. We seek to merge some of these age groups together to obtain a more tractable number of final categories. Since age is a variable with a natural ordering of outcomes, we will additionally require that if say 20- and 23-year-olds belong to the same age group, then so are the respondents aged 21 and 22.

The Agglomerative Hierarchical Clustering (AHC) algorithm starts with the finest partition of a given event tree into stages, in which every node is in a separate stage. At each step, the algorithm then finds the two stages, which when merged, provide the highest improvement in the CEG score. The algorithm continues until the coarsest partition has been reached and then the structure with the highest overall score is selected. Because of the additive decomposition of the marginal likelihood of each CEG model, the score of the CEG on $\tilde{\boldsymbol{X}} = (X_1, \dots, \tilde{X}_t, Y)$ can be maximised level by level: finding the best score for each of its underlying variables independently. The proposed approach is to compare which outcomes of X_t land in the same stages of the best scoring model and on this basis transform X_t into the coarser variable $\tilde{X}_t$. However, when the algorithm searches for the best two stages to be merged, there are typically no restrictions assumed about which positions can end up in the same stages (other than that they correspond to the same underlying random variable). Hence, the algorithm may produce stages invalidating the ordering of $\omega_1 \prec \dots \prec \omega_K$. We propose two modifications which take this constrain into account. The first approach is a brute-force search over all possible partitions, the second is a modified version of the greedy search. Both approaches consider only the step of the full AHC algorithm that merges stages associated with the variable X_t.

Let $\mathcal{T}_0$ be the staged tree on $\boldsymbol{X} = (X_1, \dots, X_t, Y)$ with the finest partition of situations into stages, i.e every situation is in a single stage. Let X_t be the variable of interest having N possible ordered outcomes $\omega_1 \prec \dots \prec \omega_N$, where N is large. Suppose that in the staged tree $\mathcal{T}_0$, there are m nodes $v_1, \dots, v_m$ on the level of X_t. With every node v_i we associate a subtree $\mathcal{T}_0(v_i)$ rooted at this node. Every node v_i has N children: $(v_{i,j})_{j=1}^N$. Every child node $v_{i,j}$ is associated with the outcome ω_j of X_t and a set of preceding events leading to its parent v_i. Fig 11 shows an example tree on three variables $\boldsymbol{X} = (X_1, X_2, Y)$, with $t = 2$, $m = 2$, and $N = 10$ (numbers chosen to be relatively small for illustration purposes). Let P be a partition of the indexing set $\{1, \ldots, N\}$ into K subsets such that every subset is non-empty and consists of only consecutive numbers. If j, k belong to the same subset of the partition P, then in every subtree $\mathcal{T}_0(v_i)$, nodes $v_{i,j}$ and v_i, k are considered to be in the same stage. The goal is to find a partition P for which merging the nodes into stages according to that partition results in maximising a given model selection score.

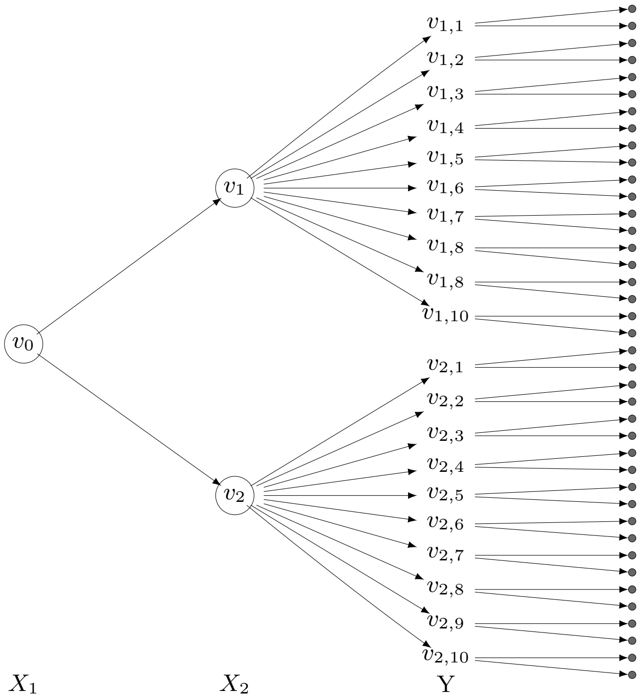


Fig. 11: Example event tree with $t = 2$ on $\boldsymbol{X} = (X_1, X_2, Y)$ where X_1 and Y are binary variables and X_2 has $N = 10$ possible outcomes. Positions $v_{1,1}, \ldots, v_{1,10}$ and $v_{2,1}, \ldots, v_{2,10}$ are in a one-to-one correspondence with outcomes $\omega_1, \ldots, \omega_{10}$ of X_2 admitting the natural ordering $\omega_1 \prec \ldots \prec \omega_{10}$.

Full search

Given a fixed size K of the final partition, the first approach is to consider a brute-force algorithm which checks every possible partition P. To find the best partition, the algorithm needs to search over $\binom{N-1}{K-1}$ possibilities, recording the score for each one, and return the highest scoring partition. This procedure is not computationally efficient, with a factorial time complexity.

Greedy algorithm for merging adjacent stages

The greedy algorithm starts with the finest partition. In every step, it iterates over the set $\{1, ..., N\}$ and finds two consecutive indices j and $j+1$ such that when in every subtree $\mathcal{T}_0(v_i)$ the nodes $v_{i,j}$ and $v_{i,j+1}$ are are merged into one stage, the model selection score is maximised. This procedure is repeated until no further improvement in the score can be obtained. An advantage of this approach is that the final size of the partition K does not need to be specified. The greedy algorithm has an $O(N)$ time complexity; significantly more efficient than the brute-force search.

6.3 Application and conclusions

Both the full search and greedy search methods have been implemented as an extension to the `stagedtrees` R package [8]. Staged trees have been fitted to discretise the age variable, optimising for accident involvement and adjusting for at most two other preceeding variables: sex and frequency of driving. The trees have been fitted on three variable vectors: $\boldsymbol{X}_1 = (X_a, Y)$, $\boldsymbol{X}_2 = (X_s, X_a, Y)$, $\boldsymbol{X}_3 = (X_f, X_s, X_a, Y)$. In the initial pre-processing step, age was split into $N = 54$ integer-valued categories. Table 3 shows the best scoring partitions for every vector of variables found with the greedy and full search methods together with the time it took to compute these values in a single-threaded process on a 2.6 GHz Intel Core i7 machine. BIC was chosen as a model selection criterion. The full search method was executed twice: once with the final number of bins K matching the size of the partition found by the greedy search K_g, and additionally for $K = K_g + 1$. The full-search method was implemented with the intention to compare it with the performance of the greedy algorithm. With BIC as the model selection criterion, stages with small number of observations are typically selected first for merging by the greedy rule. Hence, it was expected that greedy algorithm will lead to sub-optimal partitions. Yet, as can be observed from Table 3 and what was also confirmed with a series of test with different subsets of the data, partitions found by the greedy search are identical or almost identical as those found with the full-search method. In case of a discrepancy in the solutions found by the greedy and the full-search algorithms, the difference in the model selections scores between the two was negligible. Let us also observe the implications of including more variables in $\boldsymbol{X}$ preceding X_a - the age. More

	greedy		full K_g		full $K_g + 1$	
	partition	time	partition	time	partition	time
$\boldsymbol{X}_1$	$(16, 18, 19, 33)$	0.19s	$(16, 18, 20, 33)$	1.07min	$(16, 18, 19, 22, 33)$	15.8min
$\boldsymbol{X}_2$	$(16, 19, 33)$	0.27s	$(16, 19, 33)$	9.1s	$(16, 18, 19, 33)$	2.83min
$\boldsymbol{X}_3$	$(16, 33)$	1.07s	$(16, 28)$	1.62s	$(16, 20, 33)$	46.72s

Table 3: Partitions of age into stages and their computational times using the greedy and full-search methods. Numbers from the partition vectors correspond to the lower bound of each age group; e.g. partition vector $(16, 18, 19, 33)$ generates 4 age groups: 16-17, 18, 19-32, 33+.

variables appearing before X_a in the event tree lead to partitions with smaller final number of bins K_g. This is an expected consequence of the following:

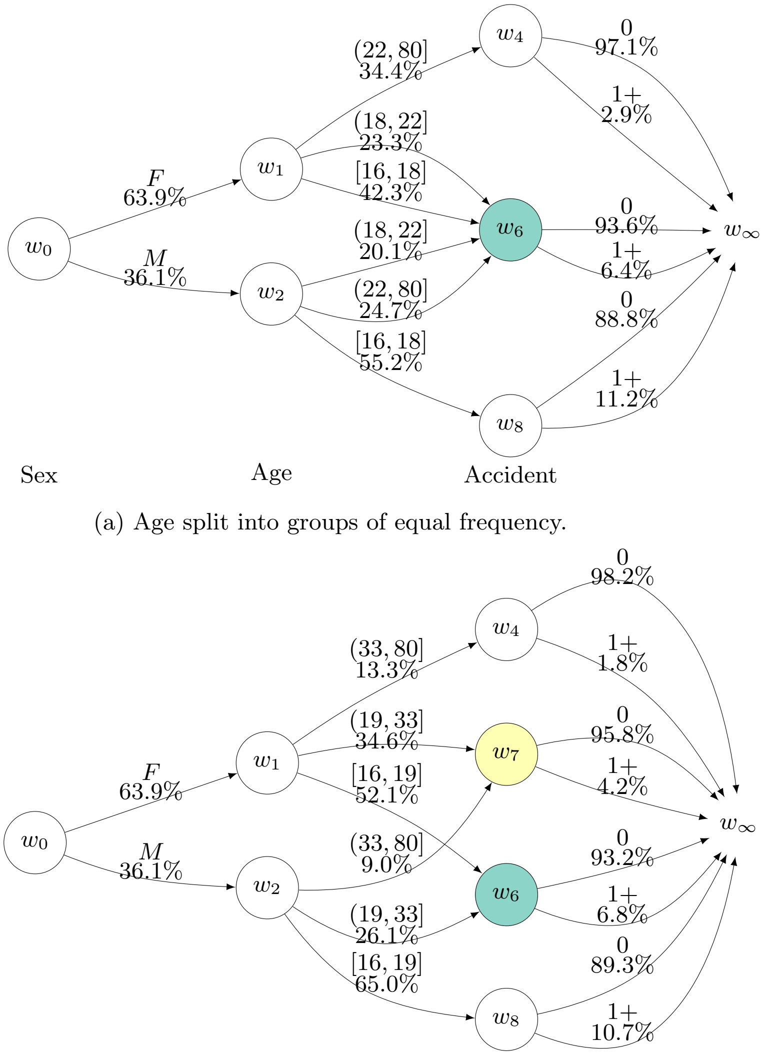
1. Preceding variables describe the data sufficiently-well; e.g.: when the frequency of driving and the sex of a respondent is known, the probability of an accident across different age groups does not vary as much as it does when only the sex of a respondent is known.
2. Deeper trees result in smaller sample sizes in the nodes considered for merging. Hence, there is less evidence for the need of more stages.

Having derived the optimised partitions for age, we can construct CEG models for each of $\tilde{\boldsymbol{X}}_1$, $\tilde{\boldsymbol{X}}_2$ and $\tilde{\boldsymbol{X}}_3$, replacing X_a with its discretised form $\tilde{X}_a$, according to the best scoring partitions. To compare the efficacy of our proposed method we also construct competing CEGs with age split into groups of equal frequencies. Table 4 shows the comparison of the final BIC and AIC scores of the models and the number of their degrees of freedom. Models were fitted with the backward hill-climbing algorithm using the `stagedtrees` package - an equivalent of the AHC method with BIC as the model selection score. [8]. CEG models fitted using the newly derived age groups scored consistently better than their equal-frequency analogues, which proves the viability of our proposed approach. Figure 12 shows the most interesting case, namely the two CEGs on $\tilde{\boldsymbol{X}}_2 = (X_s, \tilde{X}_a, Y)$; defining the thresholds between age groups according to the best scoring partition leads to one more stage on the level of accident involvement. By adjusting the thresholds of age for both sex and accident involvement, the CEG in Fig 12b has better capabilities in capturing the differences in accident involvement between female and male drivers compared to the CEG in Fig 12a. In Fig 12a all 16- to 22-year-old female and all 18- to 80-year-old male respondents have a 6.4% estimated probability of accident involvement. CEG from Fig 12b shows a more refined categorisation: women aged 16-19 and men aged 19-33 having a 6.8% probability of an accident, and women aged 19-33 and men aged 33-80 with a 4.2% probability of an accident.

	equal-frequencies			optimised partition		
	BIC	AIC	df	BIC	AIC	df
$\tilde{\boldsymbol{X}}_1$	28183.34	28140.39	6	27320.61	27277.66	6
$\tilde{\boldsymbol{X}}_2$	36806.24	36748.97	8	34633.25	34568.83	9
$\tilde{\boldsymbol{X}}_3$	48423.52	48351.94	10	42427.73	42356.14	10

Table 4: Comparison of CEG models with the age variable split into groups of equal frequency and groups according to the best scoring partition as in Table 3.

While testing the proposed above methods, another interesting observation was made. If the greedy partitioning algorithm is run on $\boldsymbol{X}_1$ for the subsets of the data with female respondents and with male respondents separately, the best scoring partition for women is given by: 16-17, 18-32 and 33-80, and for men by: 16-17, 18, 19-80. Compare this to the best scoring partition found by the greedy algorithm with the full data: 16-17, 18, 19-32, 33-80 - exactly the intersection of female and male partitions.



(a) Age split into groups of equal frequency.

(b) Age split into groupxs defined by the best scoring partition.

Fig. 12: Comparison of CEGs on $\boldsymbol{X}_2$ with different thresholds between the categories of age.

7 Discussion and further research

One of the aims of the ‘Cohort II’ study was to investigate the impact of changes to the testing regime, specifically the introduction of the hazard perception test. For this specific purpose, the framework of casual CEGs as discussed in [5] might prove very useful. For instance, we may add to our model from section 4 an extra variable describing whether a respondent took the hazard perception test and measure its total effect on accident involvement. However, before any casual inference can be done, first a reliable partition of nodes into stages should be selected, which is not always a straightforward task. All of the currently available algorithms for fitting staged event trees rely on greedy heuristics; after fitting staged trees with the hazard perception variable, the final CEG models led to ambiguous conclusions contradicting the findings from the original report of ‘Cohort II’ study. However, it was later discovered that these spurious observations were caused by the order in which certain stages were merged. Greedy heuristics led to merges between stages that resulted in biased conditional probability estimates of accident involvement. This observation brings us to the following conclusion - greedy approaches should always be applied with caution, and that findings derived from CEG models fitted with greedy algorithms should additionally be supported with appropriate hypothesis testing or at least compared against the raw conditional probability tables before presenting back to a client.

In regards to the methods from section 6, we note that while the presented results may look promising, these should mainly be treated as a proof of concept and further work needs to be done to improve on the robustness of these methods. In particular, with deeper trees, it is not unusual that the final subtrees considered may contain many unobserved categories, for which special treatment is needed. A possibility might be to use a Bayesian approach and apply a prior distribution over the event tree before the partitioning algorithm is run. Applying a non-informative prior may additionally act as a smoothing parameter reducing the impact of noise in the data. Another aspect worth considering is the implications of including additional variables in the event tree that succeed the variable chosen for discretisation. Different partitions of the sample space for a variable situated in the middle of an event tree may influence the groupings of downstream nodes into stages and positions; hence, result in more complex or more simplified structures of the final CEGs. In this case, the final CEG score cannot be maximised independently for each of the variables, leading to a much more complex problem that can quickly become intractable.

Methods and implementation

In the initial stages of the project, trees were fitted with the `stagedtrees` R package [8]. The package implements a number of non-Bayesian greedy search algorithms, in particular backward hill-climbing method which is an analogue to the Bayesian AHC search. It also allows for applying Laplacian smoothing which gives equivalent results to introducing a non-informative prior distribution over the edges of the event trees. In more advanced stages of the project, I additionally relied on the Python code developed by A. Shenvi as presented in [12]. It uses the Agglomerative hierarchical clustering (AHC) algorithm with a Bayes factor score for model selection. Partitions of the nodes into stages found by the AHC method coincided with those previously found with the `stagedtrees` package. Due to the specific requirements of my research, I extended both the R and Python implementations for fitting staged trees and CEGs with further functionalities. Contributions to the software include:

– Adding new methods to the `stagedtrees` R package for fitting staged trees with an ordinal variable as discussed in section 6
– Extending the Python code with additional parameters for plotting the event trees, staged trees and CEGs.
– Implementing a pipeline for exporting the DOT graphs generated with `PyDotPlus` Python package to documents typeset in LaTeXand PGF/TikZ.

The code used throughout this project together with the documentation is publicly available on my GitHub.

Learning outcomes

This research project was an invaluable learning experience allowing me to gain a hands-on experience with statistical modelling and Bayesian inference. I began the project with no prior university-level education on Bayesian statistics or graphical models. The book on Chain Event Graphs by Prof. J.Q. Smith [5] provided an excellent introduction to Bayesian networks and Bayesian inference on discrete statistical models, including the prior-to-posterior conjugate analyses and approaches to model selection. The project introduced me to a wide range of topics outside the scope of the standard degree; moreover, it inspired my choices for the future university modules.

Acknowledgements

I would like to thank my supervisor, Prof. Jane Hutton, for her support throughout the project. Our meetings always resulted in stimulating discussions leaving me with new interesting ideas to think about and motivating me to continue the work. I would also like to thank Aditi Shenvi for sharing her work and experiences with Chain Event Graphs.

References

1. L. M. Barclay, R. A. Collazo, J. Q. Smith, P. A. Thwaites, and A. E. Nicholson. The dynamic chain event graph. *Electronic Journal of Statistics*, 9(2):2130 – 2169, 2015.
2. L. M. Barclay, J. L. Hutton, and J. Q. Smith. Chain Event Graphs for Informed Missingness. *Bayesian Analysis*, 9(1):53 – 76, 2014.
3. W. M. Campion and D. Rubin. Multiple imputation for nonresponse in surveys. *Journal of Marketing Research*, 26:485, 1989.
4. R. Collazo and J. Smith. *The Dynamic Chain Event Graph*. PhD thesis, University of Warwick, Department of Statistics, 07 2017.
5. R. A. Collazo, C. Goergen, and J. Q. Smith. *Chain Event Graphs*. CRC Press, 1 edition, 2017.
6. U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI*, 1993.
7. G. Freeman and J. Smith. Bayesian MAP model selection of chain event graphs. *Journal of Multivariate Analysis*, 102(7):1152–1165, 2011.
8. M. L. Gherardo Varando, Federico Carli and E. Riccomagno. *stagedtrees: staged event tree probability models*. R package version 2.2.0.
9. L. Kurgan and K. Cios. Caim discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16(2):145–153, 2004.
10. R. N. Marco Scutari, Tomi Silander. *bnlearn: Bayesian network structure learning, parameter learning and inference*. R package version 3.3.0.
11. W. P., T. S., G. G., and J. E. *Cohort II: a Study of Learner and Novice Drivers, 2001-2005*, volume 1 of *Road Safety Research Report*. Department for Transport, England, 2008.
12. A. Shenvi and J. Q. Smith. Constructing a chain event graph from a staged tree. In *PGM*, 2020.