

Falsification-Aware Causal Structure Discovery for Wildfire Severity under Spatial and Environmental Confounding in Central Chile

Efraín Campusano*, Marcela Quezada-Araya, Hernán Astudillo, Pablo Schwarzenberg, Paulo Quinsacara and Ernesto Vivanco

Universidad Andres Bello, Viña del Mar, Chile

Abstract

Wildfire incident databases often combine variables recorded at different stages of an event: pre-existing terrain and spatial context, meteorological descriptors associated with the incident, and burned-area variables that summarize final impact. Treating all of these fields as if they had the same temporal and causal role can produce strong statistical relationships that are valid within the database but difficult to interpret physically. We examine 6,610 wildfire records from the Maule Region of Chile using a cleaned dataset of 22 event-level variables. Before graph learning, the data matrix is audited for relationships induced by the recording process; several burned-area subtotals and total burned area are shown to be exact algebraic functions of their components. We then apply a greedy directed acyclic graph search with a Gaussian Bayesian Information Criterion (BIC) score and no predefined tier-based orientation constraints. The selected graph contains 22 nodes and 69 directed edges and separates a smaller terrain-meteorology-spatial-context subsystem from a dense vegetation and impact-accounting subsystem. Because nominal codes and the circular wind-direction variable are approximated within a Gaussian local score, because important common causes are unobserved, and because the sample is conditioned on recorded fires, the learned arrows are treated as exploratory dependency hypotheses rather than established physical causal effects. Graph-level compatibility is evaluated using 1,000 topology-preserving variable-to-node relabelings. The observed Local Markov Condition violation fraction is 0.1414 (27/191), compared with a null mean of 0.5423 ± 0.0432 ; the parent-association violation fraction is 0.0000 (0/69), compared with 0.4550 ± 0.0607 . Both add-one empirical p-values are 0.000999. These results show that the semantic placement of variables is substantially more compatible with the observed dependencies than random relabelings of the same topology, while neither the global falsification test nor the current data establish edge-level causal orientation. The resulting DAG is therefore used primarily as an exploratory database audit and as a structured basis for future, independently validated causal analyses.

Keywords

Wildfire, dependency structure, causal discovery, database audit, graph falsification, burned-area severity, spatial confounding, Chile

1. Introduction

Wildfire risk in central Chile is shaped by processes that operate simultaneously at climatic, landscape, and human scales. The Maule Region is a particularly relevant case because Mediterranean climatic conditions coexist with extensive forestry plantations, agricultural land, rural settlements, and heterogeneous terrain. Long dry summers favor fuel desiccation, while prolonged hydroclimatic deficits can further reduce vegetation moisture and increase the probability that an ignition develops into a severe event. The 2010–2015 central-Chile megadrought illustrates how persistent water deficits altered the regional environmental background [1]. The extreme 2017 fire season subsequently showed how drought, fire weather, vegetation exposure, and human pressures can combine to produce unusually large impacts [2, 3].

ICAIW 2026: Workshops at the 9th International Conference on Applied Informatics 2026, October 13–17, 2026, Orlando, United States

*Corresponding author.

✉ e.campusanomelendez@uandresbello.edu (E. Campusano); m.quezadaaraya@uandresbello.edu (M. Quezada-Araya); hernan.astudillo@unab.cl (H. Astudillo); pablo.schwarzenberg@unab.cl (P. Schwarzenberg); p.quinsacaracofre@uandresbello.edu (P. Quinsacara); e.vivancotapia@uandresbello.edu (E. Vivanco)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

These mechanisms motivate predictive modeling, but prediction is not the only question of interest. A predictive model is free to exploit any stable association that improves performance. A causal analysis instead asks whether a relationship can be interpreted as part of a mechanism and, eventually, whether an intervention on one variable would change another. This distinction is especially important in operational incident databases. A single wildfire record can contain variables that exist before ignition, variables that describe conditions around the event, administrative fields that proxy geography, and impact variables that are only known after the fire. If all of them are treated as equivalent predictors, strong statistical relationships can be recovered for very different reasons.

The CONAF incident records analyzed here illustrate this problem directly. The database includes terrain and meteorological descriptors together with vegetation-specific burned areas, subtotals, and total burned area. Several of the latter fields are exact algebraic aggregates of the detailed damage components. A graph-learning algorithm should detect those relationships because they are genuine dependencies in the recorded data. However, they should not be interpreted in the same way as a possible environmental pathway from weather or terrain to fire severity. The graph therefore has to be read both statistically and semantically.

Causal graphical models are useful for this purpose because they force the temporal and analytical role of each variable to be made explicit. A directed acyclic graph (DAG) can encode a factorization of the joint distribution and, under additional assumptions, hypotheses about interventions, confounding, mediation, and conditional independence [4, 5]. Causal discovery attempts to infer aspects of this structure from observational data, but the result is not automatically a physical causal mechanism. Markov equivalence, latent common causes, measurement error, selection, and model misspecification limit what can be identified from incident records alone [6, 7].

This paper extends the earlier exploratory wildfire analysis in three substantive ways. First, the final reported experiment uses a 22-node BIC-based greedy DAG search without hard tier-orientation constraints. The learned DAG is interpreted as an exploratory dependency structure and database-audit object; variable roles are used after discovery to prevent statistical directions from being promoted automatically to physical causal claims. Second, the event matrix is audited before graph interpretation, including the severity distribution and the exact accounting identities among burned-area components, subtotals, and total area. Third, graph criticism is expanded to $B = 1,000$ topology-preserving variable relabelings. These null graphs retain the same structure and edge count while changing which variable occupies each node, allowing us to test whether the learned semantic organization is unusually compatible with the data. This falsification is intentionally graph-level: it does not measure bootstrap stability of individual adjacencies or orientations.

The study addresses the following research question:

What dependency structure among spatial context, terrain, meteorological conditions, burned-area composition, and wildfire severity is supported by the Maule incident records, and how strongly does that structure withstand topology-preserving falsification?

The scope of this study is intentionally limited. Our aim is not to present the learned graph as a definitive causal explanation of wildfire occurrence or propagation. The analytical dataset contains only documented wildfire events and does not include matched observations from periods or locations without fire. Consequently, every learned dependency is conditional on inclusion as a recorded wildfire; if inclusion is jointly influenced by environmental, geographic, accessibility, reporting, or severity-related factors, conditioning on recorded-fire status can induce collider/selection bias. In addition, several potential common causes are absent or only partially represented, including independently measured fuel moisture, ignition cause, road accessibility, suppression capacity, and detailed pre-fire land cover. Causal sufficiency is therefore not assumed. For these reasons, the graph is treated primarily as an exploratory database audit and as a structured set of hypotheses for later analyses, not as an actionable physical causal model.

The remainder of the paper develops this objective. Section 2 establishes the wildfire context and introduces the causal concepts that guide the analysis, while Section 3 describes the dataset and the

checks carried out before structure learning. Section 4 presents the DAG search procedure and the falsification strategy. Section 5 reports the main structural and empirical results. Section 6 considers their implications for wildfire-severity analysis and identifies where causal interpretation must remain cautious. Section 7 discusses the study’s main limitations and outlines the validation steps required before progressing toward causal-effect estimation. Finally, Section 8 summarizes the main findings and their implications.

2. Background and Causal Framework

2.1. Wildfire analytics and the causal interpretation gap

Wildfire research encompasses multiple interconnected aspects of fire behavior and management, from the conditions that lead to ignition and the processes governing fire spread to the extent of the affected area, the resulting damage, and the allocation of prevention and response resources. Machine learning models are widely used because they can capture nonlinear interactions among weather, topography, vegetation, land use, and human activity [8]. In Chile, spatial analyses have shown that wildfire occurrence and affected area are geographically structured and associated with land cover, roads, settlements, and accessibility [9, 10]. Bayesian-network studies in wildfire risk further demonstrate the value of representing dependencies explicitly, especially when observational data are combined with expert knowledge [11, 12].

The causal problem is different from feature ranking. A commune code may be predictive because it summarizes many unmeasured local properties, but changing the code is not a physical intervention. Likewise, the affected area of scrub may be almost perfectly related to total burned area because it contributes arithmetically to the total. Such a variable can be useful descriptively while being unsuitable as an upstream explanatory exposure. This motivates an explicit causal vocabulary.

2.2. Structural causal models and interventions

A structural causal model represents each variable as the output of a mechanism driven by its direct causes and unobserved influences. For variable X_j ,

$$X_j = f_j(\text{Pa}_G(X_j), U_j), \quad (1)$$

where $\text{Pa}_G(X_j)$ is the parent set of X_j in graph G and U_j contains exogenous influences not represented by the measured parents [4, 5]. Equation 1 should be read as a data-generating statement. In a wildfire application, severity could conceptually depend on weather, terrain, pre-fire fuels, accessibility, suppression, and unmeasured factors. The present dataset observes only part of this system, so Equation 1 also makes the latent-confounding limitation explicit.

An intervention changes one mechanism rather than merely conditioning on an observed value. Pearl’s $\text{do}(\cdot)$ operator writes this as

$$\text{do}(X_k = x) : \quad X_k \leftarrow x. \quad (2)$$

The incoming causes of X_k are conceptually removed and X_k is fixed at x , while the other structural equations remain unchanged [4]. This is why $p(Y \mid \text{do}(X = x))$ and $p(Y \mid X = x)$ need not be equal. The latter compares naturally occurring observations and can retain differences produced by common causes of X and Y . The former describes an intervention distribution.

The potential-outcomes framework expresses the same scientific goal through counterfactual outcomes. If $Y(x)$ denotes wildfire severity under exposure level x , a causal contrast between two exposure levels can be written

$$\tau(x_1, x_0) = \mathbb{E}[Y(x_1) - Y(x_0)]. \quad (3)$$

For a binary treatment this corresponds to an average treatment effect; for a continuous environmental exposure it can compare two scientifically meaningful levels [13, 14]. For example, a future study could

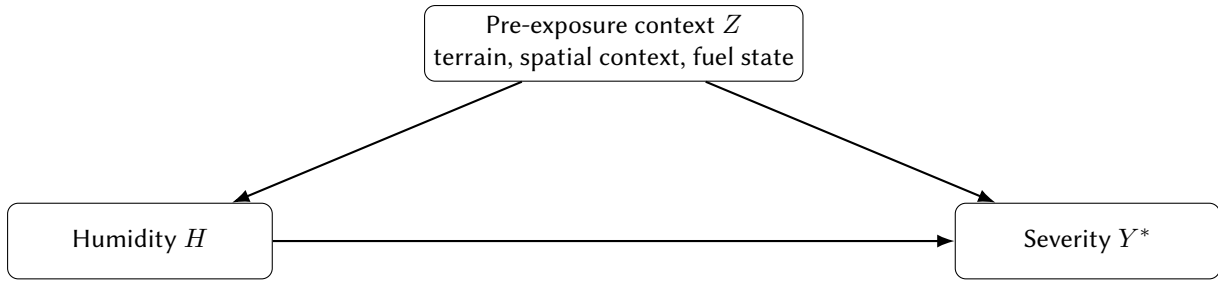


Figure 1: Conceptual back-door structure for a future humidity–severity analysis. It illustrates the identification logic and is not presented as the uniquely identified subgraph of the current wildfire DAG.

compare expected severity under two pre-ignition humidity levels. The present paper does not estimate such an effect; it establishes a graph-based foundation for deciding which variables could be exposures, confounders, or descendants.

2.3. Back-door adjustment and a wildfire-specific estimand

If a set of pre-exposure covariates Z blocks all non-causal back-door paths from exposure X to outcome Y , and Z contains no descendants of X , the intervention distribution can be identified by

$$p(y \mid \text{do}(x)) = \sum_z p(y \mid x, z)p(z), \quad (4)$$

with integration replacing the sum for continuous variables [4, 14]. Equation 4 means that the outcome is first evaluated among observations that are comparable with respect to the confounders and then averaged over the target population. Its validity depends on the adjustment set being sufficient; no statistical algorithm can compensate for an important unmeasured common cause simply by applying the formula.

For wildfire severity, let

$$Y^* = \log(1 + \text{total_area_ha}) \quad (5)$$

be the transformed burned-area endpoint and let H be relative humidity measured in a clearly defined pre-ignition window. A future intervention-oriented contrast could be

$$\tau_H(h_1, h_0) = \mathbb{E}[Y^* \mid \text{do}(H = h_1)] - \mathbb{E}[Y^* \mid \text{do}(H = h_0)]. \quad (6)$$

Equation 6 is used to state more clearly the type of causal question that could be examined in a later stage of the study. The intention is not to suggest that relative humidity can be directly manipulated across a landscape. Instead, the equation describes the contrast we would like to estimate from observational data if the required causal assumptions were considered reasonable. In that setting, variables describing spatial context, terrain, and independently measured pre-fire fuel conditions could be considered as possible adjustment variables. Burned-area components observed after the event, however, should generally be excluded from such an adjustment set because they form part of the fire outcome or occur downstream of it.

Figure 1 illustrates this point using a simple confounding structure. The path $H \leftarrow Z \rightarrow Y^*$ shows how humidity and fire severity may appear associated because both are related to a common pre-event factor represented by Z . If Z adequately captures that source of confounding, conditioning on it removes the non-causal path while preserving the pathway $H \rightarrow Y^*$ that is relevant to the effect of interest.

2.4. DAG factorization and score-based discovery

A DAG $G = (V, E)$ over $X_1, \dots, X_p$ represents the factorization

$$p(x_1, \dots, x_p) = \prod_{j=1}^p p(x_j \mid x_{\text{Pa}_G(j)}). \quad (7)$$

Under the Causal Markov condition, the parents of a variable contain the information needed to separate it from its non-descendants in the graph. The fidelity assumption states that the conditional independence relationships represented in the graph must be consistent with the relationships observed in the data [6, 5]. However, even when this assumption and the previous ones are considered reasonable, observational data are often insufficient to identify a single causal structure. In general, it is possible to obtain several DAGs that represent the same statistical relationships between variables. For this reason, distinguishing between these equivalent structures requires incorporating additional assumptions, prior domain knowledge, or external information.

Score-based causal discovery addresses this problem by comparing candidate graph structures according to how well they explain the data while penalizing unnecessary complexity. Greedy Equivalence Search (GES) constructs the graph structure iteratively. First, it adds relationships between variables when these improve the model score, and then removes relationships whose removal further improves the score [7]. FGES follows the same idea but incorporates optimizations that allow this search to be performed more efficiently when working with a large number of variables [15]. In our analysis, we used a greedy strategy with similar characteristics, evaluating each structure using a Gaussian BIC score. Furthermore, we discarded any changes that would generate a directed cycle, thus ensuring that the resulting structure remained a directed graph.

For graph G , BIC is

$$\text{BIC}(G) = -2\ell(\hat{\theta}_G) + k_G \log n, \quad (8)$$

where $\ell(\hat{\theta}_G)$ is the maximized log-likelihood, k_G is the number of free parameters, and n is the sample size [16]. In the convention used here, lower values are preferred. The first criterion measures how well the model represents the data, while the second penalizes overly complex models. This balance is especially important when working with large observational samples, as many weak relationships between variables may emerge. To avoid adding unnecessary connections to the graph, a new relationship is only retained if it sufficiently improves the model fit to justify the increased complexity.

2.5. Why falsification is required

The observed difference may arise from an unobserved common cause, an artifact introduced by data encoding, or a genuine direct relationship. A BIC score alone cannot distinguish among these possibilities, so we evaluate additional implications of the learned graph. Permutation methods provide a nonparametric empirical reference distribution for such diagnostics without requiring strong parametric assumptions [17, 18, 19].

In our null model, the graph topology and edge count are held fixed while wildfire variables are randomly reassigned to node positions. The comparison therefore controls for structural properties such as graph density and degree sequence. Any advantage of the learned graph must instead arise from the particular alignment between the variables and their positions in the graph.

3. Data and Empirical Setting

3.1. CONAF event records

The dataset contains 6,610 forest fires recorded in the Maule Region between September 29, 2011, and June 21, 2021 [20]. After data cleaning, 22 event-level variables were retained, with no missing values in the final analytical dataset. Date-related variables and geographic coordinates were excluded

Table 1

Analytical roles of the 22 wildfire variables.

Block	Representative variables	Interpretation
Spatial context	cut_reg_prov_com	Administrative proxy for latent geographic, land-use, accessibility, management, and reporting differences.
Terrain	slope, exposition, topography	Pre-event physiographic context.
Meteorology	temperature, humidity_percent, wind_speed, wind_direction	Event-context hazard descriptors.
Detailed impact	pine classes, eucalyptus, other plantations, wooded, scrub, pastureland, agricultural, waste	Area affected by the observed fire; downstream outcome composition.
Aggregates	three subtotals, total_area_ha	Derived accounting totals and recorded severity.

before causal structure learning because they were not included as nodes in the model. However, the administrative variable `cut_reg_prov_com`, which identifies the territorial location of each event, was retained. This variable was used to represent general differences between geographic areas and not as a direct replacement for the original coordinates. Table 1 groups the 22 variables according to their analytical roles. This distinction is important because the variables correspond to different stages of a wildfire event. Terrain characteristics exist prior to ignition, whereas the meteorological variables describe environmental conditions associated with the event. However, the available records do not provide sufficient temporal detail to establish their precise timing relative to ignition. The administrative code is therefore treated as a spatial-context proxy rather than as a physical intervention variable. In contrast, vegetation-specific burned areas and their corresponding aggregates describe impacts measured after the fire.

Because the dataset contains only recorded fires, the graph describes dependency structure conditional on being an observed wildfire event. It cannot identify the probability that a location or day experiences a fire in the first place. More importantly for interpretation, recorded-fire status is itself a selection variable: when inclusion or documentation depends on multiple causes (for example severity, accessibility, geography, or reporting practices), conditioning on that status can open non-causal paths between those causes. The present DAG must therefore be read as a dependency model within the selected incident population. A causal occurrence study would require non-fire comparison units or an explicit space–time risk set.

3.2. Descriptive profile

Wildfire severity is highly right-skewed. The mean total burned area is 50.19 ha, but the median is only 0.50 ha. The 95th percentile is 15.50 ha and the 99th percentile is 190.98 ha, whereas the reported sample maximum is 159,812.58 ha. This maximum is an extreme tail observation rather than a representative event, which explains its large separation from the 99th percentile. The value is retained in the descriptive profile, while the $\log(1 + x)$ transformation in Equation 5 reduces its leverage in the Gaussian local regressions. Temperature has mean 25.47 and median 26.20 in the recorded units; relative humidity has mean 41.91% and median 40%; wind speed has mean 5.59 and median 4.10. The cleaned data contain 30 administrative codes and three topography classes. The impact fields are also sparse in different ways. Scrubland and grassland showed significant area affected by fires: 59.20% and 61.15% of events, respectively. In contrast, mature pine trees showed area affected in only 2.47% of cases, and other plant species in 1.29%. These differences should be considered when interpreting the graph. A node with many connections does not necessarily represent a physically larger variable. In some cases, its high connectivity may simply be due to how the fire damage was classified and recorded in the dataset.

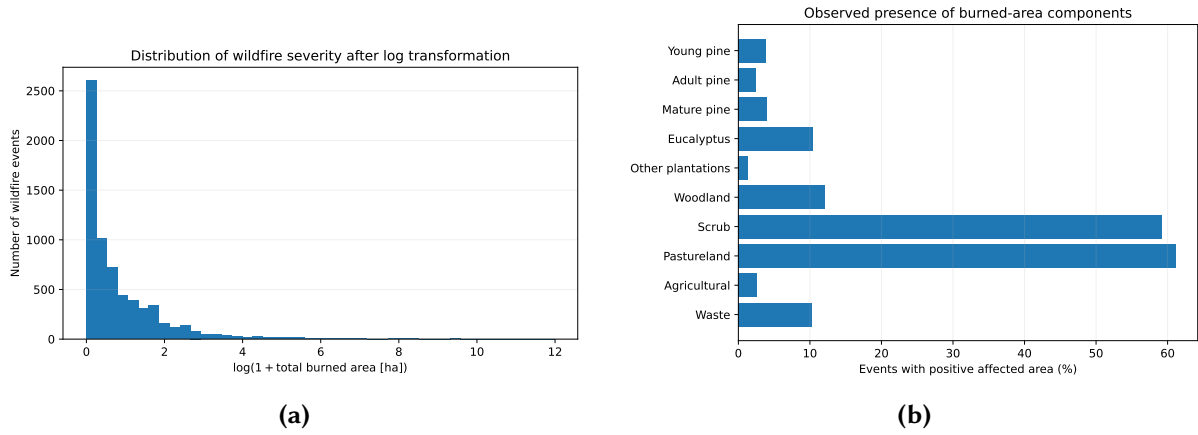


Figure 2: Descriptive profile of wildfire severity and burned-area composition. (a) Distribution of severity after the $\log(1 + \text{area})$ transformation used for score-based structure learning. (b) Percentage of wildfire events with positive affected area in each detailed burned-area component.

Figure 2(a) shows the distribution after applying the severity transformation in Equation 5. The transformation does not make extreme events disappear; instead, it compresses the long right tail so that small, medium, and very large incidents can be represented on the same scale. This is relevant for the Gaussian BIC score because the untransformed maximum is several orders of magnitude larger than the median and could otherwise dominate local regressions. Figure 2(b) complements this view by showing the percentage of events with positive affected area in each detailed burned-area component, making the strong sparsity differences among impact variables explicit.

3.3. Deterministic burned-area identities

The most important data-audit result is that the aggregate impact fields are exact arithmetic functions of the detailed components. For every one of the 6,610 events,

$$S_{\text{plant}} = P_{\text{child}} + P_{\text{adult}} + P_{\text{senior}} + E + O, \quad (9)$$

$$S_{\text{natural}} = W + S + G, \quad (10)$$

$$S_{\text{other}} = A + R, \quad (11)$$

$$T = S_{\text{plant}} + S_{\text{natural}} + S_{\text{other}}. \quad (12)$$

Here T is total burned area; the remaining symbols denote the corresponding plantation, natural-vegetation, agricultural, and waste components. Maximum absolute residuals are below 2×10^{-11} ha and no row exceeds a tolerance of 10^{-8} ha.

Equation 12 explains why the impact subgraph is expected to be dense. These are not merely correlated measurements; the aggregates are defined from their components. Consequently, edges inside this block primarily reflect outcome bookkeeping. A driver-focused causal model should retain total area as the principal severity endpoint and replace affected-area composition with independently measured pre-fire vegetation or fuel variables.

4. Methodology

4.1. Preprocessing and graph search

Burned-area variables are transformed with $\log(1 + x)$ and continuous variables are standardized before scoring. The reported 22-node graph retains the original semantic meaning of the variables. Administrative and topographic codes and exposition are represented as single coded nodes, while wind direction is represented by its recorded angular code. A Gaussian local BIC score is not measurement-scale-correct for nominal categories and does not model circularity in wind direction; using it for these

Table 2

Configuration of the extended discovery experiment.

Element	Value	Purpose
Events	6,610	One row per recorded wildfire.
Nodes	22	Context, terrain, meteorology, and impact.
Impact transform	$\log(1 + x)$	Reduce extreme right-tail leverage.
Score	Gaussian BIC	Penalized local likelihood.
BIC multiplier	1.0	Standard complexity penalty.
Maximum parents	7	Limit unstable high-degree parent sets.
Hard orientation tiers	None	Separate statistical discovery from semantic interpretation.
Final graph	22 nodes, 69 edges	Reported dependency DAG.

nodes is therefore a deliberate approximation of the present exploratory audit, not a claim that their conditional distributions are Gaussian continuous variables. Consequently, score differences and edge orientations involving these coded variables are interpreted cautiously and are not used here to estimate intervention effects. Mixed-type scoring and circular representations such as sine/cosine encoding are reserved for a later sensitivity analysis and are not part of the reported experiment.

For node X_j with parent set $\text{Pa}(X_j)$, the Gaussian local regression produces residual sum of squares RSS_j . Up to model-independent constants, the local BIC contribution is

$$\text{BIC}_j = n \log \left(\frac{\text{RSS}_j}{n} \right) + k_j \log n. \quad (13)$$

Candidate edge additions, removals, and reversals are considered while maintaining acyclicity. A move from G to G' is favorable when

$$\Delta \text{BIC} = \text{BIC}(G') - \text{BIC}(G) < 0. \quad (14)$$

Equation 14 makes the search rule transparent: an edge must improve fit enough to compensate for the additional model complexity. The final configuration uses the standard BIC penalty multiplier 1.0 and a maximum of seven parents per node. No hard terrain–weather–impact orientation tiers are imposed in the reported run; causal interpretation is applied after learning.

4.2. Topology-preserving falsification

The learned graph is compared with $B = 1,000$ random variable-to-node relabelings. Each null graph retains the original topology, edge count, degree sequence, and parent-set sizes; only the assignment of wildfire variables to node positions is changed. This design tests whether the learned semantic placement of variables is more consistent with the graph’s implied dependencies than a random assignment under the same structural complexity. It is a global falsification test of variable-to-topology alignment; it is not a resampling analysis of edge-selection or orientation stability.

For diagnostic statistic S , the add-one empirical p-value is

$$p_{\text{emp}} = \frac{1 + b}{1 + B}, \quad (15)$$

where b is the number of null graphs at least as favorable as the observed graph [17]. With $B = 1,000$, the smallest possible value is $1/1001 = 0.000999$. This finite resolution is reported explicitly rather than writing $p = 0$ when no null replicate exceeds the observed statistic.

Two graph implications are examined. First, the Local Markov Condition (LMC) requires

$$X_j \perp\!\!\!\perp \text{ND}_G(j) \setminus \text{Pa}_G(j) \mid \text{Pa}_G(j), \quad (16)$$

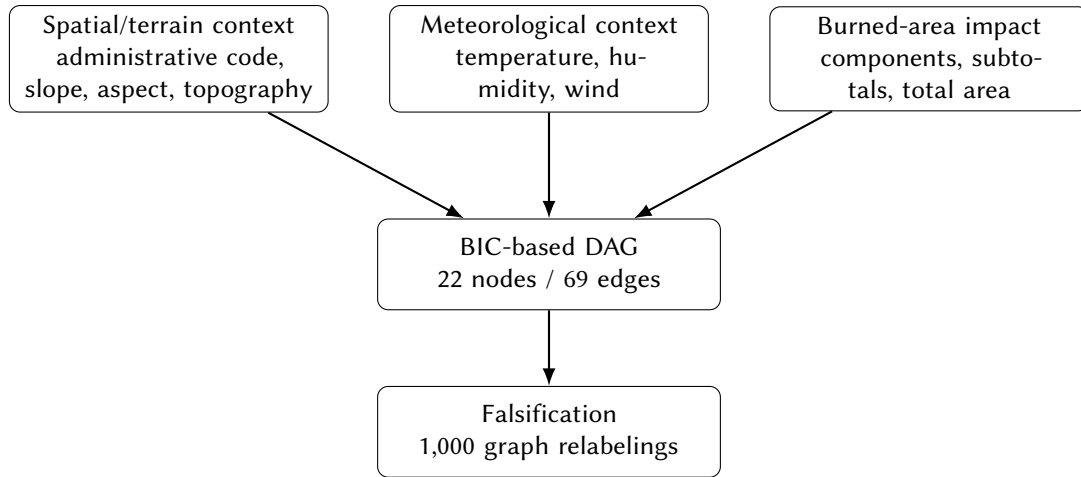


Figure 3: Analytical role of the variable blocks. Roles guide interpretation but were not imposed as hard orientation tiers in the reported search.

where $\text{ND}_G(j)$ denotes non-descendants. In practical terms, once the direct parents of a node are known, its remaining non-descendants should provide no additional information under the DAG model. We test these implied conditional independencies and summarize the proportion that are rejected; lower violation fractions indicate greater compatibility with the learned graph.

Second, for a directed edge $X_i \rightarrow X_j$, the parent-association diagnostic considers

$$X_i \perp\!\!\!\perp X_j \mid \text{Pa}_G(j) \setminus \{X_i\}. \quad (17)$$

If Equation 17 is supported, the selected parent should retain a meaningful conditional association with the child after accounting for the child's other parents. If little conditional association remains, the relation is counted as a violation of this diagnostic. Again, lower violation fractions indicate better compatibility. Neither diagnostic establishes causal orientation; together, they assess whether both the included and excluded relationships are empirically consistent with the learned graph.

5. Results

5.1. Learned graph structure

The final BIC-selected DAG contains 22 nodes and 69 directed edges, but its structure is far from homogeneous. Most high-degree nodes are associated with vegetation-impact or aggregate fields, whereas terrain, meteorological, and administrative variables form a smaller contextual subsystem. This separation is the central structural finding: much of the strongest dependency pattern in the incident database appears to arise from the composition and accounting of fire impacts rather than from a simple physical pathway linking weather conditions to wildfire severity.

Figure 3 summarizes this analytical interpretation. Although all variable blocks enter the same structure-learning procedure, they do not carry the same causal meaning after discovery. Context variables may plausibly precede the event, meteorological variables characterize hazard conditions, and burned-area fields primarily describe downstream impacts. The search is allowed to discover all statistical dependencies; the interpretation layer prevents post-event accounting relations from being promoted automatically as intervention pathways.

The most connected variables are `subtotal_natural_vegetation` (degree 11); `others_plants`, `wooded`, and `subtotal_plantations` (10 each); `pine_adult`, `pastureland`, and `pine_child` (9 each); and `eucalyptus` (8). Every node in this list belongs to the impact subsystem. Their centrality therefore indicates a tightly coupled accounting representation rather than evidence that these variables are upstream causes of wildfire occurrence.

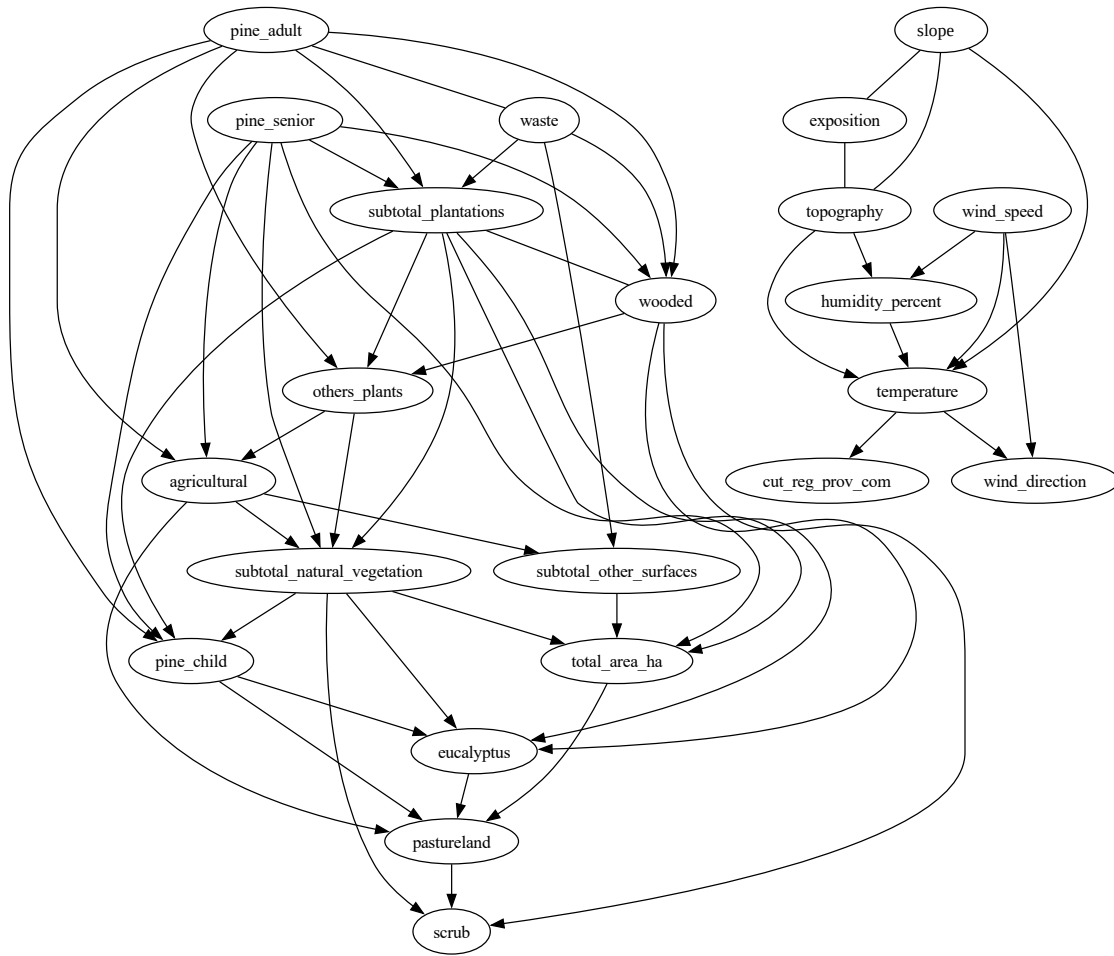


Figure 4: Full learned 22-node wildfire dependency graph obtained from the Maule incident data. The left-hand side of the graph is dominated by burned-area components and accounting aggregates, while terrain and meteorological variables are concentrated on the right. The direction of the edges should be interpreted as a structure-learning hypothesis rather than as established causation. Their interpretation also depends on the timing of the variables, the deterministic relationships among accounting fields, and the assumptions described throughout the analysis.

5.2. Contextual subsystem

Terrain, meteorological, and administrative variables form a smaller contextual subsystem that may help prioritize variables for later causal analysis. Within this subsystem, the learned graph connects slope and other physiographic descriptors with temperature, relative humidity, wind speed, and wind direction. In the present study these links are interpreted only as conditional-dependence hypotheses. Their orientation is not established causally because the event-level weather timing is imprecise, the score approximates nominal and circular variables, the sample is selected on recorded fires, and relevant common causes may be unobserved. The administrative code is not a physical driver; it is a proxy for latent spatial differences in forestry composition, accessibility, land management, human activity, suppression capacity, and reporting practices.

Meteorological timing also matters. The event-level matrix associates weather with the incident but does not establish a sufficiently precise temporal sequence between ignition, early spread, and measurement. For a future causal estimate such as Equation 6, meteorological exposure should therefore be rebuilt from an independent source using a pre-specified pre-ignition window.

Table 3
Highest-degree nodes in the learned DAG.

Variable	Degree	Interpretive role
subtotal_natural_vegetation	11	Aggregate impact node.
others_plants	10	Detailed vegetation-impact component.
wooded	10	Natural-vegetation impact component.
subtotal_plantations	10	Plantation aggregate.
pine_adult	9	Plantation impact component.
pastureland	9	Natural-vegetation impact component.
pine_child	9	Plantation impact component.
eucalyptus	8	Plantation impact component.

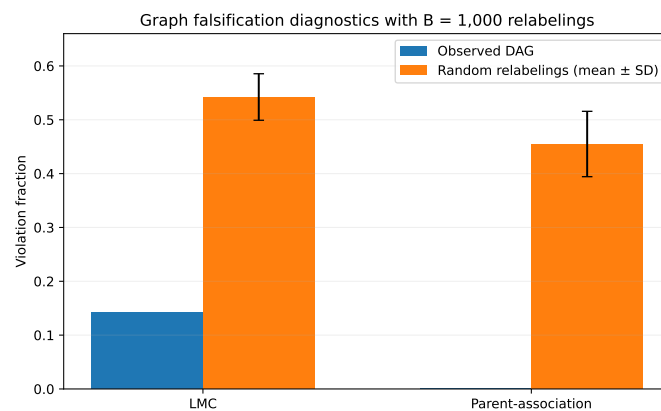


Figure 5: Observed graph diagnostics versus $B = 1,000$ topology-preserving variable relabelings. Lower violation fractions indicate greater compatibility with the graph implications. Error bars show one standard deviation of the relabeling distribution.

Table 4
Falsification summary.

Diagnostic	Observed	Relabeling mean $\pm$ SD	p_{emp}
LMC violation fraction	0.1414 (27/191)	0.5423 ± 0.0432	0.000999
Parent-association violation fraction	0.0000 (0/69)	0.4550 ± 0.0607	0.000999

5.3. Falsification results

The observed graph is sharply separated from the topology-preserving null. Its LMC violation fraction is 0.1414, corresponding to 27 violations among 191 tested implications. Across 1,000 random variable-to-node relabelings, the mean violation fraction is 0.5423 ± 0.0432 . None of the null assignments is at least as favorable as the observed graph, so Equation 15 gives $p_{\text{LMC}} = 0.000999$.

The parent-association result is even more separated. The observed graph has 0 violations among 69 parent-child relations, while the relabeling distribution has mean violation fraction 0.4550 ± 0.0607 . Again, no null graph is at least as favorable, yielding $p_{\text{PA}} = 0.000999$.

These results provide stronger evidence about graph-level compatibility than the BIC score alone. Because the null graphs preserve topology, density, and edge count, the comparison specifically tests whether the observed alignment between wildfire variables and graph positions is more compatible with the data than random semantic assignments. The result supports the learned graph as a nontrivial dependency model of the recorded wildfire events. It does not test whether a particular adjacency would recur after resampling, does not establish orientation stability, and does not show that any individual arrow is a true causal relationship.

6. Discussion

6.1. What the graph supports

The learned DAG is used here primarily as an exploratory database audit. It summarizes conditional dependencies in the CONAF incident table and exposes how strongly the recorded impact-accounting variables organize the data. The topology-preserving relabeling test shows that the observed variable-to-node alignment is substantially more compatible with the data than random assignments, but this is a global property of the selected graph. It is not evidence that the graph is an actionable physical causal model.

Four limitations prevent such an interpretation: Gaussian scoring is misspecified for nominal and circular variables; relevant common causes are unobserved; conditioning on documented wildfires can induce collider/selection bias; and the relabeling experiment does not quantify edge-level stability. Section 7 details these threats. Consequently, contextual variables are treated as hypothesis-generating, whereas the impact subsystem is principally an audit of database construction and outcome bookkeeping.

6.2. Why compositional outcomes change causal interpretation

The accounting relationships in Equation 12 are particularly important when interpreting the learned structure. Total burned area is obtained from the three subtotals, which are themselves constructed from more detailed burned-area components. Because of these exact relationships, a structure-learning algorithm has a strong statistical reason to connect these variables. Those connections are meaningful in terms of the dataset, but they do not necessarily describe separate physical processes operating during the wildfire.

This distinction suggests that future work would benefit from using two complementary graph specifications. An *audit graph* could retain the complete set of impact variables in order to examine redundancy, internal consistency, and possible sources of information leakage. A second, more focused *driver graph* could treat `total_area_ha` as the primary severity outcome while replacing post-fire burned-area components with variables that characterize vegetation or fuel conditions before ignition. This formulation would make it easier to examine the contributions of weather, terrain, and landscape conditions without allowing deterministic accounting relationships to dominate the structure-learning process.

6.3. Spatial confounding and administrative context

Administrative variables need to be interpreted with care. A commune code is not itself a physical mechanism; its role is to capture geographic differences that are not otherwise represented in the dataset. Communes in the Maule Region vary in climate, elevation, forestry activity, road accessibility, land management, ignition pressure, suppression capacity, and in the ways incidents are detected and recorded. The administrative code may therefore stand in for several unmeasured processes at the same time.

There is a trade-off in retaining this variable. It helps represent broad geographic heterogeneity that would otherwise be absent from the model, but it also combines potentially distinct mechanisms within a single category. Conditioning on the commune code may therefore mask variation that would become visible at a finer spatial scale. A useful next step would be to compare results from the full dataset with analyses conducted within individual communes or across broader geographic strata. If the association between weather conditions and wildfire severity persists across these subsets, this would provide stronger evidence that it is not primarily explained by between-area differences. Methods that explicitly allow for latent common causes, such as FCI or RFCI, would also be useful because they do not require all relevant confounders to be directly observed [21].

6.4. From discovery to effect estimation

The present graph is best viewed as a screening and audit tool that can inform, but cannot by itself identify, a later effect-estimation analysis. Consider, for example, the humidity contrast defined in Equation 6. Before attempting to estimate that contrast, humidity would need to be measured from an independent source and assigned to a clearly defined period before ignition. A candidate adjustment set could then be constructed using both the graph and knowledge about the temporal ordering of the variables. Terrain and spatial context are plausible candidates because they are available before the exposure and may influence both local atmospheric conditions and fire severity.

The burned-area components should be treated differently. Since they are measured after the event and form part of the final severity outcome, including them in an adjustment set could block relevant causal pathways or introduce post-treatment bias. The same reasoning applies when temperature or wind is chosen as the exposure of interest. The causal question should be defined in terms of an environmental condition measured at a meaningful stage before or during the early development of the fire, rather than simply using whichever variable happens to be available in the incident record. In this sense, the graph is valuable because it forces a distinction between the name of a variable and the causal role that variable can reasonably play.

6.5. Implications for predictive leakage and data engineering

These distinctions also matter when the same database is used for prediction. A model intended to predict final burned area should not use post-event subtotals or burned-area components that are already part of the calculation of `total_area_ha`. Doing so would give the model access to information that would not be available at the moment when the prediction is supposed to be made, producing target leakage and an unrealistically optimistic estimate of predictive performance.

One practical use of the exploratory graph is to distinguish variables according to when and how they arise in the wildfire record. Some describe the context in which the event occurred, others may represent potential exposures or downstream consequences, and some are deterministic transformations of the target itself. These distinctions matter in operational datasets because variables recorded before, during, and after an incident are often stored together in the same table. If those differences are ignored, a predictive model can achieve apparently strong performance by using information that would not be available at the time a real forecast is made.

7. Threats to Validity and Validation Roadmap

The first limitation concerns model and construct validity. The administrative code is only a proxy for broader spatial processes, the meteorological records do not establish a precise temporal relationship with ignition, and the vegetation-impact variables describe what burned rather than the fuel available before the event. In addition, nominal categories and the circular wind-direction variable are represented within a Gaussian local scoring framework. This is a misspecified likelihood for those measurement scales and is acceptable here only as an exploratory approximation; it prevents strong interpretation of score differences or directions involving these nodes. A confirmatory analysis should use mixed-type local models and a circular-aware wind representation.

Second, causal sufficiency cannot be assumed. Potentially relevant unobserved common causes include drought conditions, fuel moisture, pre-fire biomass, ignition mechanisms, road access, suppression response, and land-management practices. Accordingly, a directed edge may reflect a latent common cause rather than a direct mechanism. Selection bias is also structurally relevant: the analysis conditions on recorded-fire status, and that selection variable can act as a collider when documentation or inclusion is influenced by more than one cause. This can induce associations that would not be present in an unselected space–time population. The absence of missing values in the cleaned analytical matrix does not remove this source of bias.

Table 5

Recommended validation before intervention-oriented claims.

Experiment	Purpose	Preferred output
Outcome simplification	Remove deterministic accounting dominance	Driver graph with <code>total_area_ha</code> as principal severity endpoint.
Bootstrap resampling	Quantify finite-sample edge stability	Adjacency and orientation frequencies.
BIC/parent sensitivity	Test score dependence	Graph overlap, SHD, and degree stability.
PC-Stable comparison	Test discovery-family dependence	Stable adjacencies shared across methods.
FCI/RFCI sensitivity	Address latent confounding	Robust ancestral/partially oriented relations.
Spatial stratification	Test geographic mixture	Pooled-versus-stratum agreement.
Independent pre-fire data	Improve temporal/construct validity	Fuel, land-cover, weather, and access variables defined before ignition.
Held-out testing	Assess transfer	Graph implications evaluated on unseen periods or areas.

Third, the 1,000-relabeling experiment evaluates graph-level compatibility but does not assess the stability of individual edges across different samples or modeling choices. No bootstrap edge-frequency analysis is reported in this paper. Individual adjacencies and orientations therefore remain unstable hypotheses and should not receive edge-level causal interpretation. Bootstrap selection frequencies, sensitivity to the BIC specification and parent-limit settings, PC-Stable comparisons, and FCI/RFCI analyses under latent confounding are part of the validation roadmap rather than evidence claimed by the present experiment [22, 21].

Table 5 summarizes the next validation stage.

The present graph should not be used for operational interventions, suppression-policy decisions, or claims that one burned-area component physically causes another. Those conclusions either require variables not present in the analytical matrix, rely on edge orientations whose stability has not been established, or confuse outcome accounting with causal mechanisms. Its defensible use at this stage is exploratory: auditing the incident database, identifying redundancy and leakage risks, and prioritizing hypotheses for better designed follow-up studies.

8. Conclusion

This study presents a falsification-aware exploratory structure analysis and database audit of 6,610 official wildfire incidents from the Maule Region of Chile. A 22-node BIC-based greedy DAG with 69 directed edges reveals two qualitatively different forms of organization: a contextual terrain–meteorology–spatial component and a dense vegetation/impact-accounting component. Direct verification of the recorded variables shows that the impact aggregates are exact arithmetic functions of detailed burned-area measurements, explaining why those nodes dominate graph centrality and why their arrows cannot be read as independent pre-event causes.

The relabeling test nevertheless shows nontrivial graph-level organization: the observed LMC violation fraction is 0.1414 versus 0.5423 ± 0.0432 under the null, while the parent-association violation fraction is 0 versus 0.4550 ± 0.0607 ; both add-one empirical p-values are 0.000999. These results support the semantic alignment of variables with the selected topology, not edge-level stability or causal orientation.

Accordingly, the DAG is not presented as an actionable physical causal model. Gaussian-score misspecification for nominal and circular variables, unobserved common causes, conditioning on recorded fires, imprecise meteorological timing, deterministic outcome accounting, and the absence of bootstrap edge-stability analysis constrain its interpretation. Its defensible contribution is an exploratory database

audit and hypothesis-prioritization framework; intervention-oriented claims require independently measured pre-fire variables, mixed-type and circular-aware modeling, explicit treatment of selection and latent confounding, and edge-level stability assessment.

Declaration on Generative AI

No generative artificial intelligence tools were used to generate, modify, optimize, debug, execute, or validate any of the computational code employed to obtain the experimental results reported in this article. All data preprocessing, model implementation, training, calibration, evaluation, metric computation, and figure generation associated with the reported results were carried out through the authors' computational workflow using conventional software libraries. Consequently, none of the numerical results, experimental outputs, or model predictions presented in this work were produced through generative AI-assisted programming.

References

- [1] R. D. Garreaud, C. Alvarez-Garreton, J. Barichivich, J. P. Boisier, D. A. Christie, M. Galleguillos, C. LeQuesne, J. McPhee, M. Zambrano-Bigiarini, The 2010–2015 megadrought in central chile: Impacts on regional hydroclimate and vegetation, *Hydrology and Earth System Sciences* 21 (2017) 6307–6327. doi:10.5194/hess-21-6307-2017.
- [2] D. M. J. S. Bowman, A. Moreira-Muñoz, C. A. Kolden, R. O. Chávez, A. A. Muñoz, F. Salinas, A. González-Reyes, R. Rocco, F. de la Barrera, G. J. Williamson, N. Borchers, L. A. Cifuentes, J. T. Abatzoglou, F. H. Johnston, Human–environmental drivers and impacts of the globally extreme 2017 chilean fires, *Ambio* 48 (2019) 350–362. doi:10.1007/s13280-018-1084-1.
- [3] X. Úbeda, P. Sarricolea, Wildfires in chile: A review, *Global and Planetary Change* 146 (2016) 152–161. doi:10.1016/j.gloplacha.2016.10.004.
- [4] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2 ed., Cambridge University Press, 2009.
- [5] J. Peters, D. Janzing, B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press, 2017.
- [6] P. Spirtes, C. Glymour, R. Scheines, *Causation, Prediction, and Search*, 2 ed., MIT Press, 2000.
- [7] D. M. Chickering, Optimal structure identification with greedy search, *Journal of Machine Learning Research* 3 (2002) 507–554.
- [8] P. Jain, S. C. P. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, M. D. Flannigan, A review of machine learning applications in wildfire science and management, *Environmental Reviews* 28 (2020) 478–505. doi:10.1139/er-2020-0019.
- [9] I. Díaz-Hormazábal, M. E. González, Análisis espacio-temporal de incendios forestales en la región del maule, chile, *Bosque* 37 (2016) 147–158. doi:10.4067/S0717-92002016000100014.
- [10] G. Azócar de la Cruz, G. Alfaro, C. Alonso, et al., Modeling the ignition risk: Analysis before and after megafire on maule region, chile, *Applied Sciences* 12 (2022) 9353. doi:10.3390/app12189353.
- [11] J. F. Carriger, M. Thompson, M. G. Barron, Causal bayesian networks in assessments of wildfire risks, *Integrated Environmental Assessment and Management* 17 (2021) 1168–1178. doi:10.1002/ieam.4443.
- [12] V. Sevinç, O. Küçük, M. Göltas, A bayesian network model for prediction and analysis of possible forest fire causes, *Forest Ecology and Management* 457 (2020) 117723. doi:10.1016/j.foreco.2019.117723.
- [13] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* 66 (1974) 688–701. doi:10.1037/h0037350.
- [14] M. A. Hernán, J. M. Robins, *Causal Inference: What If*, Chapman & Hall/CRC, 2020. Open-access textbook.
- [15] J. Ramsey, M. Glymour, R. Sanchez-Romero, C. Glymour, A million variables and more: The fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with

- an application to functional magnetic resonance images, *International Journal of Data Science and Analytics* 3 (2017) 121–129. doi:10.1007/s41060-016-0032-z.
- [16] G. Schwarz, Estimating the dimension of a model, *The Annals of Statistics* 6 (1978) 461–464. doi:10.1214/aos/1176344136.
- [17] P. I. Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3 ed., Springer, 2005.
- [18] D. Freedman, D. Lane, A nonstochastic interpretation of reported significance levels, *Journal of Business & Economic Statistics* 1 (1983) 292–298. doi:10.1080/07350015.1983.10509354.
- [19] M. Scutari, A. Brogini, Bayesian network structure learning with permutation tests, *Communications in Statistics – Theory and Methods* 41 (2012) 3233–3243. doi:10.1080/03610926.2011.593284.
- [20] Corporación Nacional Forestal (CONAF), Sistema de determinación de causas y origen de incendios forestales, Official wildfire cause and origin records, Chile, n.d. Source system for the incident data analyzed in this study.
- [21] J. Zhang, Causal reasoning with ancestral graphs, *Journal of Machine Learning Research* 9 (2008) 1437–1474.
- [22] D. Colombo, M. H. Maathuis, Order-independent constraint-based causal structure learning, *Journal of Machine Learning Research* 15 (2014) 3741–3782.