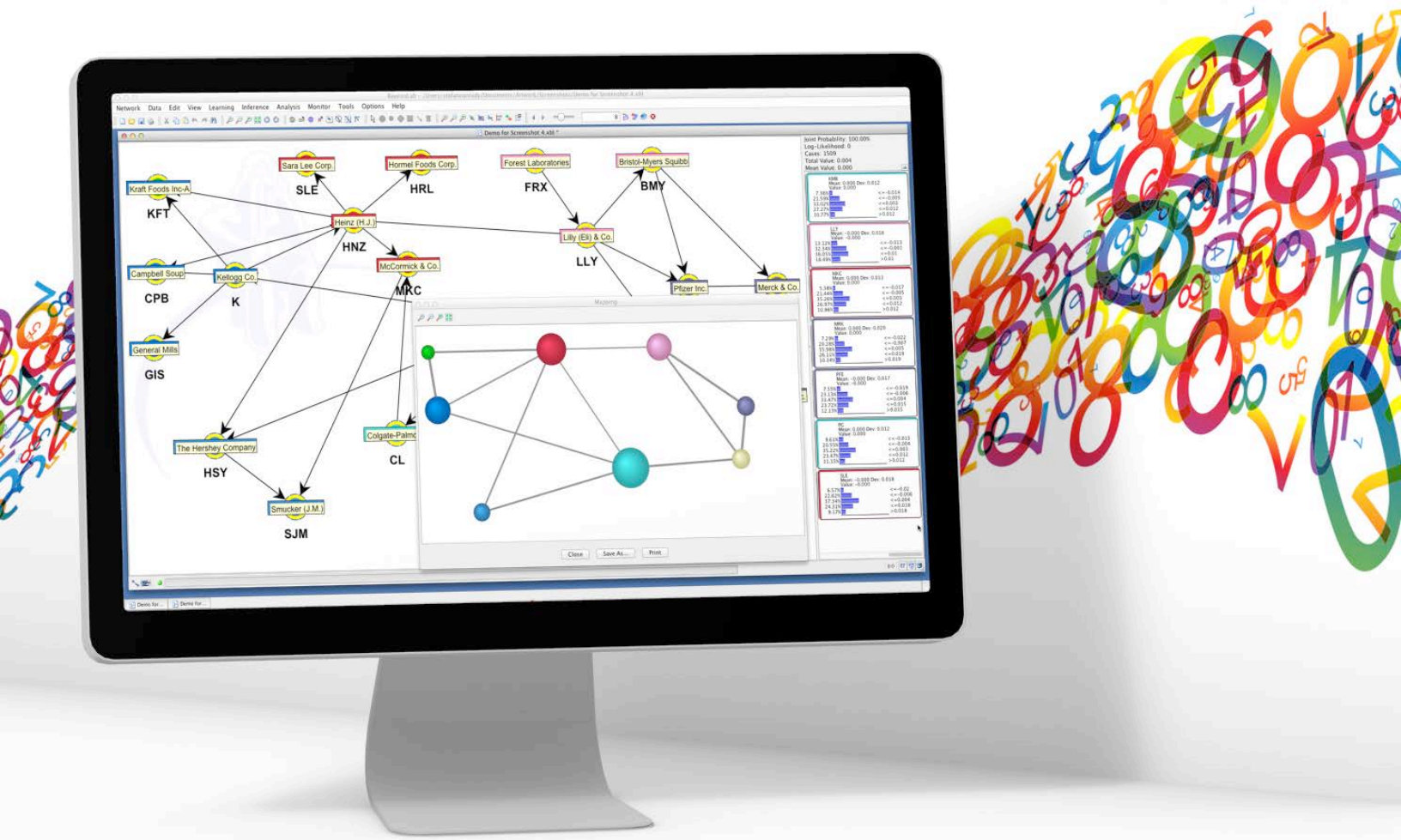


Vehicle Size, Weight, and Injury Risk

High-Dimensional Modeling and Causal Inference with Bayesian Networks

Stefan Conrady, stefan.conrady@bayesia.us
Dr. Lionel Jouffe, jouffe@bayesia.com

June 20, 2013

BAYESIA
Your Decision Partner

Table of Contents

Introduction

Objective

This paper's intent is to illustrate how Bayesian networks and BayesiaLab can help overcome certain limitations of traditional statistical methods in high-dimensional problem domains. We consider the vehicle safety discussion in the recent Final Rule,[1] issued by the Environmental Protection Agency (EPA) and the National Highway Traffic Safety Administration (NHTSA) on future CAFE[2] standards, as an ideal topic for our demonstration purposes.

Although this paper is meant to focus on technique as opposed to the subject matter itself, our findings will inevitably generate new insights. However, it is not our intention to challenge the judgement of the EPA/NHTSA Final Rule. Rather, we plan to take an independent look at the overall problem domain while considering the rationale presented in the Final Rule. Instead of merely replicating the existing analyses with different tools, we will draw upon a broader set of variables and use alternative methods to create a complementary view of some aspects of this problem domain. Extending beyond the traditional parametric methods employed in the EPA/NHTSA studies, we want to show how Bayesian networks can provide a powerful framework for forecasting the impact of regulatory intervention. Ultimately, we wish to use Bayesian networks for reasoning about consequences of actions not yet taken.

Admittedly, we will restate a number of the original research questions in order to better suit our expository requirements. Even though a macro view of this domain was required by EPA/NHTSA, i.e. societal costs and benefits, we believe that we can employ Bayesian networks particularly well for understanding high-dimensional dynamics at the micro level. Consequently, we examine this domain at a "higher resolution" by using additional accident attributes and finer measurement scales. Primarily for explanatory clarity, we also restrict our study to more narrowly defined contexts, i.e. vehicle-to-vehicle collisions, as opposed to all motor vehicle accidents. We also need to emphasize that all of our considerations exclusively relate to vehicle

[1] When referencing the EPA/NHTSA Final Rule, we refer to the version of the document signed on August 28, 2012, which was submitted to the Federal Register. However, when referring to the overall rationale presented in the Final Rule, we implicitly include all supporting studies that informed the Final Rule.

[2] The Corporate Average Fuel Economy (CAFE) are regulations in the United States, first enacted by the U.S. Congress in 1975, and intended to improve the average fuel economy of passenger cars and light trucks.

safety. We do not address any of the environmental justifications given in the EPA/NHTSA Final Rule. In that sense, we only focus on a small portion of the overall problem domain.

This paper is meant to portray a prototypical research workflow, presenting an alternating sequence of questions and answers. As part of this discourse, we gradually introduce a number of Bayesian network-specific concepts, but, each time, we only cross the proverbial bridge when we come to it. In the beginning chapters, we strive to provide a large amount of detail, including step-by-step instructions with many screenshots for using BayesiaLab. As we progress through this study, in later chapters, we try omitting some technicalities in favor of presenting the bigger picture of Bayesian networks as powerful reasoning framework.

We hope that readers can follow along by replicating the entire workflow on their own computers. For this purpose, a free, fully-functional evaluation version of BayesiaLab, valid for 30 days, can be requested via info@bayesia.us. Also, the entire preprocessed source data set is available for download from our server: www.bayesia.us/white_papers/data/NASS_data.csv.

Background

In October 2012, the Environmental Protection Agency (EPA) and the National Highway Traffic Safety Administration (NHTSA) issued the Final Rule, "2017 and Later Model Year Light-Duty Vehicle Greenhouse Gas Emissions and Corporate Average Fuel Economy Standards."

One of the most important concerns in the Final Rule was its potential impact on vehicle safety. This should not be surprising as it is a commonly held notion that larger and heavier vehicles, which are less fuel-efficient, are generally safer in accidents.

This belief is supported by the principle of conservation of linear momentum and Newton's well-known laws of motion. In collisions of two objects of different mass, the deceleration force acting on the heavier object is smaller.

Secondly, larger vehicles typically have longer crumple zones that extend the time over which the velocity change occurs, thus reducing the deceleration. Vehicle manufacturers and independent organizations have observed this many times in crash tests under controlled laboratory conditions.

It is also known that vehicle size and weight[3] are key factors for fuel economy. More specifically, the energy required to propel a vehicle over any given distance is a linear function of the vehicle's frontal area and mass. Thus, a reduction in mass directly translates into a reduced energy requirement, i.e. lower fuel consumption

Therefore, at least in theory, a conflict of objectives arises between vehicle safety and fuel economy. The question is, what does the real world look like? Are smaller, lighter cars really putting passengers at substantially greater risk of injury or death? One could hypothesize that so many other factors influence the probability and severity of injuries, including highly advanced restraint systems, that vehicle size may ultimately not determine life or death.

Given that the government, both at the state and the federal level, has collected records regarding hundreds of thousands of accidents over decades, one would imagine that modern data analysis can produce an in-depth understanding of injury risk in real-world vehicle crashes.

This is precisely what EPA and NHTSA did in order to estimate the societal costs and benefits of the proposed new CAFE rule. In fact, a large portion of the 1994-page Final Rule[4] is devoted to discussing vehicle safety. Based on their technical and statistical analyses, they conclude that there is a safety-neutral compliance path with the new CAFE standards that *includes* mass reduction.

General Considerations

To provide motivation and context for our proposed workflow, we will briefly discuss a number of initial thoughts regarding the EPA/NHTSA Final Rule. As an introduction to the technical discussion, we will first bring up a number of general considerations about the problem domain that will influence our approach.

1. Active Versus Passive Safety

The EPA/NHTSA studies have used "fatalities by estimated vehicle miles travelled (VMT)" as the principal dependent variable. This measure thus reflects all contributing as well as mitigating factors with regard to fatality risk. This includes human characteristics and behavior, environmental conditions, and vehicle characteristics and behavior (e.g. small passenger car with ABS and ESP). In fact, the fatality risk is a function of *one's own* attributes as well as the attributes of *any other participant* in the accident. In order to model the impact of vehicle weight reduction at the society-level, one would naturally have to take all of the above into account.

---

[3] "Weight" and "mass" are used interchangeably throughout this paper.

[4] We refer to the version of the document signed on August 28, 2012, which was submitted to the Federal Register. Page numbers refer to this version only.

As opposed to a society-level analysis, we are approaching this domain more narrowly by looking at the risk of injury only as a function of vehicle characteristics and accident attributes. We believe that this approach helps isolating vehicle crashworthiness, i.e. a vehicle's *passive safety performance*, as opposed to performing a joint analysis of crash propensity *and* crashworthiness. This implies that we omit the potential relevance of vehicle attributes and occupant characteristics with regard to *preventing* an accident, i.e. *active safety performance*. It would be quite reasonable to include the role of vehicle weight in the context of active safety. For instance, the braking distance of a vehicle is, among other things, a function of vehicle mass. Similarly, occupant characteristics most certainly affect the probability of accidents, with younger drivers being a well-known high-risk group.

As a result of drivers' characteristics and vehicles' behavior, at least a portion of victims (and their vehicles) "self-select" themselves through their actions to "participate" in an accident. Speaking in epidemiological terms, our study may thus be subject to a self-selection bias. This would indeed be an issue that would have to be addressed for society-level inference. However, this potential self-selection bias should not interfere with our demonstration of the workflow while exclusively focusing on passive safety performance.

2. Dependent Variable

The EPA/NHTSA studies use a binary response variable, i.e. fatal vs. non-fatal, in order to measure accident outcome. In the narrower context of our study, we believe that a binary response variable may not be comprehensive enough to characterize the passive safety performance of a vehicle.

Also, survival is not only a function of the passive safety performance of a vehicle during an accident, but it is also influenced by the quality of the medical care provided to the accident victim after the accident.

While it is widely held belief among experts that vehicle safety has much improved over the last decade, the recent study by Glance at al. (2012) reports that, given the same injury level, there has also been a significant reduction in mortality of trauma patients since 2002.

> *"In-hospital mortality and major complications for adult trauma patients admitted to level I or level II trauma centers declined by 30% between 2000 and 2009. After stratifying patients by injury severity, the mortality rate for patients presenting with moderate or severe injuries declined by 40% to 50%, whereas mortality rates remained unchanged in patients with the least severe or the most severe injuries."*

Given that the fatality data that was used to inform the EPA/NHTSA Final Rule was collected between 2002 and 2008, we speculate that identical injuries could have had different outcomes, i.e. fatal versus non-fatal, as a function of the year when the injury occurred. Thus, we find it important to use an outcome vari-

able that characterizes the severity of injuries sustained during the accident, as opposed to only counting fatalities.

3. Covariates

Similar to the binary fatal/non-fatal classification, other key variables in the EPA/NHTSA studies are also binned into two states, e.g. two weight classes {Cars<2,950 lbs., Cars>2,950 lbs.}. While the discretization of variables will also become necessary in our approach with Bayesian networks, we hypothesize that using two bins may be too "coarse" as a starting point. By using two intervals only, we would implicitly make the assumption of linearity in estimating the effect of vehicle weight on the dependent variable.

Furthermore, we speculate that a number of potentially relevant covariates can be added to provide a richer description of the accident dynamics. For instance, in a collision between two vehicles, we presume the angle of impact to be relevant, e.g. whether an accident is a frontal collision or a side impact. Also, specifically for two-vehicle collisions, we consider that the mass of both vehicles is important, as opposed to measuring this variable for one vehicle only. We will attempt to address these points with our selection of data sources and variables.

4. Consumer Response

The "law of unintended consequences" has become an idiomatic warning that an intervention in a complex system often creates unanticipated and undesirable outcomes. One such unintended consequence might be the consumers' response to the new CAFE rule.

The EPA/NHTSA Final Rule notes that all statistical models suggest mass reduction in small cars would be harmful or, at best, close to neutral, and that the consumer choice behavior given price increases is unknown. Also, the EPA/NHTSA Final Rule has put great emphasis on preventing vehicle manufacturers from "downsizing" vehicles as a result of the CAFE rule: "in the agencies' judgment, footprint-based standards [for manufacturers] discourage vehicle downsizing that might compromise occupant protection."[5]

However, EPA/NHTSA Final Rule does not provide an impact assessment with regard to future consumer choice in response to the new standards. Given that the Final Rule states that vehicle prices for consumers will rise significantly, "between $1,461 and $1,616 per vehicle in MY 2025,"[6] as a direct consequence of the CAFE rule, one can reasonably speculate that consumers might downsize their vehicles.

[5] EPA Final Rule, p. 214

[6] EPA Final Rule, p. 123

Rather, the Final Rule states: "Because the agencies have not yet developed sufficient confidence in their vehicle choice modeling efforts, we believe it is premature to use them in this rulemaking."[7] We speculate that this may limit one's ability to draw conclusions with regard to the overall societal cost.

Unfortunately, we currently lack the appropriate data to build a consumer response model that would address this question within our framework. However, in terms of the methodology, we have presented a vehicle choice modeling approach in our white paper, *Modeling Vehicle Choice and Simulating Market Share* (Conrady and Jouffe, 2010).

Technical Considerations

1. Assumption of Functional Forms

Given the familiar laws of physics that are applicable to collisions, one could hypothesize about certain functional forms for modeling the mechanisms that cause injuries of vehicle passengers. However, a priori, we cannot know whether any such assumptions are justified. Because this is a common challenge in many parametric statistical analyses, one would typically require a discussion regarding the choice of functional form, e.g. justifying the assumption of linearity.

We are not in a position to reexamine the choice of functional forms in the EPA/NHTSA studies. However, our proposed approach, learning Bayesian networks with BayesiaLab, has the advantage that no specification of any functional forms is required at all. Rather, BayesiaLab's knowledge discovery algorithms use information-theoretic measures to search for any kind of probabilistic relationships between variables. As we will demonstrate later, we can capture the relationship between injury severity and angle of impact, which is clearly nonlinear.

2. Interactions and Collinearity

All of the studies supporting the EPA/NHTSA Final Rule use a broad set of control variables in their regression models. However, none of the studies use interaction effects between these covariates. As such, an assumption is implicitly made that the covariates are all independent. However, examining the relationships between the covariates reveals that strong correlations do indeed exist, which violates the assumption of independence. In fact, collinearity is highlighted numerous times, e.g. "NHTSA considered the near multicollinearity of mass and footprint to be a major issue in the 2010 report and voiced concern about inaccurately estimated regression coefficients."[8]

---

[7] EPA Final Rule, p. 310

[8] Kahane, p. xi

The nature of learning a Bayesian network does automatically take into account a multitude of potential relationships between all variables and can even include collinear relationships without problem. We will see that countless relevant interactions between covariates exist, which are essential to capture the dynamics of the domain.

3. Causality

This last point is perhaps the most challenging one among the technical issues. The EPA/NHTSA studies use *statistical* models for purposes of *causal* inference. *Statistical* (or observational) inference, as in *"given that we observe,"* is not the same as *causal* inference, as in *"given that we do."* Only under strict conditions, and with many additional assumptions, can we move from the former to the latter.[9] Admittedly, causal inference from observational data is challenging and can be controversial. All the more it is important to clearly state the assumptions and why they might be justified.

With Bayesian networks we want to present a framework that allows researchers to explore this domain in a "causally correct" way, i.e. allowing—with the help of human domain knowledge—to disentangle "statistical correlation" and "causal effects."

[9] See Pearl (2009).

Exploratory Analysis

Data Overview

In order to better understand the nature of accidents, the National Automotive Sampling System (NASS) Crashworthiness Data System (CDS) was established as a nationwide crash data collection program sponsored by the U.S. Department of Transportation. The National Center for Statistics and Analysis (NCSA), part of the National Highway Traffic Safety Administration (NHTSA), started the data collection for the NASS program in 1979. Data collection is accomplished at 24 geographic sites, called Primary Sampling Units (PSUs). These data are weighted to represent all police reported motor vehicle crashes occurring in the USA during the year involving passenger cars, light trucks and vans that were towed due to damage. All data are available publicly from NHTSA's FTP server.[10]

Data Set for Study

We use the following subset of files from the 11-file dataset published by NHTSA:

- ACC: Accident Record (accident.sas7bdat)
- GV: General Vehicle Record (gv.sas7bdat)
- OA: Occupant Assessment Record (oa.sas7bdat)
- VE: Exterior Vehicle Record (ve.sas7bdat)

We have joined the records of these tables via their unique identifiers and then concatenated all files from 1995 through 2011 into a single table. This table contains records regarding approximately 200,000 occupants of 100,000 vehicles involved in 37,000 accidents. Each record contains more than 400 variables, although there is a substantial amount of missing values.

The comprehensive nature of this dataset is ideal for our exploration of the interactions of crash-related variables and their ultimate impact on passenger safety.

[10] ftp://ftp.nhtsa.dot.gov/NASS/

Notation

- All variables names/labels follow the the format "*DatasetAbbreviation_VariableName*", e.g. *GV_SPLIMIT* for the variable *Speed Limit* from the dataset *General Vehicle Record*.
- All variable and node names are italicized.
- Names of BayesiaLab-specific features and functions are capitalized and printed in bold. Such terms can be looked up in the BayesiaLab Library: library.bayesia.com.

Data Filters and Variable Selection

To start with our exploration of this problem domain, we narrow our focus by selecting subsets of variables and records:

1. We restrict our analysis to horizontal vehicle-to-vehicle collisions, with the vehicle under study being MY2000 or later.[11] In this context, we only examine the condition of the driver. Also, we exclude collisions involving large trucks (GVWR>10,000 lbs.[12]) and motorcycles. For data consistency purposes, we filter out unusual and very rare accident types, e.g. accidents with principal deformation to the underside of the vehicle. We apply these filters primarily for expositional simplicity. However, we do recognize that this limits our ability to broadly generalize the findings.
2. However, no records are excluded solely due to missing values. BayesiaLab offers advanced missing values processing techniques, which we can leverage here. This is an important point as the vast majority of records contain some missing values. In fact, if we applied a traditional casewise/listwise deletion, most records would be eliminated from the database.
3. Furthermore, many of the 400+ variables provide a level of detail that by far exceeds the scope of this paper. Thus, we limit our initial selection to 19 variables that appear a priori relevant and are generally in line with the variables studied in the EPA/NHTSA research.
4. In addition to the variables defined in the NASS/CDS database, we introduce *GV_FOOTPRINT* as a variable that captures vehicle footprint.[13] This new variable is computed as the product of

[11] We consider MY2000 a reasonable cutoff point, as by then second-generation airbags were mandatory for all passenger vehicles.

[12] Gross Vehicle Weight Rating

[13] "Footprint is defined as a vehicle's wheelbase multiplied by its average track width – in other words, the area enclosed by the points at which the wheels meet the ground." Final Rule, p. 69.

VE_WHEELBAS (Wheelbase) and *VE_ORIGAVTW (Average Track Width)*, which are recorded in the original database.

The following table summarizes the variables included in our study.

Variable Name	Long Name	Units/States	Comment
GV_CURBWGT	Vehicle Curb Weight	kg	
GV_DVLAT	Lateral Component of Delta V	km/h	
GV_DVLONG	Longitudinal Component of Delta V	km/h	
GV_ENERGY	Energy Absorption	J	
GV_FOOTPRINT	Vehicle Footprint	m^2	calculated as WHEELBAS x ORIGAVTW
GV_LANES	Number of Lanes	count	
GV_MODELYR	Vehicle Model Year	year	
GV_OTVEHWGT	Weight Of The Other Vehicle	kg	
GV_SPLIMIT	SpeedLimit	mph	converted into U.S. customary units
GV_WGTCDTR	Truck Weight Code	missing = Passenger Vehicle	
		6,000 and less	
		6,001 - 10,000	
OA_AGE	Age of Occupant	years	
OA_BAGDEPLY	Air Bag System Deployed	Nondeployed	
		Bag Deployed	
OA_HEIGHT	Height of Occupant	cm	
OA_MAIS	Maximum Known Occupant AIS	Not Injured	AIS Probability of Death
		Minor Injury	0%
		Moderate Injury	1-2%
		Serious Injury	8-10%
		Severe Injury	5-50%
		Critical Injury	5-50%
		Maximum Injury	100% (Unsurvivable)
		Unknown	Missing Value
OA_MANUSE	Manual Belt System Use	Used	
		Not Used	
OA_SEX	Occupant's Sex	Male	
		Female	
OA_WEIGHT	Occupant's Weight	kg	
VE_GAD1	Deformation Location (Highest)	Left	
		Front	
		Rear	
		Right	
VE_PDOF_TR	Clock Direction for Principal Direction of Force (Highest)	Degrees	Transformed variable, rotated 135 degrees counterclockwise

OA_MAIS (Maximum Known Occupant AIS)

The outcome variable *OA_MAIS* represents the *Maximum Known Occupant AIS*. "AIS" stands for "Abbreviated Injury Scale." The Abbreviated Injury Scale is an anatomically-based, consensus-derived global severity scoring system that classifies each injury by body region according to its relative importance on a 6-point ordinal scale (1=minor and 6=maximal).

Coordinate System for Variable PDOF1 (Principal Direction of Force)

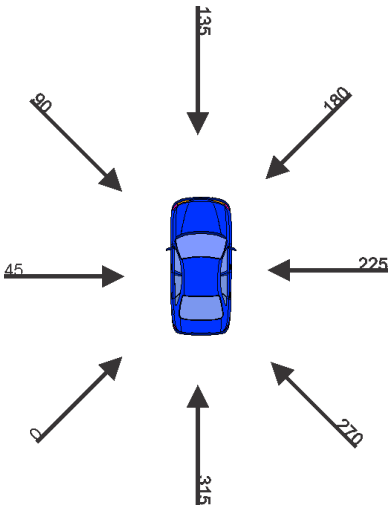
Most of the variables' scales and units are self-explanatory, perhaps with the exception of *GV_PDOF1 (Principal Direction of Force)*. This variable records the direction of the highest force acting on the vehicle during the accident. A frontal collision, i.e. in the direction of travel from the perspective of the vehicle under study, would imply *PDOF1=0*. Conversely, a rear impact, would mean *PDOF1=180*, etc.

Given the requirements of data discretization as part of the data import process (see next chapter), we rotate the coordinate system by 135 degrees counterclockwise. The values in this new coordinate system are recorded in the transformed variable *GV_PDOF1_TR*. This rotation prevents that frequently occurring, similar values (e.g. frontal collisions at 355°, 0°, and 5° on the original scale) are split into different bins due to the natural break at 0°.

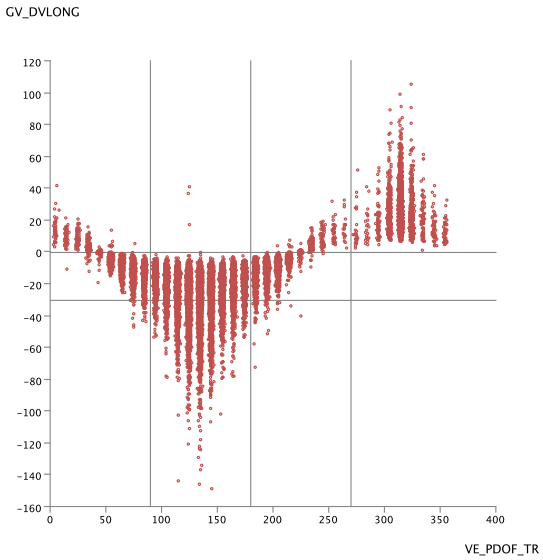
To make it easier to interpret the values of the transformed variable *GV_PDOF1_TR*, we briefly illustrate our new coordinate system. For instance, a *GV_PDOF1_TR=45* now means that the vehicle under study collided on the driver's side and that the impact was perpendicular to the direction of travel. A 135° impact represents a direct frontal collision, e.g. with oncoming traffic. Conversely, a rear-end collision is represented by 315° angle.

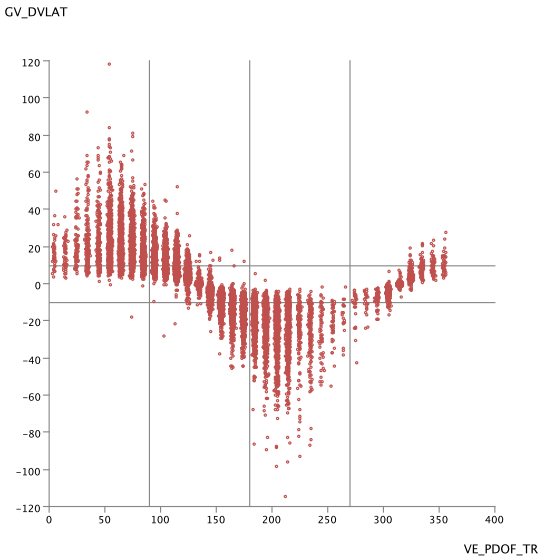
This coordinate system may become more intuitive to understand when it is viewed in quadrant form, in clockwise direction:

- 0°-90°: Impact from left
- 90°-180°: Frontal impact
- 180°-270°: Impact from right
- 270°-360°: Rear impact



To confirm the plausibility of this transformed variable, we can plot *GV_DVLAT* and *GV_DVLONG* as a function of *GV_PDOF1_TR*. We would anticipate that a full frontal impact, i.e. *GV_PDOF1_TR=135*, would be associated with the highest decrease in velocity, i.e. *GV_DVLONG≪0*. At the same time, we would expect the lateral Delta V to be near zero, i.e. *GV_DVLAT≈0*. This is indeed what the following plots confirm.[14] They also illustrate the respective signs of Delta V as a function of the impact angle.





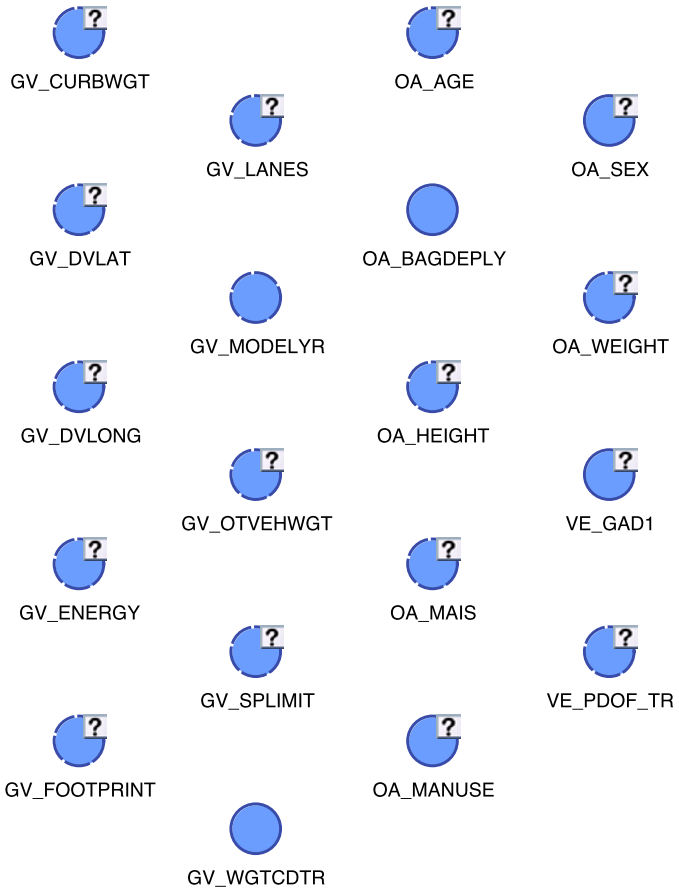
Data Import

The first step in our process towards creating a Bayesian network is importing the raw data into BayesiaLab. The import and discretization processes have been described extensively in some of our previous white papers, so we will omit the details here.[15] However, we should note that we adjust some of the bins that were found by BayesiaLab's automatic discretization algorithms, so they reflect typical conventions regarding this domain. For instance, if the discretization algorithm proposed a bin threshold of 54.5 for GV_*SPLIMIT*, we would change this threshold to 55 (mph) in order to be consistent with our common understanding of speed limits.

[14] Scatterplots of this kind can be produced in BayesiaLab with the **Charts** function (**Data | Charts**).

[15] http://bayesia.us/index.php/whitepapers.

Upon completion of the import process, we obtain an initially unconnected network, which is shown below.[16] All variables are now represented by nodes, one of the core building blocks in a Bayesian network. A node can stand for any variable of interest. Discretized nodes are shown with a dashed outline, whereas solid outlines indicate discrete variables, i.e. variables with either categorical or discrete numerical values. Once the variables appear in this new form in a graph, we will exclusively refer to them as nodes.

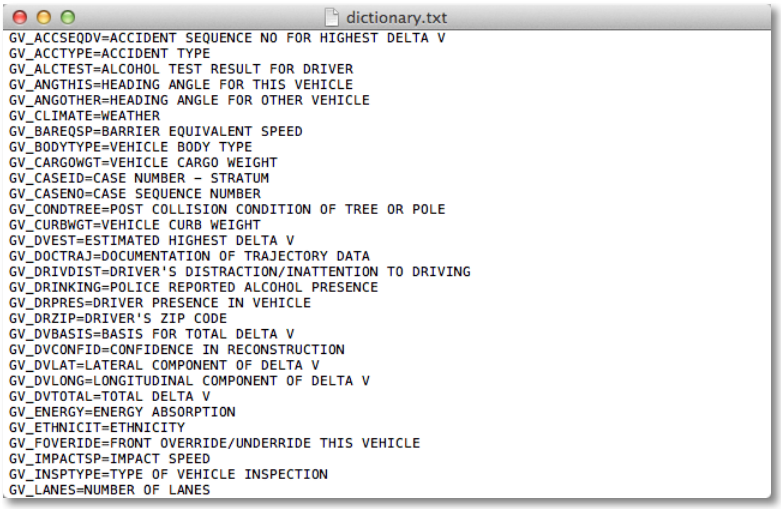


At this point, it is practical to add **Node Comments** to the **Node Names** that are displayed under each node by default. In BayesiaLab, **Node Comments** are typically used for longer and more descriptive titles, which can be turned on or off, depending on the desired view of the graph. Here, we associate a dictionary of the

[16] Throughout this paper we will mostly apply the **Staggered Layout**, instead of the default **Grid Layout**. These option are available under **View | Layout**.

complete, long variable names[17] with the **Node Comments**, while the more compact variables names of the original dataset remain as **Node Names**.[18]

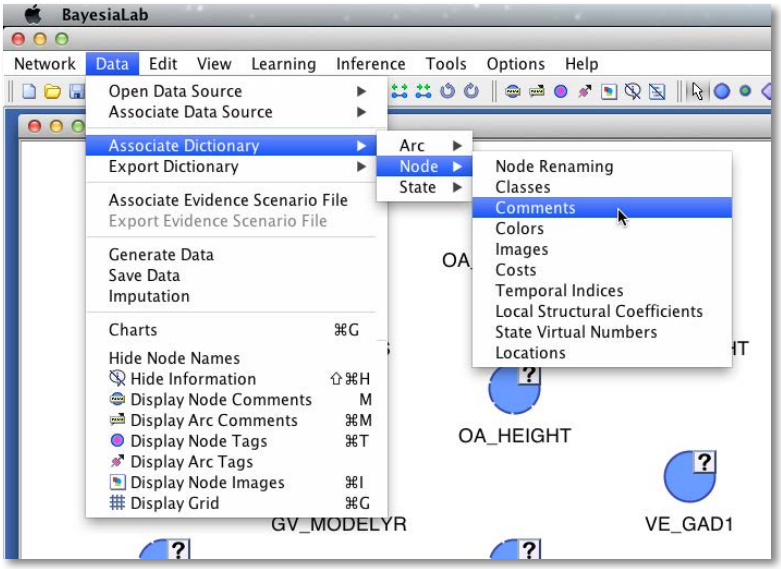
The syntax for this association is rather straightforward: we simply define a text file which includes one **Node Name** per line. Each **Node Name** is followed by the equal sign ("="), or alternatively TAB or SPACE, and then by the long variable description, which will serve as the **Node Comment**.



This file can then be loaded into BayesiaLab via **Data | Associate Dictionary | Node | Comments**.

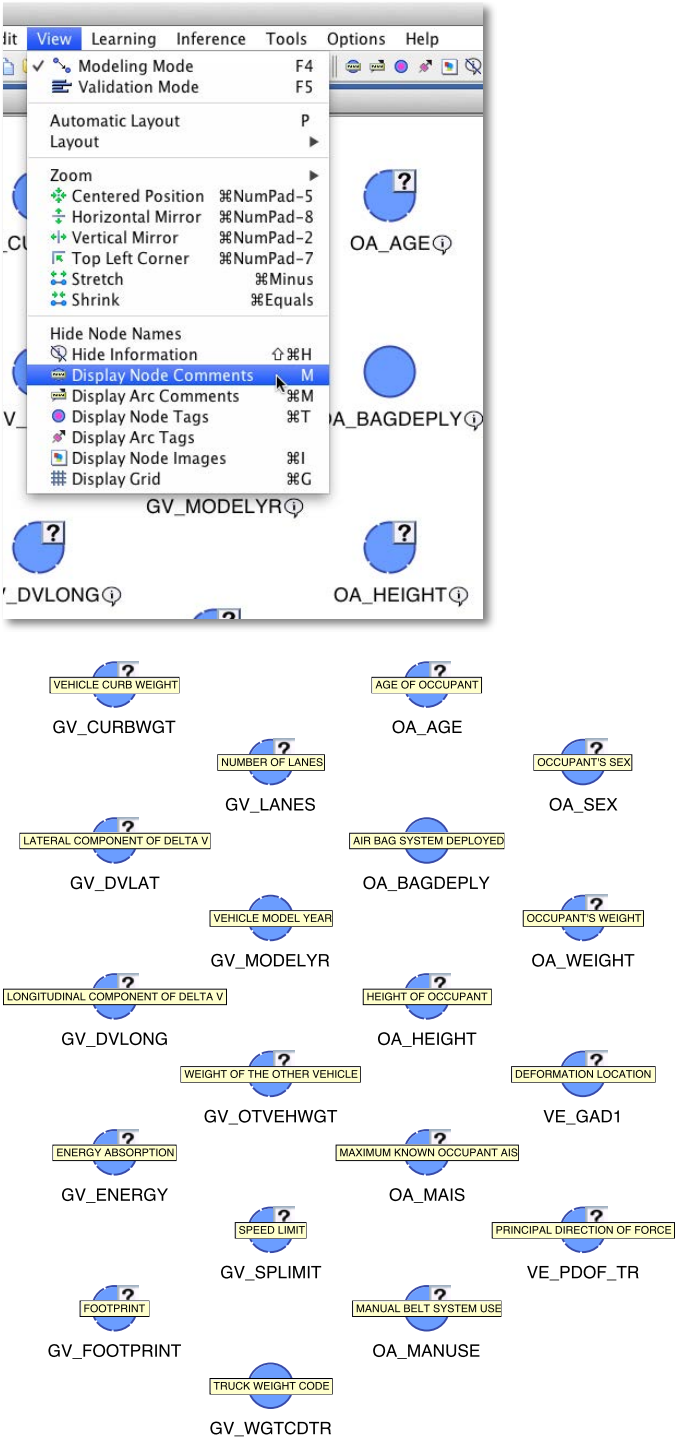
[17] The original SAS files do include long variable names in addition to the standard ones, which are limited to 8 characters.

[18] To maintain a compact presentation, we will typically use the original variable name when referencing a particular node.



Once the **Node Comment** are associated, a small call-out symbol will appear next to each **Node Name**. This indicates that **Node Comments** are available for display.

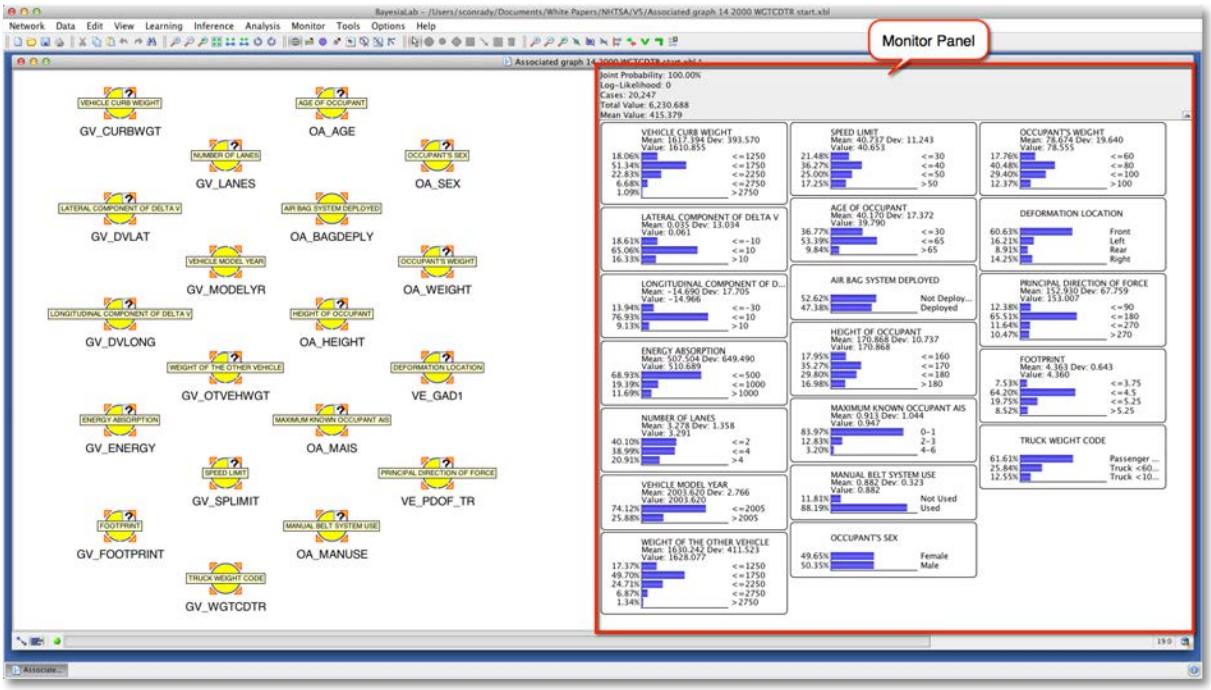
As the name implies, selecting **View | Display Node Comments** (or alternatively pressing the keyboard shortcut "M") will reveal the long variable names.



Node Comments can be displayed either for all nodes or only for selected ones. Given the sometimes cryptic nature of the original variable names, we will keep the more self-explanatory **Node Comments** turned on for most graphs.

Initial Review

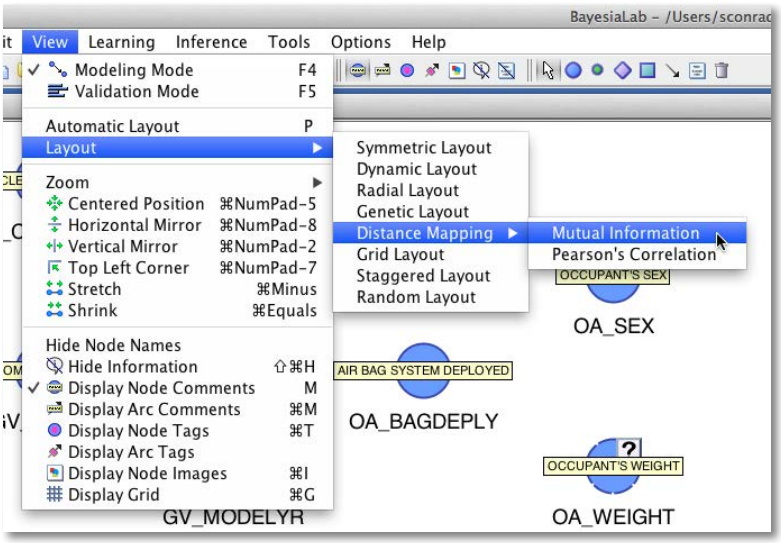
Upon data import, it is good practice to review the probability distributions of all nodes. The best way to get a complete overview is to switch into the **Validation Mode** (shortcut "F5"), selecting all nodes (Ctrl+A) and then double-clicking on any one of them. This brings up the **Monitors** for each node within the **Monitor Panel**.



Each **Monitor** contains a histogram representing the marginal probability distributions of the states of its associated node. This allows us to review the distributions and compare them with our own domain understanding. For instance, the gender mix, *OA_SEX*, is approximately at the expected uniform level, and other nodes appear to have reasonable distributions, too.

Distance Mapping

Going beyond these basic statistics of individual nodes, we can employ a number of visualization techniques offered by BayesiaLab. Our starting point is **Distance Mapping** based on **Mutual Information: View | Layout | Distance Mapping | Mutual Information.**



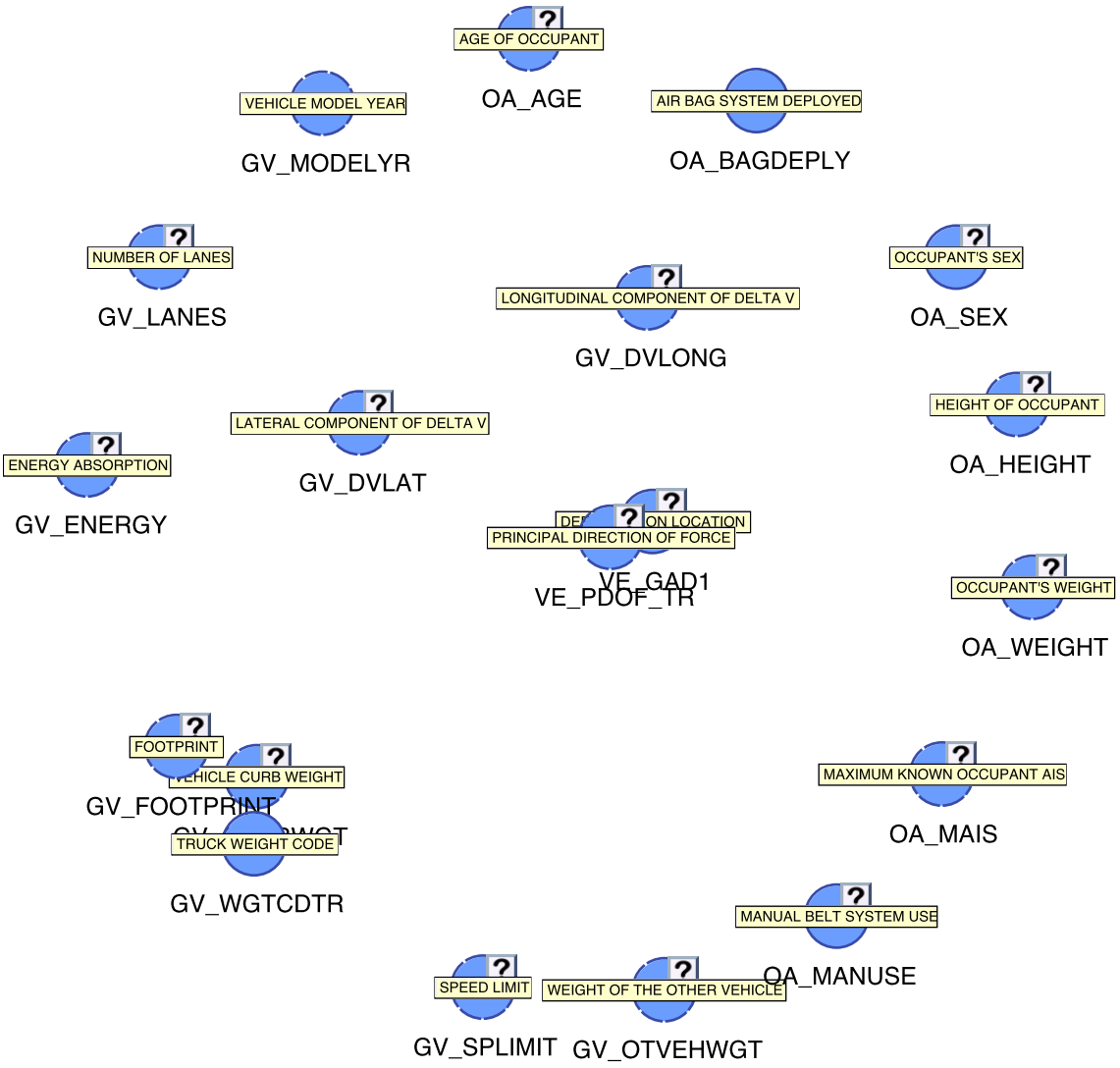
As the name of this layout algorithm implies, the generated layout is determined by the **Mutual Information** between each pair of nodes.

Mutual Information measures the information that X and Y share: it measures how much knowing one of these variables reduces our uncertainty about the other. For example, if X and Y are independent, then knowing X does not provide any information about Y and vice versa, so their **Mutual Information** is zero. At the other extreme, if X and Y are identical then all information conveyed by X is shared with Y: knowing X determines the value of Y and vice versa.

Formal Definition of Mutual Information

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

In the **Distance Mapping** graph, the distance between nodes is inversely proportional to their **Mutual Information.**

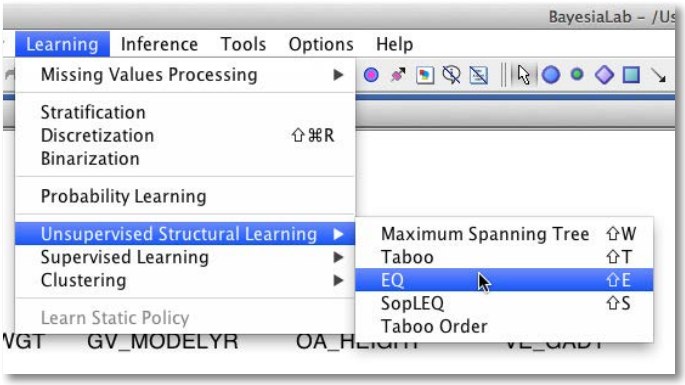


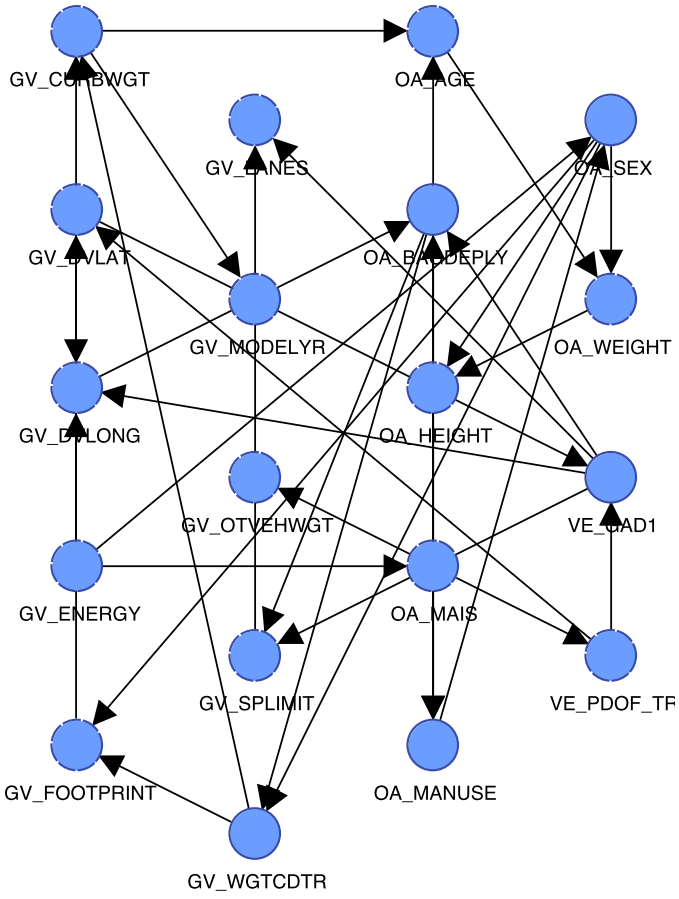
Among all nodes in our example, there appear to be several clusters which can be intuitively interpreted. *VE_PDOF_TR* and *VE_GAD1*, in the center of the above graph, reflect impact angle and deformation location, two geometrically connected metrics. *GV_FOOTPRNT*, *GV_CURBWGT* and *GV_WGTCDTR* are closely related to each other and to the overarching concept of vehicle size. Quite literally, knowing the state of a given node reduces our uncertainty regarding the states of the nearby nodes.

Unsupervised Learning

When exploring a new domain, we usually recommend performing **Unsupervised Learning** on the newly imported database. This is also the case here, even though our principal objective is targeted learning, for which **Supervised Learning** will later be the main tool.

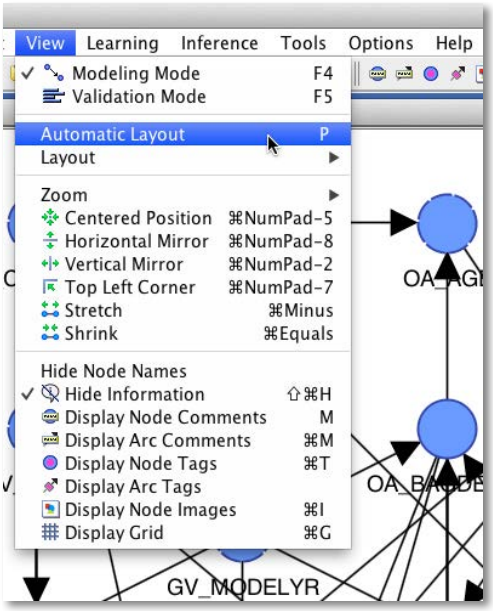
Learning | Unsupervised Structural Learning | EQ initiates the **EQ Algorithm,** which is suitable for the initial review of the database.[19]



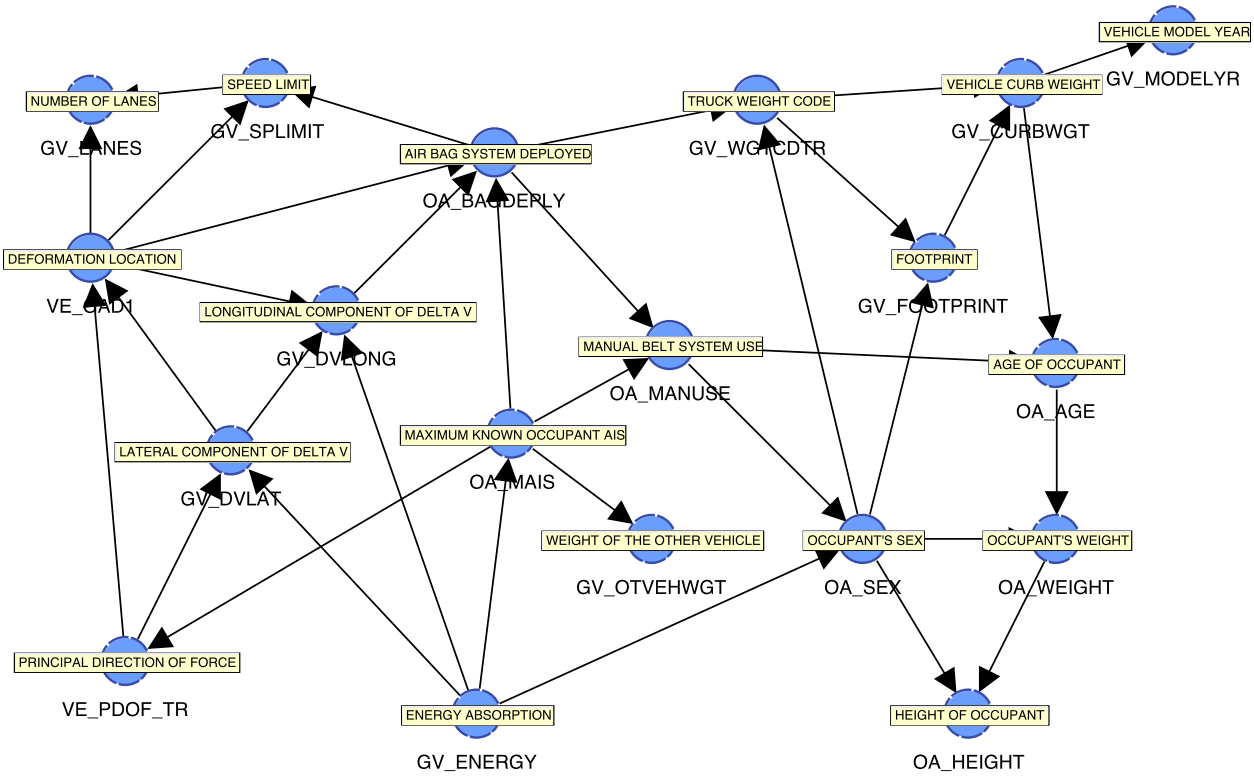


[19] The very fast **Maximum Weight Spanning Tree** algorithm is recommended for larger databases with significantly more variables.

In its raw form, the crossing arcs make this network somewhat tricky to read. BayesiaLab has a number of layout algorithms that can quickly disentangle such a network and produce a much more user-friendly format. We can select **View | Automatic Layout** or alternatively use the shortcut "P".



Now we can visually review the learned network structure and compare it to our own domain knowledge. This allows us to do a "sanity check" of the database and the variables, and it may highlight any inconsistencies.



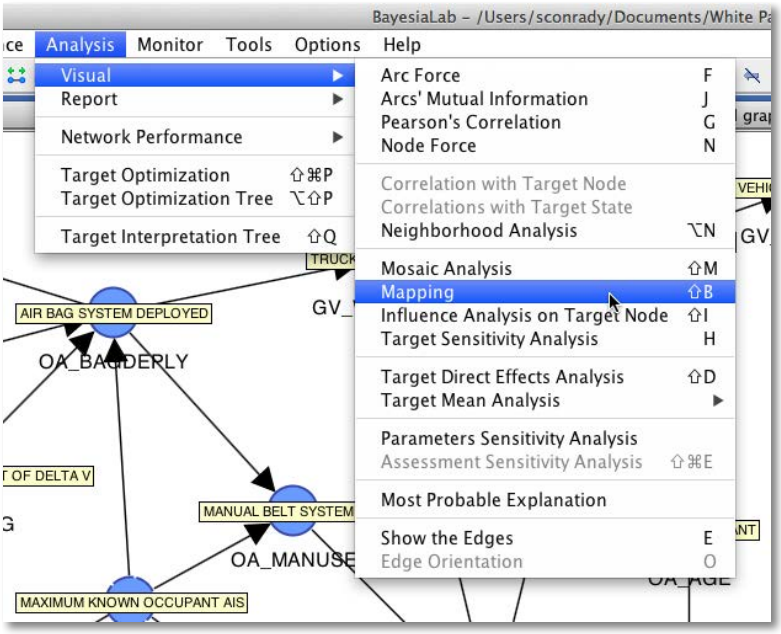
Indeed, in our first learning attempt, we immediately find 34 arcs between the 19 variables included in the model, so interactions appear to be manifold.

Although it is tempting, we must not interpret the arc directions as causal directions. What we see here, by default, are merely statistical associations, not causal relations. We would have to present a significant amount of theory to explain why Bayesian networks always must have directed arcs. However, this goes beyond the scope of this presentation. Rather, we refer to the literature listed in the references and our other white papers.

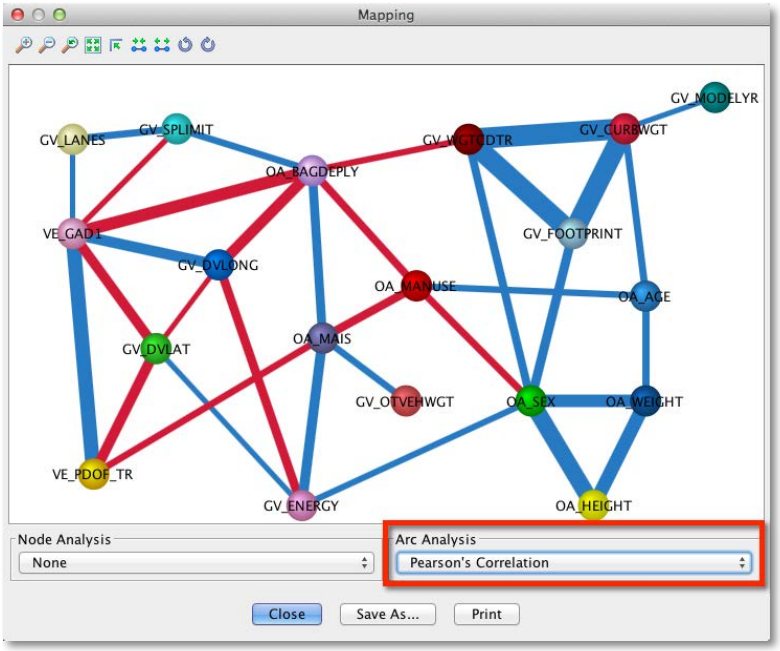
Mapping

Beyond qualitatively inspecting the network structure, BayesiaLab allows us to visualize the quantitative part of this network. To do this, we first need to switch into the **Validation Mode** by clicking on the button in the lower lefthand corner of the **Graph Panel**, or by using the "F5" key as a shortcut.

From within the **Validation Mode**, we can start the **Mapping** function: **Analysis | Visual | Mapping**

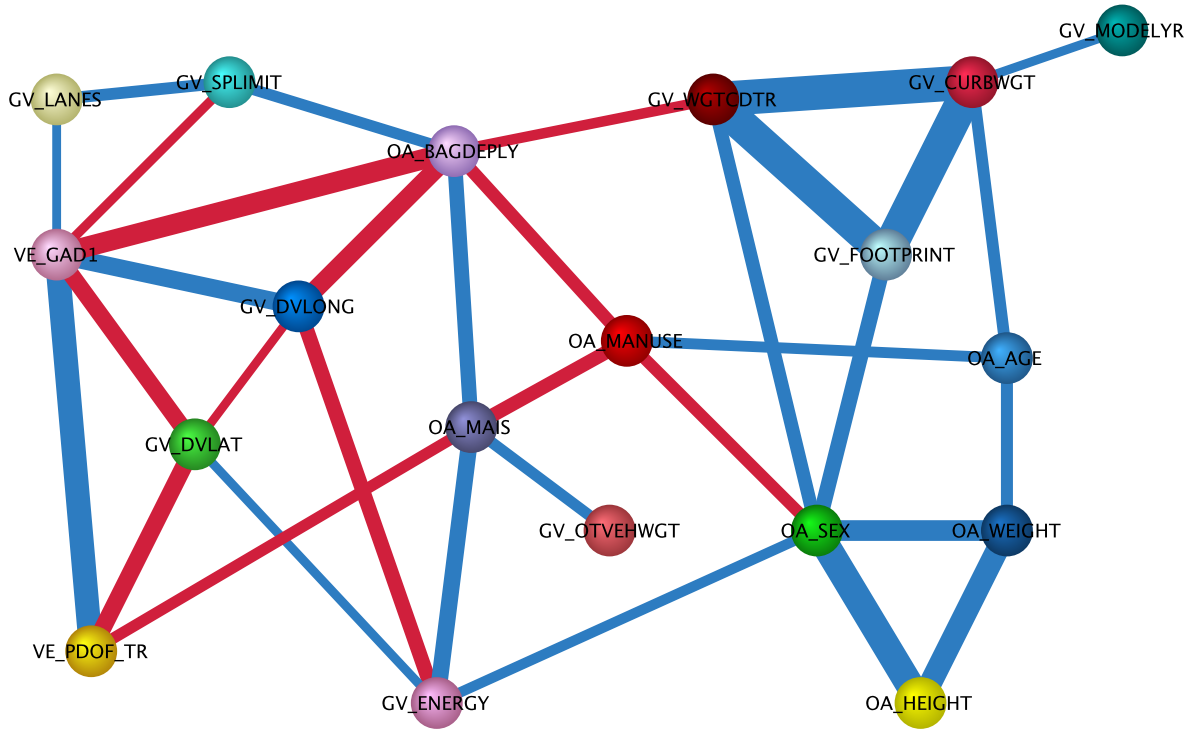


The **Mapping** window opens up and presents a new view of the graph.



The **Mapping** window features drop-down menus for **Node Analysis** and **Arc Analysis**. However, we are only interested in **Arc Analysis** at this time and select **Pearson's Correlation** as the metric to be displayed.

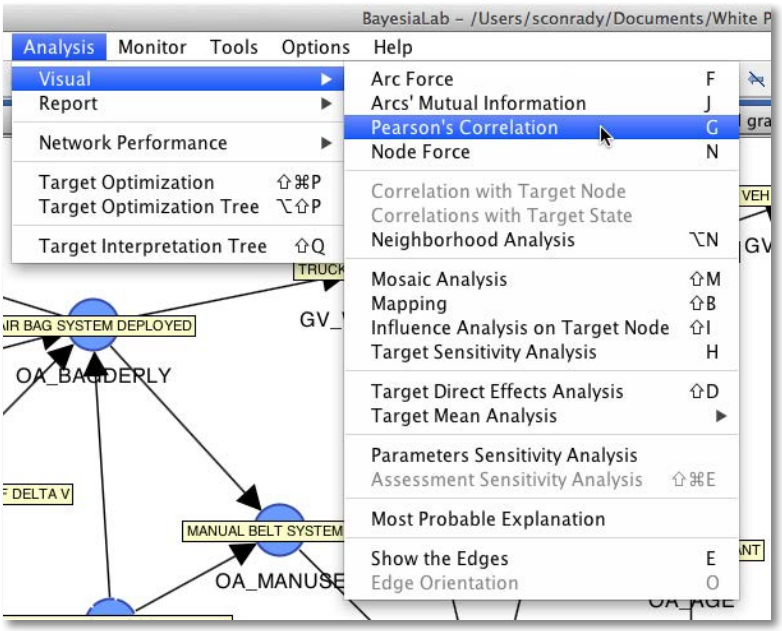
Arc Analysis: Pearson's Correlation

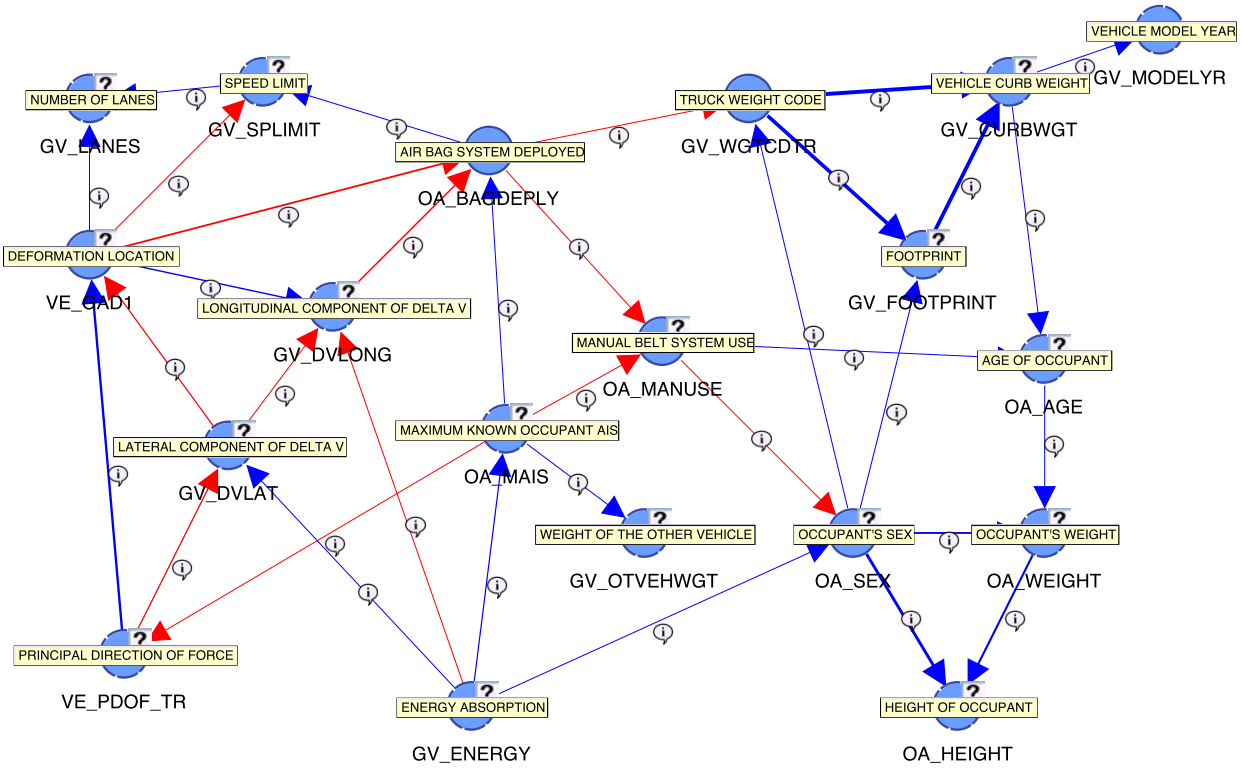


The thickness of the arcs, beyond a fixed minimum size,[20] is now proportional to the **Pearson Correlation** between the nodes. Also, the blue and red colors indicate positive and negative correlations respectively.

BayesiaLab can also visualize the same properties in a slightly different format. This is available via **Analysis | Visual | Pearson's Correlation.**

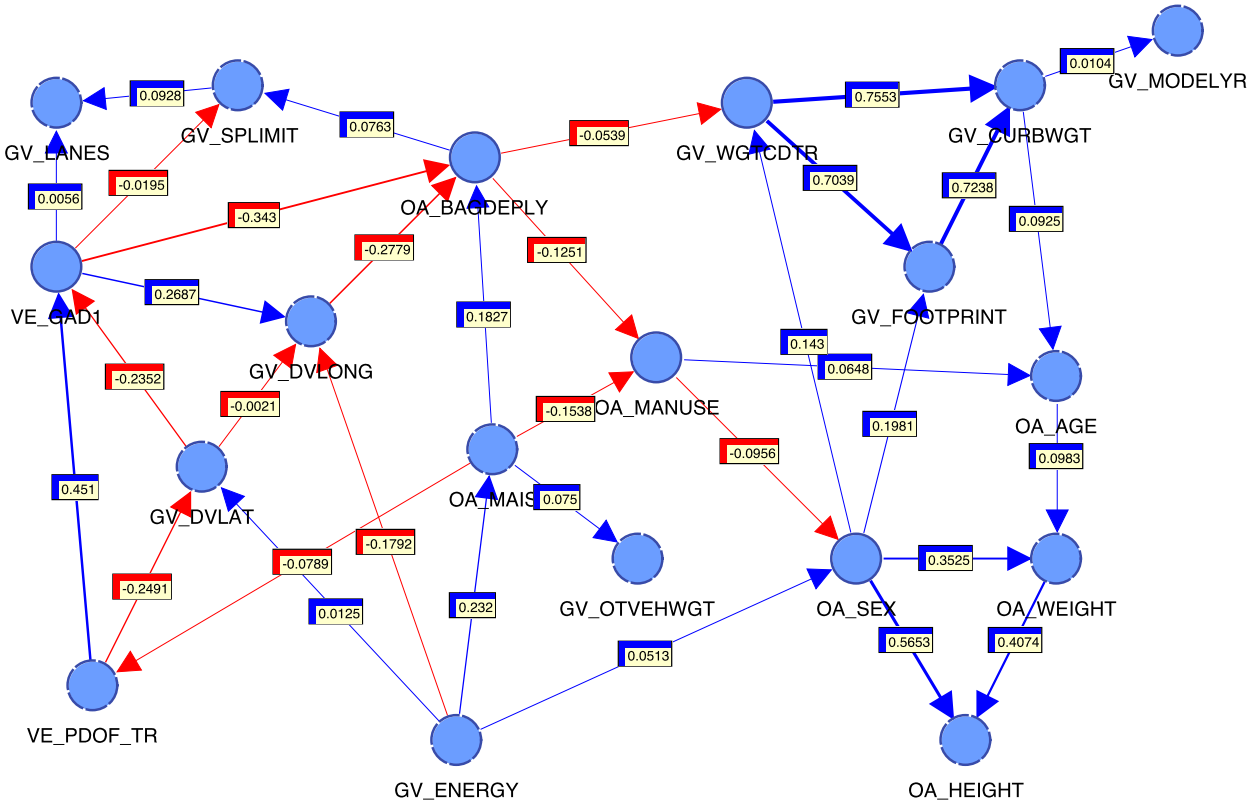
[20] The minimum and maximum sizes can be changed via **Edit Sizes** from the **Contextual Menu** in the **Mapping Window.**





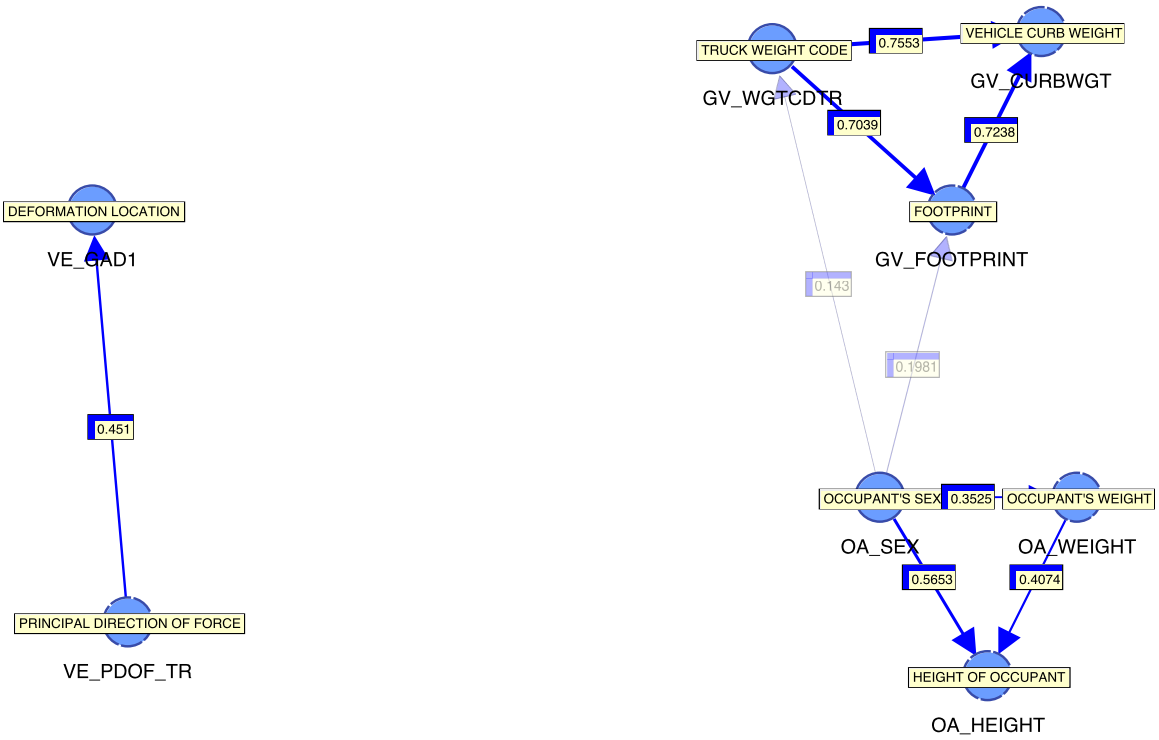
Here, too, the arc thickness is proportional to **Pearson's Correlation**. Additionally, callouts indicate that further information can be displayed. We opt to display this numerical information via **View | Display Arc Comments**. This function is also available via a button in the menu:

To emphasize the **Arc Comments**, as opposed to the **Node Comments**, we have turned off the latter in the following graph.



The multitude of numbers presented in this graph can still be overwhelming. We may wish to "tune out" weaker connections to focus on the more important ones. The slider control within the menu bar allows us to interactively change the threshold below which connections should be excluded from display.

At this setting, only nodes are shown that are connected with an absolute correlation coefficient of 0.34 or higher. The remaining nodes and arcs are shown in the graph below.

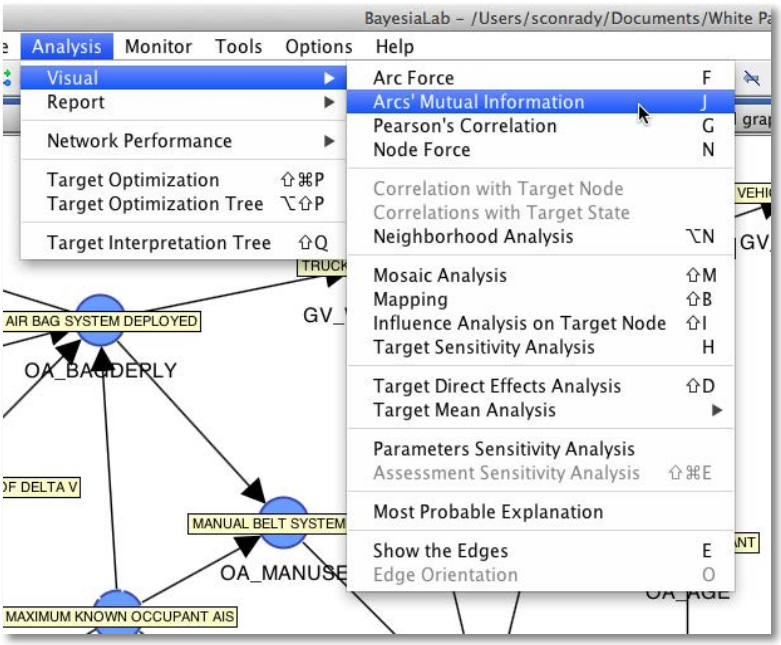


There are a number of nodes that stand out as highly correlated, in particular *GV_CURBWGT, GV_WGTCDTR* and *GV_FOOTPRNT*. This is plausible as these nodes can be understood as proxies for overall vehicle size. Strong relationships also exist between *OA_HEIGHT, OA_WEIGHT* and *OA_SEX*, which is consistent with our general knowledge that men, on average, are taller and heavier than women.

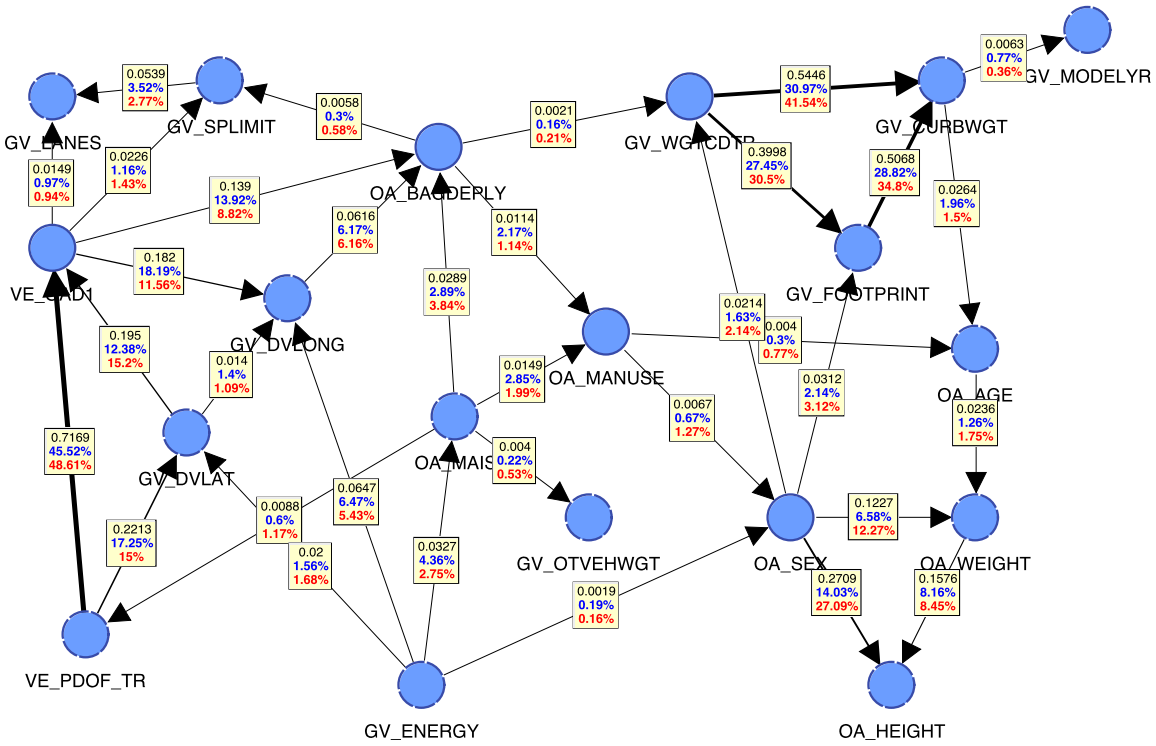
Note that BayesiaLab can compute **Pearson's Correlation** for any pair of nodes with *ordered* states, regardless of whether they are numerical or categorical (e.g. Monday, Tuesday, etc.). However, the computed values are only meaningful if a linear relationship can be assumed. For some of the node pairs shown above, this may not be an unreasonable hypothesis. However, in the case of *VE_GAD1* (VE_GAD1∈{Left, Front, Rear, Right}) and *VE_PDOF_TR* (*VE_PDOF_TR*∈[0,360]) it would not be sensible to interpret the relationship as linear. Rather, the computed correlation is purely an artifact of the random ordering of states of *VE_GAD1*. However, we will see in the next section that strong (albeit nonlinear) links do exist between these variables.

Mutual Information

An alternative perspective on the relationships can be provided by displaying **Arc's Mutual Information**, which is a valid measure regardless of variable type, i.e. including the relationships between (not-ordered) categorical and numerical variables.



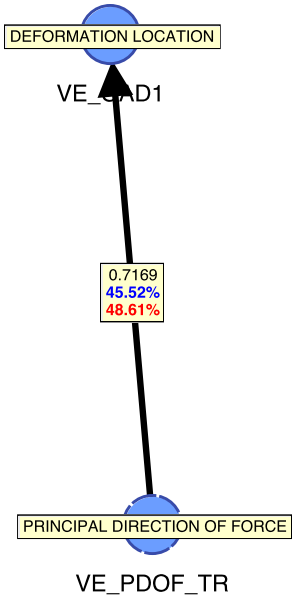
As before, we can bring up the numerical information by clicking the **Display Arc Comments** button.

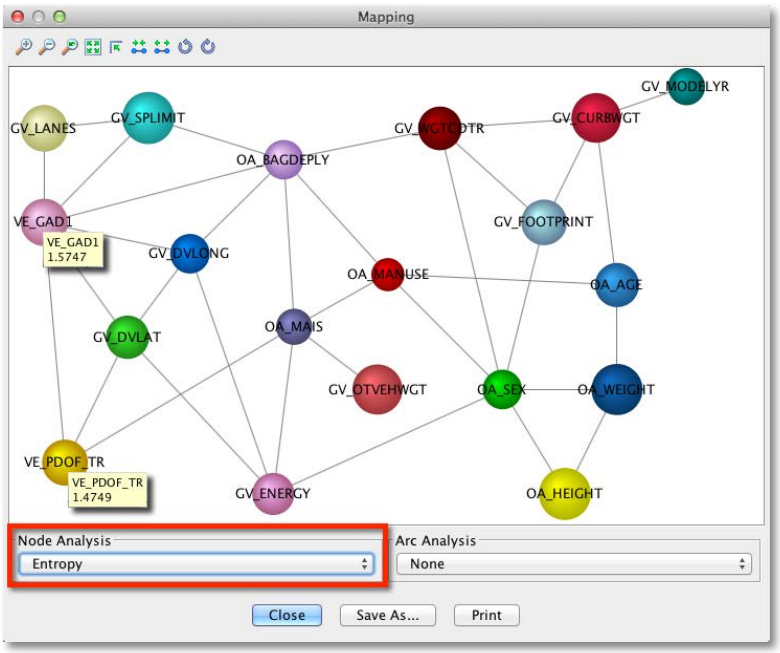


Mutual Information I(X,Y) measures how much (on average) the observation of a random variable Y tells us about the uncertainty of X, i.e. by how much the entropy of X is reduced if we have information on Y. **Mutual Information** is a symmetric metric, which reflects the uncertainty reduction of X by knowing Y as well as of Y by knowing X.

We can once again use the slider in the menu bar to adjust the threshold for the display of arcs. Moving the slider towards the right, we gradually filter out arcs that fall below the selected threshold.

In our example, knowing the value of *VE_PDOF_TR* on average reduces the uncertainty of the value of *VE_GAD1* by 0.7916 bits, which means that it reduces its uncertainty by 45.52% (shown in blue, in the direction of the arc). Conversely, knowing *VE_GAD1* reduces the uncertainty of *VE_PDOF_TR* by 48.61% (shown in red, in the opposite direction of the arc). Given that **Mutual Information** is a symmetric metric, this implies that the marginal uncertainty, (or **Entropy**) of *VE_GAD1* is higher than that of *VE_PDOF_TR*. We can easily confirm this by directly visualizing the entropy via the **Mapping** function. As opposed to our previous use of this function, we now select **Entropy** from the **Node Analysis** drop-down menu. The node size is now proportional to the nodes' **Entropy**. We can obtain the exact **Entropy** values by hovering with the mouse pointer over the respective nodes.





Bayesian Network Properties

It is necessary to emphasize that, despite the visual nature of a Bayesian network, it is not a visualization of *data*. Rather, it is the *structure* that is visualized. So, what we see is the *model*, not the *data*. The Bayesian network is meant to be a generalization of the underlying data, rather than a "bit-perfect" replica of the data. Theoretically, and at a huge computational cost, a fully-connected Bayesian network can produce a perfect fit. However, that would bring us back to nothing more than the raw data, instead of generating an interpretable abstraction of the data.

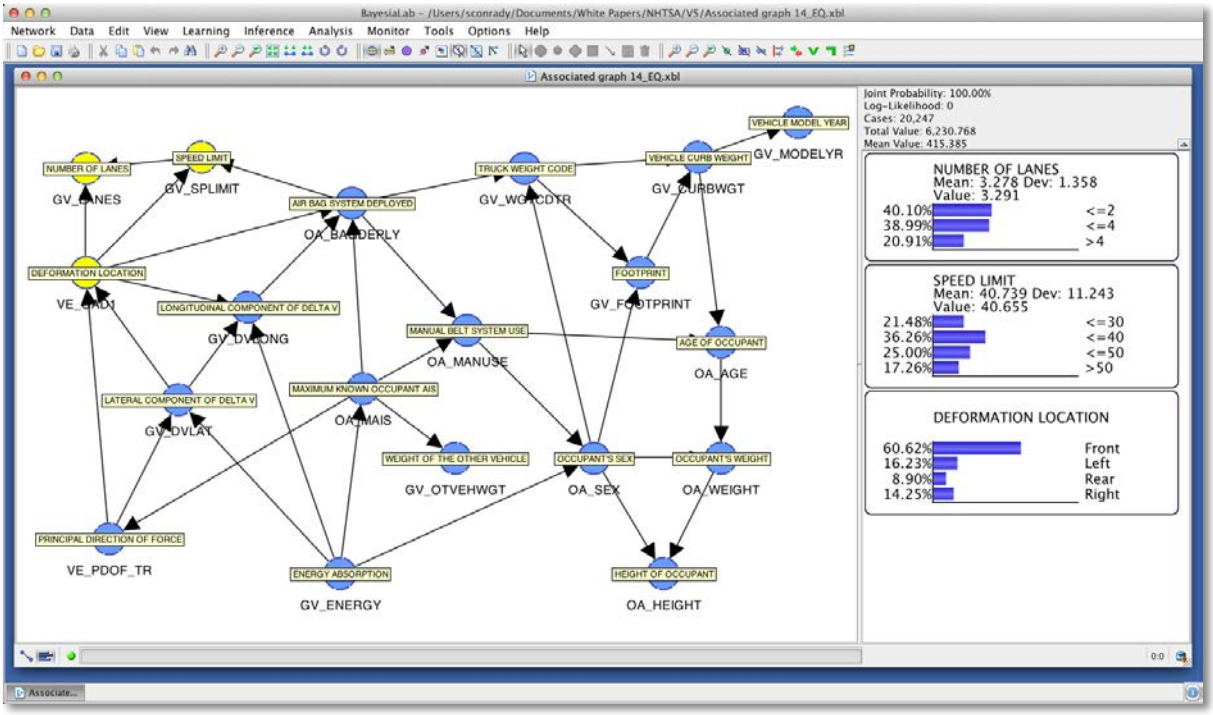
Omnidirectional Inference

Any network that we see here is a fully specified and estimated model that can be used for inference. A particularly important property is what we call "omnidirectional inference." While traditional statistical models usually contain one dependent and many independent variables, that distinction is not necessary in a Bayesian network. In fact, all variables can be treated equivalently, which is particularly interesting for exploratory research.

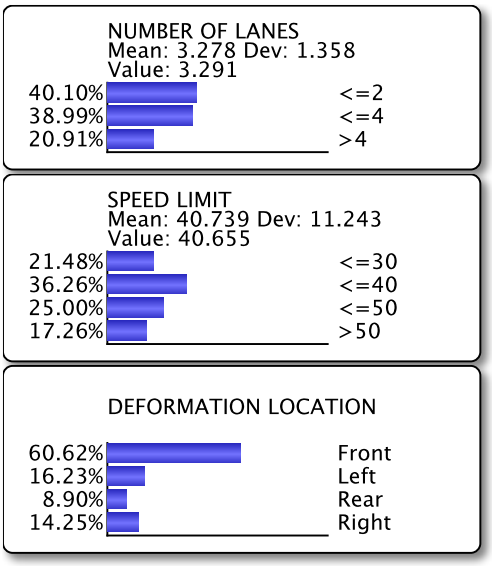
To gain familiarity with all the interactions learned from the data, we will experiment with omnidirectional inference and run various exploratory queries on different subsets of the model.

Example 1: Number of Lanes, Deformation Location and Speed Limit

In **Validation Mode**, double-clicking on an individual node, or on a selected set of nodes, brings up the corresponding **Monitors** on the righthand-side **Monitor Panel**. Conversely, double-clicking again would remove them.

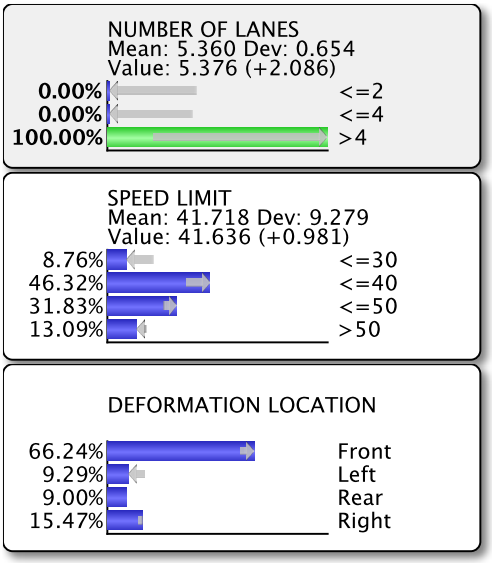


For instance, we show the **Monitors** for *GV_LANES*, *GV_SPLIMIT*, and *VE_GAD1*. Small histograms will show us the marginal distributions of those variables.



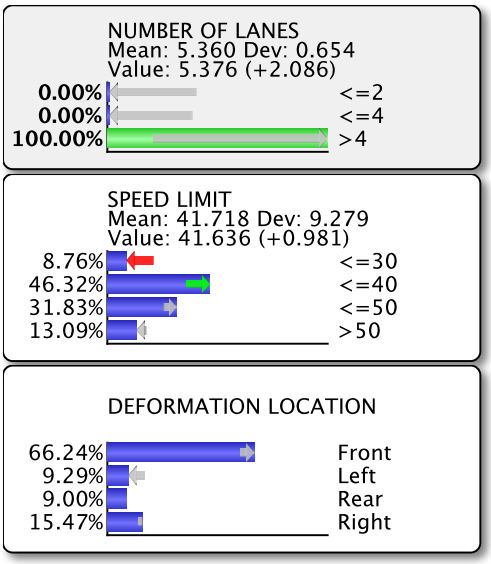
We can see that of all accidents 38.99% occur on roads with 3 or 4 lanes, or that 17.26% happen in areas with speed limits greater than 50 mph.

We might now want to ask the question, "what is the distribution of impact angles for accidents that happen on roads with more than 4 lanes?" We can use the network to answer this query by "setting evidence" via the **Monitor** for *GV_LANES*. In BayesiaLab, this simply requires a double-click on the ">4" bar of this **Monitor**. Upon setting the evidence, the *GV_LANES>4* bar turns green, and we can now read the posterior probability distributions of the other nodes.



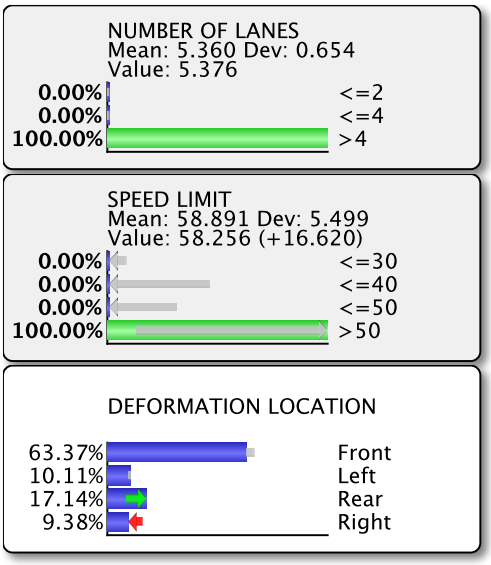
We see, for instance, that the share of left-side collisions *(VE_GAD1=Left)* has dropped from 16.23% to 9.29%. However, we can observe another change. Given that we are focusing on roads with 5 or more lanes, now only 8.76% have a speed limit of 30 mph or below. In the marginal distribution, this share was 21.48%. The little gray arrows indicate the amount of change versus the previous distribution.

The **Maximum Variation of Probabilities** can be highlighted with red and green arrows by clicking the corresponding button in the menu.



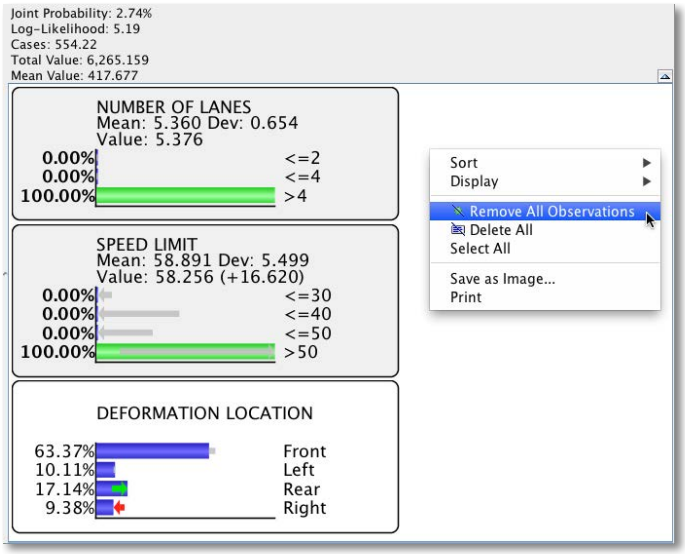
It is now obvious that one piece of evidence, i.e. setting *GV_LANES>4*, has generated multiple updates to other variables' distributions as if we had multiple dependent variables. In fact, *all* variables throughout the network were updated, but we only see the changes of distributions of those nodes that are currently shown in the **Monitor Panel**.

We now set a second piece of evidence, *GV_SPLIMIT>50*.



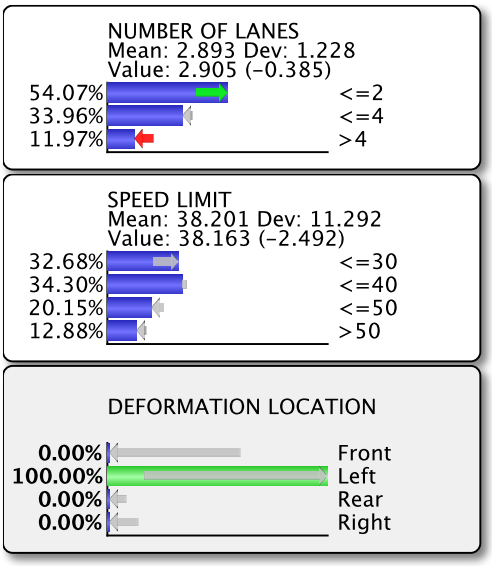
As a result, we see a big change in *VE_GAD1=Rear*, i.e. rear impacts; their probability jumps from 9.00% to 17.14%. Again, this should not be surprising as roads with more than 4 lanes and with speed limits of 50 mph or higher are typically highways with fewer intersections. Presumably, less cross-traffic would cause fewer side impacts.

Before we proceed, we remove all evidence and reset the **Monitors** to their marginal distribution. This can be done by right-clicking on the background of the **Monitor Panel** and selecting **Remove All Observations** from the **Contextual Menu**.



Alternatively, clicking the **Remove all Observations** button on the **Main Menu** does the same.

Once all evidence is removed, we can set new evidence. More specifically, we want to focus on side impacts on the driver's side only, which can be expressed as *VE_GAD1=Left*.

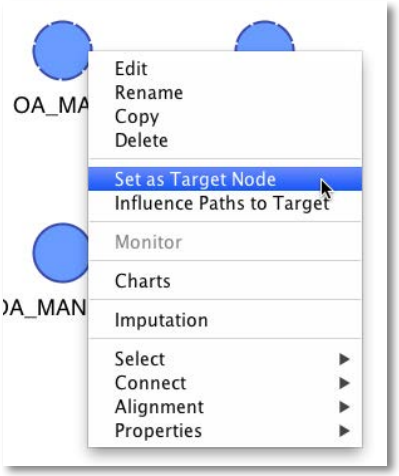


Now the probability of *GV_LANES>4* has decreased and the probability of *GV_SPLIMIT<=30* has increased. One can speculate that such kinds of collisions occur in areas with many intersections, which is more often the case on minor roads with fewer lanes and a lower speed limit.

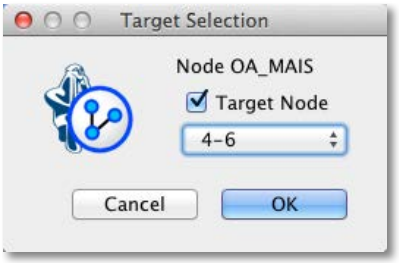
This demonstrates how we can reason backwards and forwards within a Bayesian network, using any desired combination of multiple dependent and independent variables. Actually, we do not even need to make that distinction. We can learn a single network and then have a choice regarding the nodes on which to set evidence and the nodes to evaluate.

Modeling Injury Severity with Supervised Learning

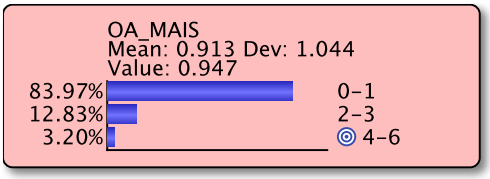
While **Unsupervised Learning** is an ideal way to examine multiple interactions within our domain for exploratory purposes, the principal task at hand is explaining injury severity *(OA_MAIS)* as a function of the other variables. For this purpose, BayesiaLab offers a number of **Supervised Learning** algorithms that focus on a target variable. We set *OA_MAIS* as the **Target Node** via the node's **Contextual Menu**, thus defining it as the principal variable of interest.



Furthermore, we designate *OA_MAIS=4-6* as the **Target State** of the **Target Node**, which will subsequently allow us to perform certain analyses with regard to this particular state, i.e. the most serious injuries.



The **Monitor** for *OA_MAIS* reminds us of the marginal distribution of this variable, which shall serve as a reference point for subsequent comparisons.

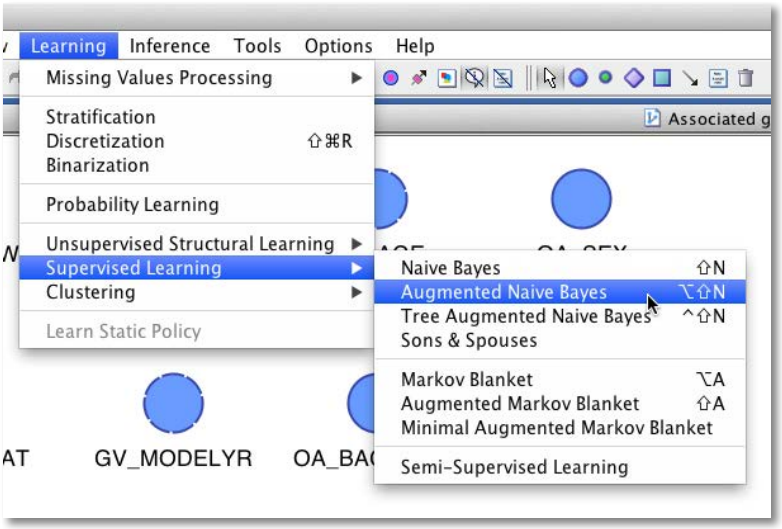


Note that pressing "T", while double-clicking a state within a **Monitor,** also allows setting the **Target Node** and the **Target State**.

Augmented Naive Bayes Learning

Now that the **Target Node** is defined, we have an array of **Supervised Learning** algorithms available. Given the small number of nodes, variable selection is not an issue and hence this should not influence our choice of algorithm. Furthermore, the number of observations does not create a challenge in terms of computational effort.[21] With these considerations, and without going into further detail, we select the **Augmented Naive Bayes** algorithm. The "augmented" part in the name of this algorithm refers to the additional unsupervised search that is performed on the basis of the given naive structure.[22]

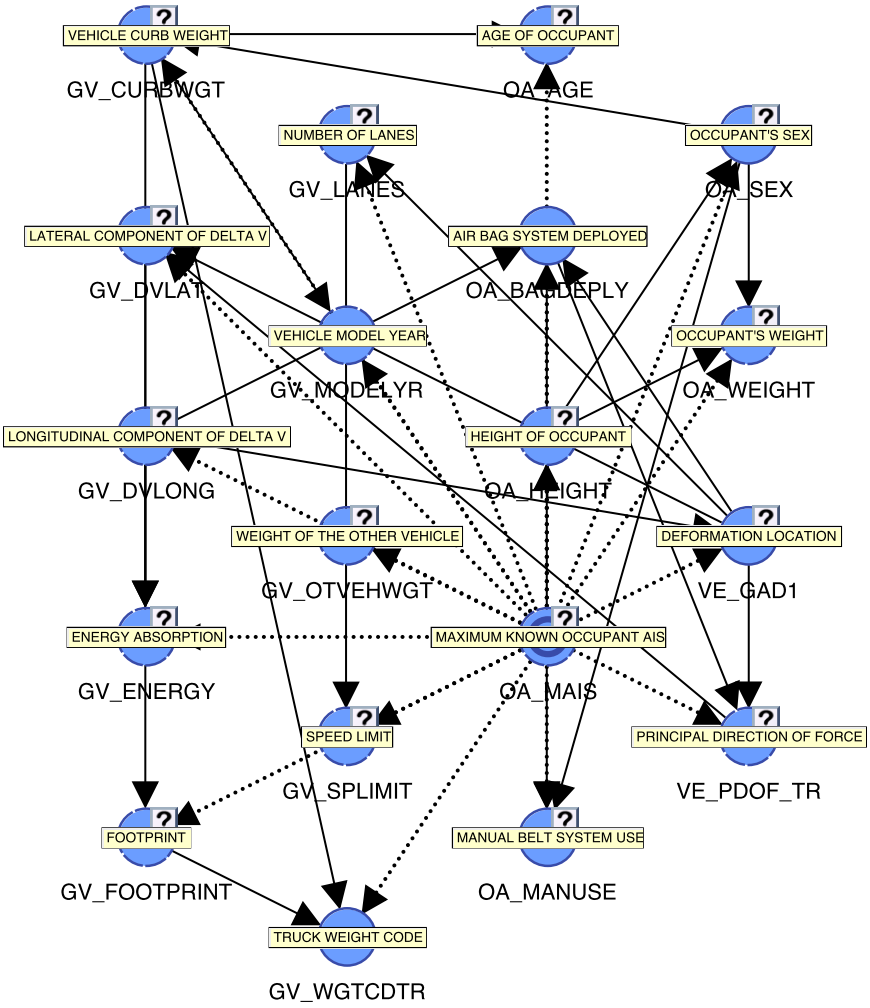
We start the learning process from the menu by selecting **Learning | Supervised Learning | Augmented Naive Bayes**.



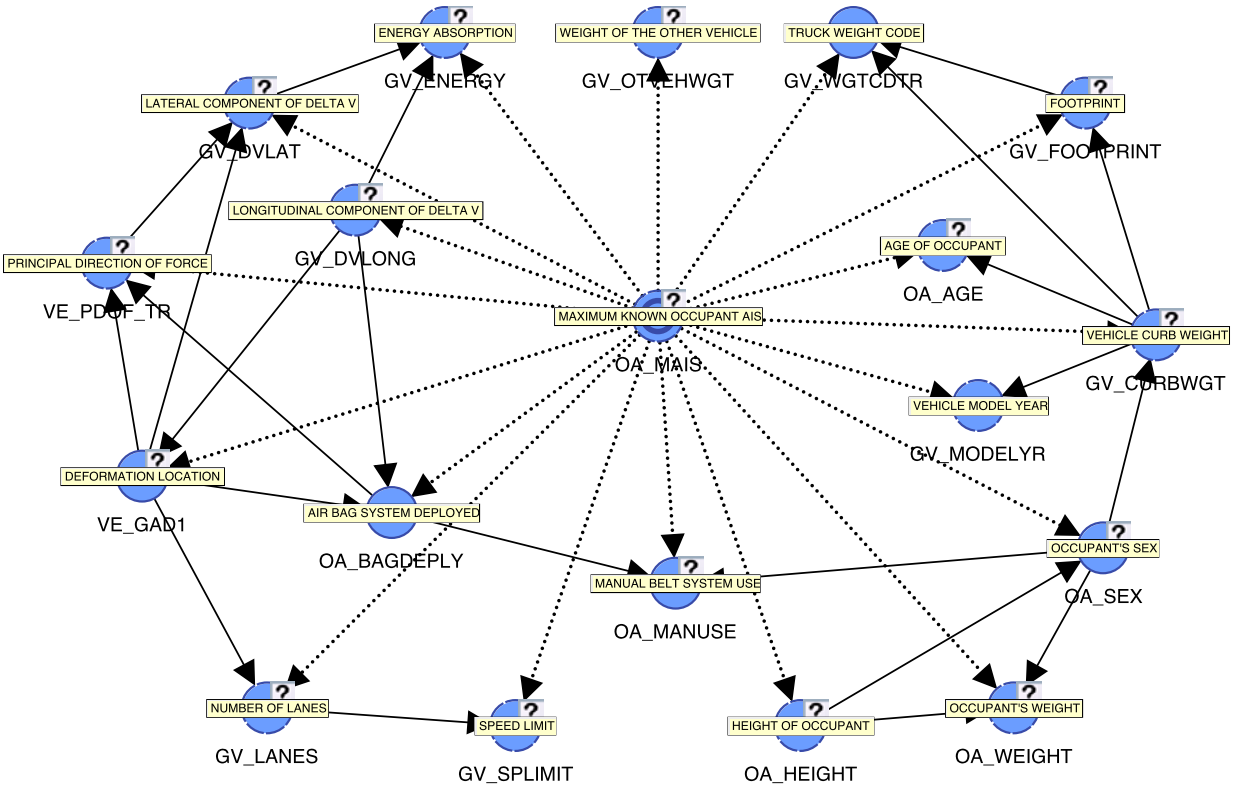
Upon completion of the learning process, BayesiaLab presents the following new network structure.

[21] Thousands of nodes or millions of records would prompt us to consider a more parsimonious approach.

[22] "Naive" refers to a network structure in which the **Target Node** is connected directly to all other nodes. Such a **Naive Bayes** structure is generated by specification, rather than by machine-learning.

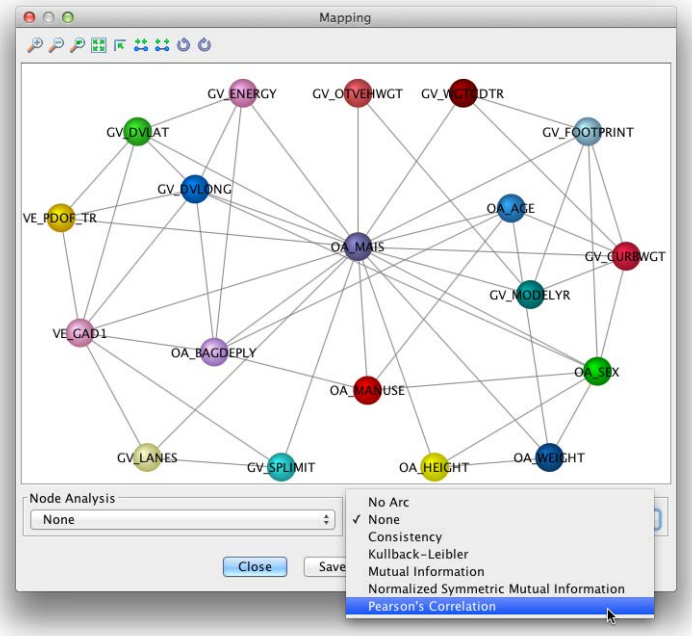


Once again, we apply the **Automatic Layout** algorithm for a clearer view of the network and immediately turn on **Display Node Comments**:

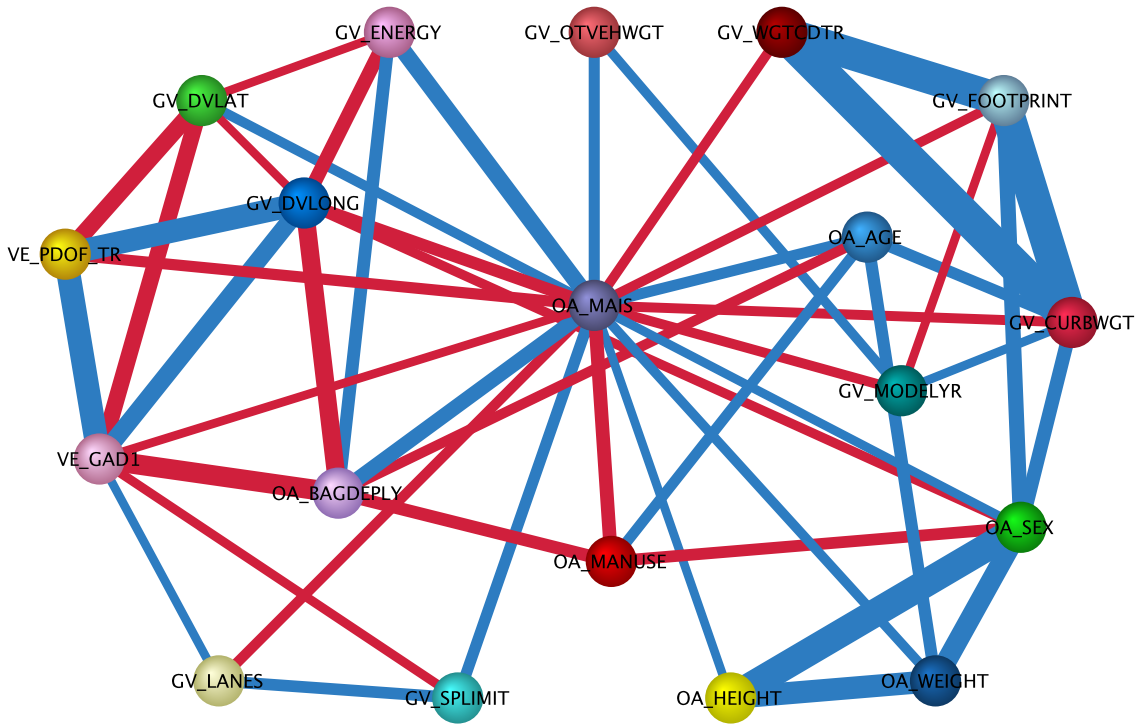


The predefined **Naive Bayes** structure is highlighted with the dotted arcs, while the augmented arcs (from the additional **Unsupervised Learning**) are shown with solid arcs.

Once the network is learned, bringing up the **Mapping** function and selecting **Pearson's Correlation** for the **Arc Analysis** provides an instant survey of the dynamics in the network.



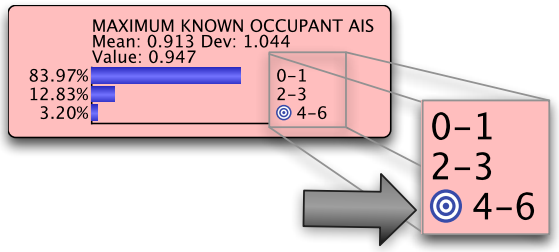
Arc Analysis: Pearson's Correlation



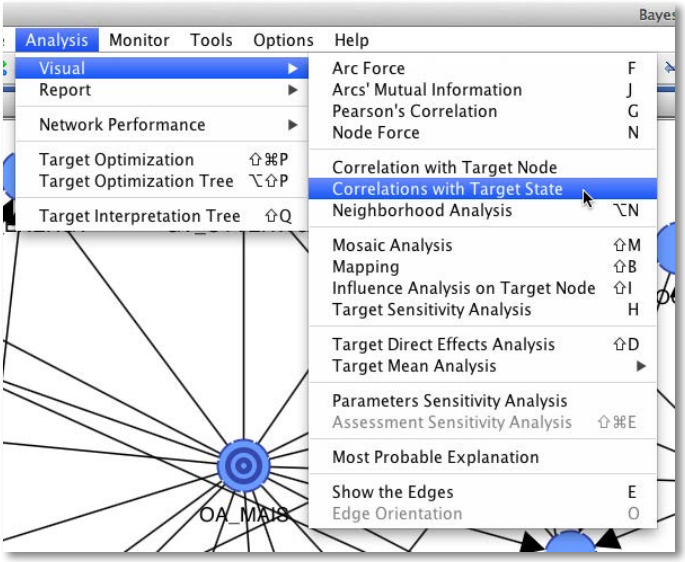
As always, the caveat applies that **Pearson's Correlation** should only be interpreted as such where the assumption of linearity can be made.

BayesiaLab offers another visualization function that helps in identifying the basic patterns of the relationships between the nodes and the **Target State** of the **Target Node**.

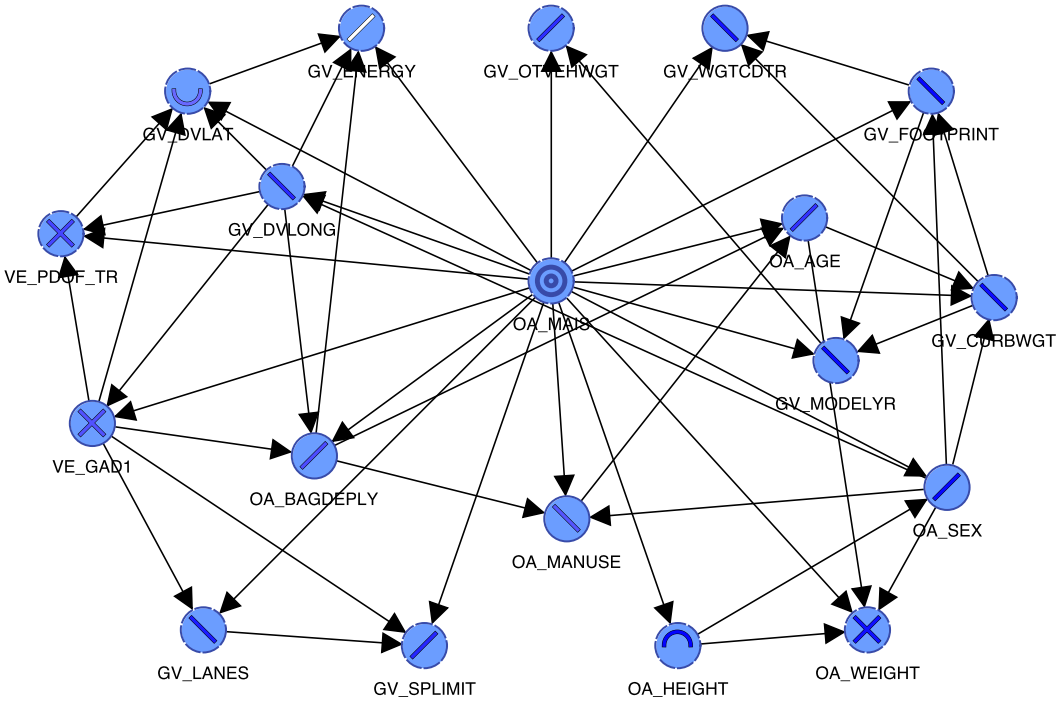
As we had previously specified the **Target State** as *OA_MAIS=4-6* (indicated by the "bullseye"), we can now perform an analysis with regard to this particular state.



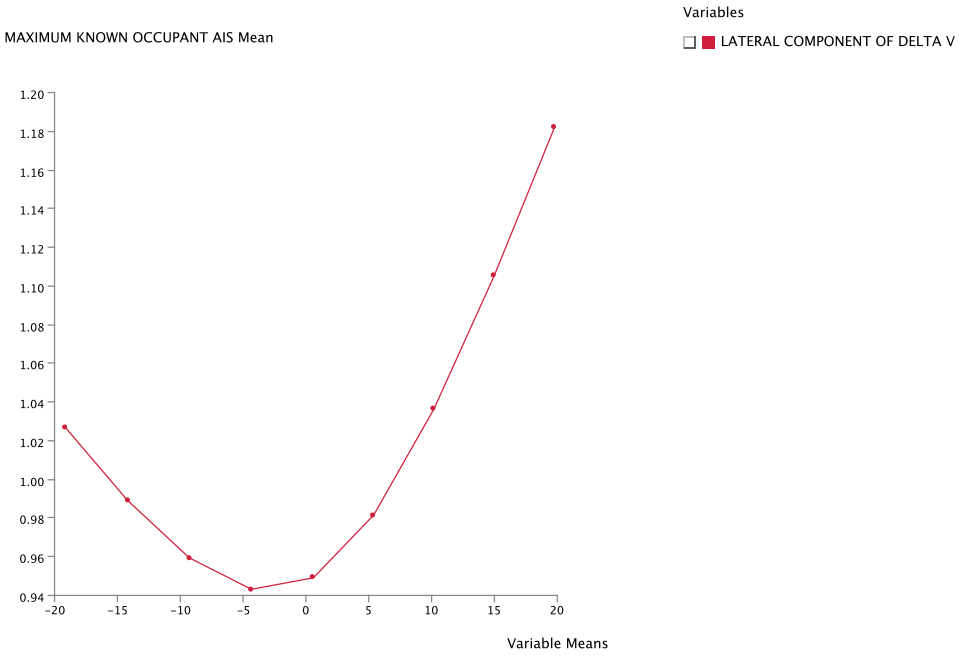
From the main menu, we select **Analysis | Visual | Correlations with the Target State:**



The symbols in each node provide a succinct overview of the basic relationship patterns with the **Target State**. "X" marks the nodes for which none of the basic patterns match. Also, the color of the symbol in each node reflects the **Binary Mutual Information** with regard to the **Target State** of the **Target Node**. White indicates the highest, dark blue represents the lowest values.



For example, this graph indicates that the relationship between *GV_DVLAT* and *OA_MAIS* follows a U-shape. Performing a **Target Mean Analysis** (**Analysis | Visual | Target Mean Analysis | Standard**) can quickly confirm this:



This is just one more of many possible network views which are all meant to facilitate our understanding of the manifold dynamics. Counterintuitive relationships would certainly become obvious at this point and should prompt a review of all steps taken thus far.

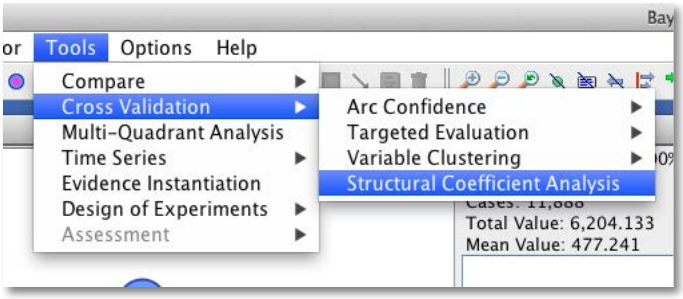
Structural Coefficient Analysis

Even if we find everything to be reasonable, we will need to ask the question whether this model does include all important interactions. Did we learn a reliable model that can be generalized? Is the model we built the best one among all the possible networks? Is the model possibly overfitted to the data?

Some of these questions can be studied with BayesiaLab's **Structural Coefficient Analysis**. Before we delve into this function, we first need to explain the **Structural Coefficient** (**SC**). It is a kind of significance threshold that can be adjusted by the analyst and that influences the degree of complexity of the induced network.

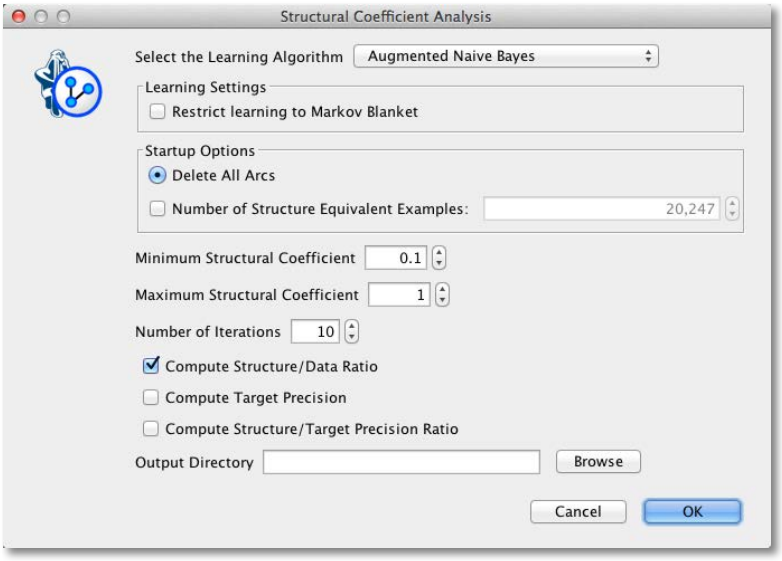
By default, this **Structural Coefficient** is set to 1, which reliably prevents the learning algorithms from overfitting the model to the data. In studies with relatively few observations, the analyst's judgment is needed for determining a possible downward adjustment of this parameter. On the other hand, when data sets are large, increasing the parameter to values greater than 1 will help manage the network complexity.

BayesiaLab can systematically examine the question of optimal complexity with the **Structural Coefficient Analysis**. In our example, we wish to know whether a more complex network—while avoiding overfit—would better capture the dynamics of our domain. **Structural Coefficient Analysis** generates several metrics that can help in making this trade-off between complexity and fit: **Tools | Cross Validation | Structural Coefficient Analysis**.

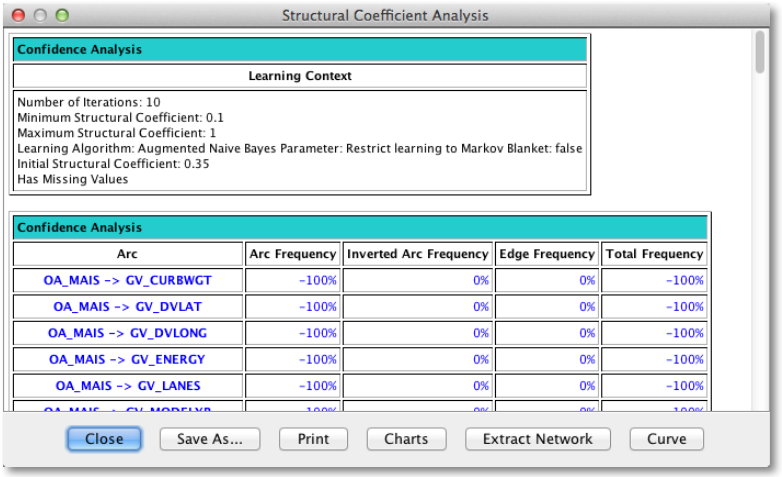


BayesiaLab prompts us to specify the range of the **Structural Coefficient** to be examined and the number of iterations to be performed. It is worth noting that the **Minimum Structural Coefficient** should not be set to 0, or even close to 0. A value of 0 would lead to learning a fully connected network, which can take a long time depending on the number of variables, or even exceed the memory capacity of the computer running BayesiaLab.

Number of Iterations determines the steps to be taken within the specified range of the **Structural Coefficient**. We leave this setting at the default level of 10. For metrics, we select **Compute Structure/Data Ratio**, which we will subsequently plot.

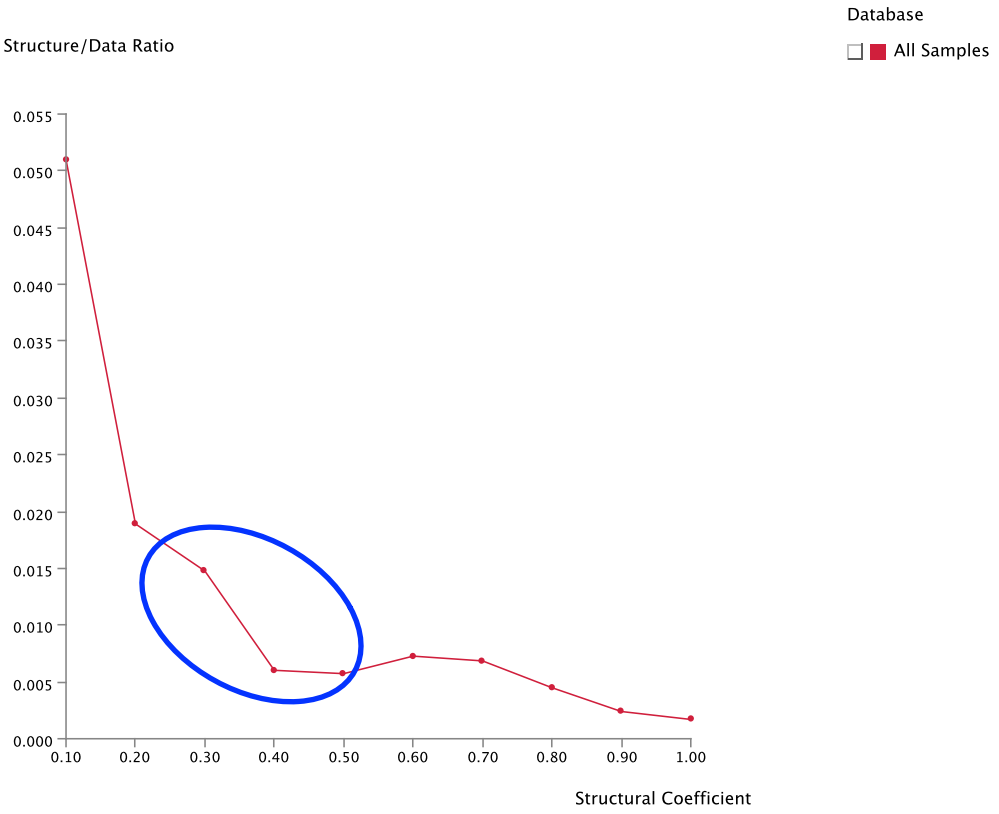


The resulting report shows how the network changes as a function of the **Structural Coefficient**. For other analyses based on **Cross-Validation,** this report can be used to determine the degree of confidence we should have in any particular arc in the structure.



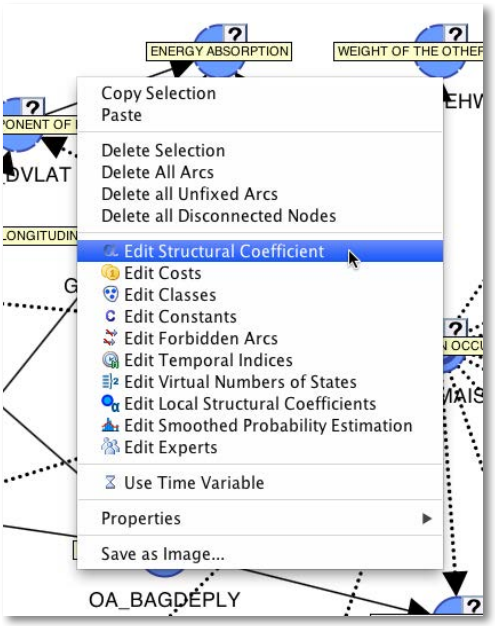
Our overall objective here is to determine the correct level of network complexity for representing the interactions with the **Target Node** without the overfitting of data. By clicking **Curve** we can plot the **Structure/**

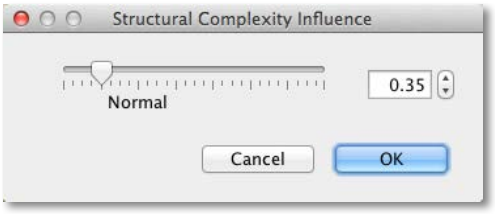
Data Ratio (y-axis) over the **Structural Coefficient** (x-axis).



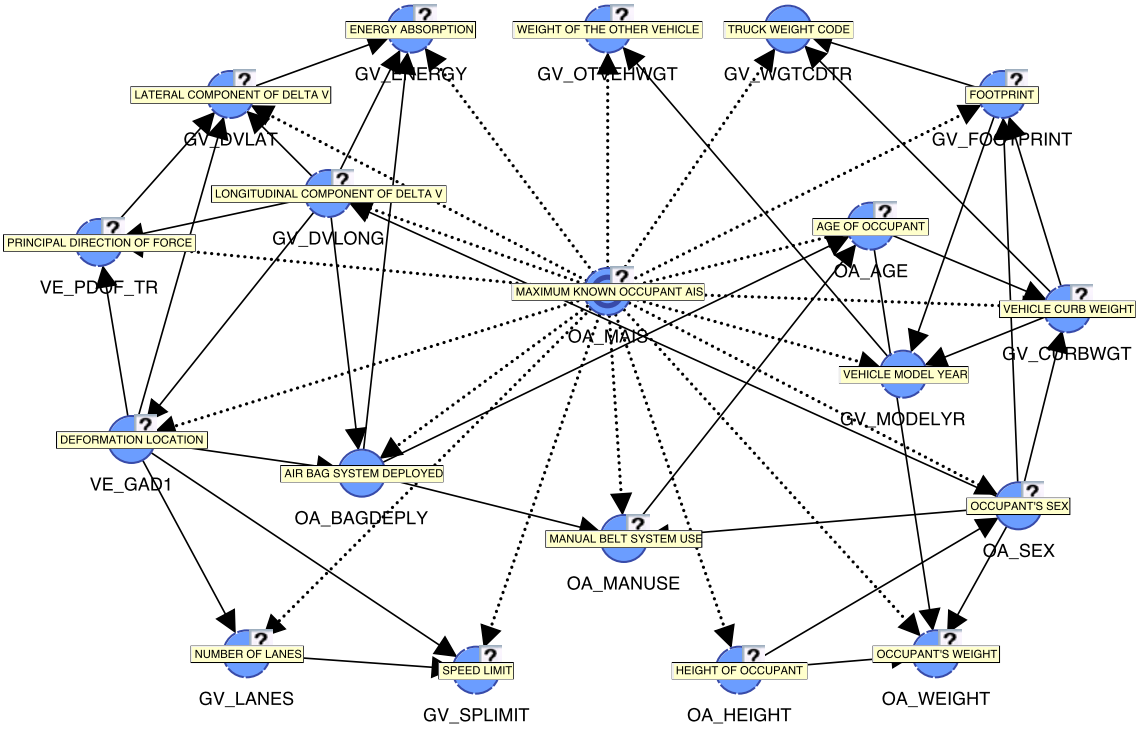
Typically, the "elbow" of the L-shaped curve above identifies a suitable value for the **Structural Coefficient** (SC). The visual inspection suggests that an SC value of around 0.35 would be a good candidate. Further to the left of this point, e.g. **SC**≈0.1, the complexity of the model increases much more than the likelihood of the data given the model. This means that arcs would be added without any significant gain in terms of better representing the data. That is the characteristic pattern of overfitting, which is what we want to avoid.

Given the results from this **Structural Coefficient Analysis**, we can now relearn the network with an **SC** value of 0.35. The **SC** value can be set by right-clicking on the background of the **Graph Panel** and then selecting **Edit Structural Coefficient** from the **Contextual Menu**, or via the menu, i.e. **Edit | Edit Structural Coefficient**.

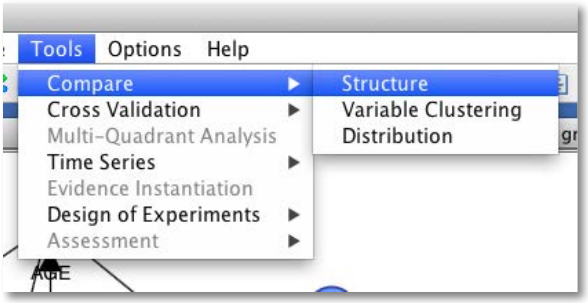




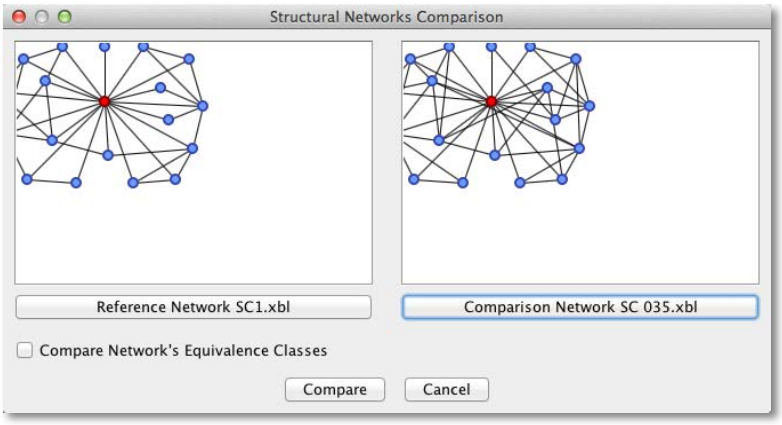
Once we relearn the network, using the same **Augmented Naive Bayes** algorithm as before, we obtain a more complex network.



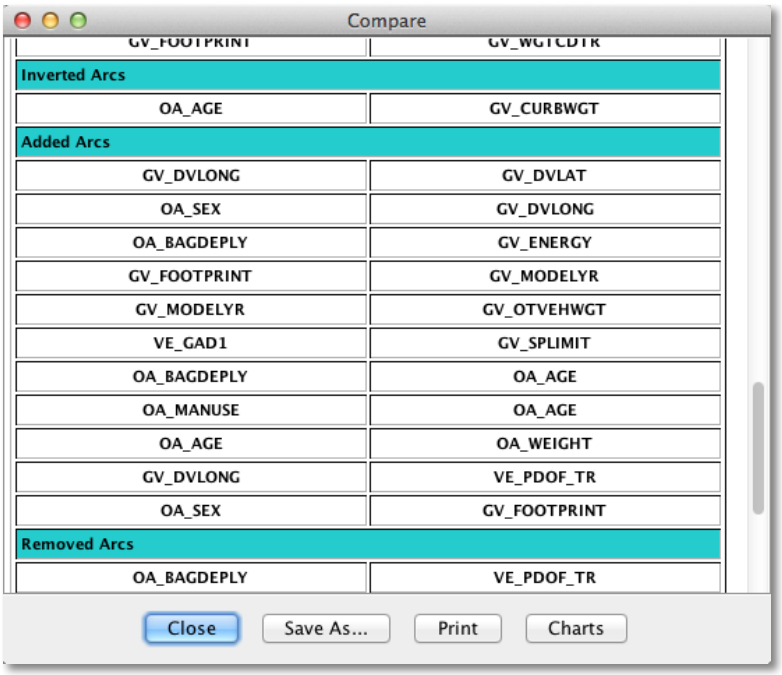
If we save the original network and the new network under different file names, we can use **Tools | Compare | Structure** to highlight the differences between both.



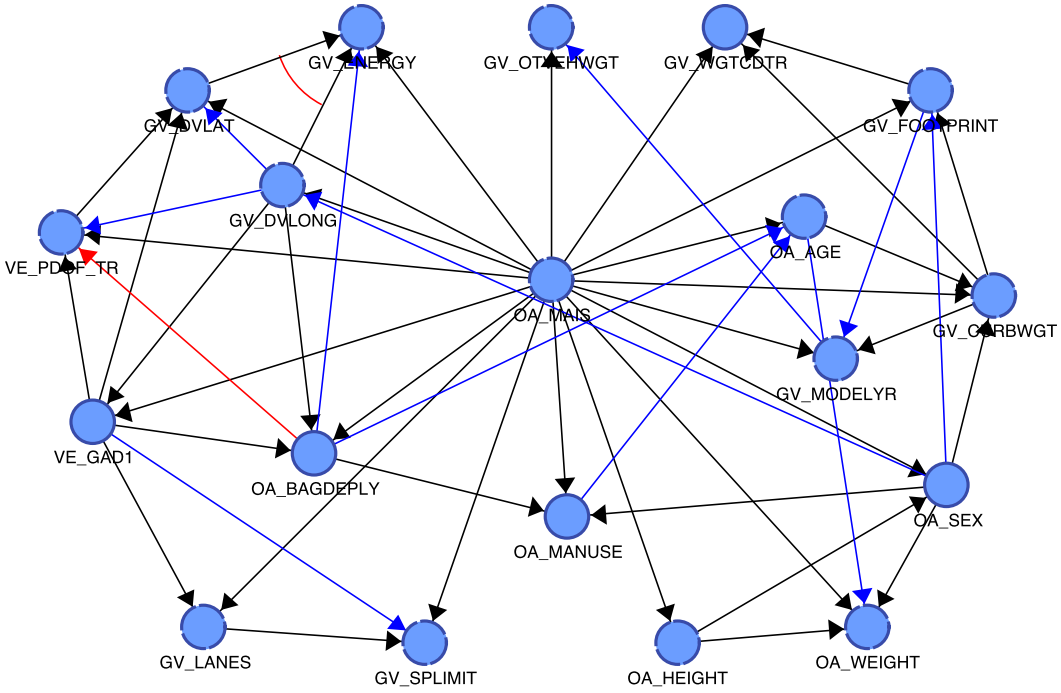
We choose the original network (SC=1) as the **Reference Network** and the newly learned network (SC=0.35) as the **Comparison Network**.



Upon selection, a table provides a list of common arcs and those arcs that have been added in the **Comparison Network**:



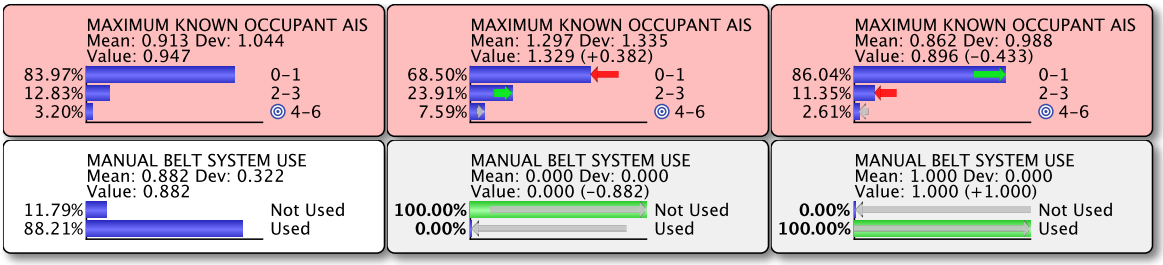
Clicking **Charts** provides a visual representation of these differences. The additional arcs, compared to the original network, are now highlighted in blue. Conversely, deleted arcs are shown in red.



Example 2: Seat Belt Usage

Similar to the analysis we performed earlier on *GV_LANES*, *GV_SPLIMIT*, and *VE_GAD1*, we will now examine the **Target Node**, *OA_MAIS*. We select and double-click the nodes *OA_MAIS* and *OA_MANUSE (Manual Belt System Use)* to bring them up in the **Monitor Panel**.

Initially, the **Monitors** show *OA_MAIS* and *OA_MANUSE* with their marginal distributions (1st column from left). We now set evidence on *OA_MANUSE* to evaluate the changes to *OA_MAIS*. Our experience tells us that not wearing a seatbelt is associated with an increased risk of injury in an accident.

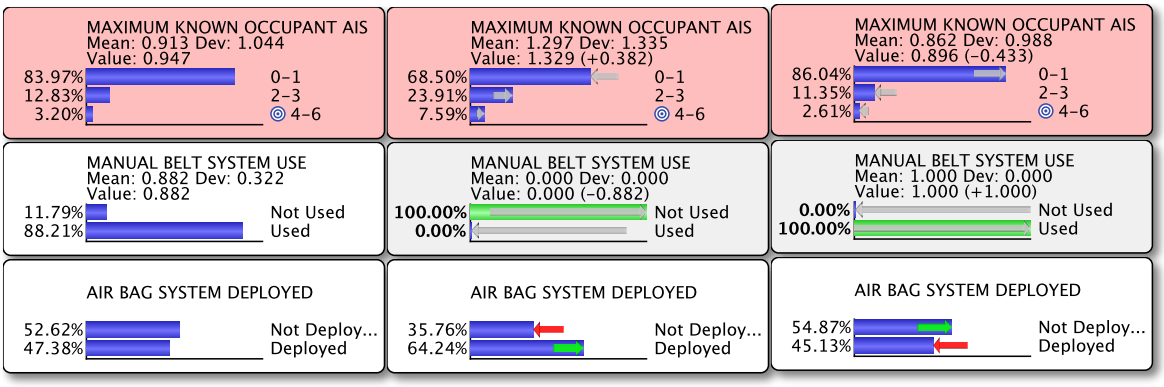


Indeed, this is precisely what we observe. For example, for *OA_MANUSE=Not Used* the probability of no/minor injury is 68.50% (2nd column). On the other end of the injury spectrum, the probability of serious injuries *(OA_MAIS=4-6)* is 7.59%.

The situation is much better for *OA_MANUSE=Used* (3rd column). The probability of no/minor injury is higher (+17.5 percentage points), and the probability of serious injury is much lower (-5 percentage points). These results appear intuitive and are in sync with our domain knowledge.

However, does this confirm that wearing a seat belt reduces the risk of sustaining a serious injury by roughly two-thirds? Not yet.

We can further examine this by including the variable *OA_BAGDPLY (Air Bag System Deployed)*. The bottom-left **Monitor** shows the marginal distribution of *OA_BAGDPLY*.



To compare the conditions *OA_MANUSE=Not Used* and *OA_MANUSE=Used*, we first set evidence on *Not Used* (2nd column, 2nd row). The posterior distribution of *OA_MAIS* has an expected value of 1.329 and the probability of *OA_BAGDPLY=Deployed* has increased to 64.24%.[23]

Setting the evidence *OA_MANUSE=Used* (3rd column, 2nd row) changes the expected value of *OA_MAIS* to 0.896, a decrease of 0.433, but it is also associated with a much different posterior distribution of *OA_BAGDPLY*, which now has a lower probability for *Deployed*. How should we interpret this?

As it turns out, many airbag systems are designed in such a way that their deployment threshold is adjusted when occupants are not wearing seat belts.[24] This means that not wearing a seat belt *causes* the airbag to be triggered differently. So, beyond the *direct* effect of the seat belt, whether or not it is worn *indirectly* influences the injury risk via the trigger mechanism of the airbag system.

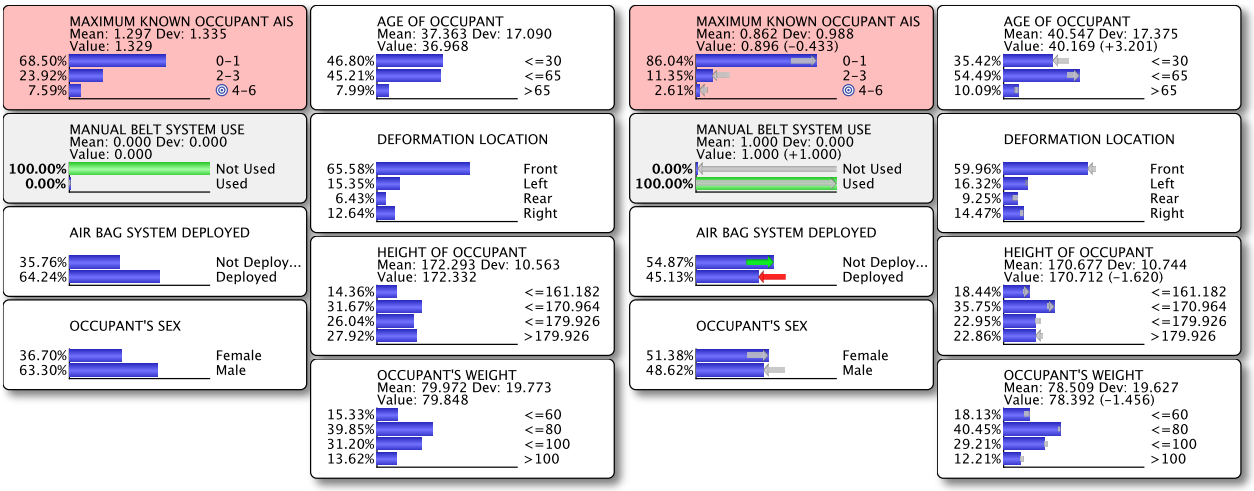
[23] *OA_MAIS* has an ordinal scale, rather than a numerical scale. As such, we need to be careful with interpreting the expected values (means) of *OA_MAIS*.

[24] Bosch Automotive Handbook (2011), p. 933.

Covariate Imbalance

Beyond the link between seat belt and airbag, there are numerous other relevant relationships. For instance, seat belt users are more likely to be female, they are older and they are, for some unknown reason, less likely to be involved in a frontal crash, etc.

Similarly to the earlier example, the evidence set on *OA_MANUSE* is propagated omnidirectionally through the network, and the posterior distributions of all nodes are updated. The **Monitors** below show the difference between the evidence *OA_MANUSE=Not Used* (left set of panels) and *OA_MANUSE=Used (right)*.



This highlights that seat belt users and non-users are quite different in their characteristics and thus not directly comparable. So, what is the benefit of the seat belt, if any?

In fact, this is a prototypical example of the challenges associated with observational studies. By default, observational studies, as the one here, permit only *observational inference*. However, performing *observational inference* on *OA_MAIS* given *OA_MANUSE* is of limited use for the researcher or the policymaker. How can we estimate causal effects with *observational* data? How can we estimate the (hopefully) positive effect of the seat belt?

Within a traditional statistical framework, a number of approaches would be available to address such differences in characteristics, including stratification, adjustment by regression, and covariate matching (e.g. Propensity Score Matching). For this tutorial, we will proceed with a method that is similar to covariate matching; however, we will do this within the framework of Bayesian networks.

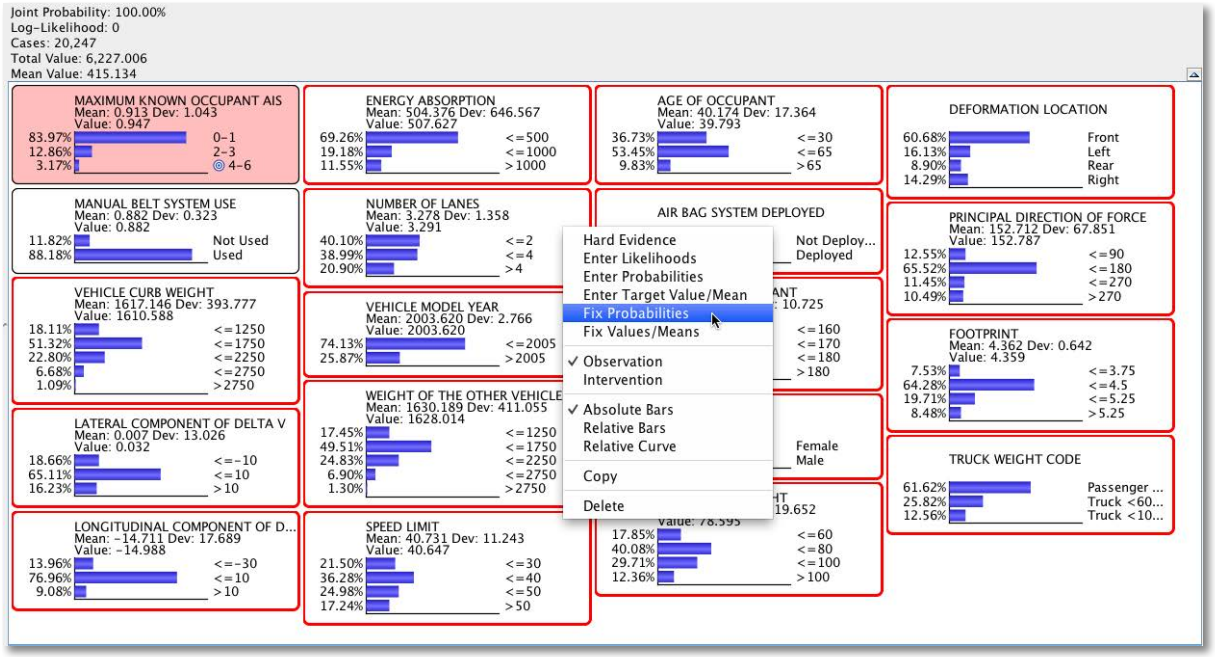
Likelihood Matching with BayesiaLab[25]

We will now briefly introduce the **Likelihood Matching** (LM) algorithm, which was originally implemented in the BayesiaLab software package for "fixing" probability distributions of an arbitrary set of variables, thus allowing to easily define complex sets of soft evidence. The LM algorithm searches for a set of likelihood distributions, which, when applied on the joint probability distribution (JPD) encoded by the Bayesian network, allows obtaining the posterior probability distributions defined (as constraints) by the user.

This allows us to perform matching across all covariates, while taking into account all their interactions, and thus to estimate the direct effect of *OA_MANUSE*. We will now illustrate a manual approach for estimating the effect; in the next chapter we will show a more automated approach with the **Direct Effects** function.

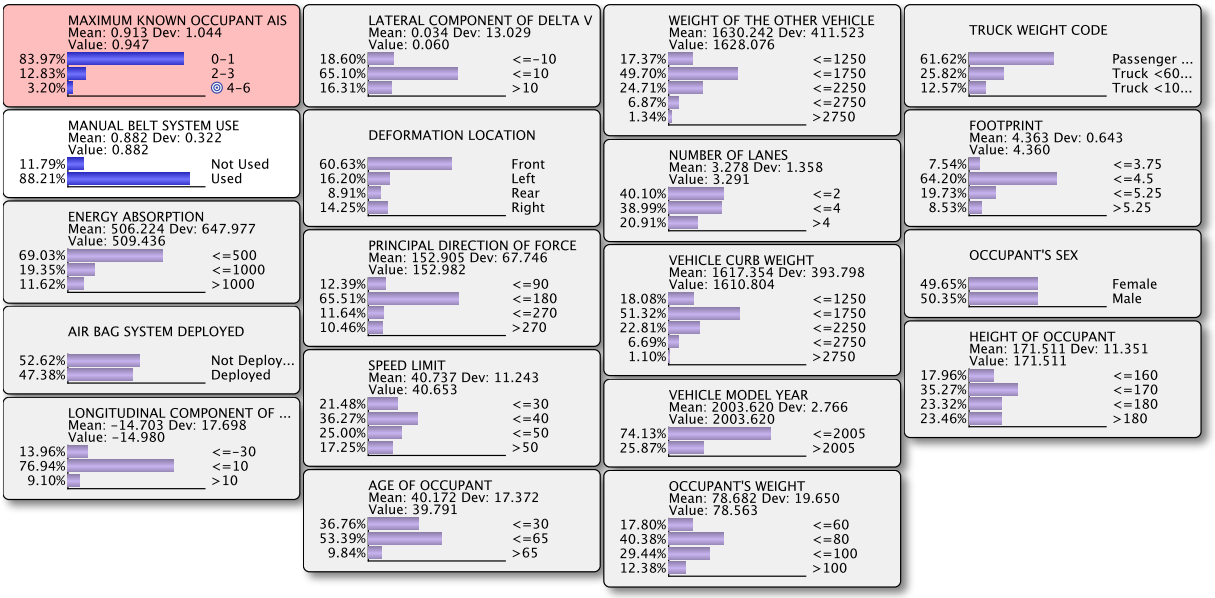
Fixing Distributions

We start with the marginal distributions of all nodes. Next, we select all covariate nodes and then right-click on any one of them. From the **Contextual Menu** we pick **Fix Probabilities**.

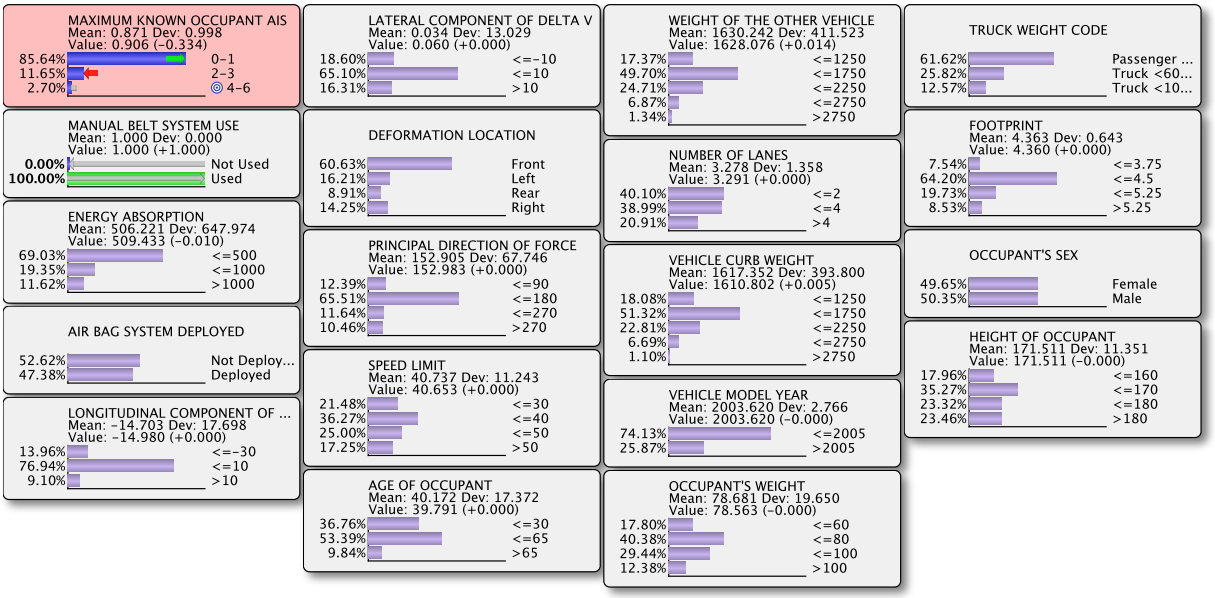


This "fixes" the (marginal) distributions of these covariate nodes. Their new fixed condition is indicated by the purple color of the bars in the **Monitors**.

[25] Likelihood Matching with BayesiaLab is explained in more detail in Conrady and Jouffe (2011)



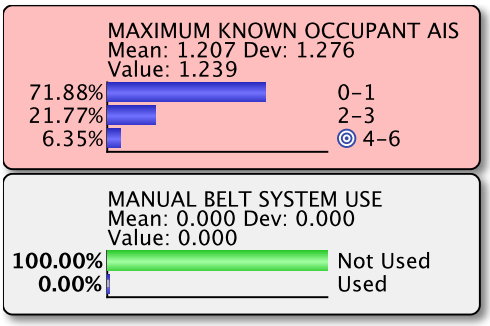
If we now set the evidence on *OA_MANUSE=Used*, we will see no significant changes in these fixed covariate distributions. The **Likelihood Matching** algorithm has indeed obtained materially equivalent distributions.



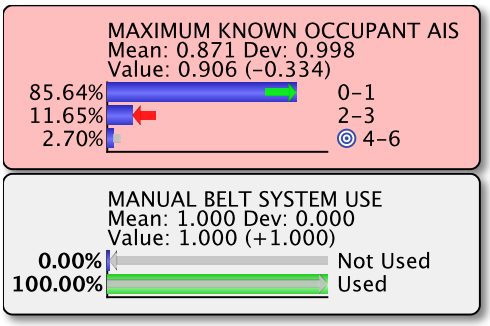
Causal Inference

Given that the covariate distributions remain fixed, we can now exclusively focus on the **Target Node** *OA_MAIS*:

OA_MANUSE=Not Used



OA_MANUSE=Used



If we make the assumption that no other unobserved confounders exist in this domain, we will now be able to give the change in the distribution of *OA_MAIS* a causal interpretation. This would represent the difference between forcing *all drivers to wear a seat belt* versus forcing *all of them not to do so*.

More formally we can express such an intervention with Pearl's do-operator[26], which reflects the active setting of a condition (or intervention) versus merely observing a condition:

Observational Inference:	`P(OA_MAIS=4-6|OA_MANUSE=Used)=2.61%`[27]
	`P(OA_MAIS=4-6|OA_MANUSE=Not Used)=7.59%`
Causal Inference:	`P(OA_MAIS=4-6|do(OA_MANUSE=Used))=2.70%`[28]
	`P(OA_MAIS=4-6|do(OA_MANUSE=Not Used))=6.35%`

[26] Pearl (2009).

[27] In words: the probability of *OA_MAIS* taking on the value "4-6" given ("|") that *OA_MANUSE* is **observed** as "Used."

[28] In words: the probability of *OA_MAIS* taking on the value "4-6" given ("|") that *OA_MANUSE* is **actively set (by intervention) to** "Used."

From these results we can easily calculate the causal effect:

```
P(OA_MAIS=4-6|do(OA_MANUSE=Used))-P(OA_MAIS=4-6|do(OA_MANUSE=Not Used))=
=-3.65%
```

We conclude that this difference, -3.65 percentage points, is the "seat belt effect" with regard to the probability of serious injury. Analogously, the effect for moderate injuries is -10.12 percentage points. Finally, for no/minor injuries, there is a positive effect of 13.79 percentage points.

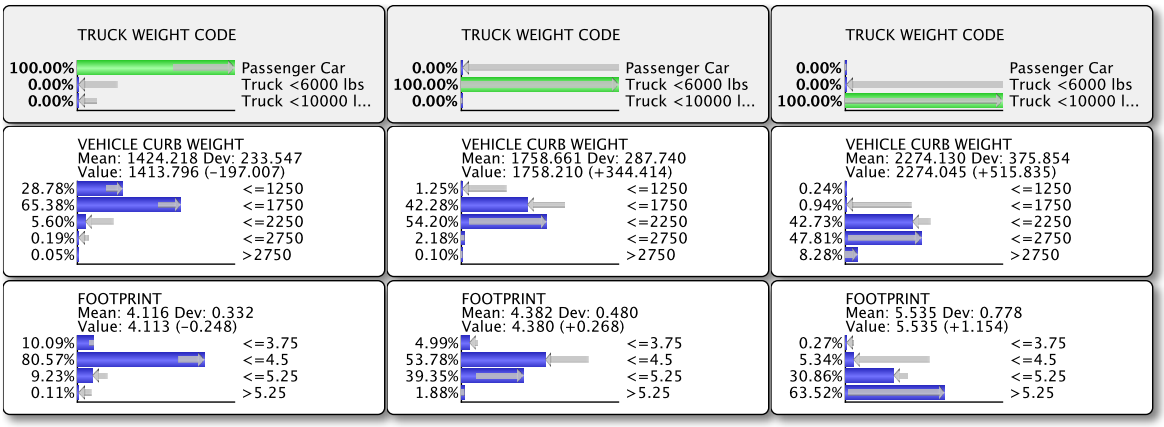
With "causal inference" formally established and implemented through this workflow, we can advance to the principal question of this study, the influence of vehicle size and weight on injury risk.

Effect of Weight and Size on Injury Risk

Our previous example regarding seat belt usage illustrated the challenges of estimating causal effects from observational data. It became clear that the interactions between variables play a crucial role. Furthermore, we have shown how **Likelihood Matching**, under the assumption of no unobserved covariates, allows us to estimate the causal effect from observational data.

Lack of Covariate Overlap

Before we continue with the effect estimation of vehicle weight and size, we will briefly review the distributions of the nodes under study, *GV_WGTCDTR*, *GV_CURBWGT* and *GV_CURBWGT* as a function of *GV_WGTCDTR (Truck Weight Code)*. We can use this code to select classes of vehicles, including *Passenger Car, Truck <6,000 lbs.*, and *Truck <10,000 lbs.*



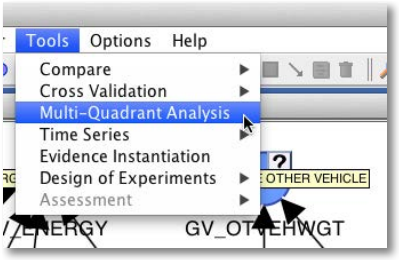
We can see that the distributions of *GV_CURBWGT* and *GV_FOOTPRINT* are very different between these vehicles classes. For *GV_WGTCDTR=Passenger Car*, we have virtually no observations in the two highest bins of *GV_CURBWGT* (left column), whereas for large trucks *(GV_WGTCDTR= Truck <10,000 lbs.)* the two bottom bins are almost empty (right column). We see a similar situation for *GV_FOOTPRINT*.

This poses two problems: Firstly, we now have a rather "coarse" discretization of values within each class of vehicles, which could potentially interfere with estimating the impact of small changes in these variables. Secondly, and perhaps more importantly, the lack of covariate overlap may prevent the **Likelihood Matching** algorithm from converging on matching likelihoods between variables. As a results, we might not be able carry out any causal inference. Hence, we must digress for a moment and resolve this issue in the following chapter on **Multi-Quadrant Analysis**.

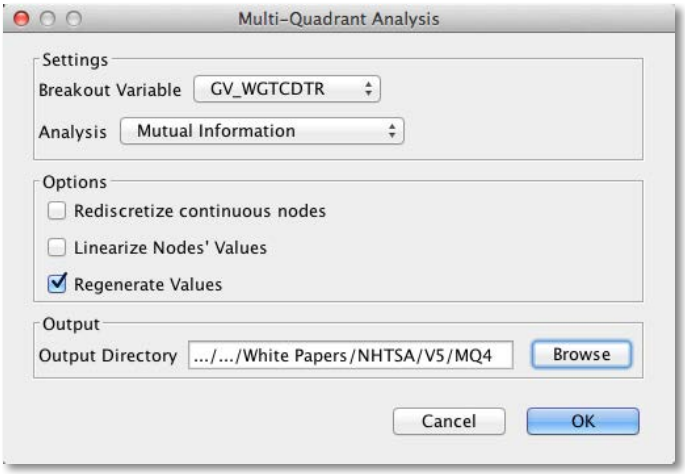
Multi-Quadrant Analysis

BayesiaLab offers a way to conveniently overcome this fairly typical problem. We can perform a **Multi-Quadrant Analysis** that will automatically generate a Bayesian network for each vehicle subset, i.e. *Passenger Car*, *Truck <6,000 lbs.*, and *Truck <10,000 lbs*.

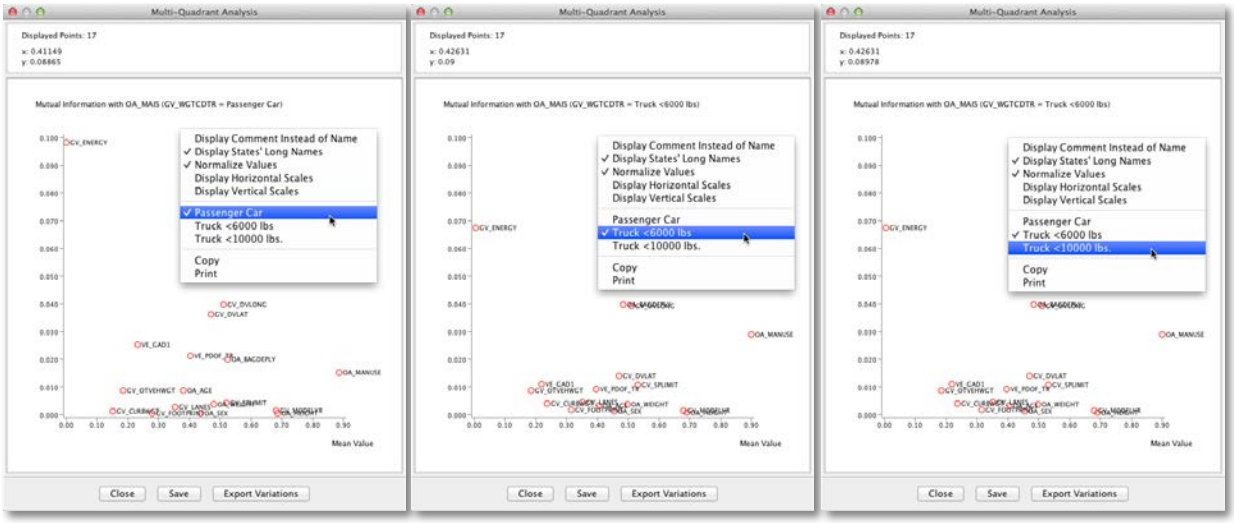
The **Multi-Quadrant Analysis** can be started from the menu via **Tools | Multi-Quadrant Analysis**:



We now need to set several options: The **Breakout Variable** (or selector variable) must be set to *GV_WGTCDTR*, so we will obtain one network for each state of this node. By checking **Regenerate Values**, the values of each state of each (continuous) node are re-computed from the data of each subset database, as defined by the **Breakout Variable**. **Rediscretize Continuous Nodes** is also an option in this context, but we will defer this step and later perform the rediscretization manually. Finally, we set an output directory where the newly generated networks will be saved.



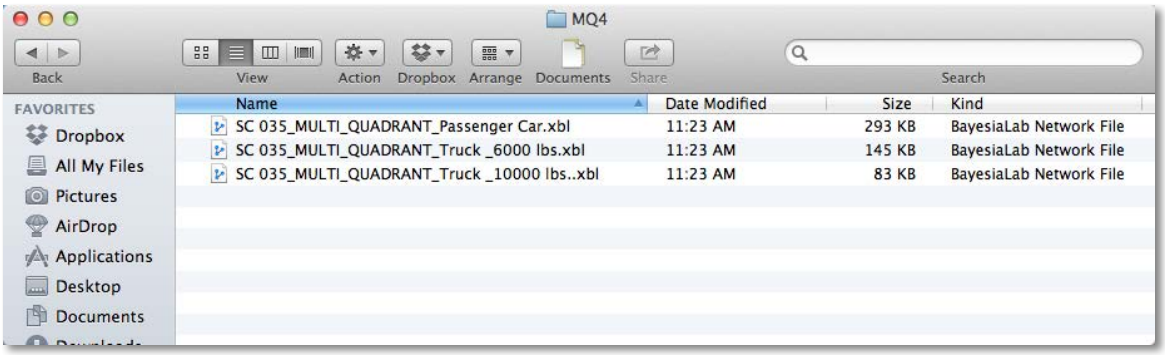
Upon completion of this step, BayesiaLab presents a set of plots, one for each state of the **Breakout Variable**, *GV_WGTCDTR*.



These plots allow us to compare the **Mutual Information** of each node in the network with regard to the **Target Node**, *OA_MAIS*. These plots can be very helpful in determining the importance of individual variables in the context of their respective subset.

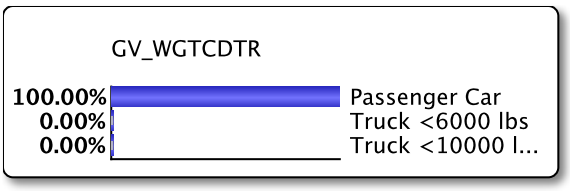
In our example, we will omit going into further detail here and instead inspect the newly generated networks. We can simply open them directly from the previously specified location. The new file names follow this format:

`original file name & "_MULTI_QUADRANT_" & Breakout Variable State & ".xbl"`



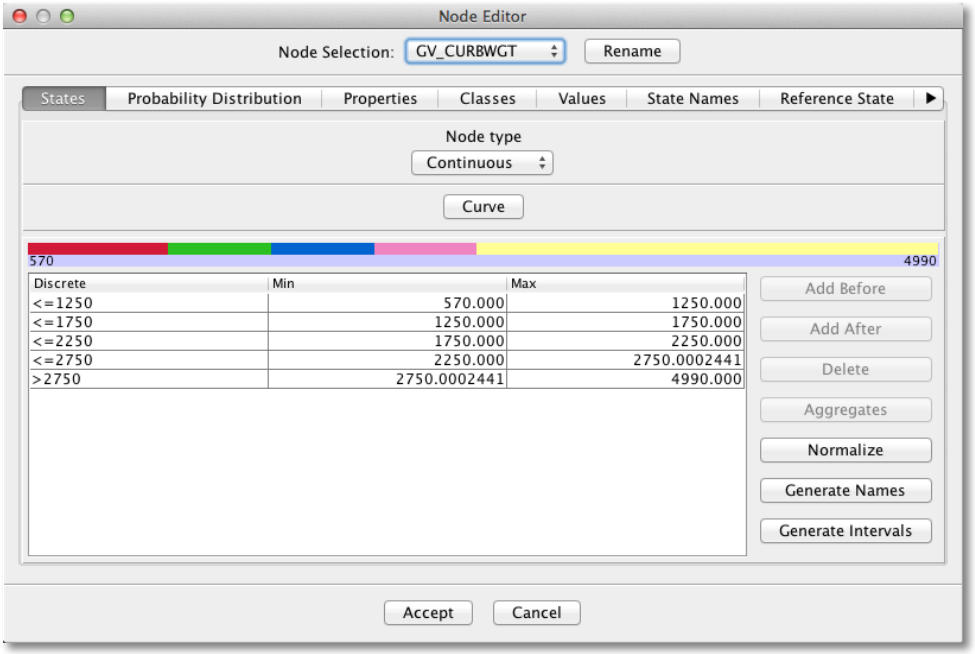
Vehicle Class: Passenger Car

When opening the *Passenger Car* file, we will notice that the structure is exactly the same as in the network that applied to the whole set. However, examining the **Monitors** will reveal that the computed probabilities now apply to the *Passenger Car* subset only.

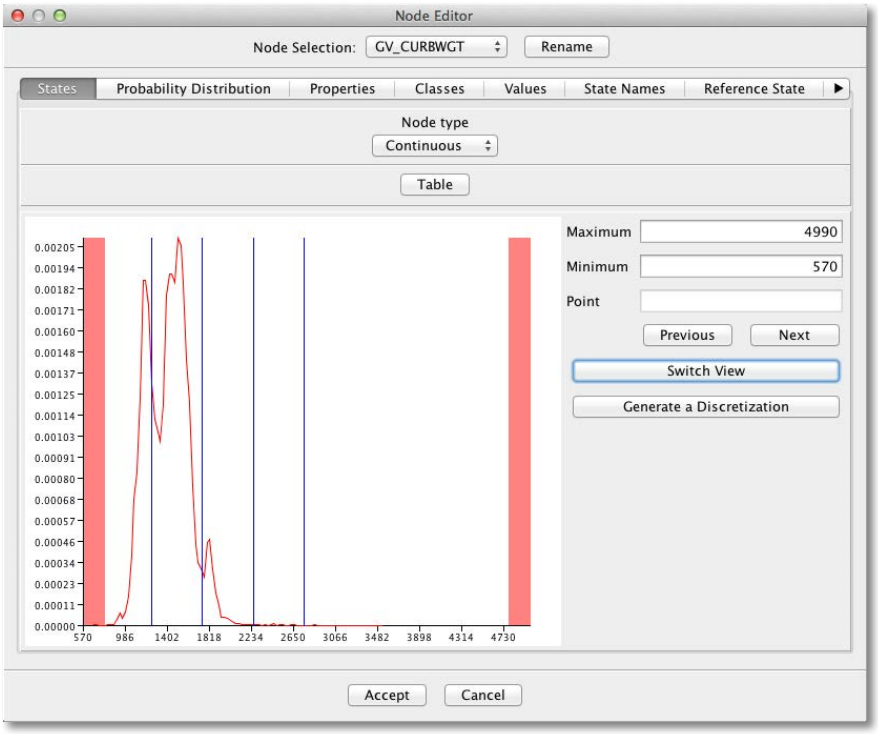


Since *GV_WGTCDTR* now only contains a single state at 100%, i.e. the same for each record in the *Passenger Car* subset, we can go ahead and remove this node from the network. We first select the node and then hit the delete key.

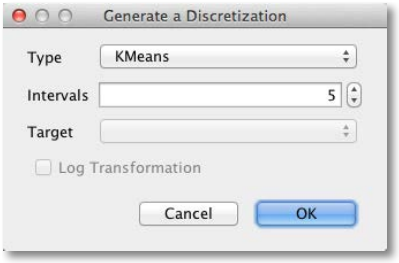
Now we can focus our attention on the two variables that prompted the **Multi-Quadrant Analysis**, namely *GV_CURBWGT* and *GV_FOOTPRINT*. By double-clicking on *GV_CURBWGT* we bring up the **Node Editor**, which allows us to review and adjust the discretization of this particular node.



We could directly change the thresholds of the bins in this table. However, it is helpful to bring up the probability density functions (PDF) by clicking on **Curve**.

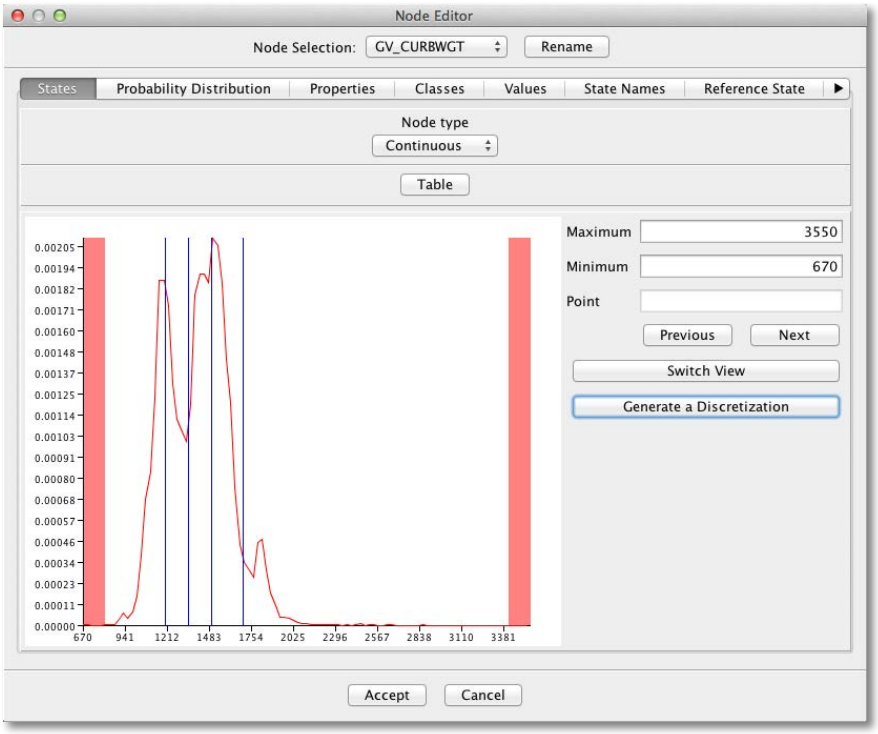


We can now use one of BayesiaLab's discretization algorithms to establish new bins that are more suitable for the distribution of *GV_CURBWGT* within the *Passenger Car* subset. Here, we will use the **K-Means** algorithm with five intervals.[29]



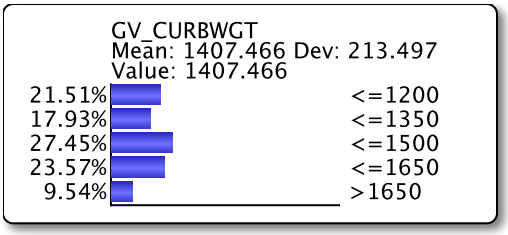
Upon discretization, the new bin intervals are highlighted in an updated PDF plot.

[29] Several of our other white papers provide recommendations regarding the choice of discretization algorithms and intervals.

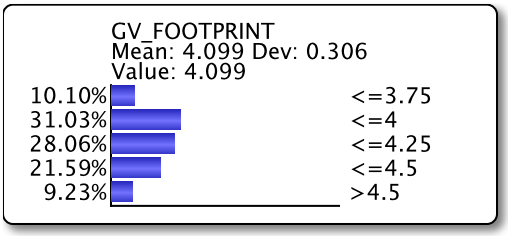


Instead of automatically determining the bins, we could also manually specify them. In our case, we take the bins proposed by the **K-Means** algorithm and round them to the nearest 50 kg: 1200.611816 → 1200; 1346.44043 → 1350, etc.

The final distribution can be seen in the **Monitor** below.



We apply the same approach for *GV_FOOTPRINT* and present the result in the following **Monitor**.



Both discretizations now adequately capture the underlying distributions of the respective variables within the subset *Passenger Car*.

We now proceed to our principal question of how *GV_CURBWGT* and *GV_FOOTPRINT* affect *OA_MAIS*. It might be tempting to immediately apply the same process as with the seat belt effect estimation, i.e. fix all covariates with **Likelihood Matching** and then simulate the response of *OA_MAIS* as a function of changing values of *GV_CURBWGT* and *GV_FOOTPRINT*. Indeed, we wish to keep everything else the same in order to determine the exclusive **Direct Effects** of *GV_CURBWGT* and *GV_FOOTPRINT* on *OA_MAIS*.

However, fixing must not apply to all nodes. The reason being, following the definition of kinetic energy,

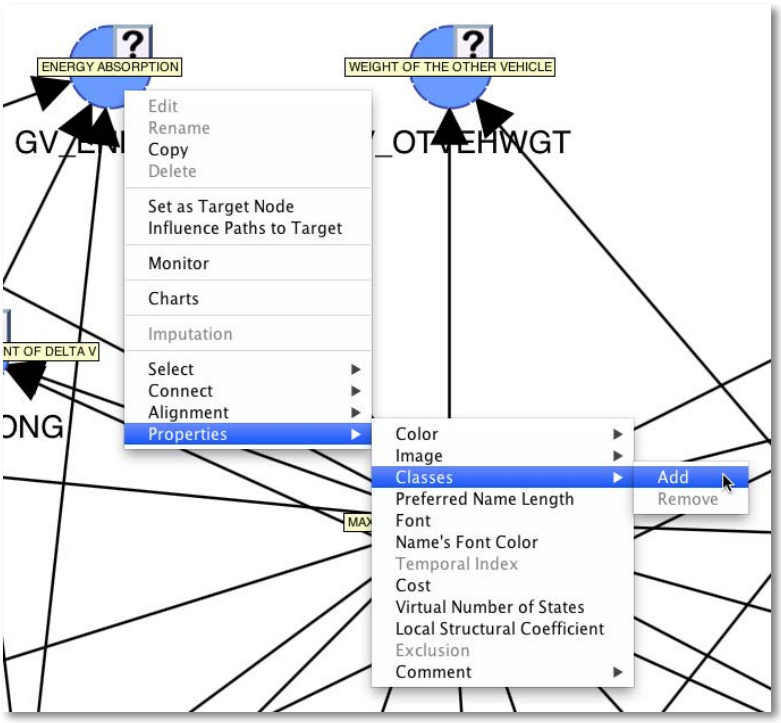
$E = \frac{1}{2} m \cdot v^2$, *GV_ENERGY* is a function of *GV_CURBWGT, GV_DVLONG* and *GV_DVLAT*.

This means that *GV_ENERGY* cannot be maintained at a fixed distribution as we test varying values of *GV_CURBWGT* and *GV_FOOTPRINT*.

It is important to understand that the relationship of *GV_CURBWGT* versus *GV_ENERGY* is different than, for instance, *GV_CURBWGT* versus *GV_DVLONG* (or *GV_DVLAT*). The node *GV_ENERGY* must be allowed to "respond" in our simulation, as in observational inference, while *GV_DVLONG* or *GV_DVLAT* have to be kept at a fixed distribution.

Non-Confounders

BayesiaLab offers a convenient way to address this requirement. We can assign a special pre-defined **Class** to variables like *GV_ENERGY*, called **Non-Confounders**. We can assign a **Class** by right-clicking the node *GV_ENERGY* and selecting **Properties | Classes | Add** from the **Contextual Menu**.



Now, we select the **Predefined Class** *Non_Confounder* from the drop-down menu.

An icon in the bottom right corner of the screen indicates that a **Class** has been set.

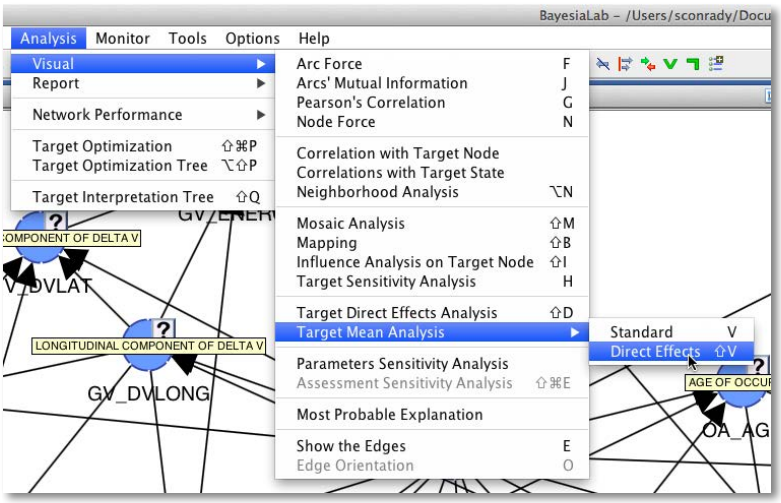
Direct Effects

The principal question of this study now adds one further challenge in that we are looking for the effect of a continuous variables, *GV_CURBWGT*, as opposed to the effect of a binary variable, such as *OA_MANUSE*. This means that our **Likelihood Matching** algorithm must now find matching confounder distributions for a range of values of *GV_CURBWGT* and *GV_FOOTPRINT*.

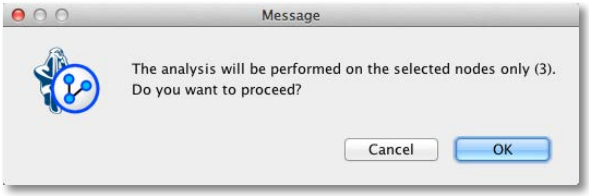
As opposed to manually fixing distributions via the **Monitors**, we now use BayesiaLab's **Direct Effects** function that performs this automatically. Thus, all confounder nodes will be fixed in their distributions, excluding the nodes currently being manipulated, the **Target Node** and the **Non-Confounder** nodes.

Although we are primarily interested in the effect of *GV_CURBWGT* and *GV_FOOTPRINT* on *OA_MAIS*, we will also consider another weight-related node, i.e. *GV_OTVEHWGT*, which represents the curb weight of the other vehicle involved in the collision. Given the principle of conservation of linear momentum, we would expect that—everything else being equal—*increasing* the weight of one's own vehicle would *decrease* injury risk, while an *increase* in the weight of the collision partner would *increase* injury risk.

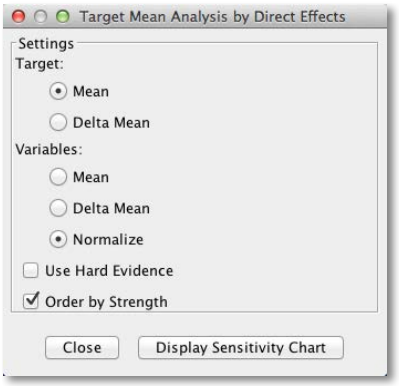
We can select these three nodes in the network and then perform a **Target Mean Analysis by Direct Effect**. (**Analysis | Visual | Target Mean Analysis | Direct Effects**).



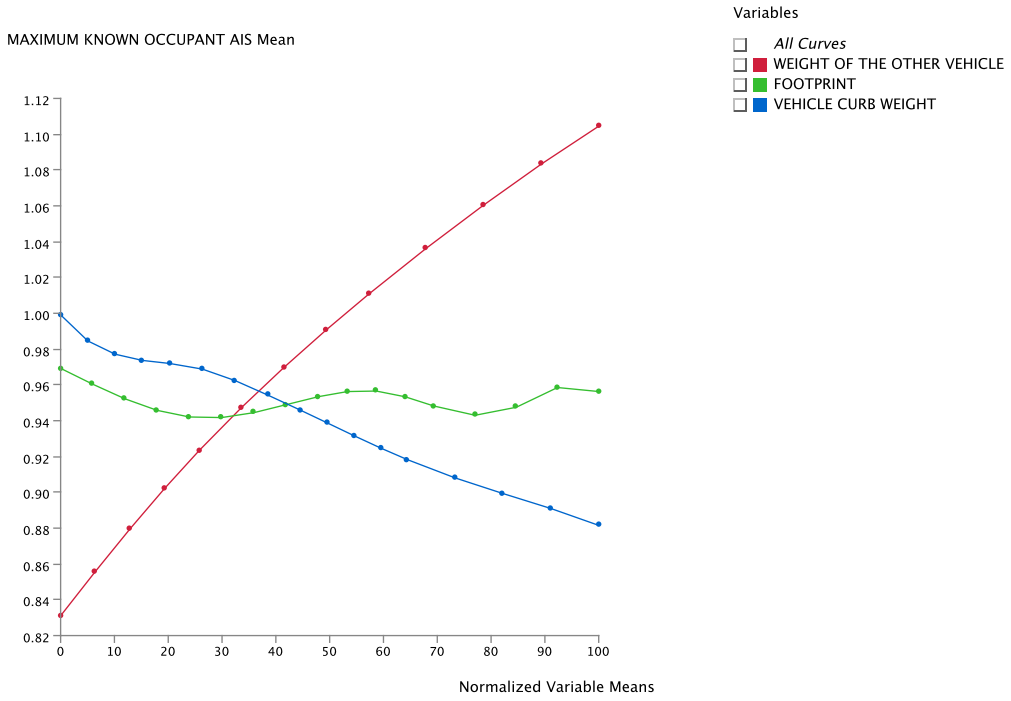
Given that only a subset of nodes was selected, we are prompted to confirm this set of three nodes.



In our case, we have different types of variables, e.g. *GV_FOOTPRINT*, measured in m^2, and *GV_CURBWGT* and *GV_OTVEHWGT*, measured in kg. As such, they could not be properly presented in a single plot, unless their x-ranges were normalized. Thus, it is helpful to select **Normalize** from the options window.

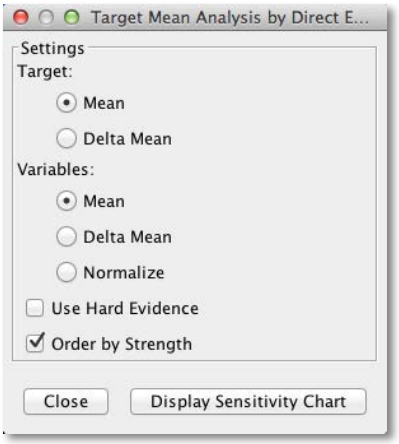


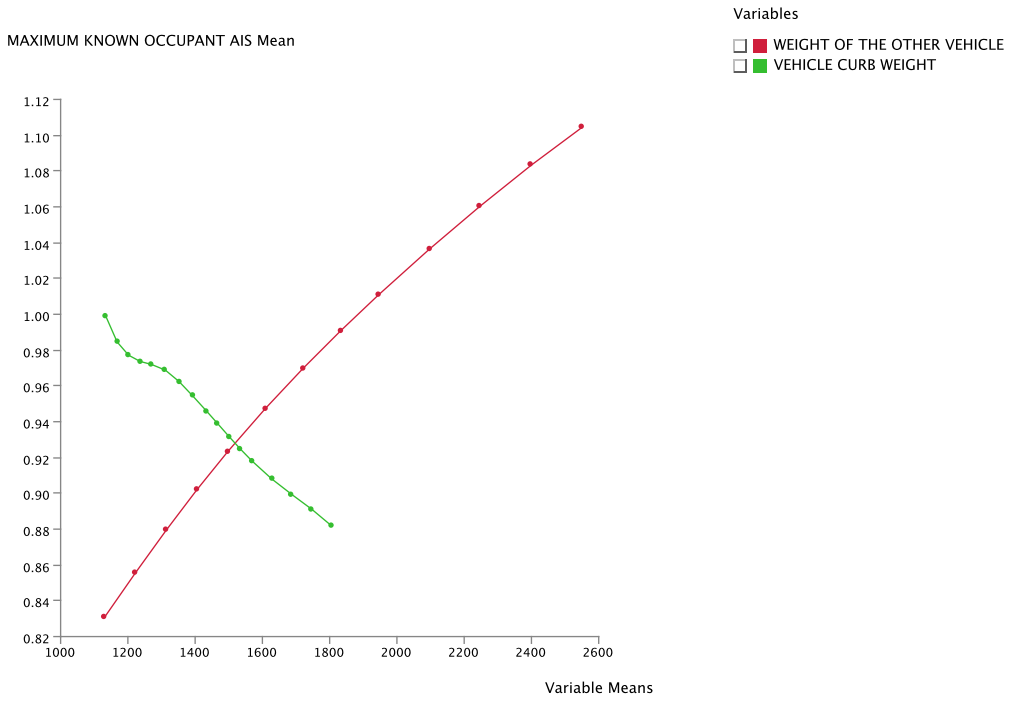
Once this is selected, we can continue to **Display Sensitivity Chart**.



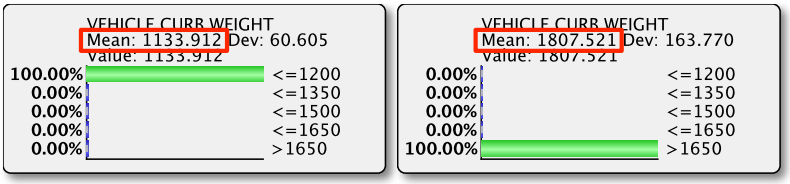
The **Direct Effects** curves of *GV_CURBWGT* and *GV_OTVEHWGT* have the expected slope. However, *GV_FOOTPRINT* appears "flat," which is somewhat counterintuitive.

Alternatively, we can exclusively focus on *GV_CURBWGT* and *GV_OTVEHWGT*, which allows us to look at them on their original scales in kilogram. Thus, we select **Mean** instead of **Normalize** as the display option.



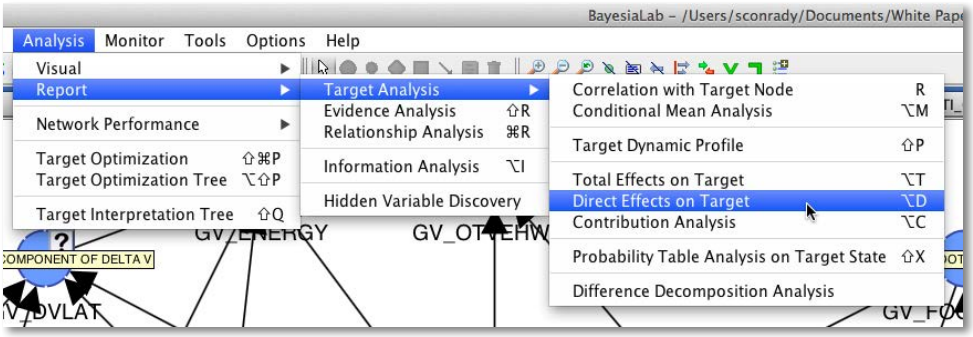


Note that the *GV_CURBWGT* range is smaller than that of *GV_OTVEHWGT*, which was not apparent on the normalized scale. With the **Multi-Quadrant Analysis**, we had previously generated specific networks for each vehicle class. As a result, the range of *GV_CURBWGT* in each network's database is delimited by the maximum and minimum values of each vehicle class. Furthermore, the maximum and minimum values that can be simulated within each network are constrained by the highest and lowest values that the states of the node *GV_CURBWGT* can represent. We can easily see these limits in the **Monitors**.

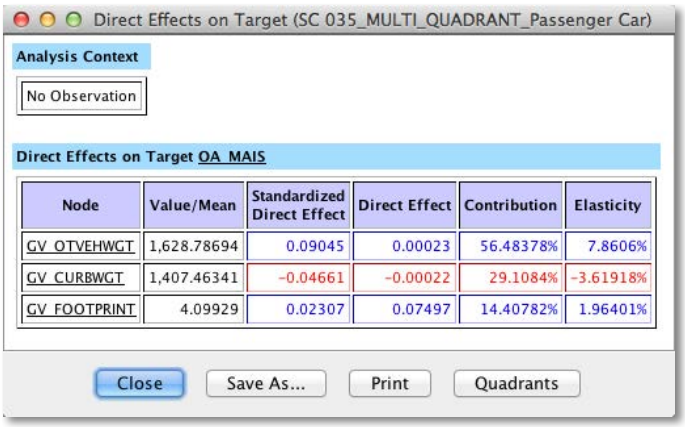


For *Passenger Car*, *GV_CURBWGT* ∈ [1133.912, 1807.521]. This also means that we cannot simulate outside this interval and are thus unable to make a statement about the shape of the **Direct Effects** curves outside these boundaries.

However, within this interval, the simulated **Direct Effects** curves confirm our a priori beliefs based on the laws of physics regarding the effects of vehicle mass. As both curves appear fairly linear, it is reasonable to have BayesiaLab estimate the slope of the curves around their mean values. We do this by selecting **Report | Target Analysis | Direct Effects on Target**.



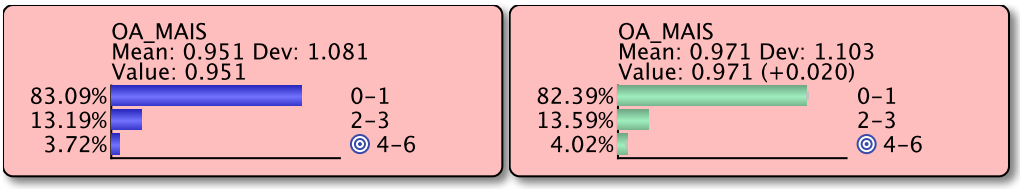
A results window shows the estimated **Direct Effects** in table format.[30]



[30] We can also save this results table in HTML format, which allows subsequent manipulation with a spreadsheet editor.

These **Direct Effects** can be interpreted similarly to coefficients in a regression. This means that increasing *GV_CURBWGT* by 100 kg would bring about a 0.02 decrease in the expected value of *OA_MAIS*. On a scale of 0-6 scale, this may seem like a minute change.

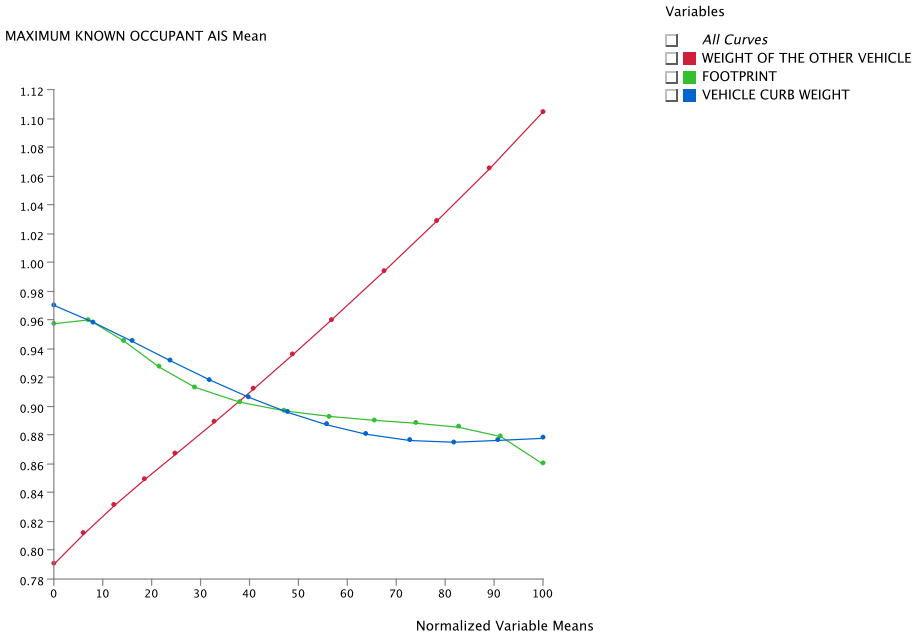
However, a negligible change in the expected value (or mean) can "hide" a substantial change in the distribution of *OA_MAIS*, which is illustrated below. An increase of 0.02 in the expected value translates in to an 8% (or 0.3 percentage points) increase in the probability of a serious or fatal injury.



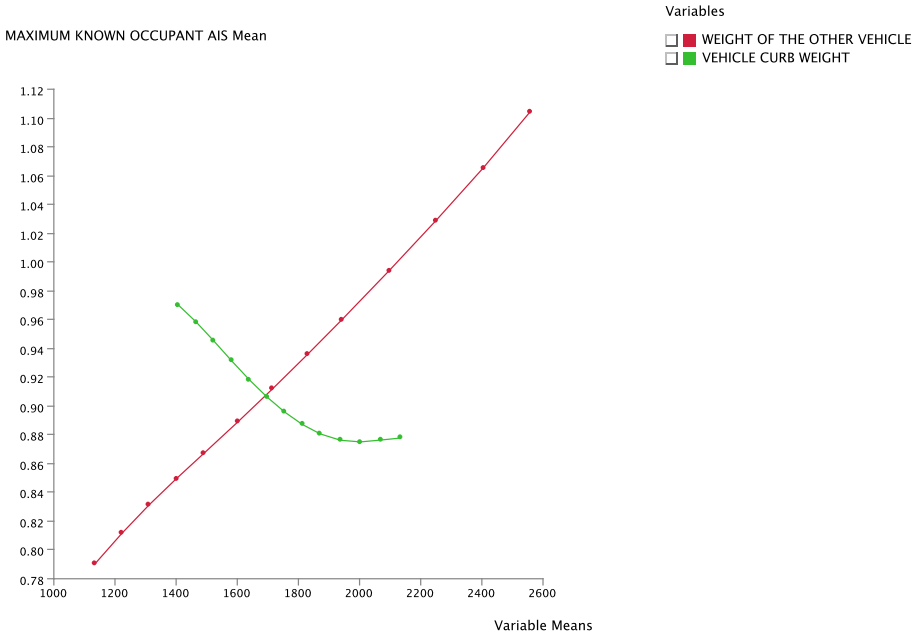
In the example, the changes in probability of serious injury are in the order of magnitude of one tenth of one percent. These numbers may appear minute, but they become fairly substantial when considering that roughly 10 million motor vehicle accidents occur in the United States every year. For instance, an increase of 0.1 percentage points in the probability of serious injury would translate into 10,000 human lives that are profoundly affected.

Vehicle Class: Trucks (<6,000 lbs.)

We now repeat the above process for trucks with GVWR of 6,000 lbs. or less. The first plot, with a normalized x-axis, shows the effect of *GV_CURBWGT*, *GV_OTVEHWGT* and *GV_FOOTPRINT* on *OA_MAIS*.



The next plot displays only *GV_CURBWGT* and *GV_OTVEHWGT* on a kilogram-scale.



For *GV_CURBWGT* and *GV_OTVEHWGT* the following table yields results consistent with what we found for the *Passenger Car* subset. In contrast to that class of vehicles, here we find that the **Direct Effects** curve for *GV_FOOTPRINT* has a clearly negative slope.

Direct Effects on Target (SC 035_MULTI_QUADRANT_Truck _6000 lbs)

Analysis Context

No Observation

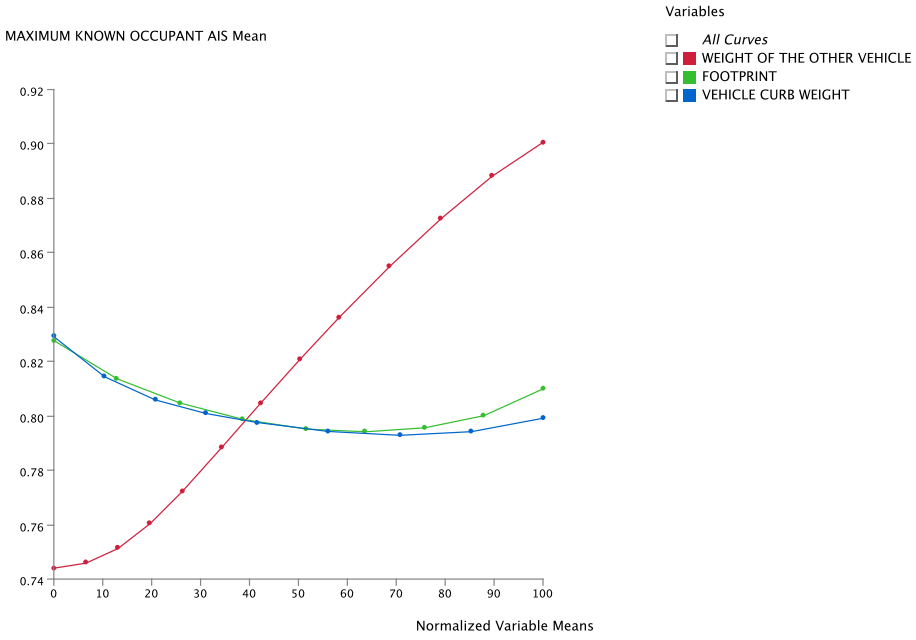
Direct Effects on Target OA_MAIS

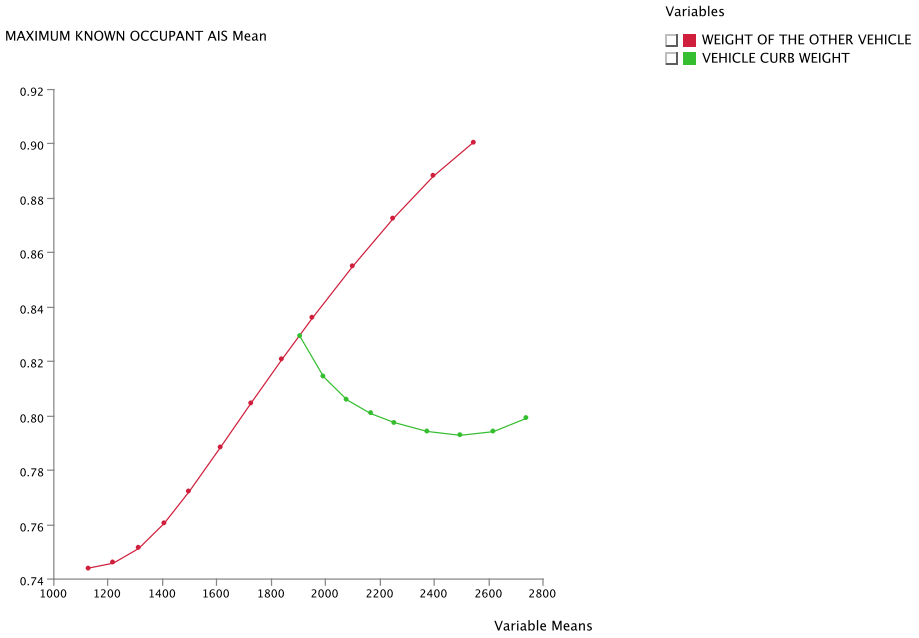
Node	Value/Mean	Standardized Direct Effect	Direct Effect	Contribution	Elasticity
GV_OTVEHWGT	1,624.20775	0.09068	0.00021	62.09664%	7.20467%
GV_CURBWGT	1,769.86362	-0.04256	-0.00016	29.14247%	-2.85105%
GV_FOOTPRINT	4.3838	-0.01279	-0.0267	8.7609%	-0.90737%

Close | Save As... | Print | Quadrants

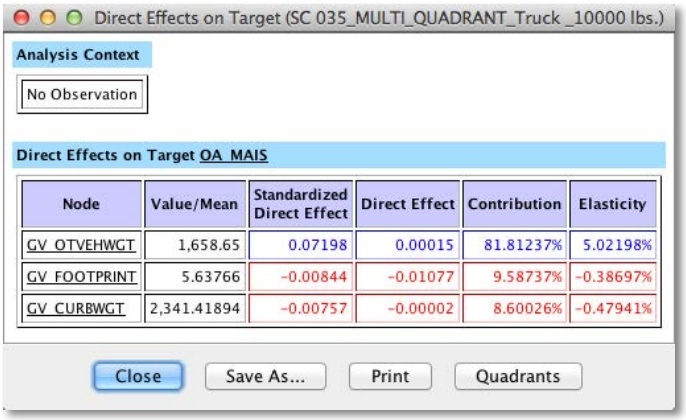
Vehicle Class: Trucks (<10,000 lbs.)

We apply this approach analogously for the final vehicle class, i.e. trucks with a GVWR between 6,000 and 10,000 lbs.





The **Direct Effects** curve for *GV_OTVEHWGT* is entirely consistent with the previous results. However, the curves of *GV_CURBWGT* and *GV_FOOTPRINT* appear nonlinear. As a result, the linearized **Direct Effects** reported in the following table for those nodes must be used with caution.



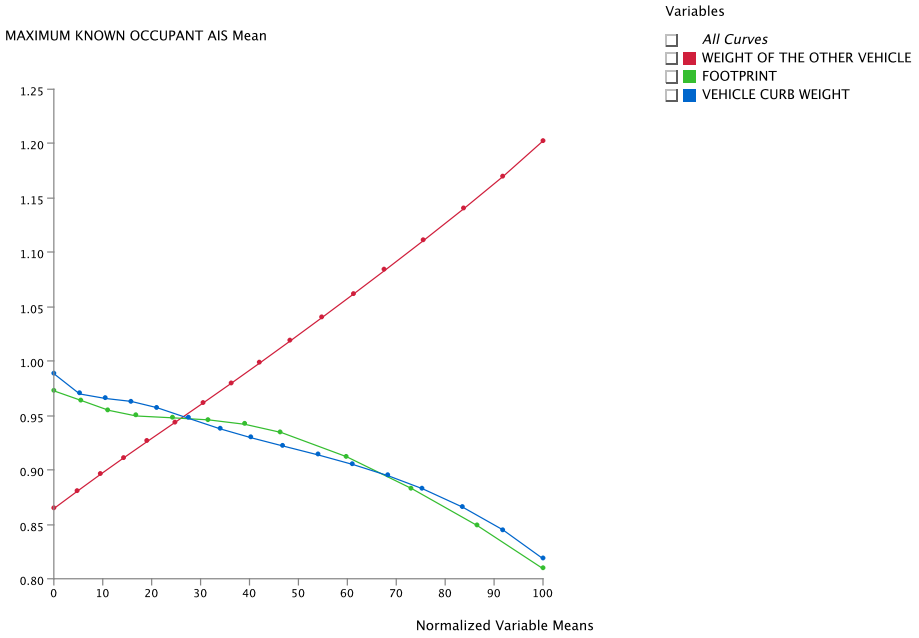
If we were to exclusively rely on this table, we might be tempted to think that the effect of *GV_CURBWGT* on *OA_MAIS* is very small. The plot has shown us that this is not so. Rather, in the range between 2,000 kg and 2,500 kg, a material difference in *OA_MAIS* can be observed.

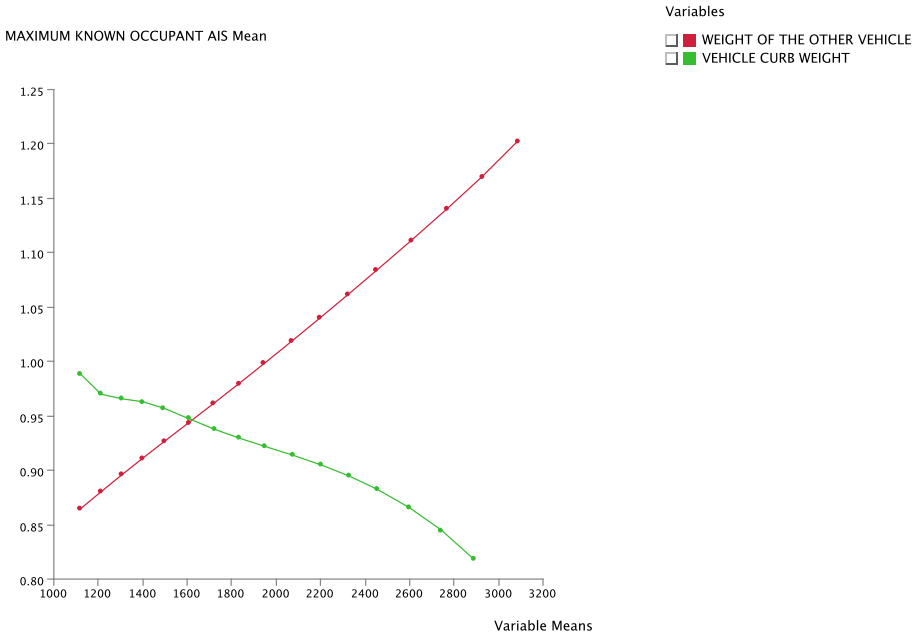
Entire Vehicle Fleet

At beginning of this section on effect estimation, we made a case for **Multi-Quadrant Analysis** due to lack of complete covariate overlap. Consequently, we could only estimate the effects of size and weight by vehicle class. However, the big picture remains of interest. What do these effects look like across all vehicle classes? Despite our initial concerns regarding the lack of overlap, would it perhaps be possible to "zoom out" to assess these dynamics for the entire vehicle fleet? We will now attempt to do just that and perform inference at the fleet level. There is unfortunately no hard-and-fast rule that tells us in advance what amount of overlap is sufficient to perform inference correctly. However, BayesiaLab contains a built-in safeguard and alerts us when **Likelihood Matching** is not possible.

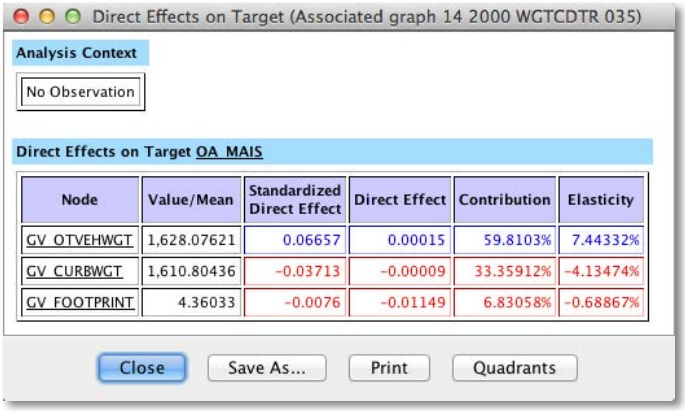
Keeping the above caveats in mind, we return to our original network, i.e. the one we had learned before applying the **Multi-Quadrant Analysis**. We now wish to see whether the class-specific effects are consistent with fleet-level effects.

Our fleet-level model turns out to be adequate after all, and the **Likelihood Matching** algorithm does return the desired results.





The above **Direct Effects** plots provide a rather clear and nearly linear picture, suggesting that computing the slope of the curves would be appropriate.



Node	Value/Mean	Standardized Direct Effect	Direct Effect	Contribution	Elasticity
GV_OTVEHWGT	1,628.07621	0.06657	0.00015	59.8103%	7.44332%
GV_CURBWGT	1,610.80436	-0.03713	-0.00009	33.35912%	-4.13474%
GV_FOOTPRINT	4.36033	-0.0076	-0.01149	6.83058%	-0.68867%

For reference, we have summarized all **Direct Effects** results in the following table. As stated earlier, due to nonlinearities apparent in some of the plots, the **Direct Effects** should only be interpreted in the context of the corresponding graphs.

Category	Node	Value/Mean	Standardized Direct Effect	Direct Effect
Passenger Cars	GV_OTVEHWGT	1628.79	0.090	0.00023
	GV_CURBWGT	1407.46	-0.047	-0.00022
	GV_FOOTPRINT	4.10	0.023	0.07497
Trucks (<6,000 lbs.)	GV_OTVEHWGT	1624.21	0.091	0.00021
	GV_CURBWGT	1769.86	-0.043	-0.00016
	GV_FOOTPRINT	4.38	-0.013	-0.02670
Trucks (<10,000 lbs.)	GV_OTVEHWGT	1658.65	0.072	0.00015
	GV_CURBWGT	2341.42	-0.008	-0.00002
	GV_FOOTPRINT	5.64	-0.008	-0.01077
Entire Fleet	GV_OTVEHWGT	1,628.08	0.067	0.00015
	GV_CURBWGT	1,610.80	-0.037	-0.00009
	GV_FOOTPRINT	4.36033	-0.008	-0.01149

Reducing Vehicle Size versus Reducing Vehicle Weight

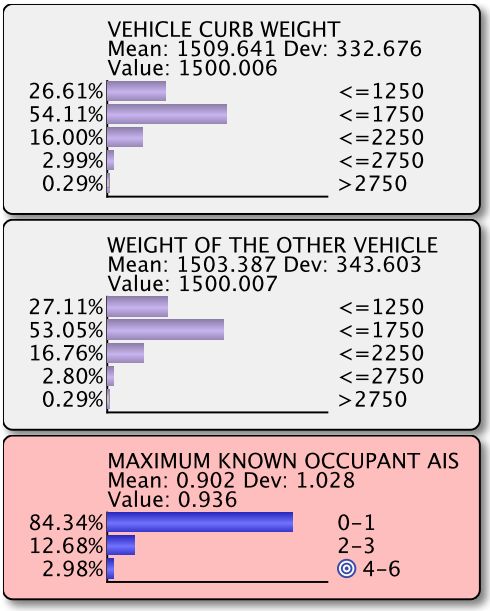
In the EPA/NHTSA Final Rule, great emphasis is put on the distinction between *reducing vehicle size* versus *reducing vehicle weight*. While the former is generally seen as harmful with regard to passenger safety, estimating the potential impact of the latter was the ultimate objective of their studies.

Despite the moderately high correlation between *GV_CURBWGT* and *GV_FOOTPRNT* (0.72), a Bayesian network can clearly distinguish between their individual information contribution towards the **Target Node**. The issue of collinearity, which the EPA/NHTSA Final Rule frequently mentions as a problem, is of no concern in our approach. However, having established their separate effects, we cannot support the notion that the *GV_CURBWGT* effect is generally "flat" given a fixed level of *GV_FOOTPRNT*. Rather, decreasing *GV_CURBWGT* and *GV_FOOTPRNT* appear as distinct effects, both individually having an impact on *OA_MAIS*.

Simulating Interventions

Thus far, we have only computed the impact of reducing *GV_CURBWGT* while leaving the distribution of *GV_OTVEHWGT* the same. This means we are simulating as to what would happen, if a new fleet of lighter fleet of vehicles were introduced, facing the older, heavier fleet. This simplification may be reasonable as long as the new fleet is relatively small compared to the existing one. However, in the long run, the distribution of *GV_OTVEHWGT* would inevitably have to become very similar to *GV_CURBWGT*.

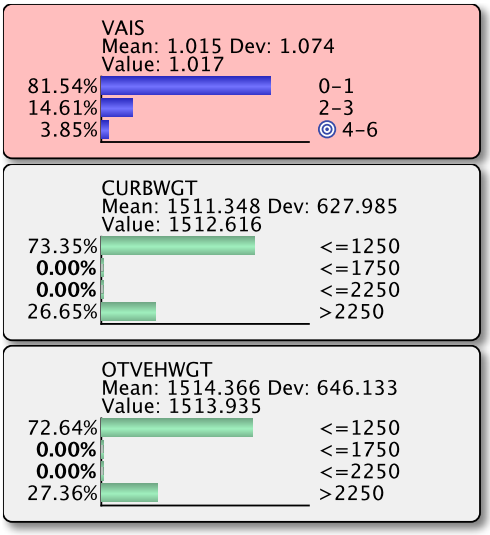
In order to simulate such a long-run condition, any change in the distribution of *GV_CURBWGT* would have to be identically applied to *GV_OTVEHWGT*. This means, once we set evidence on *GV_CURBWGT* (and fixed its distribution) we need to set the same distribution on *GV_OTVEHWGT* (and fix it).



Note that we are now performing causal inference, i.e. "given that we do" versus "given that we observe." Instead of *observing* the injury risk of lighter vehicles (and all the attributes and characteristics that go along with such vehicles), we do make them lighter (or rather force them by mandate), while keeping all the attributes and characteristics of drivers and accidents the same.

For instance, reducing the average weight of all vehicles to 1,500 kg appears beneficial. The risk of serious or fatal injury drops to 2.98%. Considering these simulated results, this may appear as a highly desirable scenario. This would theoretically support a policy that "lightens" the overall vehicle fleet my government mandate. However, we must recognize that such a scenario may not be feasible as functional requirements of vans and trucks could probably not be met within mandatory weight restrictions.

Alternatively, we could speculate about a scenario in which there is a bimodal distribution of vehicle masses. Perhaps all new passenger cars would become lighter (again by government mandate), whereas light trucks would maintain their current weights due to functional requirements. Such a bimodal scenario could also emerge in the form of a "new and light" fleet versus an "old and heavy" existing fleet. We choose a rather extreme bimodal distribution to simulate such a scenario, consciously exaggerating to make our point.



As opposed to a homogenous scenario of having predominantly light vehicles on the road, this extreme bimodal distribution of vehicles increases the injury risk versus the baseline. This suggests that a universally lighter fleet, once established, would lower injury risk, while a weight-wise diverse fleet would increase risk.

These simulations of various interventions highlight the challenges that policymakers face. For instance, would it be acceptable—with the objective of long-term societal safety benefit—to mandate the purchase of lighter vehicles, if this potentially exposed individuals to an increased injury risk in the short term? What are the ethical considerations with regard to trading off "probably more risk now for some" versus "perhaps less risk later for many"? It appears difficult to envision a regulatory scenario that will benefit society as a whole, without any adverse effects to some subpopulation, at least for some period of time. However, it clearly goes beyond the scope of our paper to elaborate on the ethical aspects of weighing such risks and benefits.

Summary

1. This paper was developed as a case study for exhibiting the analytics and reasoning capabilities of the Bayesian network framework and the BayesiaLab software platform.

2. Our study examined a subset of accidents with the objective of understanding injury drivers at a detail level, which could not be fully explored with the data and techniques used in the context of the EPA/ NHTSA Final Rule.

3. For an in-depth understanding of the dynamics of this problem domain, it was of great importance to capture the multitude of high-dimensional interactions between variables. We achieved this by machine-learning Bayesian networks with BayesiaLab.

4. On the basis of the machine-learned Bayesian networks, BayesiaLab's **Likelihood Matching** was used to estimate the exclusive **Direct Effects** of individual variables on the outcome variable. The estimated effects were generally consistent with prior domain knowledge and the laws of physics.

5. Simulating domain interventions, e.g. the impact of regulatory action, required carrying out causal inference, using **Likelihood Matching**. We emphasized the distinction between observational and causal inference in this context.

6. With regard to this particular collision type, we conclude that injury risk remains a function of both mass and size. More specifically, as a result of decomposing the individual effects of vehicle size and weight, the notion of "mass reduction being safety-neutral given a fixed footprint" could not be supported in this specific context at this time.

Conclusion

Machine-learning Bayesian networks from historical accident data using the BayesiaLab software platform allows us to comprehensively and compactly capture the complex dynamics of real-world vehicle crashes. With the domain encoded as a Bayesian network, we can "embrace" the high-dimensional interactions and leverage them for performing observational and causal inference. By employing Bayesian networks, we provide an improved framework for reasoning about vehicle size, weight, injury risk, and ultimately about the consequences of regulatory intervention.

References

2017 and Later Model Year Light-Duty Vehicle Greenhouse Gas Emissions and Corporate Average Fuel Economy Standards. Final Rule. Washington, D.C.: Department of Transportation, Environmental Protection Agency, National Highway Traffic Safety Administration, October 15, 2012. https://federalregister.gov/a/2012-21972.

2017 and Later Model Year Light-Duty Vehicle Greenhouse Gas Emissions and Corporate Average Fuel Economy Standards. Department of Transportation, Environmental Protection Agency, National Highway Traffic Safety Administration, August 28, 2012.

A Dismissal of Safety, Choice, and Cost: The Obama Administration's New Auto Regulations. Staff Report. Washington, D.C.: U.S. House of Representatives Committee on Oversight and Government Reform, August 10, 2012.

Bastani, Parisa, John B. Heywood, and Chris Hope. U.S. CAFE Standards - Potential for Meeting Light-duty Vehicle Fuel Economy Targets, 2016-2025. MIT Energy Initiative Report. Massachussets Institute of Technology, January 2012. http://web.mit.edu/sloan-auto-lab/research/beforeh2/files/CAFE_2012.pdf.

Chen, T. Donna, and Kara M. Kockelman. "THE ROLES OF VEHICLE FOOTPRINT, HEIGHT, AND WEIGHT IN CRASH OUTCOMES: APPLICATION OF A HETEROSCEDASTIC ORDERED PRO-BIT MODEL." In Transportation Research Board 91st Annual Meeting, 2012. http://www.ce.utexas.edu/prof/kockelman/public_html/TRB12CrashFootprint.pdf.

"Compliance Question - Will Automakers Build Bigger Trucks to Get Around New CAFE Regulations?" Autoweek. Accessed September 9, 2012. http://www.autoweek.com/article/20060407/free/60403023.

Conrady, Stefan, and Lionel Jouffe. "Causal Inference and Direct Effects - Pearl's Graph Surgery and Jouffe's Likelihood Matching Illustrated with Simpson's Paradox and a Marketing Mix Model," September 15, 2011. http://bayesia.us/index.php/causality.

———. "Modeling Vehicle Choice and Simulating Market Share with Bayesian Networks" (December 18, 2010). http://bayesia.us/index.php/market-share-simulation.

"Crashworthiness Data System - 2009 Coding and Editing Manual." U.S. Department of Transportation National Highway Traffic Safety Administration, January 2009.

"Crashworthiness Data System - 2010 Coding and Editing Manual," n.d.

"Crashworthiness Data System - Analytical User's Manual 2009 File." National Center for Statistics and Analysis National Highway Traffic Safety Administration U.S. Department of Transportation. Accessed May 25, 2013. http://www-nrd.nhtsa.dot.gov/Pubs/NASS09.pdf.

Effectiveness and Impact of Corporate Average Fuel Economy: CAFE Standards. Joseph Henry Press, 2003.

"Federal Motor Vehicle Safety Standards; Occupant Crash Protection." National Highway Traffic Safety Administration, 1998. http://www.nhtsa.gov/cars/rules/rulings/AAirBagSNPRM.

Final Regulatory Impact Analysis Corporate Average Fuel Economy for MY 2012-MY 2016 Passenger Cars and Light Trucks Office. Washington, D.C.: U.S. Department of Transportation National Highway Traffic Safety Administration, n.d.

"Final Rulemaking to Establish Light-Duty Vehicle Greenhouse Gas Emission Standards and Corporate Average Fuel Economy Standards - Regulatory Impact Analysis." Office of Transportation and Air Quality U.S. Environmental Protection Agency, April 2010. www.epa.gov/otaq/climate/regulations/420r10009.pdf.

Gabler, Hampton C., and William T. Hollowell. "The Aggressivity of Light Trucks and Vans in Traffic Crashes." SAE Transactions 107 (1999): 1444–1452.

Glance, Laurent G, Turner M Osler, Dana B Mukamel, and Andrew W Dick. "Outcomes of Adult Trauma Patients Admitted to Trauma Centers in Pennsylvania, 2000-2009." Archives of Surgery (Chicago, Ill.: 1960) 147, no. 8 (August 2012): 732–737. doi:10.1001/archsurg.2012.1138.

Hampton, Carolyn E., and Hampton C. Gabler. "Evaluation of the Accuracy of NASS/CDS Delta-V Estimates from the Enhanced WinSmash Algorithm." Annals of Advances in Automotive Medicine / Annual Scientific Conference 54 (January 2010): 241–252.

Kahane, Charles J. Relationships Between Fatality Risk, Mass, and Footprint in Model Year 2000-2007 Passenger Cars and LTVs - Final Report. Washington, D.C.: National Highway Traffic Safety Administration, August 2012.

Kahane, Charles Jesse. Vehicle Weight, Fatality Risk and Crash Compatibility of Model Year 1991-99 Passenger Cars and Light Trucks, 2003. http://trid.trb.org/view.aspx?id=661597.

Lund, Adrian. "The Relative Safety of Large and Small Passenger Vehicles." presented at the NHTSA Mass-Size Safety Symposium, Washington, D.C., February 25, 2011.

National Automotive Sampling System (NASS) General Estimates System (GES) 2010 Coding and Editing Manual. Washington, D.C.: U. S. Department of Transportation National Highway Traffic Safety Administration National Center for Statistics and Analysis, December 2011.

National Automotive Sampling System (NASS) General Estimates System (GES) Analytical Users - 2010 File. Washington, D.C.: U. S. Department of Transportation National Highway Traffic Safety Administration National Center for Statistics and Analysis, 2010.

National Automotive Sampling System (NASS) General Estimates System (GES) Analytical Users Manual 1988-2010. Washington, D.C.: U. S. Department of Transportation National Highway Traffic Safety Administration National Center for Statistics and Analysis, December 2011.

"News Release - NEW CRASH TESTS DEMONSTRATE THE INFLUENCE OF VEHICLE SIZE AND WEIGHT ON SAFETY IN CRASHES; RESULTS ARE RELEVANT TO FUEL ECONOMY POLICIES." Insurance Institute for Highway Safety, April 14, 2009. http://www.iihs.org/news/2009/iihs_news_041409.pdf.

Pearl, Judea. "Causal Inference in Statistics: An Overview." Statistics Surveys 3, no. 0 (2009): 96–146. doi:10.1214/09-SS057.

Radja, Gregory A. National Automotive Sampling System – Crashworthiness Data System, 2011 Analytical User's Manual. Office of Data Acquisition National Center for Statistics and Analysis National Highway Traffic Safety Administration, October 2012.

Singh, Harry. Mass Reduction for Light-Duty Vehicles for Model Years 2017–2025 - Final Report. U. S. Department of Transportation National Highway Traffic Safety Administration, n.d.

Spirtes, Peter. "Introduction to Causal Inference." The Journal of Machine Learning Research 99 (2010): 1643–1662.

Status Report - The Risk of Dying in One Vehicle Versus Another. Insurance Institute for Highway Safety, March 19, 2005.

Tschöke, Helmut, and Hanns-Erhard Heinze. "Einige Unkonventionelle Betrachtungen Zum Kraftstoffverbrauch von PKW." Magdeburger Wissenschaftsjournal (2001): 1–2.

Wenzel, Tom. "An Analysis of the Relationship Between Casualty Risk Per Crash and Vehicle Mass and Footprint for Model Year 2000-2007 Light-Duty Vehicles" (2011). http://www.epa.gov/otaq/climate/documents/lbnl-2012-phase-2.pdf.

Contact Information

Bayesia USA

312 Hamlet's End Way
Franklin, TN 37067
USA
Phone: +1 888-386-8383
info@bayesia.us
www.bayesia.us

Bayesia Singapore Pte. Ltd.

20 Cecil Street
#14-01, Equity Plaza
Singapore 049705
Phone: +65 3158 2690
info@bayesia.sg
www.bayesia.sg

Bayesia S.A.S.

6, rue Léonard de Vinci
BP 119
53001 Laval Cedex
France
Phone: +33(0)2 43 49 75 69
info@bayesia.com
www.bayesia.com

Copyright