Predicting Significance Activity Given Tsunami, Magnitude, and Location

1 Project Description

1.1 Problem Importance and Description

Our project seeks to understand the factors contributing to the significance of earthquakes. In particular, we want to focus on these variables to learn whether earthquakes will be significant enough such that they will cause more damage and consequences. Understanding these relationships is important for quick categorization of earthquake severity and a more effective emergency response system.

1.2 Importance to Probabilistic Reasoning or Learning

Knowing the magnitude, the location, and the tsunami occurrence of an earthquake allows us to quickly predict the potential damage the earthquake brings to minimize potential risks, whether in response time or risks in personally evaluating the site itself. A learned belief network provides $P(Significance|Magnitude, Tsunami, State)$ and by modeling the conditional relationships, we can quantify how big, how strong, or where an earthquake must be to meaningfully increase significance score.

2 Data Sourcing and Pre-processing

2.1 Data Source

Our dataset was taken from the Kaggle dataset by Alessandro Lobello [1], *The Ultimate Earthquake Dataset from 1990–2023*.

The raw dataset includes time in milliseconds, place (e.g., “7 km W of Cobb, California”), status (reviewed, automatic, manual), tsunami occurrence (0 or 1), significance (an integer from 0 to 2910), data_type (earthquake, chemical explosion, quarry blast, etc.), magnitude, state, longitude, latitude, depth, and date.

For our project, we will be focusing on the following core variables:

- **Magnitude**: A numeric value from –10 to 10 from the Richter scale measuring the strength of an earthquake.
- **Significance**: The impact level of the earthquake on a scale from 0–2910.
- **State**: The location where the earthquake occurred.
- **Tsunami Occurrence**: A deterministic binary variable indicating whether the USGS flagged a tsunami occurrence.

2.2 Pre-processing the Data

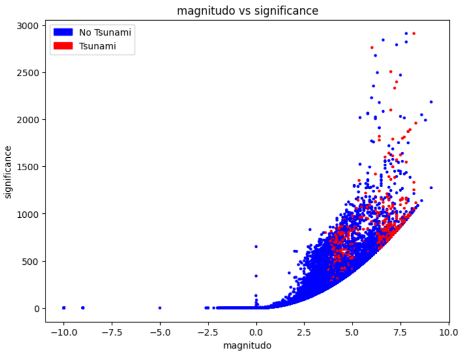

Figure 1: Magnitude and Significance Correlation

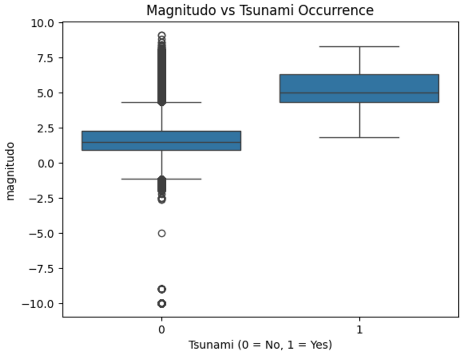

Figure 2: Magnitude and Tsunami Occurrences

Figure 1 shows how earthquake significance increases sharply with magnitude, with both tsunami and non-tsunami events following a similar upward trend. It highlights the strong nonlinear relationship between magnitude and impact severity, which is why we determined that magnitude and significance were not independent of each other.

Additionally, the boxplot shown in Figure 2 compares magnitude distributions for tsunami vs. non-tsunami events, showing that tsunami-associated earthquakes tend to have noticeably higher magnitudes. It reinforces magnitude as a major factor in tsunami generation.

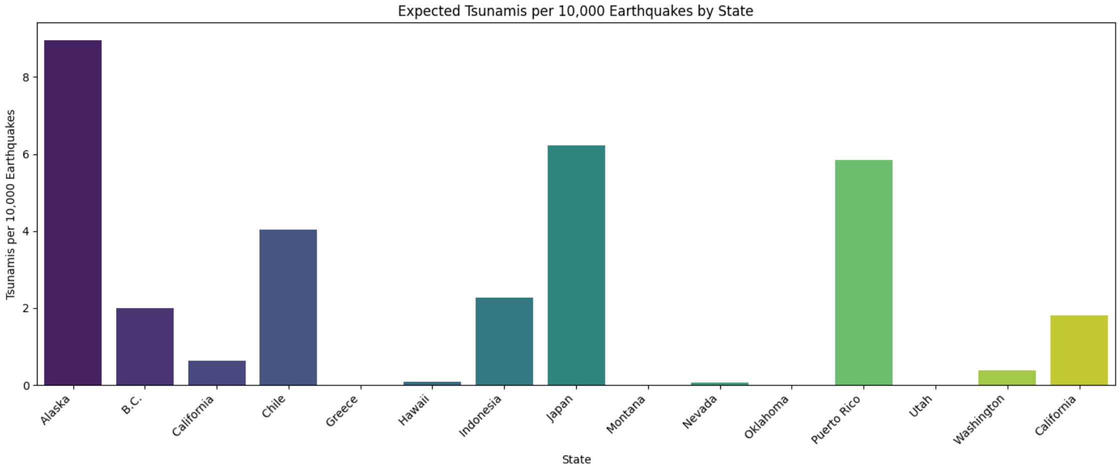

Figure 3: Expected Tsunamis per 10,000 of the top 15 States

This bar chart shows which regions experience the highest tsunami rates relative to their total number of earthquakes. Coastal and subduction-zone areas like Alaska, Japan, and Puerto Rico stand out with significantly higher tsunami likelihood.

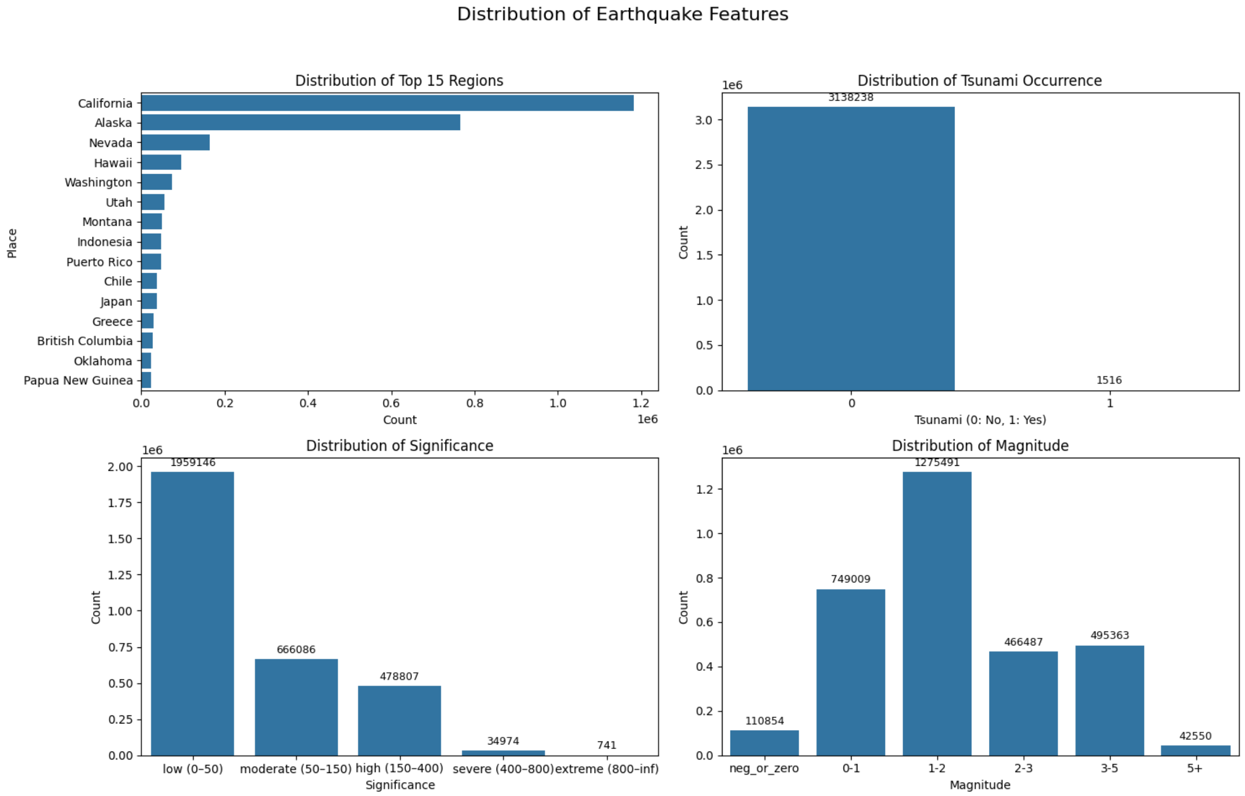

Figure 4: Distribution of all features being used

These distributions summarize the key variables in the dataset, showing the most active regions, the imbalance in tsunami events, and the overall spread of magnitudes and significance scores. They help motivate feature binning choices and illustrate patterns the BN later captures.

We applied several pre-processing steps to clean and standardize the dataset before learning the Bayesian network:

- Using the **data_type** feature, we filtered rows where `data_type != ‘‘earthquake’’`, removing all events not caused by earthquakes such as ice quakes, mining explosions, sonic booms, and quarry blasts.
- Using the **status** feature, we filtered rows where `status != ‘‘reviewed’’`, removing uncertain or low-quality automatic detections.
- We dropped all rows with missing values in the core attributes needed for modeling so that all remaining data points form complete instances.
- We cleaned and consolidated the **state** variable. The dataset contains hundreds of noisy state/location strings, many representing the same place due to:
  - spacing inconsistencies (e.g., “ California” vs. “California”),
  - abbreviations (e.g., “USA” vs. “United States”, “OR” vs. “Oregon”),
  - directional and regional variants referring to the same location (e.g., “Southern Alaska”, “Gulf of Alaska”, “Alaska region”),
  - tectonic or oceanic phrasing for the same geographic area (e.g., “New Zealand”, “New Zealand region”, “North Island of New Zealand”).

  We applied a multi-stage canonicalization pipeline involving directional stripping, region reduction, border/island normalization, synonym and abbreviation resolution, and curated manual overrides to transform all of these into canonical place names.
- After cleaning the state column, we removed locations with fewer than 100 earthquake records to prevent extremely sparse CPT entries and unstable probability estimates.

- We retained this canonicalized **Place** variable for interpretability, but also constructed a second, complementary geographic representation using the earthquake **latitude** and **longitude** coordinates. First, we sampled 50,000 events and ran **HDBSCAN** with a haversine metric to identify dense seismic subregions. We then trained a k-nearest neighbors classifier on these sample clusters to assign a fine-grained **Region_id** to all earthquakes based solely on their coordinates.
- Since HDBSCAN produces over a hundred micro-clusters, we consolidated them into 15 coherent large-scale **MacroRegion** categories. This was done by computing the centroids of each Region_id and running **K-Means** clustering on these centroids. The resulting MacroRegion variable provides a clean, data-driven representation of global seismic zones (e.g., Alaska subduction region, Japan trench region, California transform region, Mid-Atlantic ridge region).
- We encoded all discrete variables, including **Place**, **Region_id**, and **MacroRegion**, as numeric values for use in Bayesian network factor tables and inference algorithms.
- We created a discretized **MagnitudeClass** variable since Bayesian networks require finite discrete state spaces. We used bins: {<0, 0–1, 1–2, 2–3, 3–5, 5+}.
- We discretized **Significance** using quantile-based binning (`qcut`) to ensure balanced class counts. The resulting **SignificanceClass** categories correspond to data-driven ranges approximating: {low, medium, high, extreme}, with the actual bin boundaries extracted from the quantiles.

3 Modeling and Inference

3.1 Probabilistic Model

We represented data dependencies as a Belief Network (BN), which is made up of a directed acyclic graph (DAG) and conditional probability tables (CPTs). For this project, we fixed the DAG using prior domain knowledge and research into data dependencies. Our focus is on learning the CPTs in order to make predictions.

In our model, we have four parameter nodes: Magnitude, which represents the energy released by the earthquake; State, which represents where in the world the earthquake occurred; Tsunami, which is a binary value indicating whether or not the earthquake caused a tsunami; and Significance, which represents the overall impact and magnitude of the earthquake.

3.2 Assumptions

We chose to evaluate two different DAG setups for this problem. We evaluate two DAG structures: one that includes the tsunami variable and one that doesn't. The tsunami-informed DAG captures causal pathways where ocean proximity and high magnitude interact to produce tsunamis, which then influence significance. The simplified DAG removes tsunamis to address sparsity and assess whether magnitude and state alone carry enough predictive power. One is estimating tsunami seen in figure 5, and the other is without tsunami data seen in figure 6.

Given the factors Magnitude, State, Tsunami, and Significance, we chose to represent the probabilities as $P(Magnitude)$ and $P(State)$. For Tsunami, we use $P(Tsunami|Magnitude, State)$. This assumption makes sense because a tsunami only occurs if the earthquake releases enough energy and if the state is located near the ocean. Finally, to represent Significance, we use $P(Significance|Magnitude, State, Tsunami)$. Significance reflects the overall impact of the earthquake, and both a tsunami and the population density of the area (for example, a major city versus a rural region) influence that impact. Magnitude also indicates the potential level of destruction. For the DAG in Figure 6, we generate a similar probability with one small adjustment. We still treat Magnitude and State as root nodes, so we use $P(Magnitude)$ and $P(State)$. However, instead of modeling tsunamis, we remove that variable entirely. The tsunami count data was highly skewed, and there

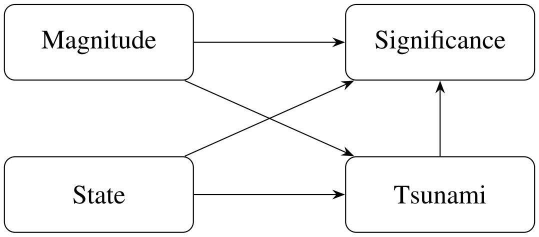

Figure 5: Full DAG (with tsunami) for BN Generation

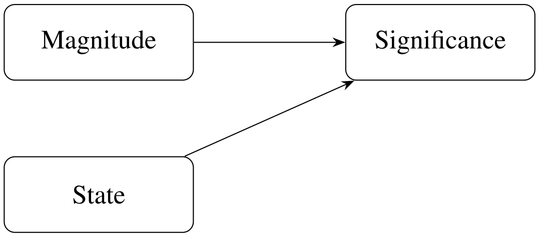

Figure 6: Small DAG (without tsunami) for BN Generation

were very few instances of tsunamis, which could cause issues for predicting significance because MLE often performs poorly with sparse data. Using $P(Significance|Magnitude, State)$ is still a reasonable choice because the location of the earthquake, whether it hits a large city or a remote field, heavily influences significance, and magnitude provides most of the remaining information.

In addition to comparing the two DAGs, we also experimented with different ways of encoding the state variable. We compared using U.S. state or country labels versus using k-means clusters based on latitude and longitude. This produces a model of the form $P(Significance|Magnitude, longitude + latitude bins, Tsunami)$. This approach offers more fine-grained spatial information, but it is harder to incorporate because some bins may end up with very few data points.

3.3 Inference

Our approach to learning the CPTs is to use maximum likelihood estimation (MLE), since our dataset contains complete instances (no missing values for any points), and we have fixed, predefined DAGs as shown in Figures 5 and 6. MLE assumes our data is independent and identically distributed. Using this approach, we can use our dataset to count the occurrences of different events to estimate the conditional probabilities. The counts used and probabilities being estimated for each DAG can be shown in Tables 1a and 1b, respectively.

Table 1: Conditional Probability Tables for Full DAG and Smaller DAG

Probability	**Formula Using MLE**
$P(\text{Magnitude})$	$\frac{1}{T}\text{count}(\text{Magnitude})$
$P(\text{State})$	$\frac{1}{T}\text{count}(\text{State})$
$P(\text{Tsunami}\|\text{Magnitude, State})$	$\frac{\text{count}(\text{Tsunami, Magnitude, State})}{\text{count}(\text{Magnitude, State})}$
$P(\text{Significance}\|\text{Magnitude, State, Tsunami})$	$\frac{\text{count}(\text{Significance, Tsunami, Magnitude, State})}{\text{count}(\text{Tsunami, Magnitude, State})}$

(a) Conditional probability table formulas for DAG in Figure 5

Probability	**Formula Using MLE**
$P(\text{Magnitude})$	$\frac{1}{T}\text{count}(\text{Magnitude})$
$P(\text{State})$	$\frac{1}{T}\text{count}(\text{State})$
$P(\text{Significance}\|\text{Magnitude, State})$	$\frac{\text{count}(\text{Significance, Magnitude, State})}{\text{count}(\text{Magnitude, State})}$

(b) Conditional probability table formulas for DAG in Figure 6

4 Results + Discussion

4.1 Qualitative Analysis

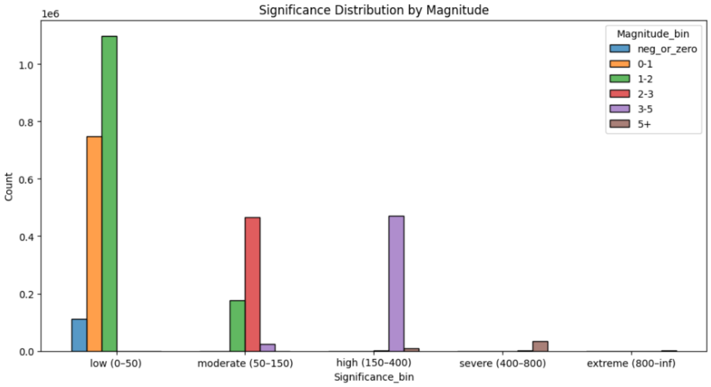

Figure 7: Significance Distribution of all Predicted Earthquake Magnitudes

From our bins specified in `Magnitude_bin`, we show the distributions in significance levels across our bins. Figure 7 shows our final magnitude bins, which we see show significance levels rising with increasing magnitude. This monotonic probability shift confirms that the model is internally consistent and that the learned CPTs are stable and logically structured. Even without a labeled test set for formal numerical metrics, this monotonic confidence behavior serves as a strong qualitative evaluation signal that the learned distributions behave realistically.

A previous iteration of our `Magnitude_bin` bins were showing severe and extreme earthquakes that had magnitudes of 2.0-3.0 on the Richter Scale.

For magnitudes between 0–1, the probability mass is concentrated almost entirely in the low significance category, indicating that weak earthquakes rarely cause meaningful societal impact. As magnitude increases into the 1–2 and 2–3 bins, probability mass begins shifting into the medium and high significance ranges, reflecting moderate damage and increased human impact.

Nearly all probability mass is allocated to extreme significance for magnitudes greater than 3+, suggesting that if earthquakes surpass this threshold, severe societal repercussions become statistically dominating. This behavior confirms the structural accuracy of our Bayesian Network and is consistent with actual seismic risk models.

From our DAGs specified in Section 3.2, we were able to provide the following results for the models' predictive behaviors in order to draw connections from our results to real-world interpretations.

To reiterate the difference between the two graphs, the larger DAG utilizes a deterministic tsunami value. This DAG will be referred to as "Full DAG". In some CPTs figures, we also compute the marginal over Tsunami to make it more comparable to the next DAG, which we took out the Tsunami variable from. This DAG will be referred to as "Smaller DAG".

Table 2: Significance CPT for Full DAG

Significance Magnitude	low (0–50)	moderate (50–150)	high (150–400)	severe (400–800)	extreme (800–inf)
neg_or_zero	0.9999	0.0001	N/A	N/A	N/A
0–1	1.0000	0.0000	N/A	N/A	N/A
1–2	0.8957	0.1043	0.0000	N/A	N/A
2–3	N/A	0.9959	0.0037	0.0004	N/A
3–5	N/A	0.2226	0.7401	0.0345	0.0029
5+	N/A	N/A	0.0957	0.6174	0.2870

(a) California Significance CPT (Tsunami = 0)

Significance Magnitude	low (0–50)	moderate (50–150)	high (150–400)	severe (400–800)	extreme (800–inf)
1–2	1.0000	N/A	N/A	N/A	N/A
3–5	N/A	N/A	0.3140	0.5289	0.1570
5+	N/A	N/A	N/A	0.3478	0.6522

(b) California Significance CPT (Tsunami = 1)

For all CPTs, we specifically modeled California, as it was the region with the most earthquakes. Across both models for both tsunami occurrence situations, we see a rapid growth in increasing earthquake magnitude, which is to be expected. However, We see a larger jump in the extreme category for Tsunami = 1, indicating that there is a higher severity when there are tsunamis present.

In this Full DAG, having tsunami occurrence as a variable is critical in visualizing the amplifying factor it has on significance. In contrast with the non-tsunami CPT with similar magnitudes, the tsunami occurrence CPT shows a significantly higher probability of extreme significance. This aligns with real-world disaster outcomes, where tsunamis often cause the largest loss of life and infrastructure damage.

Table 3: Significance CPT for California (Smaller DAG)

Significance Magnitude	low (0–50)	moderate (50–150)	high (150–400)	severe (400–800)	extreme (800–inf)
neg_or_zero	0.9999	0.0001	N/A	N/A	N/A
0-1	1.0000	0.0000	N/A	N/A	N/A
1-2	0.8957	0.1043	0.0000	N/A	N/A
2-3	N/A	0.9959	0.0037	0.0004	N/A
3-5	N/A	0.2205	0.7361	0.0391	0.0043
5+	N/A	N/A	0.0797	0.5725	0.3478

The Smaller DAG in Table 4 removes tsunami entirely but still captures the general magnitude → significance trend. However, it fails to differentiate catastrophic tsunami-driven disasters from non-tsunami earthquakes of similar magnitude. Therefore, having a separation between tsunami and non-tsunami earthquakes is a necessary explanatory variable rather than just redundant noise.

Our Full DAG also shows a clearer separation between moderate and extreme categories compared to the Smaller DAG, demonstrating that including tsunami as a parent variable improves the model's ability to distinguish catastrophic events rather than simply categorize large earthquakes.

As a result, the Full DAG more accurately captures the dynamics of actual disasters, especially in coastal areas like California and other non-landlocked areas, where tsunamis are the most likely to be present.

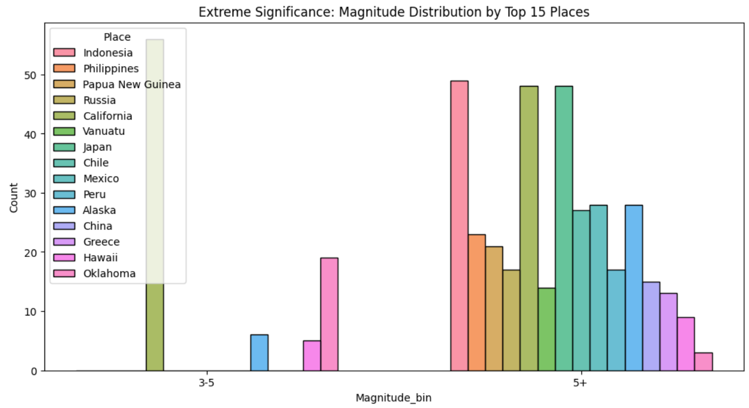

Figure 8: Magnitude Distribution of the top 15 Extreme Regions

From our results, we conclude that the Full DAG, including the occurrence or non-occurrence of a tsunami, more accurately captures the dynamics of actual disasters, especially in coastal areas like California and other non-landlocked areas, where tsunamis are the most likely to be present.

To expand upon this insight from the Full DAG, we also took the Extreme Significance category and plotted the magnitude distributions in Figure 8 across the top 15 regions with the highest Significance values. This highlights the regions where locations tend to experience stronger earthquakes, which offers insight into global seismic hot spots. Knowing these regions and being able to more accurately predict their Significances is critical in understanding the potential risks that an earthquake might bring and the according precautions the people in those regions would have to take.

4.2 Scalability

Expanding the dataset to include earlier years or additional global sources would likely improve generalization by exposing the model to a broader range of earthquake patterns. However, earthquake data is not extremely large by ML standards, meaning the computational gains from more data may eventually flatten out once the variety of magnitude/location combinations saturates. If the time horizon is expanded, we would need to account for the changing importance of different cities/region, which will also be discussed in Convergence.

4.3 Convergence

One major challenge to convergence in our model is that earthquake significance is not a static property but a context-dependent variable. In particular, the importance of a location changes over time due to urban development, population density shifts, infrastructure growth, and socioeconomic factors. Because our dataset spans more than three decades (1990–2023), this temporal drift introduces non-stationarity into the data distribution.

As a result, the conditional relationships learned by the Bayesian Network may shift over time, making it more difficult for the model to converge to a single stable mapping between features and significance. Without explicitly modeling time-dependent effects, the learned CPTs represent an average over evolving dynamics rather than a fixed physical process. This limits strict convergence guarantees but still allows the model to capture meaningful long-term probabilistic trends.

Rather than being a supervised classifier with labeled test outputs, this project is essentially a probabilistic inference model. Standard metrics like accuracy, precision, or F1-score are not directly applicable since significance is created probabilistically from learnt CPTs rather than from ground truth labels per individual sample.

Rather, distributional realism, structural consistency, monotonic probability behavior, interpretability of the learnt CPTs, and sensitivity to additional causative variables like tsunami are used to assess the quality of the model. In risk-based inference systems, where uncertainty modeling is the main goal, these assessment dimensions are regarded as best practices for Bayesian network modeling.

5 Conclusion

5.1 Limitations

While the Bayesian Network captured meaningful relationships between magnitude, state, tsunami occurrence, and earthquake significance, the model has several limitations that affect how well it generalizes. One major issue is that the dataset's column descriptions were vague—especially the "significance" field, which was not well-defined in the source data and required external investigation to learn that it reflects felt intensity and human impact. This uncertainty influences how confidently we can interpret or model the variable. Compared to other earthquake datasets that include detailed intensity measurements from multiple instruments, our dataset lacked finer-grained attributes that could have improved model fidelity. Additionally, discretizing continuous variables led to information loss, and the large number of geographic labels created sparsity issues that produced unstable CPT estimates for many states. The tsunami variable was highly imbalanced, causing some probability tables to collapse into near-deterministic outcomes. Finally, the fixed DAG structure enforced certain independence assumptions that may not align with real seismic processes, limiting how well the network captures deeper causal relationships.

5.2 Future Study

Future work could address these limitations by incorporating richer, better-documented earthquake datasets that include explicit variable definitions and multiple intensity measurements from different instruments. These additional features would allow the model to represent severity more accurately and reduce ambiguity in how significance is interpreted. Regularizing CPTs with Bayesian priors or smoothing techniques would help reduce overconfidence in sparse categories, especially for regions with very few tsunami events. Another improvement is to cluster states into broader tectonic zones to reduce sparsity and generate more stable probability estimates. Adding new seismic features—such as depth, fault mechanism, rupture type, or plate boundary classification—could greatly enhance the expressive power of the network. Exploring data-driven structure learning may also reveal alternative DAG structures that better capture empirical dependencies. Finally, evaluating predictive accuracy through MAP inference, sampling-based uncertainty estimates, or extending the BN into a dynamic or spatiotemporal model could support more realistic forecasting and decision-making applications.

6 Reflections and Contributions

6.1 Suggestions for Future Students

Watch out for potential data leakage because it can directly affect feature selection, network design, and result validity. For example, using a future data to predict an event that happened in the past would be unreasonable. This can be checked when designing and analyzing the Belief Network.

Always explore the entire dataset before settling with which dataset to use, which includes checking the range of each numeric feature, checking the size of the dataset, understanding the meaning behind each feature, etc. Additionally, if the dataset contains qualitative or text-based features, the pre-processing step could be troublesome.

6.2 Team contributions

Rohan Arcot
Rohan contributed to the project by developing the data preprocessing and geographic cleaning pipeline. He implemented the full place-name normalization system that standardizes inconsistent earthquake location strings using layered transformations, including directional stripping, region reduction, island and border normalization, synonym and abbreviation resolution, and targeted overrides for edge cases. Rohan also built the geospatial clustering framework that uses HDBSCAN on sampled coordinate data, followed by k-nearest neighbor propagation and K-Means consolidation, to convert raw latitude–longitude inputs into 15 coherent macro-regions.

Nina Ervin
Nina contributed to the project by coding maximum likelihood estimation. She also wrote the final report section methodology (Milestone 2) and modeling and Inference (Milestone 3-4). Nina has learned more about how to turn data into a solvable probability problem, as well as how to create DAGs in LaTeX.

Audrey Liang:
Audrey contributed to the project by writing the problem description and section 2.1, data sourcing. She also wrote and revised section 4.1 for the figures and tables in Section 4. Additionally, she created the data visualizations for

the Significance Data Visualizations for both full and smaller DAGs, pivot tables of the results and formatted the final LaTeX documents. This project taught Audrey more efficient ways of modeling probabilistic data for better interpretability.

Hantian Lin:
Hantian contributed to the project by writing the project description and choosing the dataset. Hantian learned to analyze the dataset and think about dataset features more critically. Writing the report out of the dataset also helped to practice organizing a study.

Taguhi Yenokyan:
Taguhi contributed by implementing and refining all visualizations. She corrected the data passed into the plotting functions by replacing `mle_data` with the appropriate `Significance_CPT` and completed the CPT-based logic for visualizing $P(\text{Significance} \mid \text{Magnitude, State, Tsunami})$, ensuring all pseudocode requirements were met (filtering by state and tsunami, grouping by magnitude bins, and displaying significance distributions). She validated model outputs and explained cases where certain states produced degenerate distributions. In addition, Taguhi wrote the Conclusion and Limitations sections, created the descriptions for all inserted visualizations, and prepared and submitted the final assignment on behalf of the team.

6.3 Generative-AI Disclosure

Generative-AI was used in reviewing and interpreting the noisy 856 “unique” location strings in the dataset as many of these entries contained inconsistent phrasing, nested directional language, historical naming variations, and geographically ambiguous terminology. GenAI was helpful in proposing consolidation ideas and identifying patterns that would be difficult to detect manually. AI served as a NLP-pattern-recognition aid while all final normalization, rules, mappings, and code implementations were designed, validated, and written by us.

It was also used in creating a function to efficiently convert the plotted CPT tables to LaTeX code.

7 References

References

[1] Lobello, A. (2023). *The Ultimate Earthquake Dataset from 1990–2023* [Data set]. Kaggle. https://www.kaggle.com/datasets/alessandrolobello/the-ultimate-earthquake-dataset-from-1990-2023