CAUSAL INFERENCE USING BAYESIAN NETWORK FOR SEARCH AND RESCUE

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Statistics

by

Amanda Elyse Belden

June 2024

COMMITTEE MEMBERSHIP

TITLE: Causal Inference Using Bayesian Network for Search and Rescue

AUTHOR: Amanda Elyse Belden

DATE SUBMITTED: June 2024

COMMITTEE CHAIR: Hunter Glanz, Ph.D.
Professor of Statistics

COMMITTEE MEMBER: Franz Kurfess, Ph.D.
Professor of Computer Science and Software Engineering

COMMITTEE MEMBER: Soma Roy, Ph.D.
Professor of Statistics

ABSTRACT

Causal Inference Using Bayesian Network for Search and Rescue

Amanda Elyse Belden

People who are considered missing have much higher probabilities of being found dead compared to those who are not considered missing in terms of Search and Rescue (SAR) missions. Dementia patients are incredibly likely to be declared missing, and in fact after removing those with dementia the probability of the mission being regarded as missing person case is only about 10%. Additionally, those who go missing are much more likely to be on private land than on protected areas such as forests and parks. These and similar associations can be represented and investigated using a Bayesian network that has been trained on Search and Rescue mission data. By finding associations between factors that affect these missions, SAR teams can find patterns in historical cases and apply them to future cases in order to narrow down their search areas, improve their plans, and hopefully lead to lower search times and fewer deaths and unsolved cases. Causal inference allows causal relationships to be determined, telling SAR teams that they can make current decisions based on these learned relationships and their decisions will cause the change that they expect based on the Bayesian network.

Keywords: Bayesian networks, causal inference, Search and Rescue

ACKNOWLEDGMENTS

I would like to give my thanks to my committee, Dr. Franz Kurfess, Dr. Hunter Glanz, and Dr. Soma Roy for providing me with guidance and support on this thesis. I would also like to thank the NYS Forest Rangers and Dr. Robert Koester for collating and providing me with access to this data. Without Dr. Koester's work, this thesis would not have been possible. Finally, I would like to thank my family and friends for their support and encouragement during my time at Cal Poly.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Suppose a young man goes for a hike in the Adirondack Park in New York. A couple hours into his hike, he realizes that the trail he has been following does not really exist anymore, and when he tries to find his way back to it, he realizes he cannot find it at all. His phone has battery, but no service to call for help, and it is quickly becoming darked and colder. What happens to him?

Search and Rescue (SAR) are the first responders for lost and missing people. Their job is to find people quickly and safely before their injuries or circumstances can kill them. Unfortunately, as in the case with the lost hiker, it can be incredibly difficult to find a single someone in a very large area with many obstacles that prevent them from being seen or detected. Additionally, many Search and Rescue teams are made up partially or fully by volunteers, leading to less training and less resources available to help in their search. So, knowing they cannot search every square foot of the park, how do they decide where to search for the hiker, knowing that they only have a maximum of a couple of days to find him before he could die?

SAR teams rely on past knowledge from hundreds and thousands of past searches. By recognizing patterns in historical missions, they can determine the most likely area to search for the current missing subject based on the similarities to the historical missions.

Each SAR mission can be classified into three main categories, a search, a rescue, or a recovery. A search is when the location of the subject is unknown, and this is normally paired with subsequent rescue or recovery. A rescue is when the location of the subject is known or has been learned, such as by a phone GPS, but the subject is unable to get themselves out of danger. This could be that they were injured or stuck in a hazardous situation, such as building collapse or at the bottom of a cliff. A recovery is when the team is looking for or retrieving a dead body. Searches and rescues typically have large time pressures and require immediate efforts to help the subject. Recoveries have less time pressure and can be delayed if necessary for better conditions. However, the response to each of these types of mission are always highly dependent on the specific scenario and can look very similar or different given the situation at hand.

For example, suppose the SAR team are notified that the hiker has gone missing two hours after he was supposed to return home. The hiker's friends say he is an experienced hiker but was trying a new trail today. They also learn that he has an old knee injury that can flare up during strenuous or prolonged activity. Based on this information, the SAR team can determine a much smaller search area than the entirety of the Adirondack. They have learned that he is within a couple hours walk of one specific trail, he will likely avoid difficult terrain due to his knee, and that he will likely be moving slowly if at all as his knee becomes more painful.

This information gives the SAR team a starting area to search, and as they explore the specific area, they might find clues or similarities to other missions that allow them to make better predictions of where to search for the hiker. They might know, for example, that based on his experience with hiking he might to try to continue walking, but due to his injury he will try to follow the first linear feature he finds, such as a river, path, trail, or similar (Koester 185).

In recent years, SAR personnel have become interested in using computer-based algorithms to improve their ability to determine potential search areas and hopefully narrow down the possible locations missing people might be found. This could allow for an increased success rate in finding missing people, cut down on the time it takes to find people, leading to improved conditions in which missing people are found, and use fewer resources to conduct searches.

However, instead of just trying to predict the location someone is found, which would be incredibly difficult to do with any accuracy, instead it could be possible to use computer models to find patterns in historical missions which can then be applied to future cases, as the SAR teams do already, but just based on their experiences.

In this thesis, the aim is to find relationships between factors about missing people and factors about Search and Rescue missions, to determine whether there are factors that can cause negative or positive impacts on the Search and Rescue outcomes. To do this, a Bayesian network will be created based on SAR mission data to illustrate the relationships between factors, and causal inference will be employed to try to discover the causal effects between these relationships.

Chapter 2

BACKGROUND

2.1 Motivation and Context

Search and Rescue (SAR) is the process and organization of searching for and giving aid to those who have found themselves lost or otherwise unable to find their way home. All over the world, Search and Rescue teams are critical first responders, providing lifesaving aid, and finding themselves in some of the most dangerous situations imaginable to do so. SAR operations range from disaster-related incidents such as wildfires and floods, to runaway children or patients with dementia missing in the middle of cities, to lost or injured hikers in the middle of nowhere. Each operation is unique and requires extensive training and experience to effectively find and save people's lives. SAR members must be incredibly skilled and versatile, from basic first aid and the ability to hike for hours, to being able to operate helicopters, navigate swift waters, and climb and rappel rocky terrains, to being able to profile and psychologically analyze missing subjects from very little information. The failure at any one of these skills, can be the difference between life and death for their subjects. At the heart of every mission, is the SAR's team absolute willingness to do anything it takes to find their subject.

From the moment a SAR team is notified of a mission, they are racing against a clock to find their subject before they die. For missions involving water or other extreme conditions, they could have just a few hours to conduct their operations. For every mission, acting as quickly as possible, yet as accurately as possible, is incredibly important. However, since every mission is so unique and dependent on so many, possibly unknown, factors this can be an incredibly challenging task. How is it possible to find one person in a 6-million-acre park, in less than a couple of days? In less than a couple of hours?

SAR teams combine their use of physical resources, which are often limited, with their immense knowledge and experience of previous missions to develop search areas, where they determine the subject is most likely to be found. They take into account every single detail they can learn about the subject, not only what they were last seen doing and where, but also their personality, their background, their physical condition, and so much more. Based on all of this information, they try to determine what their subject will

do when faced with being lost or injured in the middle of nowhere and use this to inform where they search first. One of the most influential books on this topic is Dr. Robert J Koester's "Lost Person Behavior: A Search and Rescue Guide on Where to Look for Land, Air, and Water." In Chapter 8, the book goes into incredible detail of the behavior of missing subjects, splitting subjects into different categories and analyzing how each of these groups differ in their actions and in how they are found based on as many previous missions as they were able at the time. Using previous SAR operation outcomes and subject categories, they can apply this information to their current mission to determine likelihoods and probabilities of their subject's potential actions and even where they might be found. The more experience and similar missions they are able to base their search area on, the better their probabilities will become, and thus the better the outcomes of their mission.

To this point, finding associations between types of missions, missing persons behaviors, and operations in general can improve SAR efforts, leading to decreased search times and potentially saving more lives. In this thesis, the goal is introduce the use of a machine learning technique called a Bayesian network (BN) as a tool to more efficiently find and analyze these associations based on previous missions, so SAR teams can focus more of their time on rescuing people who are currently in distress, and do so based on more background knowledge than before.

2.2 Literature Review

It is important to review the methods and ideas that have preceded and informed this thesis. The following sections will explore Search and Rescue, Bayesian networks, and causal inference which will provide necessary context for the work in this thesis.

2.2.1 Search and Rescue

Search and Rescue operations take in numerous factors when determining search areas and how to deploy their resources. These factors can be classified into four general groups: location information, weather information, SAR resources available, and missing subject information. Most SAR operations are conducted by local teams, so they are typically well equipped for their specific location and weather requirements. The main considerations for how to conduct a search come from the missing subject

information, as the SAR resources available and used can be influenced by the specifics of the mission, such as if other agencies are going to be pulled in due to cross-jurisdictional search areas.

As mentioned previously, Robert Koester's book is a guide on how to conduct SAR operations depending on various subject categories. He divides subjects into approximately 30 categories, such as: Child, Dementia, Despondent, Hiker, Hunter, Water Related, etc. For each of these categories, he creates a profile, generates some statistics, and recommends initial tactics and potential investigative questions. His book's data is based on a SAR dataset called ISRID, the International Search & Rescue Incident Database, which is a collection of SAR operations from around the world, beginning in 2002. This database is confidential and since many differing SAR teams contribute to it, the data inside comes in many different formats, leading to sporadic and sparse variables that coincide between many of them. Due to this, many of Koester's statistics are based on small sample sizes and cannot be generalized to all SAR missions. The specifics of ISRID's data sources are delineated in Chapter 2 of Koester's book.

Additionally, this book covers types of lost person strategies, located in Chapter 5. There are ten strategies listed, some of which are as follows: Random Traveling, Route Traveling, Backtracking, Staying Put, Doing Nothing, and five others. This information was recorded in the ISRID database at the conclusion of searchers, and it suggests that SAR teams can try to find associations between certain types of people and what strategy they will use when they find themselves lost. These associations can then be applied to future SAR operations allowing them to better predict what path the subject might take and where to concentrate or begin their search.

Based on the information in this book, it becomes clear that different types of missing subjects and their activities and choices can have great influence on how SAR operations should be conducted. Thus, machine learning techniques that hope to discover associations between factors to improve SAR outcomes should make sure to include and emphasize missing subject information to highlight how these factors can influence mission outcomes.

2.2.2 Bayesian Networks

A Bayesian network (BN) is a type of machine learning model that can be further classified as a probabilistic reasoning model. This type of model attempts to determine probabilistic relationships among any number of variables of interest (Heckerman 33). A Bayesian network is typically represented in a graphical form; hence it being called a 'network.' The graph most often used is a directed acyclic graph (DAG), where every node in the network is one of the variables of interest, and every edge connecting the nodes is a probabilistic relationship between those two nodes. Figure 1 is an example of a directed acyclic graph.

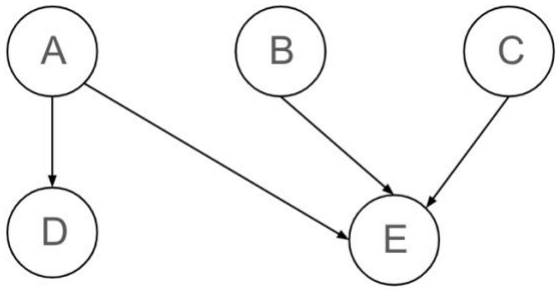

Figure 1: Example of a Bayesian network represented by a DAG

Each node in a Bayesian network can be called a 'parent' node if it has an arrow pointing away from it, such as A, B, and C. Each node can be called a 'child' node if it has an arrow pointing towards it, such as D and E. These terms can be used to describe the associations and directions of arrows between variables in a DAG.

The graph in Figure 1 shows the variables of interest A-E, and each arrow shows a probabilistic relationship between the two variables it connects. For example, the Bayesian network found a direct association between Factor A and Factor D, but no direct association between Factor D and Factor E, shown by the presence and lack of an arrow connection, respectively. The DAG is a visual representation of the Bayesian network, but the BN itself has encoded numerical probabilistic relationships between the variables shown in the visual diagram. For example, if Factor A has two levels 'Yes' or 'No' and Factor D has two levels 'Yes' or 'No' the Bayesian network would have encoded the conditional distribution between these levels, such as P(B = 'Yes' | A = 'Yes') = x and P(B = 'Yes' | A = 'No') = y. Based on the

values of x and y, it is possible to see how the value of A affects the probability of B being equal to 'Yes.' By defining these probabilistic relationships, conclusions like, "changing the level chosen for A is associated with a x-y percentage point increase in the probability of B being equal to 'Yes.' Understanding how an intervention on Factor A affects Factor B allows us to learn a causal relationship between the variables, where changing Factor A is *causing* a change in Factor B (Heckerman 70). This is because nothing else in the model or data has changed except for the level of Factor A, thus it must be the cause of the change in the probability of Factor B. The technique of intervening on a value of one factor to see how it changes the probability of another is called a 'counterfactual' and it is one method of employing causal inference.

Bayesian networks tend to be less sensitive to some missingness in the data entries and generally avoid the problem of overfitting data (Heckerman 33) compared to other machine learning models. Both of these qualities make a BN ideal for many real-world applications of datasets as they these tend to have some missing values in the data and may be slightly smaller datasets than would be ideal. Overfitting tends to be a problem in many machine learning models specifically, as the model fits too closely to the data that it was trained on but then cannot generalize to new data that it has never seen before. Bayesian networks, however, are constructed in such a way that can avoid the problem of overfitting. These models are constructed using search methods that create many potential models and model selection criteria to compare and select the model that fits the data best. Due to this model selection process, there is an efficient way to train the BN models without overfitting the data (Heckerman 56).

2.2.3 Causal Inference

Causal inference is a general idea in statistics that attempts to find causal relationships between variables using data. This is most commonly achieved by designing and conducting experiments where the treatments of one or several factors are applied onto the observational units to see what effect they have on another variable associated with those observational units. By controlling any extraneous factors as much as possible and randomizing the order of treatments to observational units, it is possible to draw cause and effect conclusions, where the treatments are causing the change in the other variable. However, when randomized treatments are not imposed by the researchers, it is typically only possible to say that there are

associations between variables because there could be other hidden or extraneous variables that are also interfering in the relationship between those two variables.

Consider this example: suppose a researcher is analyzing the relationship between ice cream sales and sunburns. If they conducted an analysis on these values, they would find a positive relationship such that as ice cream sales increase, the number of sunburns also increases. Does this mean that ice cream sales cause an increase in sunburns? That sunburns cause an increase in ice cream sales? If the researchers had conducted an experiment where they had one group buy ice cream, and the other group did not, would the ice cream group develop sunburns? No, they would not, meaning that ice cream sales do not cause sunburns, but then why are they positively associated? There is a third, hidden variable: hot weather. When there is hot weather, this causes more people to buy ice cream, and causes more people to develop sunburns. Thus, both variables increase at the same time, but the cause was a variable that was not considered as part of the data. This is called a confounding variable in statistics, and it is this that makes drawing causal conclusions from observational data very difficult, because it is hard to tell whether there are additional variables that are the real cause of the changes represented in the data.

Most data that exists in the world is observational; it was collected with no interference by researchers trying to control one or more of the variables, including the data that will be used in this thesis. So, why does it matter if it is not viable to determine causal relationships from observational data? Imagine the following scenario: a Search and Rescue mission coordinator is in the middle of a mission searching for a someone who has been missing for two days in snowy conditions. She only has the resources available to choose one of two options: add an extra search dog or add five more rangers. Which decision causes the larger increase in probability of finding the subject alive? Do either of them cause a change in this probability? Perhaps the only resource that would cause a change would be the use of a helicopter. Better funded SAR teams may have more search dogs, more rangers, and access to helicopters, leading to them having higher probabilities of finding missing subjects alive than other SAR teams. But the relationship between more dogs and rangers with finding the subject alive could be positive only because of the confounding variable of the helicopter. But the SAR coordinator needs an answer now, and her missing subject's life could depend on whether the resource chosen truly has a causal relationship with the subject's

status. In this situation, it may be obvious that any of these options are likely to have a causal relationship with finding someone faster, but there may be less obvious relationships that are still important to investigate. The ability to make good decisions based on the relationships between variables depends on whether the decision made is based on a causal relationship (Pearl 22).

A Bayesian network created by data alone does not represent causal relationships between variables, it only represents the associations between variables. In order to determine a causal effect, there must be an intervention on the variables to ascertain how changing one variable, with all else remaining equal, changes the other variable. This is done using randomized treatment imposition in experiments, but can be done on observational data using a method called do-calculus. Do-calculus can be defined as "based on our causal assumption that X is the cause variable and Y is the effect variable, if we deliberately trim the sample space of X to control (not observe) it to take the value x, the probability of Y = y will change to P(Y = y | X = x) under this assumption," (Heckerman 6). When there is an intervention on X, where it is forced to be a certain value, this is denoted as do(X = x). Heckerman goes on to state that by calculating both postintervention values of P(Y = y | do[X = x]) and P(Y = y | do[X = x−]) then the difference of these is known as the 'average causal effect (ACE)' (pg 7). Finally, he states, "Because it is conditioned on our deliberate action on X, that is, do[X = x]and do[X = x−],we can also make an inductive conclusion that X = x is a probabilistic cause of Y = y if the ACE is calculated to be positive in that the positive ACE value evidently indicates a rise in the probability of Y = y when X is changed from x− to x by our action," (Heckerman 7). This means that if the postintervention values are different, there is a true causal effect between the variables.

2.3 Data Sources

The data used in this model comes from the New York State (NYS) Department of Environmental Conservation, Division of Forest Protection. A section of the data is publicly available and is called the Wildland Search and Rescue Missions by NYS Forest Rangers. The data collection for this dataset began in 2012 and was last updated March 30, 2023 at the time of download. This included 5,414 rows and 29 columns of data.

This data is also augmented by a confidential dataset collated by Dr. Robert J Koester, also from the NYS Department of Environmental Conservation, Division of Forest Protection. The data collection for this dataset began in 2014 and ended on June 30, 2022. This data came in two sheets, one with 4,285 rows and 47 columns, the other with 4,285 rows and 17 columns.

All three sheets of data contain overlapping columns and contain multiple rows per distinct SAR operation. This is because a unique incident number is assigned to each missing subject within one SAR operation. For example, if five people went missing as a group, that search mission would be recorded five times in the dataset, once for each person.

In order for this data to be usable for the chosen modeling process, the three sheets of data needed to be unified, and the data needed to be transformed so that each SAR operation was only recorded once. This will be discussed in detail in the following section.

The variables can be classified into four main groups: missing subject related, SAR team related, location related, or weather related. For missing subject variables, this included things such as gender, age, last known activity, dementia status, etc. For SAR team variables, this included variables such as the number of resources available, such as total number of rangers, dogs, or helicopters, the time of notification, the number of agencies involved, etc. For location, this included information on the last known location, the found location, the type of terrain, the elevation, etc. Finally, for weather, there was minimum and maximum temperature, minimum and maximum snow depth, rain indicators, etc. There were many more variables available, but these can give a sense of what types of information were included in this dataset.

Chapter 3

METHODS

12.1 Data Pre-Processing

Before modeling data, it is important to ensure that the data is in a correct and appropriate format for the specific model chosen. For Bayesian networks, complex models can require a lot of computing power and data. Since this dataset is relatively small, it is important to reduce complexity as much as possible to find meaningful associations between variables. To do this, the data must be processed in such a way to reduce complexity while maintaining as much useful information as possible. To see the specifics of the data cleaning process, see the data cleaning file located in the Appendix.

12.1.1 Data Aggregation

The first step in processing this data was to combine all three datasets together to ensure that the model only needs to analyze one set of data. The datasets were merged based on common Incident Numbers and any overlapping columns there were in more than one of the datasets were removed.

The next step was to arrange the data such that each row of the dataset represented a unique SAR operation, instead of a unique missing person. This is because many of the variables will be the same regardless of the specific person in the group, such as weather, location, SAR resources used, and even the group's last known activity. To ensure data specific to the people was not lost, several variables were added to the dataset, including, the number of people in the missing group, the minimum and maximum ages in the group, and the proportion of females and proportion of males in the group. After this, there were 3004 rows or missions in the dataset.

12.1.2 Quantitative to Categorical

Since this dataset has many variables and a relatively small number of rows, it is important to reduce the complexity of the variables to ensure that the model has the ability to find relationships between the variables. To this point, any quantitative variables in the dataset were binned into discrete categories. For example the start time of the incident was binned into four hour intervals, and the subject ages were binned

into groups for every additional decade. This can help reduce the complexity of the Bayesian network and deal with sparse data with large gaps between values within each variable.

To do this, each quantitative variable was analyzed separately to determine the best number of bins or categories for that variable. All of them were limited to a maximum of 10 bins. The specific category limits were determined based on the spread of each variable, aiming to have similar counts of missions within each category. Large spikes at specific values were typically given their own bins as they could potentially represent an important value for that variable.

12.1.3 Combine Categories and One-Hot Encoding

Similarly, some variables that were already categorical also needed to be re-categorized into a smaller number of categories due to the same complexity concerns for the Bayesian network. Similar categories were combined, such as Boating and Kayaking, since these would likely elicit similar SAR operations.

Additionally, one-hot encoding was used to create indicator variables for many of the categorical variables with many levels that were not easily combined. One-hot encoding changes a variable with multiple categories into a new variable for each of those categories, with only the values of 'Yes' or 'No' in these new variables to indicate whether this category applied to them. For example, take the variable which denotes the SAR equipment used during the operation, Equipment, with categories ('Helicopter', 'Boat', 'Technical Rope', …). Each mission could have used any number of those categories, and they were recorded as a list in that column. With one-hot encoding, each of those categories became their own variable and were assigned 'Yes' or 'No' depending on if they partook in that activity. This helps reduce model complexity because a long list of equipment would be difficult for the model to analyze and associate with all of the other variables for each mission. These indicator variables with only two levels are much easier for the model to understand, and it allows the model to identify which specific pieces of equipment are important for the model to know and which pieces of equipment are not as important.

Occasionally, these indicator variables overlapped in meaning with each other, so some of them were combined into one variable, such as 'Snow Terrain Indicator' and 'Snow Indicator.' This further helped reduce complexity and the number of associations the Bayesian network would need to investigate.

12.1.4 Remove Sparse and Redundant Variables

Finally, some of these new variables had very few data points for some of the categories. If there are thousands of missions, and only a couple are based on fugitives, or on chainsaw related activities, for example, it would be very difficult for the model to find meaningful associations with these variables and any of the others. Thus, very sparse variables or categories were either recategorized as made sense or removed from the dataset. To this point, all fugitive based missions were removed from the dataset and the 'chainsaw related activity' indicator variable was dropped. Several more variables were treated similarly, as needed. Additionally, any unique identifiers were also dropped at this time.

Additionally, many of variables became redundant after becoming indicators. For example, indicators that the value was a 'NA' value or the variables that represented the incident start year, notification year, and close year were also redundant as they are obviously associated. These obvious, but not useful associations led to the elimination of all but one of the variables that were obviously associated. If the associations were less obvious, or if there was a potential for obviously associated variables to also be separately associated with other variables, then the variables were not removed at this point.

12.2 Exploratory Data Analysis

After the above process was completed, there were 3004 rows and 84 variables in the dataset. This section will investigate a few of these variables to get an understanding of the data.

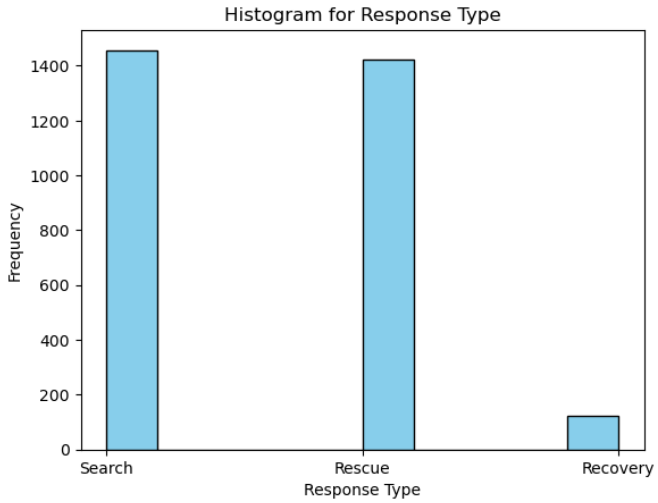

Figure 2: Histogram of Response Type

Figure 2 shows the distribution of missions in this dataset that were classified as ‘Searches,’ ‘Rescues,’ or ‘Recoveries.’ These classifications are assigned by Search and Rescue teams based on the mission’s reasoning or cause. A ‘Search’ would be when the SAR team does not know where the subject is located, a ‘Rescue’ is where they might know the subject’s location but they are in some way prevented from getting home alone, and ‘Recovery’ is where they are recovering a dead body. Most cases in this dataset are either ‘Search’ or ‘Rescue.’

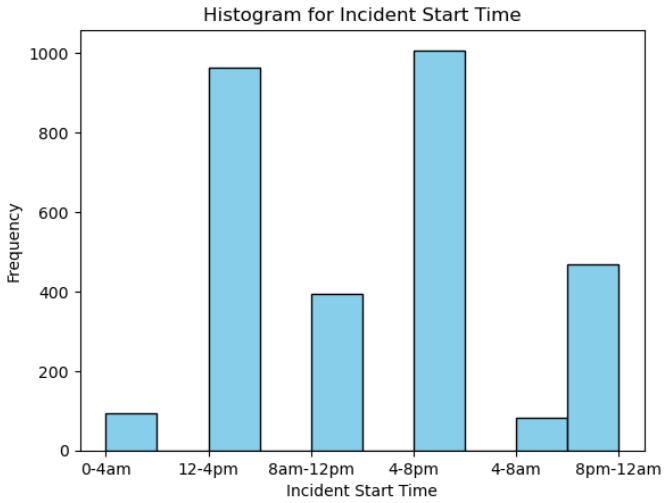

Figure 3: Histogram of Incident Start Time

Figure 3 shows the distribution of start times for each incident. As shown, there are peaks between 12-4:00pm and 4-8:00pm, indicating that there might be some relationship between these times and incidents.

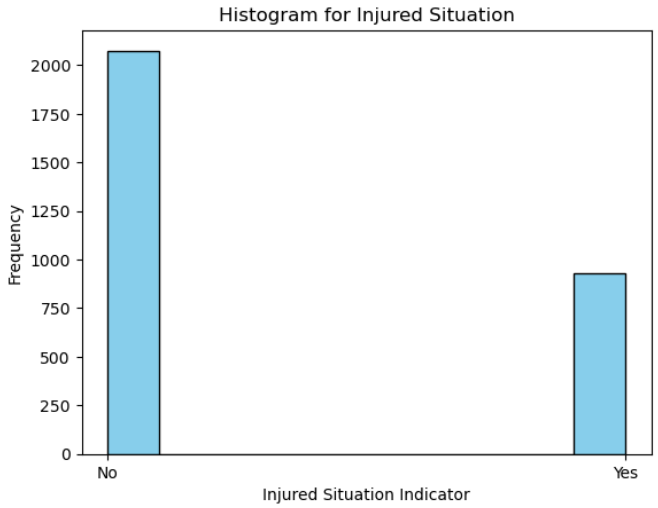

Figure 4: Histogram of Injured Situation Indicator

Figure 4 shows the distribution of incidents that were caused by an injury. Approximately a third of incidents were due to injuries.

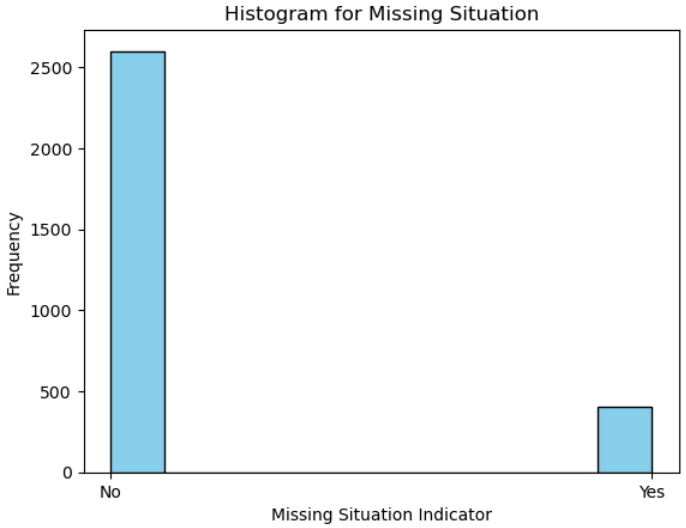

Figure 5: Histogram of Missing Situation Indicator

Figure 5 shows the distribution of incidents that were caused by one or more of the subjects involved in a mission being reported as missing. Most missions were not caused by missing subjects.

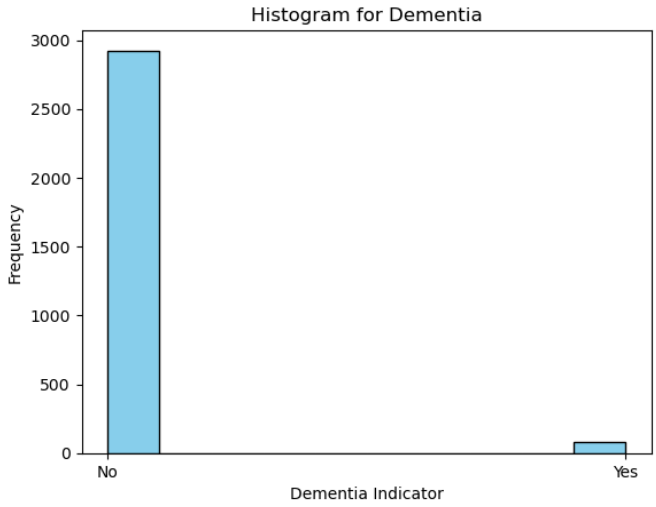

Figure 6: Histogram of Dementia Patient Indicator

Figure 6 shows the distribution of incidents where the missing subject had Dementia or Alzheimer's. Most incidents in this dataset were not about a subject with Dementia.

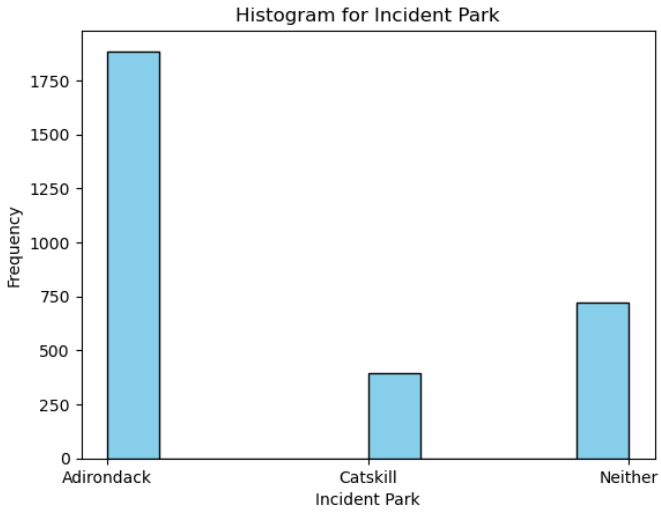

Figure 7: Histogram of Incident Park

Figure 7 shows the distribution of incidents that took place in the Adirondack Park, the Catskill Park, or neither. As shown, many cases from the Adirondack Park, which makes sense as it is incredibly large.

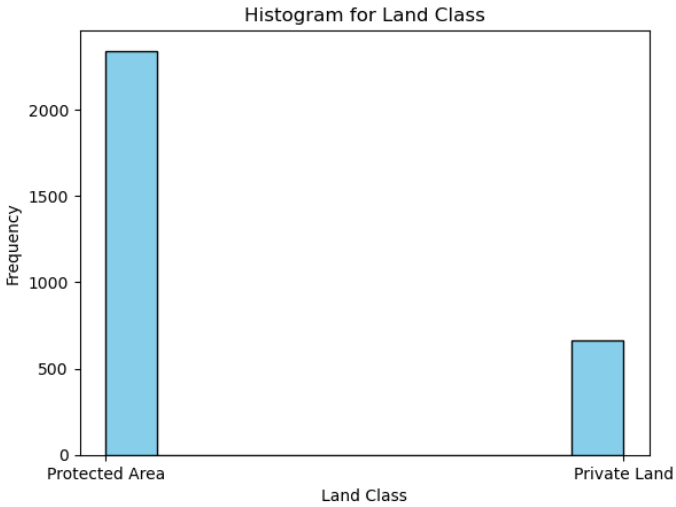

Figure 8: Histogram of Land Class

Figure 8 shows the distribution of incidents that happened on either protected areas, such as forests and parks, or on private land. Approximately three-quarters of the incidents happened in protected areas.

There are many more variables available in this dataset that will not be shown in the body of this thesis. To see more of the variables, visit the Appendix.

12.3 Bayesian Network Creation

There are many types of statistical and machine learning models that exist that can be used to learn associations between variables. Many times, the goal is to predict the value of one variable, such as predicting stock changes or predicting whether an image is a dog or a cat, which would necessitate the use of a prediction-based machine learning model. However, there also exist inferential models that focus on discovering associations between all variables in the model. A Bayesian network is a type of inferential model which represents the associations between variables as a network, where every node in the network is a variable, and every edge represents the conditional probability between the two variables it connects.

Using machine learning to create a Bayesian network allows the algorithm to consider two forms of inputs to create the final model. They can take in data, where the associations between variables are based on their mathematical associations. They can also take in a prior model, which is where the user can specify a Bayesian network that they think might represent the data. The algorithm then uses this prior model as a

starting place and uses the data to change it as needed. Some people believe the use of a prior model introduces subjectiveness to the results of the final model because each researcher could provide a different prior model and might get slightly different final models. Bayesian statisticians argue that although priors are subjective, there are many decisions made in any modeling process that are subjective, even for typical frequentist analyses. Although the prior model will slightly affect the final model, most of the information will come from the data, so Bayesians argue that this slight subjectivity does not negate the use of these models (Robert 105-106). However, it is also possible to use something called an 'uninformative' prior which essentially represents a model with no prior information so that all information comes strictly from the data. This is the type of prior used in this thesis.

12.3.1 Model Specifications

The models and code for this thesis were written in Python and the main library used to both create the Bayesian network and make inferences was *pgmpy*.

To construct the Bayesian network, an uninformative prior was used where each of the variables were considered completely unrelated to each other. This allowed all the associations found by the model to come purely from the data, with no information from previous knowledge about the variables and their relationships.

A heuristic search algorithm known as Hill-Climb was employed to find the "best" Bayesian network possible based on the data. This algorithm begins with a starting model, which is the uninformative prior described above. Then, it applies one change to each of the possible associations in the model separately and assigns each one a score. It selects the change that improved the score the most and that becomes the new model. A 'change' at any step could be an edge addition, deletion, or reversal. This process is repeated iteratively until the score stops improving, and the final network is selected (Koller & Friedman 814). The specific scoring method used by the algorithm can be determined by the user. For this thesis, the Bayesian Information Criterion score was used, which is described in the next section.

Since Hill-Climb is a greedy algorithm, it only considers a one-step change and checks if it will improve the score, so it is possible for it to become stuck at a local maximum (Koller & Friedman 815) where any

new change will reduce the score. This means that a better scoring model could technically be possible. This is because the algorithm only considers which single association is the best at any step, so it never found that adding a couple of the lower scoring associations ends up becoming a better model than adding the one strong association, for example. Thus, the final model determined by the Hill-Climb algorithm may not be the best model possible, so further model pruning may be needed.

12.3.2 Model Scoring

Since inferential models do not have a response variable, typical statistical techniques for model comparison do not work. If models have a response variable, they can be compared based on how much of the variation in the response variable the model explains. Without a response variable, it is hard to determine which model is the best because they each might explain more variability in differing variables in the model. Instead, score-based comparisons can be made. Two common scores used in model comparisons are called the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These statistics both compute the log-likelihood (based on Maximum Likelihood Estimation), which aims to maximize the conditional probability of observing the data given a specific probability distribution and its parameters (Koller & Friedman 802). They also have a penalty for complex models, where the more parameters included in the model, the larger the penalty. The difference between AIC and BIC is that BIC has a larger penalty for additional parameters, and thus prefers simpler models.

The values of AIC and BIC are not directly interpretable as they have no units or meaning alone, but they can be used to compare values between models. After training many models on the data, the model with the optimized AIC or BIC value could be considered the ‘best’ model. However, since these values are calculated using approximations, they are heuristic values and might not guarantee that the model with the optimized value is truly the ‘best’ model to exist. Additionally, since there are no units or meaning attached to these values, it could be that all of the models are very bad at representing the data, and the ‘best’ model is really just the least bad model. Due to these concerns, it is important not to treat AIC and BIC values as definitive statements on the model’s usefulness or quality. Each model suggested by these values should be carefully examined to determine if the model is a reasonable network and provides useful information.

Since the dataset used in this process has a large number of variables with comparatively few observations, this analysis will use BIC to combat model complexity. The goal with this score is to maximize the likelihood of the data given the model. Based on the formula used in the *pgmpy* library, the aim is to maximize the BIC (Note: In typical statistics-based calculations of AIC and BIC, the goal is to minimize these values as they are calculated using the negation of the likelihood. The library *pgmpy* in Python uses a formula that does not multiply the likelihood by a negative number, and thus aims to maximize the BIC (Koller & Friedman 802)). The next section will walk through the model selection and comparison process used to determine the final Bayesian network model.

12.4 Modeling Process

In order to build Bayesian networks, it is important to determine which variables are important to the model and to the audience in order to create an accurate and meaningful model. To do this, variables that are irrelevant or overlap in meaning with others need to be removed to allow the model to focus only on the relationships between important or useful variables.

After data pre-processing, the full dataset had 85 variables about 3004 SAR operations. Since a Bayesian network attempts to find a meaningful association between all pairs of variables in the dataset, having 85 variables could create a very complex Bayesian network. Additionally, since the final Bayesian network will not include any variables that did not have meaningful associations, it is possible to reduce the size of the dataset based on variables that are not found relevant by the BN. To determine which variables were least important to the modeling process, a function was built to determine how often each of the variables were selected by the Bayesian network.

The function works as follows:

1) Take a random sample of some number *k* variables from the dataset,
2) Create a Bayesian network with those sampled variables,
3) Calculate the BIC score for this Bayesian network,
4) For each variable in the network, add the BIC to each variable's running total BIC score,
5) For each variable not in the network, add a large negative number to penalize their lack of selection,

6) Repeat steps 1-4 some large number *n* times,

7) Graph the running total BIC scores for all variables.

Based on a graph of the running total BIC scores for all the variables, the variables with the smallest (closest to zero) scores were most often selected to be included in the Bayesian networks and the variables with the largest (farthest from zero) scores were least often selected. Thus, the largest scoring variables were considered unimportant relative to the other variables that scored higher. Then, the lowest five scoring variables were removed from the dataset and the function was rerun on this smaller dataset, repeating this process until the BIC scores stopped improving.

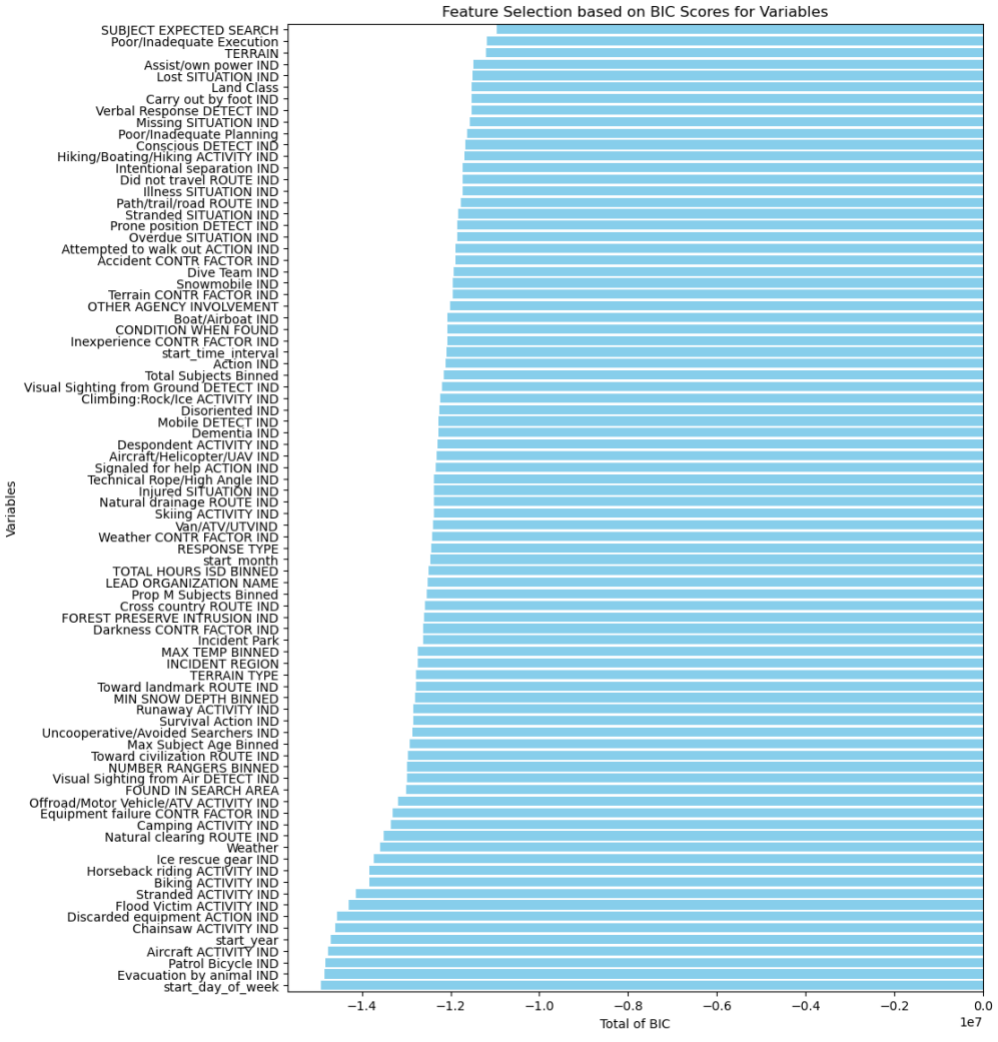

Figure 9: BIC Score Ranking for all possible variables

Since at each run of this loop, there is a random selection of *k* variables, the non-selected variables are also getting the same punishment for not being selected by the Bayesian network as the 'unimportant' variables. This is because the BIC scores were negative, and so if the non-chosen variables were allowed to have a score of 0, they would be seen as more important than variables that were truly selected by the Bayesian network. Thus, it is technically possible that some of the variables were unlucky and never chosen by the model, leading to a lower score than they would have received had they been selected. However, this is a very remote possibility due to the values of *k* and *n* chosen. Typically, this function was called on k=20 and n=150, which is a high number of repetitions and since there were 84 variables total, approximately a quarter of all the variables were chosen for each of those 150 runs. Additionally, since the function was run multiple times, any large changes that were due to very unlucky randomness would have been noticed between the runs of the function.

Figure 9 shows the ranking for 84 of the variables originally available to the model. Based on this graph, the variables with the shortest bars were most often selected by the model, and those with the longest bars were least often selected. Variables such as the year the mission started and the day of the week that the mission started were two of the least selected variables, as shown by the having the longest bars. The bottom five variables would be removed from the dataset and a new Bayesian network and BIC would be calculated from the smaller dataset. Then, this function would be rerun and more of the least important variables removed until the BIC stopped improving. This process was repeated five times until the BIC started to stabilize.

However, this function only determines which variables were consistently chosen to have associations to other variables in the model, it does not determine or analyze how important those relationships were. For example, the function would say that the incident start month and the amount of snow were important because they were selected in every model as having an association. This is a well-known relationship, however, and it is not worthwhile to use the limited amounts of data to model it. Thus, the function could only be treated as a way to find unimportant variables that never or rarely were associated with other variables. It could not be treated as a way to find important variables because some of the associations it

was considering provided unnecessary information about variables that were obviously related and uninteresting to model.

Due to this, it was important to manually inspect and prune the Bayesian network for uninteresting associations. When a variable only had one, uninteresting, relationship, it would be removed from the model and the BIC scores were compared with and without it to determine if it was uninteresting and unimportant, or uninteresting yet still important to the model. If a variable had multiple relationships, this process was typically not carried out, as usually at least one of them was important, if uninteresting. An example would be the association between mission start month and their activity being skiing, where start month had no other associations. Due to this, start month was uninteresting, and after the BIC increased after it was removed, it was also unimportant.

After the above processes were complete, the variables important to the Bayesian network were selected. To ensure the Bayesian network was consistent, the model was rerun five times on the same variables to ensure the same model was built and that the BIC did not change. This also verified that none of the associations were being chosen randomly.

Chapter 4

RESULTS

4.1 Final Model

After the above processes were carried out, the final model was selected, as visualized in Figure 10.

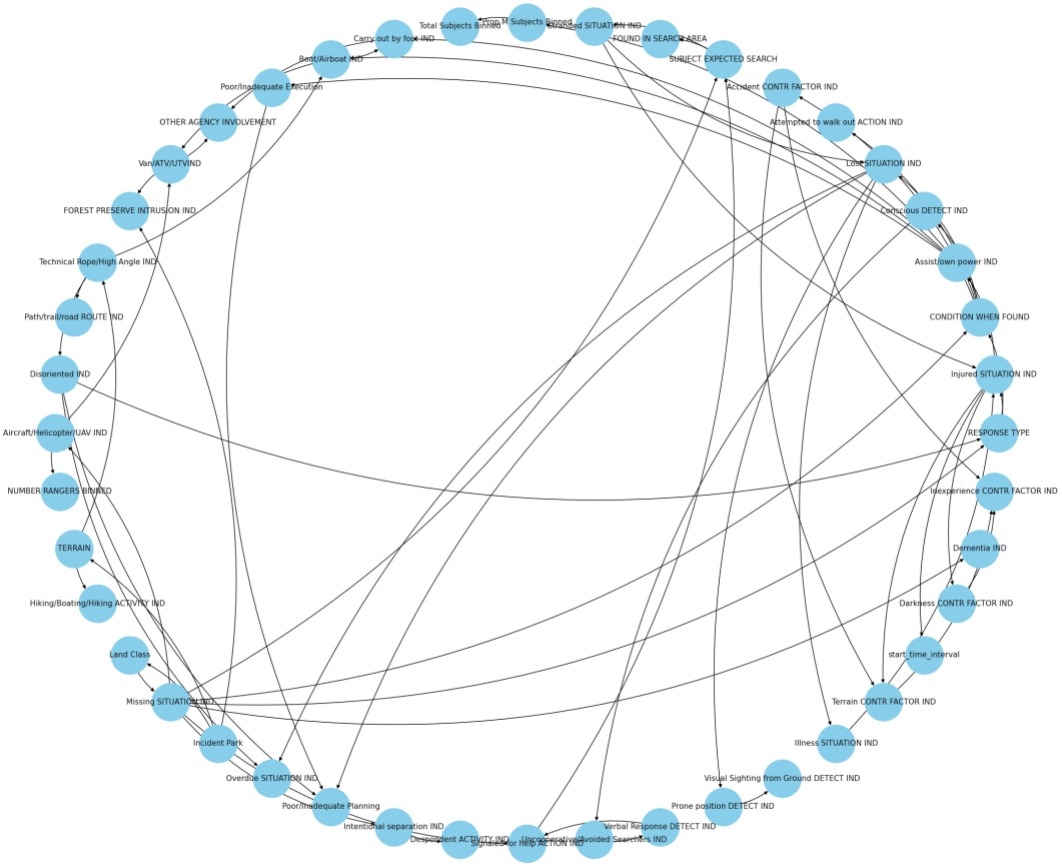

Figure 10: Final Bayesian network represented as a DAG

This final network includes 44 of the original 116 variables. The full model can be difficult to read as there are many arrows that overlap; however, it is still interesting to have a visualization of the full model. As specific associations are analyzed, a closer view of the network will be shown.

4.2 Inferences

To begin analyzing this Bayesian network, a first step is to assess the conditional distributions of associated factors. As there are so many associations in this model, it will not be possible to analyze all of them in this thesis. Select associations will be chosen to investigate further. For example, in the Bayesian network there is an association relating the *Injured Situation Indicator* variable to the *Incident Start Time Interval* variable, as shown closer in Figure 11.

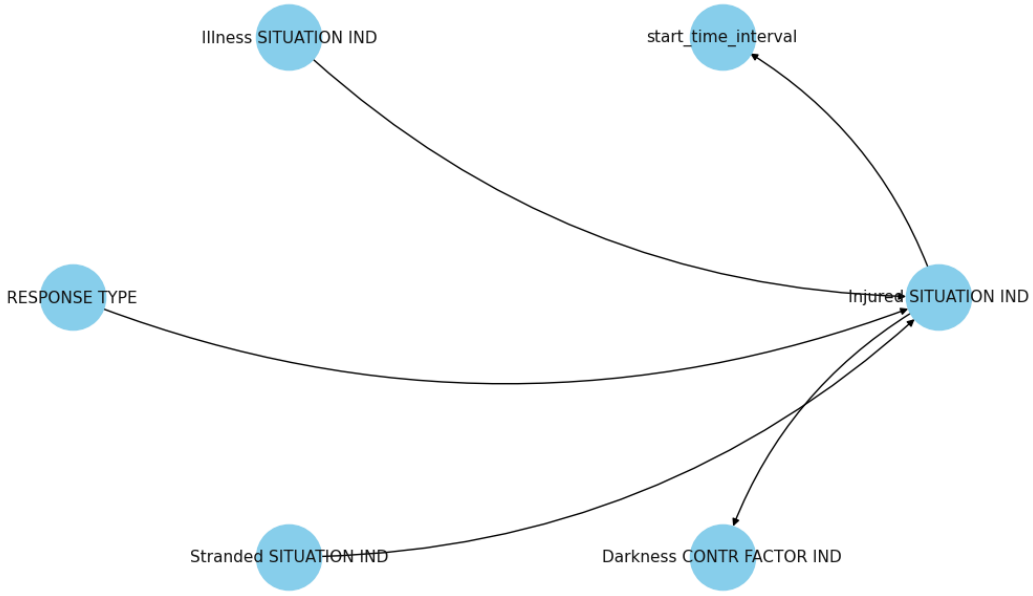

Figure 11: DAG of Injured Situation Indicator and its parents

Figure 11 is a closer view of the Bayesian network, with all of the parents of *Injured Situation Indicator* and two of its children. Table 1 shows the conditional probabilities of the *Injured Situation Indicator* value given the *Incident Start Time* values. The *Injured Situation Indicator* would be a 'Yes' if the subject turned out to be injured during the mission and 'No' if not. The *Incident Start Time* is the time at which a person has been reported to be missing or in need of aid.

Table 1 shows that incidents that begin between 8:00am to 4:00pm have a higher probability of being related to a situation due to an injured subject than if the mission begins from 4:00pm to 4:00am. Thus, incidents beginning during the daytime could tell the SAR teams to treat the incidents more similarly to how they treat incidents for injured subjects.

Table 1: Conditional Probability of Injured Situation Given the Incident Start Time

Start Time Interval	Injured Indicator	Probability
00:00 – 4:00 AM	No	0.9110
	Yes	0.0890
4:00 – 8:00 AM	No	1.0000
	Yes	0.0000
8:00 AM – 12:00 PM	No	0.4960
	Yes	**0.5040**
12:00 – 4:00 PM	No	0.5171
	Yes	**0.4829**
4:00 – 8:00 PM	No	0.7233
	Yes	0.2767
8:00 – 12:00 AM	No	1.0000
	Yes	0.0000

Table 2: Conditional Probability of Injured Situation Given Darkness Contributing Factor

Darkness Contributing Factor Indicator	Injured Indicator	Probability
Yes	No	0.9692
	Yes	**0.0308**
No	No	0.6203
	Yes	**0.3697**

Table 2 shows the conditional probability of the *Injured Situation Indicator* value given the *Darkness Contributing Factor Indicator* values. *Darkness Contributing Factor Indicator* is recorded as a 'Yes' or 'No' after the mission to indicate whether darkness contributed to the subject becoming lost. This shows that given that darkness was a contributing factor of the cause of the incident, there is a lower probability of the situation leading to the subjects needing help is being injured. This seems counterintuitive but if the SAR teams know that the reason they are looking for someone who is injured, they were likely to have been doing an activity that led to them being injured and unable to get home, instead of being lost. Since

they were doing an activity, they were likely doing so during the day, and that is when they became injured. Note that one of the parent nodes for *Injured Situation* is *Response Type*. *Response Type* is the type of SAR mission carried out, a search, rescue, or recovery. As mentioned previously, these differ based on if the subject's location is known or not from the beginning of the mission and whether they are retrieving someone who is unable to get themselves home or if they are retrieving someone who is already deceased.

Table 3: Conditional Probability of Injured Situation Given Response Type

Response Type	Injured Indicator	Probability
Recovery	No	0.5111
	Yes	**0.4889**
Rescue	No	0.2618
	Yes	**0.7382**
Search	No	0.9902
	Yes	**0.0098**

Table 3 shows that given the type of mission the SAR chooses to implement, there are large changes in the probability of the situation leading to the missing subject needing help was because they were injured. This makes more sense intuitively, if the SAR team knows that the subject was injured when they began their mission, they are most likely to conduct a rescue operation, since the subject was not lost, just unable to bring themselves back home. This ties back in with the injured situation having higher probabilities of not having to do with darkness and occurring during the daytime. A subject who is injured, but not lost was likely doing some sort of daytime activity that went awry.

Another set of associations are shown in Figure 5, focusing on the situations where the subject was missing.

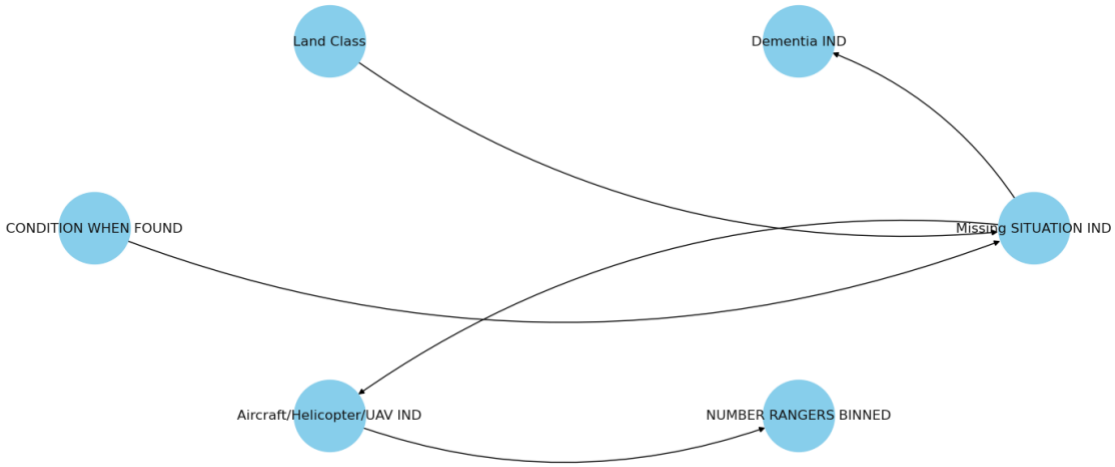

Figure 12: DAG of Missing Situation Indicator and some of its associations

Figure 12 shows only a few of the associations that *Missing Situation Indicator* has in the full Bayesian network. *Missing Situation Indicator* is recorded as 'Yes' or 'No' depending on why the person needed a SAR mission and could be recorded after the fact. Other options would be illness, injury, being lost etc. Based on this small version of the Bayesian network, Figure 12 shows that whether or not the subject's situation was that they were missing is related both to the classification of the land where the search is conducted and whether or not the subject has Dementia or Alzheimer's.

Table 4: Conditional Probability of Land Class Given Missing Situation

Missing Situation Indicator	Land Class	Probability
Yes	Private Land	**0.7619**
	Protected Area	0.2381
No	Private Land	0.1463
	Protected Area	**0.8537**

As shown in Table 4, given the person went missing, their last known location was much more likely to be on private land than in a protected area, such as a forest or park. Given the person needed SAR's aid for some other reason than being missing, such as being lost or injured, their last known location was much more likely to be in a protected area than on private land. If the person was lost or injured, it would make

sense that a SAR team was called to help them if they were in a forest or other rough terrain that typical police or ambulances could not reach. However, it is slightly surprising that so many of the missing situations occur on private land. Additionally, in Figure 12 it can be seen that *Missing Situation* is also associated with *Dementia*.

Table 5 shows that given the person had Dementia or Alzheimer's, all the cases in this dataset were because they went missing. However, given the person did not have Dementia, there was a fairly high probability, 0.8941, that the case was not due to a missing subject. This indicates that most cases that are unrelated to Dementia are not for missing subjects, but for other types of situations, such as lost or injured subjects. Thus, many of the missing subject cases are explained by whether the subject has Dementia.

Table 5: Conditional Probability of Missing Situation Given Dementia

Dementia Indicator	Missing Situation Indicator	Probability
Yes	No	0.0000
	Yes	**1.0000**
No	No	**0.8941**
	Yes	0.1059

The associations represented in the full Bayesian network can be represented similarly, using the conditional probability distributions. However, it is also possible to use causal inference to show how intervening on past missions could have changed the probabilities of different variables.

4.3 Causal Inferences

Continuing with analyzing associations found in Figure 12, the do-calculus method can be applied to find the causal effect of missing situation on the subjects condition when they are found.

Table 6: Conditional Probability of Condition When Found Given do(Missing Situation)

do(Missing Situation Indicator)	Condition When Found	Probability
do(Yes)	Deceased	**0.3071**
	No Medical Assistance Required	0.3106
	Required Medical Treatment	0.3823
do(No)	Deceased	**0.0561**
	No Medical Assistance Required	0.5118
	Required Medical Treatment	0.4321

After intervening on the missingness of the subject, there is a causal effect that people who are missing have a higher probability of being found deceased, from 0.0561 to 0.3071. If it were possible to retroactively change people's situation to be missing, all else held equal, there would be increased likelihood of those missions ending in a deceased subject. Thus, SAR teams need to focus their efforts on finding missing subjects as quickly as possible as this variable is causing higher probabilities of death.

Going back to Figure 11, there is no direct association between *Stranded Situation Indicator* and *Darkness Contributing Factor Indicator*. They are indirectly associated through *Injured Situation Indicator*. However, by applying do-calculus, it can be shown that there is some causal relationship between these variables.

Table 7: Conditional Probability of Darkness Contributing Factor Given do(Stranded Situation)

do(Stranded Situation Indicator)	Darkness Contributing Factor Indicator	Probability
do(Yes)	No	0.8438
	Yes	**0.1562**
do(No)	No	0.8871
	Yes	**0.1129**

Table 7 shows that the intervention on *Stranded Situation Indicator* has a small causal effect on the *Darkness Contributing Factor Indicator*. This increase of probability from 0.1129 to 0.1562 of darkness

being a contributing factor is caused by the subject being stranded. It is hard to say whether this effect would be statistically significant, so there are some limitations to how much trust should be placed in this causal effect.

Unfortunately, there are not any associations between these variables that provide new or momentous information. Additionally, there are a lack of actionable association, where the SAR could change one of the variables to change the outcome. Many of these associations were between variables that cannot be changed by the SAR team during the course of a mission, and thus this limits the use of this Bayesian network to only be able to look at probabilities of what the missing person might do, instead of what the SAR team can change. With more data, however, more interesting associations could possibly be found and the Bayesian network could accommodate more variables.

4.4 Scope and Concerns

The most pressing issue with this Bayesian network is the large number of variables in comparison to the relatively small number of missions. This means that the associations found by the model and the number of variables allowed in the model were limited, and the ones that were included in the model could have parameters that are unstable due to the small sample size.

There is also always a concern with Bayesian networks that a different prior graph, scoring method, and scoring criterion could lead to a different final model, which can make it hard to trust that the selected model is truly the optimal model.

Finally, causal inference based on observational data is still considered untrustworthy by many statisticians. Combined with the small sample sizes and the inability to determine statistical significance of parameters with a Bayesian network, and there are many concerns that can limit how much trust is put into the causal effects.

The scope of findings from this Bayesian network and these analyses are limited to only similar missions in New York worked by the NYS Forest Rangers or similar entities, from the years 2012 to 2023. It also only applies to non-fugitive cases, as these missions were removed prior to the analyses.

Chapter 6

CONCLUSION

Bayesian networks can be used to represent the associations between factors that affect Search and Rescue missions, allowing for SAR teams to discover potentially useful information that could improve their searches. Any improvement that can lead SAR teams to be more accurate and efficient in finding missing people can help save lives. Although there were not any shocking or actionable associations discovered in this analysis, it has shown that a Bayesian network can be used for this purpose and given more data could become a helpful tool in discovering patterns between missions and between different types of missing subjects.

Future analyses would benefit from having access to much more data. Since SAR missions are so complex, the only way to model them with any effectualness is highly dependent on having access to plenty of missions and variables. Beyond that, having SME develop a prior Bayesian network could help supplement gaps in the data and provide directionality in the causal relationships.

Additionally, information gained from this Bayesian network could be used to supplement or compared to other models, such as Bayesian networks trained on different data, or entirely different types of models, such as neural networks.

Another possible avenue of interest could be to try clustering techniques to identify similar types of missing subjects, allowing targeted plans of action to be developed for each unique type of missing subjects. This would be similar to customer segmentation analyses but applied to SAR purposes instead.

There are many more types of machine learning and deep learning models that could be applied to this area, however, as always, the usefulness of any of these models depend on the amount and quality of data available. It would be prudent to encourage SAR teams around the world to collect, store, and share their mission data, to help advance search techniques for everyone.

REFERENCES

Heckerman, David. "A tutorial on learning with Bayesian networks." Innovations in Bayesian networks: Theory and applications (2008): 33-82.

Koester, Robert J. Lost Person Behavior: A Search and Rescue Guide on Where to Look, for Land, Air and Water. dbS Productions, 2008.

Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

Lu, Yonggang, Qiujie Zheng, and Daniel Quinn. "Introducing Causal Inference Using Bayesian Networks and do-Calculus." Journal of Statistics and Data Science Education 31.1 (2023): 3-17.

Pearl, Judea. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000.

Robert, Christian P., and Christian P. Robert. "From prior information to prior distributions." The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (2007): 105-163.

APPENDIX

The code used for this thesis is saved in Jupyter notebooks and be accessed on GitHub:

https://github.com/aebelden4/MastersThesis.git

The publicly available NYS Forest Rangers data can be found at:

https://data.ny.gov/stories/s/Wildland-Search-and-Rescue-Missions-by-NYS-Forest-/3uy6-27rt/