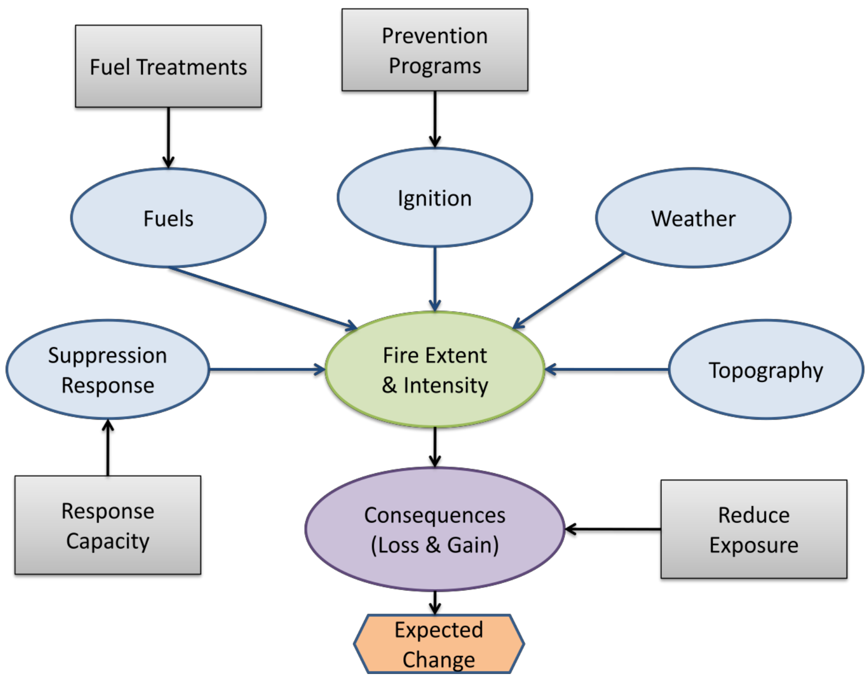
Appendix 2

The Analytical Approach, Including a Discussion of Key Datasets and Clusters

Conceptual Models

Wildland fire is a complex issue that involves multiple interacting factors spanning the natural, human, and built environments. During Phases I and II, the National Science and Analysis Team (NSAT) examined various aspects of wildland fire and developed conceptual models specific to each component. The purpose of these models was to display the interactions and relationships among different factors, such as the relationship between fuel treatments and the extent and intensity of wildfire. This conceptual model (shown at the left) shows possible options in the gray boxes that could be considered. These options are connected to the colored ellipses by "cause and effect" arrows which indicate the relationship between the options and the data. A series of options need to be considered and evaluated on effectiveness and the ability to implement in order to provide an expected change (at the bottom of the diagram). The NSAT analysis performed during Phase III, and as presented in this report, was designed to meet this requirement.



Data and Data Preparation

The NSAT identified various data sets to be used in Phase III to build analytical models consistent with the concepts articulated in Phase II. Building on these efforts, Phase III has involved an extensive effort to collect data necessary to quantify relationships and provide a rigorous examination of risk.

The types of data collected can be broadly categorized into five general types: biophysical, socioeconomic, land-use and ownership, wildfire frequency and extent, and incident response. Biophysical variables include physical measures such as precipitation, temperature, and terrain. They also include characteristics of vegetation which contribute to wildfire behavior. Socioeconomic variables describe the demographic and economic characteristics of populations and communities within each county, and also describe the distribution of homes within the wildland-urban interface (WUI). Land-use and ownership describes the mixture of public and private lands and also helps quantify the extent to which lands might be suitable for active management, e.g., by highlighting areas that have historically supported timber harvest. Variables describing wildfire frequency and extent have been gathered from various reporting systems that have been put in place by federal, state, and local fire departments. They also include data from independent monitoring systems that track wildfire using satellites and other remote devices. Finally, they include a series of modeled products from governmental and private entities. Similarly, incident response information has been gathered from many of the same reporting systems. These variables track who responded to wildfire, how long they took to arrive on site, and how long was required before the fire was contained. Information on injuries and casualties can also be found in these same reporting systems.

Before data were used in analysis, three additional steps were accomplished. The first step was one of quality control. Obvious errors in the data were corrected where it was apparent that the corrections would enhance the fidelity of the original data. In some cases, limited numbers of observations were omitted from further consideration due to obvious mistakes that could not be corrected and/or had missing information. The second step involved compiling, reformatting, and/or summarizing data to fit within a common sampling frame, at the county level scale. For some data sets (E.g. many of the social economic variables) data were originally provided at the county level and no reformatting was necessary. Other higher-resolution data were processed using GIS techniques to provide a county-level summary. Many data were also normalized to provide comparative area-based or incident-based metrics such as acres burned per hundred square miles or firefighter injuries per 1000 incidents.

The final step in data preparation involved filtering and consolidation. In this step, a preliminary correlation analysis was used to identify common patterns among the data that allowed a subset to be used to characterize conditions efficiently. That is, a smaller set of variables were identified that were highly correlated with other variables and could be used alone without significant loss of information. Statistical techniques including factor analysis and clustering were used to reduce the number of variables further by creating "super" variables that were either linear combinations of other variables (from factor analysis) or categorical groupings of counties based on their similarities (using cluster analysis). The combination of filtering and consolidation techniques allowed the total number of variables considered to be reduced by nearly two-thirds. Even so, there were over 100 variables available for potential analysis.

Modeling and Bayesian Networks

Various analytical models were constructed for the primary purpose of relating causal or contributing factors to variables which collectively index levels of risk. These risk metrics include measures of hazard such as frequency and magnitude of wildfire, any direct measures of loss or injury, and various measures related to exposure, such as the number or density of homes in the wildland-urban interface (WUI). Although hazard and loss are often combined into single measures of risk, such measures were not constructed in our analysis due in part to the county-level resolution of the original data. For example, there are homes distributed throughout the wildland urban-interface and in some cases, large wildfires are likely to occur within the county, but it is not possible to determine which portion of the county is most likely to experience wildfire or which off-site effects of wildfire might be relevant to overall impacts. Such spatial interactions are important for producing an accurate and precise estimate of risk. Lacking more specific information, a more straightforward and simple assumption is used, where the total risk is proportional to county-level hazard, exposure, and potential loss.

Many of the analytical models used in our analysis were constructed using Bayesian networks. Bayesian networks are decision analysis tools that use conditional probabilities to link variables together and express the degree of relationship between them. They provide a highly flexible modeling environment that works equally well with simple and complex problems. Here, we use a simple example using climate, fuel, and wildfire to illustrate the basics behind a Bayesian network. Consider the two graphs shown in Figure 1.

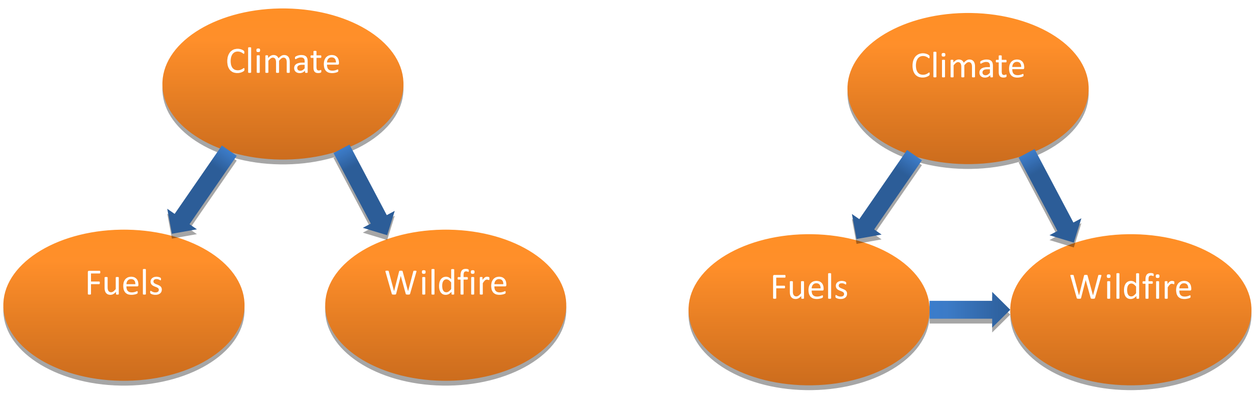


Figure 1. Simple graphical models of two possible hypotheses of the relationships among climate, vegetative fuels, and wildfire.

In the first graph on the left, it is assumed that climate affects both vegetation (fuels) and wildfire, but vegetative fuels and wildfire are independent given climate (e.g., there is no connection between fuels and wildfire that does not pass through climate). The second graph uses the same three nodes, but specifies a different relationship where vegetative fuels and wildfire are both related to climate, but vegetation has an additional direct effect on wildfire. The principal difference in the two graphs is that the first graph suggests that the manipulation of vegetation would have no measurable effect on wildfire. Only by changing climate could one expect wildfire to change. In contrast, the second graph allows for changes in vegetation to have an effect on wildfire independent of changes in climate. Importantly, quantitative models based on either graph could be based on exactly the same data, but they would have very different implications for management.

Bayesian networks begin with graphs like these, but then quantify the relationships using empirical data and/or expert opinion. Each node in the network can be represented by a single quantitative variable.

Arrows are used within the Bayesian networks to identify conditional dependencies or influences, much as the arrows in the graph above are used to relate one variable to another. The direction of the arrows are important, in that they indicate causal dependencies as well as determine how information can flow from one node to another. In this context, information is defined explicitly as that which causes a change in probability assignment. To facilitate calculation—as well as communication—continuous variables are often broken into discrete classes; discrete or categorical variables require no such modification.

As an example, consider the Bayesian network shown in Figure 2. This simple network has three nodes: *Region*, *Annual Ignitions*, and *Normalized Area Burned*. *Region* simply refers to the three regions identified within the Cohesive Strategy. *Annual Ignitions* is the mean number of outdoor fires reported per year, summed from three separate reporting systems representing federal, state, and local response units. *Normalized Area Burned* is an estimate of the expected number of acres burned in these reported incidents during a high-fire-occurrence year (i.e., the 95th percentile). This network was parameterized (trained) using data from all of the counties in the conterminous United States (lower 48 states), where each county was treated as a single observation and weighed equally regardless of area. The unconditional network (Figure 2) shows the marginal distributions of the values of each variable. One can see from the probability histograms, for example, that 33.4% of the counties are in the Northeast, 15% of the counties reported between 50 and 75 ignition points per year, and 14.3% of the counties might expect to burn 2000 or more acres (much more in some counties) in a year.

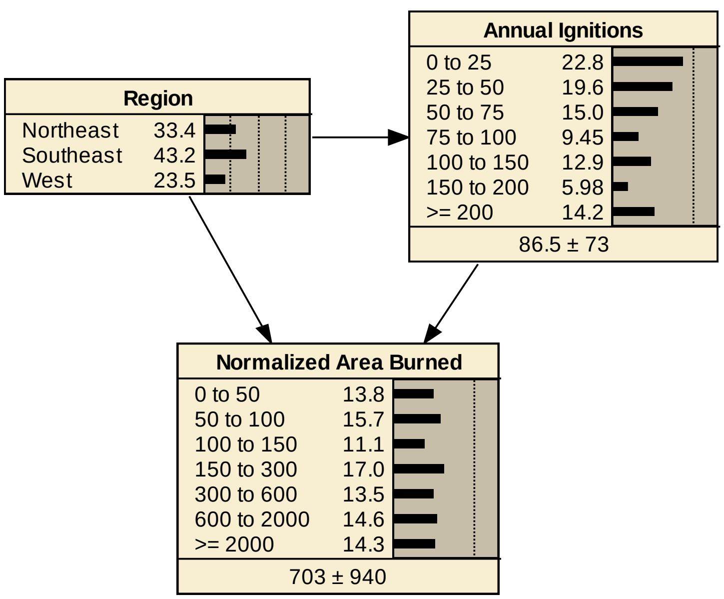


Figure 2. Simple Bayesian network illustrating the relationships among Cohesive Strategy *Region*, Annual Ignitions, and Normalized Area Burned. Probability histograms represent the percent of the counties within the conterminous United States within each class.

The Bayesian networks constructed for our analyses are necessarily more elaborate than the simple graphs depicted above, but they use the same basic concepts. For example, the network depicted in Figure 3 uses logic similar to Figure 1 regarding the relationship between climate, fuels, and wildfire, but expands that concept by using multiple nodes or variables for each component. This particular network uses three super variables (*Warmness Factor 1*, *Wetness Factor 2*, and *Terrain Factor 3*) from a factor analysis of physical attributes including seasonal precipitation and temperature, elevation, slope, and regional cluster analyses of vegetation and surface fuels. It also includes *Region*, *Annual Ignitions*, and *Normalized Area Burned* from Figures 2 and 3, and additional nodes from an independent modeling exercise, *Mean Burn Probability* and *Mean Flame Intensity*. A primary difference between the networks in Figure 3 and Figure 2 is the relationship between *Region* and *Normalized Are Burned* now passes through a series of intermediate nodes related to climate and vegetation, which allows for greater exploration of the causal factors influencing area burned by wildfires.

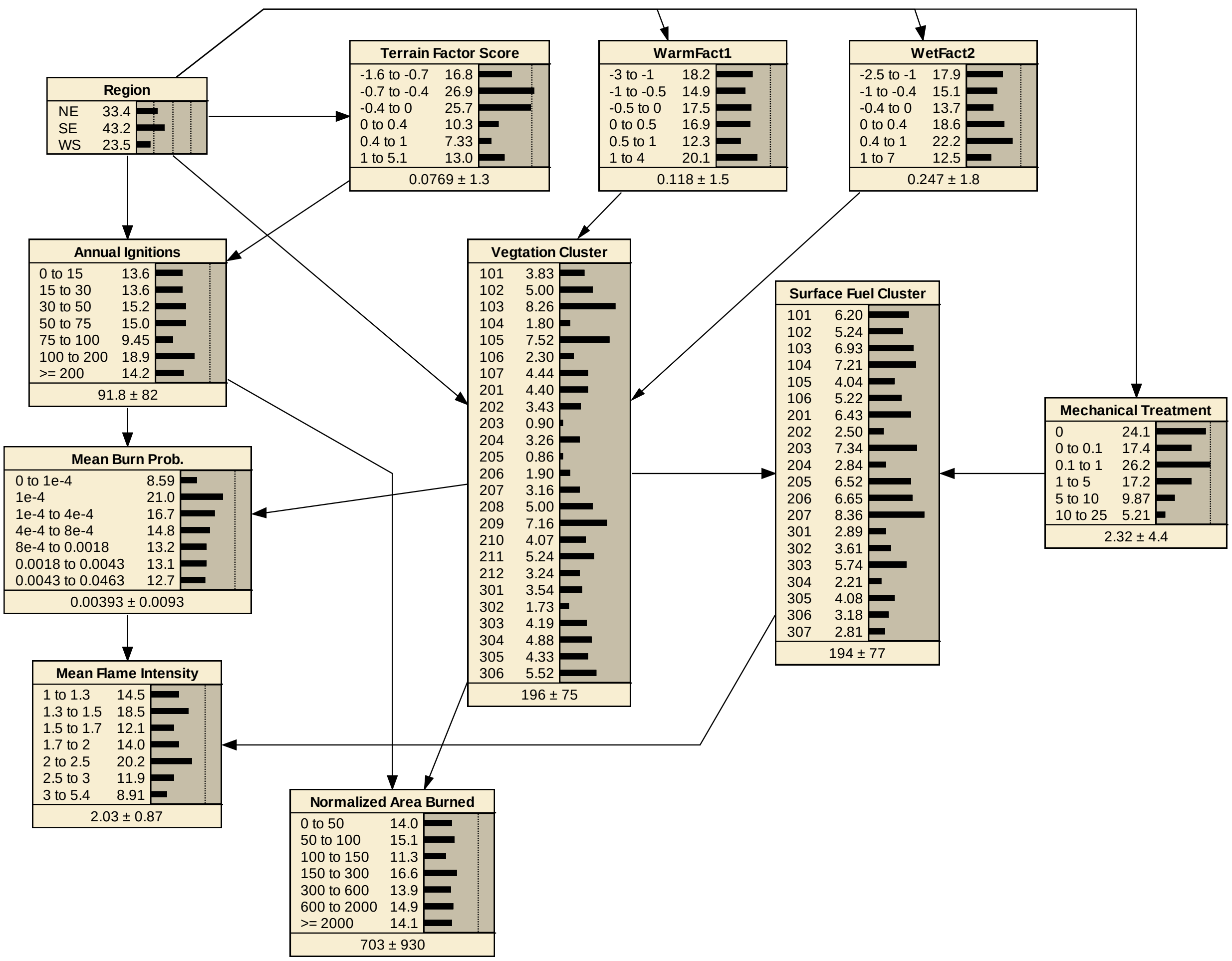


Figure 3. Bayesian network illustrating relationships among variables reflecting the physical environment, vegetation and surface fuels, mechanical treatments in forested areas, wildfire ignitions, and various measures of wildfire extent and intensity.

Construction of Cluster, Classification Trees and Combos

Included at the end of this section is a foldout (11x17) of a larger Bayesian network that shows the relationship of most of the key data variables to three new nodes called Resiliency Class, Community Cluster and Combination Class. Constructing these three new nodes became the next critical step of the analysis. From all of the work with the regions, it is apparent that the primary way that people value resources are connected to the two main drivers of Landscape Resilience and Community Resilience. To be able to compare fire impacts and possible options nationwide, the NSAT needed to characterize the nation based on these two main drivers.

The first step was to determine which of the key factors can characterize similarities and differences among counties with regard to landscape resilience and communities. This simplification, from the hundreds of variables available at this resolution, provides us with a platform to broadly address regional tradeoffs between resilience, response and human-fire interactions.

Statistical techniques including factor analysis and clustering were used to construct the Community Clusters. Cluster analysis is a categorical grouping of counties based on their similarities. A different, but comparable process of using a classification tree was used to determine the Landscape Resilience Classes.

The description of these classes and clusters are found in Appendix 3 and Appendix 4.

Once these classes and clusters were determined, the two were intersected to be able to understand how the Landscape Classes and Community Clusters intersect and impact one another. This yielded the "Combination Class" node seen in the large Bayesian network and these "Combinations" are described in detail in Appendix 5.

Modeling the Options utilizing the Combos

Using these combinations and the Bayesian network, the analysis focused on options that WFEC had created during a planning retreat held in Denver in February, 2013. These can be found in Appendix 6.

Trade-offs and Synergies

The NSAT has constructed some visualizations and comparative tools to be able to work with the WFEC the week of June 24th to be able to discuss which value drivers will be driving the prioritization of the use of resources. This discussion will need to focus on trade-offs and synergies when employing different combinations of options at both a national and regional scale. These visualizations and comparative tools can be found in Appendix 7.