Traffic Causality Analysis for Robust Road Freight

May 2025

A Research Report from the Pacific Southwest Region University Transportation Center

John Krumm, University of Southern California

Cyrus Shahabi, University of Southern California

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
PSR-23-06 TO 076	N/A	N/A

4. Title and Subtitle Traffic Causality Analysis for Robust Road Freight	**5. Report Date** May 6, 2025 **6. Performing Organization Code** N/A
7. Author(s) John Krumm, 0000-0003-4394-6704 Cyrus Shahabi, 0000-0001-9118-0681	**8. Performing Organization Report No.** PSR-23-06 TO 076
9. Performing Organization Name and Address METRANS Transportation Center University of Southern California University Park Campus, RGL 216 Los Angeles, CA 90089-0626	**10. Work Unit No.** N/A **11. Contract or Grant No.** USDOT Grant 69A3551747109 Caltrans 65A0674
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590	**13. Type of Report and Period Covered** Final report (4/2/2024 – 12/31/2024) **14. Sponsoring Agency Code** USDOT OST-R

15. Supplementary Notes
N/A

16. Abstract
Road traffic congestion is a persistent problem. Focusing resources on the causes of congestion is a potentially efficient strategy for reducing slowdowns. We present an algorithm to discover which parts of the Los Angeles highway system tend to cause slowdowns on other parts of the highway. We use time series of road speeds as inputs to our causal discovery algorithm. Finding other algorithms inadequate, we develop a new approach that is novel in two ways. First, it concentrates on just the presence or absence of events in the time series, where an event indicates the temporal beginning of a traffic slowdown. Second, we train a binary classifier to identify pairs of cause/effect locations using pairs of road locations where we are reasonably certain a priori of their causal connections, both positive and negative. We test our approach on six months of road speed data from 195 different highway speed sensors in the Los Angeles area, showing that our approach is superior to state-of-the-art baselines. We include an analysis of parts of the highway that are especially subject to slowdowns for freight traffic.

17. Key Words	18. Distribution Statement
road traffic, causality, causal discovery, traffic jams, freight	No restrictions.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	28	N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Contents

About the Pacific Southwest Region University Transportation Center

The Pacific Southwest Region University Transportation Center (UTC) is the Region 9 University Transportation Center funded under the US Department of Transportation's University Transportation Centers Program. Established in 2016, the Pacific Southwest Region UTC (PSR) is led by the University of Southern California and includes seven partners: Long Beach State University; University of California, Davis; University of California, Irvine; University of California, Los Angeles; University of Hawaii; Northern Arizona University; Pima Community College.

The Pacific Southwest Region UTC conducts an integrated, multidisciplinary program of research, education and technology transfer aimed at *improving the mobility of people and goods throughout the region.* Our program is organized around four themes: 1) technology to address transportation problems and improve mobility; 2) improving mobility for vulnerable populations; 3) Improving resilience and protecting the environment; and 4) managing mobility in high growth areas.

U.S. Department of Transportation (USDOT) Disclaimer

California Department of Transportation (CALTRANS) Disclaimer

Disclosure

Principal Investigator, Co-Principal Investigators, others, conducted this research titled, "Traffic Causality Analysis for Robust Road Freight" at the Department of Computer Science in the Viterbi School of Engineering at the University of Southern California. The research took place from April 2, 2024 to December 31, 2024 and was funded by a grant from Caltrans in the amount of $100,000. The research was conducted as part of the Pacific Southwest Region University Transportation Center research program.

Abstract

Road traffic congestion is a persistent problem. Focusing resources on the causes of congestion is a potentially efficient strategy for reducing slowdowns. We present an algorithm to discover which parts of the Los Angeles highway system tend to cause slowdowns on other parts of the highway. We use time series of road speeds as inputs to our causal discovery algorithm. Finding other algorithms inadequate, we develop a new approach that is novel in two ways. First, it concentrates on just the presence or absence of events in the time series, where an event indicates the temporal beginning of a traffic slowdown. Second, we train a binary classifier to identify pairs of cause/effect locations using pairs of road locations where we are reasonably certain a priori of their causal connections, both positive and negative. We test our approach on six months of road speed data from 195 different highway speed sensors in the Los Angeles area, showing that our approach is superior to state-of-the-art baselines. We include an analysis of parts of the highway that are especially subject to slowdowns for freight traffic.

Traffic Causality Analysis for Robust Road Freight

Executive Summary

Executive Summary: Traffic Causality Analysis for Congestion Reduction

Urban traffic congestion imposes enormous costs on the U.S. economy, with drivers losing 43 hours and $74 billion annually due to delays. The freight sector suffers similar losses while idling vehicles increase pollution. Rather than broadly investing in road restructuring, a targeted approach is needed to identify specific road sections that cause widespread congestion.

Our research introduces a novel method to identify causal relationships in traffic patterns, pinpointing road sections that regularly trigger slowdowns elsewhere in the network. Unlike traditional approaches that focus solely on persistently congested areas, our method identifies the root causes of congestion that propagate through the transportation system.

Using six months of data from California's transportation measurement system across 195 measurement stations in Los Angeles, we developed an event-based traffic causality model that significantly outperforms existing techniques. Our approach:

1. Identifies unexpected traffic slowdowns by comparing actual speeds to predicted patterns
2. Converts continuous speed data into binary slowdown events
3. Calculates the probability that slowdowns at one location cause slowdowns at another
4. Uses machine learning to classify and validate causal relationships

The advantages of our method include:

- Making no assumptions about relationships between speed values at different locations
- Eliminating confounding effects of time-based traffic patterns
- Relying on natural traffic disruption events as experiments
- Processing data more efficiently through binary event conversion
- Achieving superior accuracy and computational efficiency compared to alternatives

By identifying the most problematic road sections that cause widespread congestion, transportation departments can allocate resources more effectively. Targeted improvements at these causal hotspots would provide greater system-wide benefits than addressing symptoms alone. This approach may enable more efficient use of limited infrastructure budgets while reducing congestion costs for drivers, businesses, and the environment.

Our method can be expanded to other regions and road types, refined with additional data sources like traffic incidents and GPS data, and developed further to model complex multi-causal relationships throughout transportation networks.

Introduction

Traffic congestion cost U.S. drivers an average of 43 hours and a total of $74 billion in lost time in 2024 (Sager 2024). In 2019, the U.S. freight sector lost $74.1 billion due to slow traffic (McCarthy 2020). Idling engines also cause more pollution. To reduce these costs, it is important to direct resources efficiently for reducing traffic congestion.

One way to relieve traffic congestion is to restructure the roads, but this is expensive. We need a framework to identify the most effective places to spend. It is not enough to identify the sections of road that consistently suffer from slowdowns, because the congestion may be propagating from other roads. Instead, it is important to identify the causes of the slowdowns, especially those roads that regularly cause slowdowns on other roads. These especially causal roads are good candidates for improvement. While traffic engineers are skilled at identifying causes, the active science of causal discovery, along with detailed data on traffic speeds, can also help to find which parts of the road tend to cause traffic slowdowns.

Ideally we would be able to perform explicit experiments to determine causality. In the case of roads, we could intentionally cause a traffic slowdown and observe how it affects traffic speeds on other roads. Since this is impractical, we depend on measured data from the relevant phenomena, using it as a set of natural experiments. The science of causal discovery uses these measured signals to identify causal relationships. For traffic, we are fortunate to have rich data on traffic from sensors in the roads giving time series of traffic speeds measured at different locations. Causal discovery can show which locations on the roads tend to cause slowdowns at other locations.

The first milestone in the science of causal discovery was Granger causality, which, at its core, examines two time series, $x(t_i)$ and $y(t_i)$ (Granger 1969). We can imagine that $y(t_i)$ is the time series of road speeds at one road location, and $x(t_i)$ is the time series of road speeds at another road location. The t_i represent discrete, evenly sampled points in time. We are interested in discovering if traffic speeds at the location x "cause" traffic speeds at location y.

Granger causality models $y(t_i)$ with an autoregressive model that says the value of $y(t_i)$ can be predicted by the p past values of $y(t_i)$, where p is some positive integer:

$$y(t_i) = \sum_{j=1}^{p} \alpha_j \, y(t_{i-j}) + \epsilon_i$$

The α_j values are real coefficients that can be computed on the time series data with least squares. ϵ_i represents the error in the estimate of $y(t_i)$. If this equation accurately predicts $y(t_i)$, then the signal is somewhat independent and unlikely dependent on another signal, such as $x(t_i)$.

Causal discovery posits that a signal $x(t_i)$ might cause $y(t_i)$. The test begins with an autoregessive model that says the value of $y(t_i)$ is a linear function of its own past values *and* the past values of $x(t_i)$:

$$y(t_i) = \sum_{j=1}^{p} \alpha_j \, y(t_{i-j}) + \sum_{i=1}^{q} \beta_k \, x(t_{i-k}) + \epsilon_i$$

The α_j and β_k values are real coefficients that can be computed on the time series data with least squares, and q is a positive integer. A series of statistical tests determine the causality. The intuition is that x causes y if the past values of x significantly help in predicting values of y. Our tests showed that Granger causality failed to find obvious causal traffic pairs from recorded speed data. One limitation of Granger causality is that it only tests for a linear relationship between the measured values.

Since Granger's test, causal discovery has matured, but the basic paradigm remains the same: There is likely a causal connection from x to y if x somehow helps explain the evolution of y. More modern tests can tolerate nonlinear relationships between the variables, accommodate many more than two variables, and produce a causal graph that represents a cascade of causal relationships propagating through a network of variables.

We are focused on finding causal connections between different pairs of locations on roads using time series of traffic speeds. One rare aspect of our causality problem is that we have pairs of locations where, for some pairs, we are quite certain of the causal connection. Locations on the same road that are near each other, with the same direction of travel, with no feeder nor turnouts between them, are likely causally connected, as illustrated in Figure 1. Pairs of points on roads that are far apart are most likely not causally connected. Because we can find many examples of ground truth pairs, both positive and negative, we can apply a machine learned, binary classifier to our problem and precisely measure its performance.

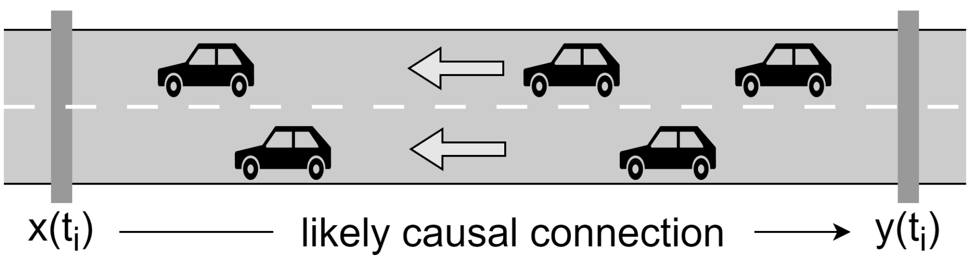


Figure 1: Road speeds at this pair of locations are likely causally connected as a slowdown at x would cause a slowdown at y after some lag in time.

We found that two modern causal discovery techniques, PCMCI (Runge 2020) and DYNOTEAR (Pamfil et al. 2020), do not perform well for our problem. Granger causality did not perform well, either. These failures are likely due to the road speed time series having a complicated, nonlinear connection even between obvious causal pairs. Our new algorithm is novel in three ways. First, we replace the time series of traffic speeds with a derived sequence of binary slowdown events. We introduce a simple method for finding the leading edges of slowdown events, and we use these events to assess causality. This means we do not need to make any assumptions nor create any models of the behavior of the raw time series, e.g. linear or nonlinear. The second innovation is that we derive a maximum likelihood method for explicitly estimating the probability that one sequence of binary events causes another sequence. Third, we can apply a machine learning classifier to the problem of causality discovery, because we have ground truth causal and noncausal pairs.

Related Work

Two representative, modern approaches to causal discovery on time series are PCMCI (Runge 2020) and DYNOTEARS (Pamfil et al. 2020). PCMCI is based on statistical tests of independence among the time series. Its goal is to build a graph of directed causal connections, where the nodes of the graph are the signals (e.g. traffic measurement stations) and the directed edges indicate a causal relationship from cause to effect. PCMCI is robust to confounding causal factors, such as if two signals have a common cause, then the two signals themselves may appear to have a causal connection. Through statistical tests, PCMCI sorts through dependencies to produce a causal graph. The relationships between the signals may be linear or nonlinear. The authors claim PCMCI is suitable for up to hundreds of time series each comprising hundreds of samples. For our traffic causality problem, we have hundreds of measurement stations with 5-minute speed samples. With six months of data, each station has over 50,000 samples, which we found exceeded the capacity of PCMCI to handle in a reasonable amount of computation time. One of the advantages of our binary, event-based representation of traffic is efficient processing with our approach. PCMCI also lacks the ability to learn from ground truth causal pairs, something that is built in to our approach.

PCMCI represents a statistical approach to causal discovery. Another category of algorithm searches directly for the entire causal graph, so-called structure learning. The directed graph is usually constrained to be acyclical, thus a directed, acyclical graph (DAG). The acyclical property ensures that no signal can be an indirect cause of itself. A significant advance in discovering causal DAGs was the paper from Zheng et al. that introduced a continous optimization criteria to enforce the acyclicity constraint, leading to high quality searches (Zheng et al. 2018). This paper spawned a cascade of related work, one of which was the DYNOTEARS algorithm for learning causal DAGs from time series data (Pamfil et al. 2020). While this work provides a foundation for us to eliminate certain edges based on prior intuition, a major pitfall of this approach is that it does not support a framework to insert edges we know to exist. The DAG restriction also may not be true for traffic, as it is conceivable that a slowdown may reinforce itself through the road network. DYNOTEARS was also very slow to run on our data.

Traffic Data

Our traffic data comes from California's Caltrans Performance Measurement System (PeMS) (California Department of Transportation 2023), which provides free downloads of traffic data from static sensors on the freeway system across all major metropolitan areas of the state. Our data was from the Los Angeles area (District 7) for the first six months of 2024 with some of the measurement stations locations shown in Figure 2. We used the "Station - 5-Minute" data, which gives, for each measurement station, timestamped five-minute aggregates of mean speed and flow for each traffic lane. Our analysis was based on mean traffic speed over all the lanes for each station, maintaining the five-minute measurement intervals for maximum temporal resolution. Typical speed data for one station for one week is shown in Figure 6.

We applied multiple filters to the measurement stations to arrive at a reduced set for analysis. First, each sensor is on a type of road from this list:

- Conventional Highway (CH)
- Collector/Distributor (CD)
- Freeway-to-Freeway Connector (FF)
- Off-Ramp (FR)
- High-Occupancy Vehicle (HV)
- On-Ramp (OR)
- Mainline (ML)

The counts of the different road types with sensors are shown in Figure 3. We chose only sensors on the "Mainline" roads to reduce the computational load in our experiments and because these roads give clear instances of very likely causal connections for testing. The "Mainline" sensors are shown on a map in Figure 2.

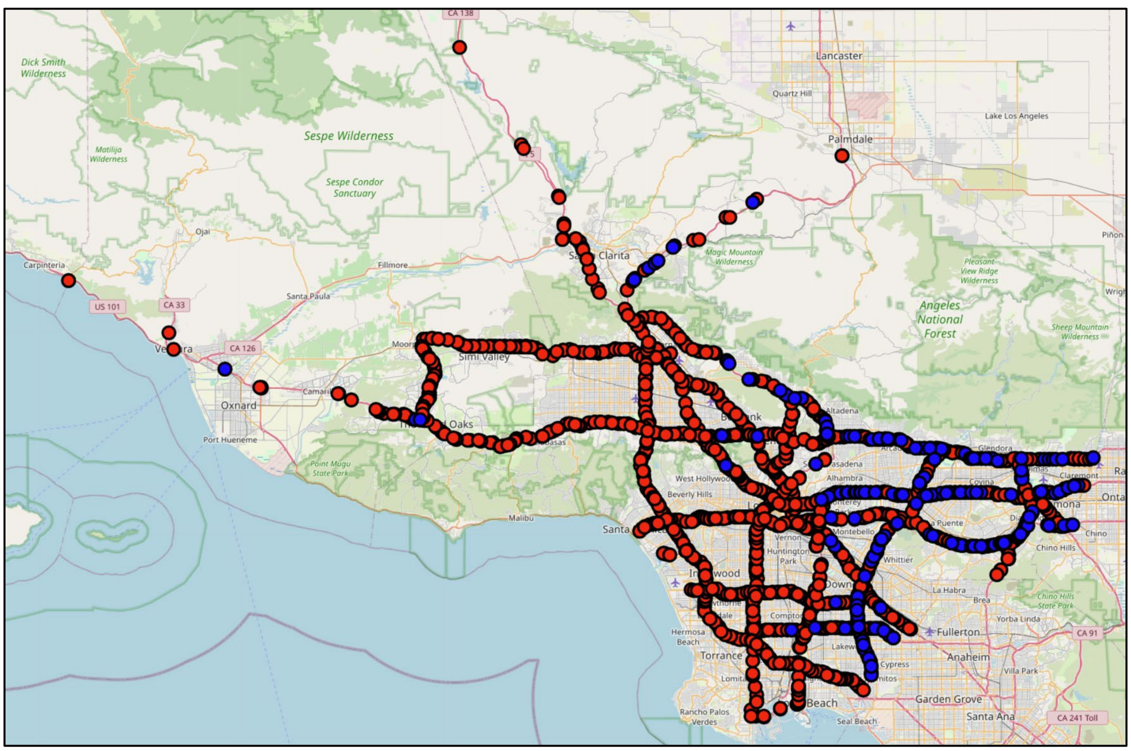


Figure 2: Locations of "Mainline" sensors in Caltrans District 7. The blue dots represent the 195 stations we used in our causality analysis, because they had sufficient speed data.

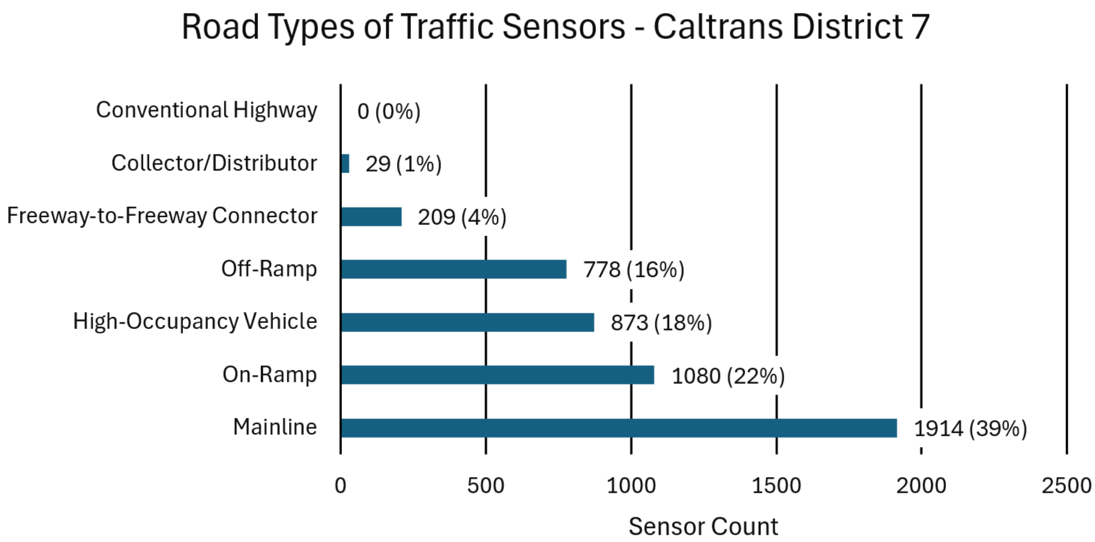


Figure 3: Counts of sensor road types in Caltrans District 7

Drive Time Data

For estimating ground truth causal pairs, we construct a distance matrix using the Bing Maps API (Microsoft 2024). To do this, for each station i, we query the API to provide us with the time it takes to travel to every station j such that $j \neq i$. Each entry D_{ij} of this drive time matrix matrix D represents the time it takes to drive from station i to station j. It is seen trivially that $D_{ii} = 0$ for all i. Because the measured drive time from the API was dependent on traffic at that time, we selected an arbitrary time (i.e. March 3rd, 2024 at midnight). This ensures that congestion does not play a strong role in the computed travel time between stations.

Ground Truth

Using intuition about road networks, we can construct a ground truth dataset to validate the results of our model. To build our ground truth set, we first select a set of stations to consider. Many of the speed values in the PeMS data are imputed due to sensor dropouts (Caltrans 2020). The imputation methods do not necessarily preserve the anomalous traffic slowdowns on which our algorithm depends, so stations with much imputed data are not good candidates for our causality analysis. Figure 4 shows how the number of sensor stations available varies with different levels of data completeness, where completness is the fraction of non-imputed data. The figure shows, for instance, that 300 stations have 75% or more non-imputed data. For our analysis, we chose 195 stations with at least 90% complete data, shown as blue dots in Figure 2.

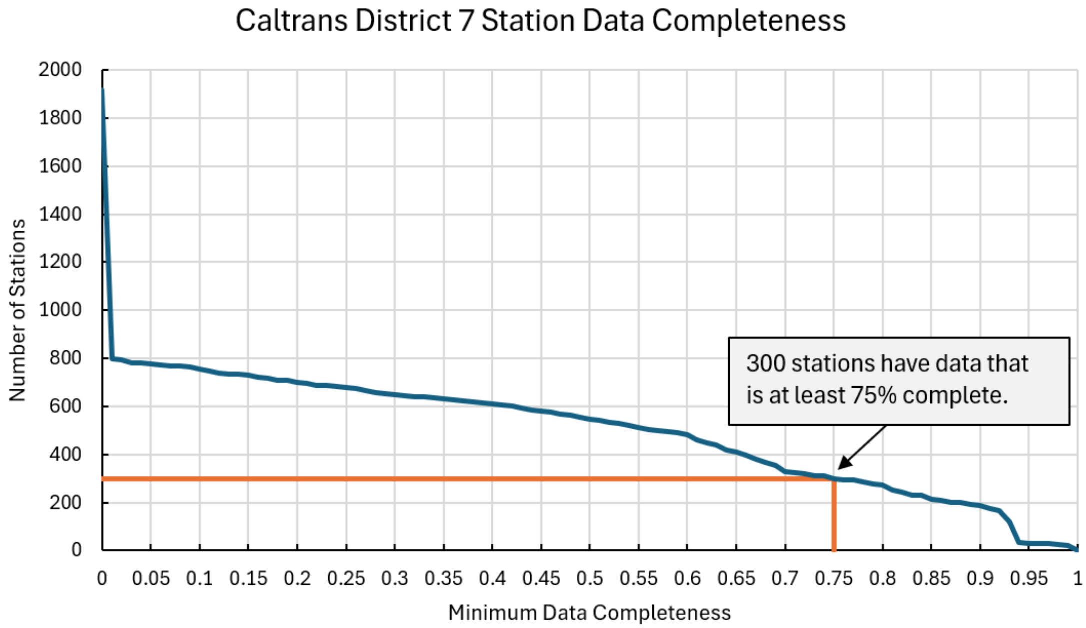


Figure 4: Counts of Mainline sensors vs data completeness in Caltrans District 7

We consider all candidate causal pairs (i, j, k), where i is the causal station, j is the affected station, and k is the determined lag. The lag k is an integer multiple of five minutes, because the time series is sampled every five minutes. This set contains ${}^{N}P_2 l_{max} = N(N-1)l_{max}$ unique edges, where l_{max} represents the total number of lags considered and $N = 195$ is the total number of stations considered.

All of these candidate causal pairs are added to our ground truth set G. In our approach, we classify candidate pairs from G as positive (causal) based on several criteria:

- **Distance required to travel**: We classify pairs of stations (i, j) as negative if their drive time is too long. Based on prior work, we infer that the maximum speed for congestion to propagate upstream from a causal point is roughly 20 kph (Fei et al. 2017). Given this and the distance to travel, we can compute the approximate time required for traffic to propagate upstream to the affected station. This distance is given to us by the drive time matrix D that we constructed. This approach allows us to compute the exact *minimum* lag that we expect a cause to propagate. This is a minimum, because it could take longer to propagate than the minimum time. For example, if stations i and j are ten kilometers apart, we say that it should take approximately 30 minutes (or a lag of six) to propagate to the affected station. For simplicity, we allow a soft threshold of one, meaning in our previous example, we consider lags of six and seven to be fair estimates of the propagation time. In this case, lags zero to five and lag eight for the two stations would be labeled negative, and six and seven would be positive. A lag of zero is considered non-causal since we assume propagation can not be instantaneous. In our experiments, we set the maximum lag $l_{max} = 8$ (40 minutes) and do not consider any stations beyond this threshold.

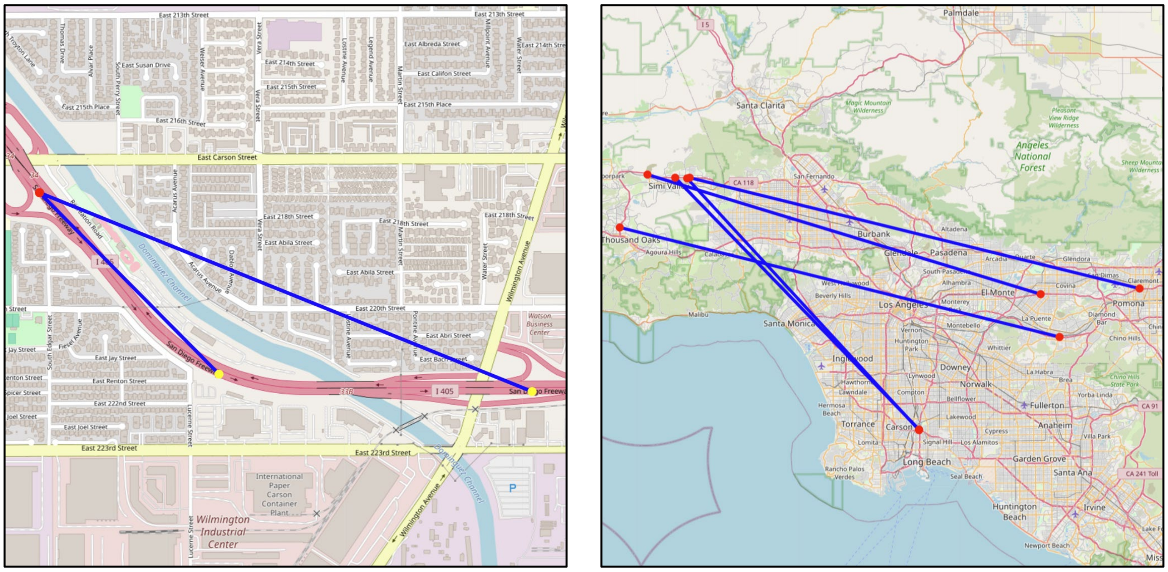
- **Highway Number**: An important consideration is how we consider stations on different major highways. For simplicity, if the travel time between the two stations D_{ij} is more than 10 minutes, we do take the pair as positive. We work under the assumption that majority of traffic events are contained to a specific road and do not spread across other roads. However, as with the other criteria, this applies only to our selection of positive ground truth pairs, and we do not automatically label pairs from different highways as negative. This is a conservative restriction, because inter-highway causality is possible. However, we are being careful to avoid including any false positives in our list of positive causality pairs.

- **Direction of Travel**: This refers to the direction of travel (i.e. northbound or southbound) on the **same** highway. All major highways typically offer two directions to travel on the same road (i.e. N/S or E/W). Thus, we say that if traffic is flowing in one direction, and there is a causal event, it should not impact any of the traffic flowing in the other direction. For example, traffic events on 110 North should not affect events on 110 West. In popular culture, a cause/effect relationship between two different highway directions is typically referred to as "rubbernecking," but is not something we can capture easily with our data, so we eliminate these stations from consideration, again being conservative.

- **Propagation Direction**: Naturally, traffic events that occur at any given station should only propagate upstream. It may seem difficult initially to determine which direction is upstream from two given stations (since roads are not perfect and can curve in many unpredictable directions), but there is a simple yet intuitive approach to resolve this issue. Given two stations i and j, we can obtain the distance to travel from i to j and the distance to travel from j to i. Call these distances D_{ij} and D_{ji}. If $D_{ij} < D_{ji}$, then we say that traffic flows naturally from i to j, and traffic events at j propagate to i (and vice versa if $D_{ij} > D_{ji}$). Note that having a longer distance does not immediately classify a candidate station pair as negative, since they could still be on different major lines.

All edges that meet these criteria are labeled positive in our dataset. For example, in Figure 5a, we see nearby stations where the causal station i is upstream from the affected station j, which is correspondingly labeled as a positive ground truth pair in our dataset. It is important when using the driving time between two stations to construct the ground truth, we use the drive time from the affected station j to the causal station i. This gives us context as to how long it would theoretically take us to see traffic from i appear at j. The remaining edges that do not meet this criteria are placed in a pool of possible negative edges, meaning they may not have a causal connection. We sort these edges in descending order by the driving time between them, D_{ij}, such that the edges with the longest driving times are near the top of the list. These are the edges with the least chance of a causal connection. When we build a set of ground truth edges, both positive (causal) and negative (non-causal), we take the entire set of positive edges based on the criteria above, and then we take an equal number of negative edges from the sorted list of edges that are far apart in terms of driving time, giving a balanced training set.

Using these criteria, and a set of 195 stations with $l_{max} = 8$, we consider $195 * 194 * 8 = 302{,}640$ total candidate edge/lag tuples, from which we label 1771 as positive ground truth edges, which is roughly 0.59% of the total possible edge/lag tuples.



(a) Positive ground truth causal pairs. *(b) Negative ground truth causal pairs.*

Figure 5: Example ground truth causal pairs. In (a), yellow indicates the location of the cause (upstream from effect), and red indicates the location of effect (downstream from cause). In (b), both causal directions are considered as negative causal pairs.

Event Based Traffic Causality

Existing methods for detecting causality in time series tend to process the time series values directly. Such is the case with PCMCI (Runge 2020) and DYNOTEARS (Pamfil et al. 2020), which we compare with in Section "Results." This leads to intense, slow processing. In contrast, our method is based on a binary time series, were each element in the time series represents the occurrence of an event or not. For traffic, the event represents the beginning of a significant, unexpected reduction in traffic speed. These unexpected slowdowns represent natural experiments whose effects we can search for in other parts of the road network.

This section describes our algorithm, beginning with how we detect unexpected slowdowns.

Detecting Traffic Slowdown Events

We will introduce some simple mathematical notation to explain our process for finding events in the speed time series, and eventually discovering causal connections between different measurement stations on the road. The speed measurement of station i at time j is $s_i(t_j)$ for $i \in [1, N]$ and $j \in [1, M]$. Here $N = 195$ measurement stations and $M = 52{,}416$ five-minute intervals in the first six months of 2024. As described above, each station i reports the average speed over all lanes every five minutes, thus t_j represents a discrete time that is a whole multiple of five minutes.

To find unexpected slowdown events at a station, we must have a notion of normal traffic speed at the station. There are several ways to do this, and they all attempt to predict what the normal traffic speed will be. We experimented with different approaches to predicting traffic speed based on historical traffic speed data. It is important that the prediction not be too good, because a 100% accurate prediction would discover nothing unexpected.

We explored two deep models for traffic speed prediction: a simple multilayer perceptron (MLP) and Long Short-Term Memory (LSTM). They both proved promising in our initial tests, but their computation time was excessive considering our six months of speed data for almost 200 traffic measurement stations to analyze. We instead chose a fast, effective method based on medians.

For a given measurement station, the median approach computes the median speed for each time slot in a generic week as its speed prediction. For example, the predicted speed at 10 am on a Monday is the median of all the speeds observed at that station at 10 am on all Mondays. Figure 6 shows one week of traffic speed data for one station. The black curve shows the measured speeds, and the orange curve shows the median week computed over all six months of data.

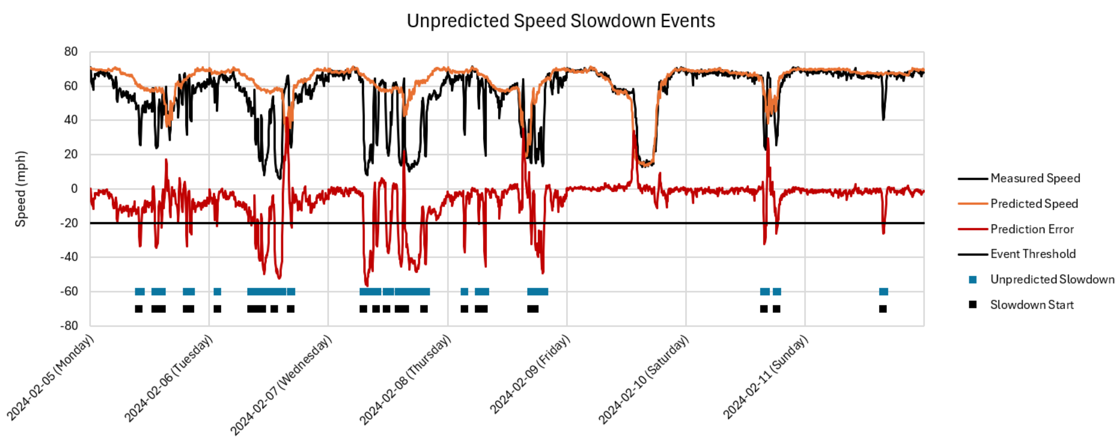


Figure 6: Traffic slowdown events are the leading edges of unpredicted slowdowns. This plot shows one week of speed data from one measurement station and its associated slowdown events.

The predicted traffic speed is $\widehat{s_i}(t_j)$, and the prediction error is $s_i(t_j) - \widehat{s_i}(t_j)$. This is shown as the red curve in Figure 6. The prediction error is negative when the actual traffic speed is slower than the prediction, indicating an unexpected slowdown. We declare a traffic slowdown when the prediction error fraction is less than a threshold α, i.e. when $\frac{s_i(t_j) - \widehat{s_i}(t_j)}{\widehat{s_i}(t_j)} < -\alpha$. We experimented with different values of α, with 0.25 being a typical value.

Comparing the prediction error to the threshold α gives a binary time series indicating the presence of an unexpected traffic slowdown, i.e.

$$u_i(t_j) = \begin{cases} \text{true} & \text{if } \dfrac{s_i(t_j) - \hat{s_i}(t_j)}{\hat{s_i}(t_j)} < -\alpha \\ \text{false} & \text{otherwise} \end{cases}$$

An example of $u_i(t_j)$ is shown as blue squares in Figure 6, and a close-up example is shown as the top signal in Figure 7. We are interested in sudden speed disruptions, so we concentrate on the leading edges of traffic slowdowns indicating the start of a slowdown, notated as $v_i(t_j)$:

$$v_i(t_j) = \begin{cases} u_i(t_j) & \text{if } j = 1 \text{ (first element)} \\ \text{true} & \text{if } u_i(t_j) = \text{true and } u_i(t_{j-1}) = \text{false} \\ \text{false} & \text{otherwise} \end{cases}$$

An example of $v_i(t_j)$ is shown as the bottom signal in Figure 7. This is essentially marking the beginning of every sequence of adjacent slowdowns with a "true". The case for $j = 1$ represents the first element.

An example of $v_i(t_j)$ is shown as the black squares in Figure 6, and a close-up example is shown as the bottom signal in Figure 7.

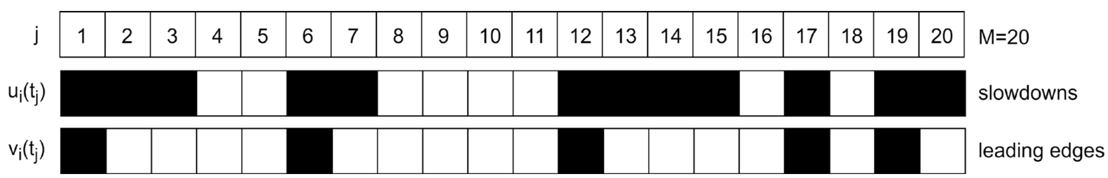


Figure 7: Black sequences of slowdowns are converted to leading edges to mark discrete events.

Probability of Cause

We can compute the probability p_c that events in one signal have caused events in another signal.

As illustrated in Figure 8, we have two binary sequences where each element of each sequence can represent the presence or absence of an event. One sequence is posited as the cause and the other as the effect, which means that events in the causal sequence sometimes cause events in the effect sequence at some constant lag in time. In Figure 8, the supposed cause is the top signal, the supposed effect is the bottom signal, and the lag is one.

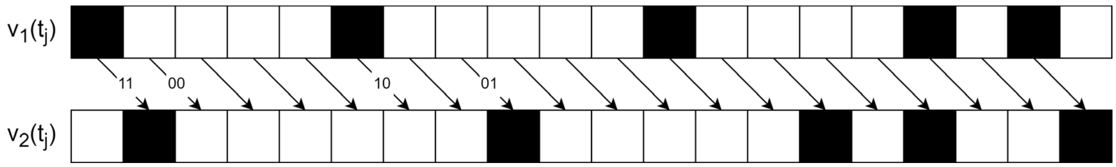


Figure 8: For a lag of one, we count the different kinds of correspondences between the hypothesized causal signal on top ($v_1(t_j)$) with the hypothesized effect signal below ($v_2(t_j)$). Examples of the four types of correspondences are marked on the arrows.

In each event sequence, an event can happen spontaneously, with no apparent cause, with probability p_s. For traffic, example spontaneous slowdowns could come from a crash, animal crossing, or distraction. The probability that an event in the causal sequence causes an event at the corresponding slot in the effect sequence is p_c. Our goal is to estimate the value of p_c in order to assess the strength of the causal connection.

There are four possible scenarios for a single corresponding pair of event slots for the cause and effect sequence. If zero (0) represents no event and one (1) represents an event, then the four possible pairs are 00, 01, 10, and 11. Assuming the pair is independent of all the other pairs, the probabilities of these four possible pairs are:

- **00**: No spontaneous event at the cause and no spontaneous event at the effect, thus $f_{00}(p_s, p_c) = (1 - p_s)(1 - p_s)$.
- **01**: No spontaneous event at the cause and a spontaneous event at the effect, thus $f_{01}(p_s, p_c) = (1 - p_s)p_s$.
- **10**: A spontaneous event at the cause and no spontaneous event at the effect and no caused event at the effect, thus $f_{10}(p_s, p_c) = p_s(1 - p_s)(1 - p_c)$.
- **11**: A spontaneous event at the cause and a spontaneous event or caused event at the effect, thus $f_{11}(p_s, p_c) = p_s(p_s + p_c - p_s p_c)$.

We note that $f_{00}(p_s, p_c) + f_{01}(p_s, p_c) + f_{10}(p_s, p_c) + f_{11}(p_s, p_c) = 1$. Our assumption about the indpendence of the pairs is bolstered by the fact that we look at only the leading edges of traffic slowdown events, not the entire extent of slowdowns. Presumably, unexpected slowdowns are independent.

For all the corresponding pairs in the two sequences, the counts of the different kinds of possible pairs are A_{00}, A_{01}, A_{10}, and A_{11}. Then the likelihood of all the counts is, assuming independence between pairs:

$$L(p_s, p_c) = (f_{00}(p_s, p_c))^{A_{00}}(f_{01}(p_s, p_c))^{A_{01}}(f_{10}(p_s, p_c))^{A_{10}}(f_{11}(p_s, p_c))^{A_{11}}$$

The log likelihood is:

$$\begin{aligned}\ell(p_s, p_c) &= \ln[(f_{00}(p_s, p_c))^{A_{00}}(f_{01}(p_s, p_c))^{A_{01}}(f_{10}(p_s, p_c))^{A_{10}}(f_{11}(p_s, p_c))^{A_{11}}] \\ &= A_{00}\ln(f_{00}(p_s, p_c)) + A_{01}\ln(f_{01}(p_s, p_c)) + \\ &\quad A_{10}\ln(f_{10}(p_s, p_c)) + A_{11}\ln(f_{11}(p_s, p_c))\end{aligned}$$

We want to find the causal probability p_c (along with the spontaneous probability p_s) that maximizes the log likelihood. We set the partial derivatives to zero, i.e. $\frac{\partial \ell}{\partial p_s} = 0$ and $\frac{\partial \ell}{\partial p_c} = 0$. The partial derivatives of $\ln(f_{ij})$ are

$$\frac{\partial \ln(f_{00})}{\partial p_s} = \frac{2}{p_s - 1}$$
$$\frac{\partial \ln(f_{00})}{\partial p_c} = 0$$
$$\frac{\partial \ln(f_{01})}{\partial p_s} = \frac{1 - 2p_s}{p_s(1 - p_s)}$$
$$\frac{\partial \ln(f_{01})}{\partial p_c} = 0$$
$$\frac{\partial \ln(f_{10})}{\partial p_s} = \frac{1 - 2p_s}{p_s(1 - p_s)}$$
$$\frac{\partial \ln(f_{10})}{\partial p_c} = \frac{1}{p_c - 1}$$
$$\frac{\partial \ln(f_{11})}{\partial p_s} = \frac{-2p_s(p_c - 1) + p_c}{p_s(p_s + p_c - p_s p_c)}$$
$$\frac{\partial \ln(f_{11})}{\partial p_c} = \frac{1 - p_s}{p_s + p_c - p_s p_c}$$

Setting

$$A_{00}\frac{\partial \ln(f_{00})}{\partial p_s} + A_{01}\frac{\partial \ln(f_{01})}{\partial p_s} + A_{10}\frac{\partial \ln(f_{10})}{\partial p_s} + A_{11}\frac{\partial \ln(f_{11})}{\partial p_s} = 0$$
$$A_{00}\frac{\partial \ln(f_{00})}{\partial p_c} + A_{01}\frac{\partial \ln(f_{01})}{\partial p_c} + A_{10}\frac{\partial \ln(f_{10})}{\partial p_c} + A_{11}\frac{\partial \ln(f_{11})}{\partial p_c} = 0$$

gives closed form expressions for the values of p_s and p_c that maximize the log likelihood, i.e.

$$p_s = \frac{A_{01} + A_{10} + A_{11}}{2(A_{00} + A_{01}) + A_{10} + A_{11}}$$
$$p_c = \frac{2A_{00}A_{11} + A_{01}(A_{11} - A_{10}) - A_{10}^2 - A_{10}A_{11}}{(2A_{00} + A_{01})(A_{10} + A_{11})}$$

To verify that these equations represent the location of the maximum, we first compute the determinant of the Hessian, i.e.

$$\begin{aligned} D &= \frac{\partial^2 \ell(p_s, p_c)}{\partial p_s^2}\frac{\partial^2 \ell(p_s, p_c)}{\partial p_c^2} - \left(\frac{\partial^2 \ell(p_s, p_c)}{\partial p_s \, \partial p_c}\right)^2 \\ &= \frac{(2A_{00} + A_{01})(A_{10} + A_{11})^3(2A_{00} + 2A_{10} + A_{10} + A_{11})}{A_{10}A_{11}(A_{01} + A_{10} + A_{11})} \end{aligned}$$

D is always positive, because $A_{ij} \geq 0$ and $A_{00} + A_{01} + A_{10} + A_{11} > 0$. Thus the critical point from the equations for p_s and p_c represents an extrema, not a saddle point. The critical point is a maximum if either $\frac{\partial^2 \ell(p_s,p_c)}{\partial p_s^2} < 0$ or $\frac{\partial^2 \ell(p_s,p_c)}{\partial p_c^2} < 0$. We find that

$$\frac{\partial^2 \ell(p_s, p_c)}{\partial p_c^2} = -\frac{(2A_{00} + A_{01})^2 (A_{10} + A_{11})^3}{A_{10} A_{11} (2A_{00} + 2A_{01} + A_{10} + A_{11})^2}$$

This is negative for the same reason that D is positive. Thus the critical point from the equations for p_s and p_c represents the unique maximum of the log likelihood, giving values for the probability of a spontaneous event p_s and the probability of a causal connection p_c.

Special Cases

Since p_s and p_c are probabilities, we know $0 \leq p_s, p_c \leq 1$. However, p_s and p_c are unconstrained in the equations for p_s and p_c. Since $\ell(p_s, p_c)$ is continuous, if the unconstrained maximum falls outside the constrained region, and because the only critical point of $\ell(p_s, p_c)$ is the unconstrained maximum, then the Extreme Value Theorem guarantees that the constrained maximum must occur somewhere on the edges of the constrained region. We look at the four edges next.

Edge $0 \leq p_s \leq 1$ **and** $p_c = 0$:
We have

$$\begin{aligned} f_{00}(p_s, p_c) &= (1 - p_s)(1 - p_s) \\ f_{01}(p_s, p_c) &= (1 - p_s)p_s \\ f_{10}(p_s, p_c) &= p_s(1 - p_s)p_s \\ f_{11}(p_s, p_c) &= p_s^2 \end{aligned}$$

Substituting these expressions into the log likelihood of the equation for $\ell(p_s, p_c)$, computing the derivative with respect to p_s, setting it to zero, and solving gives

$$\begin{aligned} p_{s,p_c=0} &= \frac{A01 + A10 + 2A11}{2(A00 + A01 + A10 + A11)} \\ p_c &= 0 \end{aligned}$$

Edge $0 \leq p_s \leq 1$ **and** $p_c = 1$:
We have

$$\begin{aligned} f_{00}(p_s, p_c) &= (1 - p_s)(1 - p_s) \\ f_{01}(p_s, p_c) &= (1 - p_s)p_s \\ f_{10}(p_s, p_c) &= 0 \\ f_{11}(p_s, p_c) &= p_s \end{aligned}$$

Following the same steps as above gives

$$p_{s,p_c=1} = \frac{A01 + A11}{2(A00 + A01) + A11}$$
$$p_c = 1$$

Edge $p_s = 0$ and $0 \leq p_c \leq 1$:
In this case, the derivative of the log likelihood with respect to p_c is zero, meaning there is no solution for the extreme value. Practically, when $p_s = 0$, there are no spontaneous events in the candidate causal signal, which means there is no opportunity to observe an event in the effect signal that was caused by the candidate causal signal. Thus it is impossible to detect causality in the case $p_s = 0$, and we can say nothing about p_c in this case.

Edge $p_s = 1$ and $0 \leq p_c \leq 1$:
This is similar to the case above in that the derivative is zero with respect to p_c. When we have $p_s = 1$, this means that every point in both signals is considered an event, and thus it is impossible to infer anything about p_c in this case.

Algorithm for Causality Probability

The derivations above suggest this algorithm for computing p_c, the probability of causality.

1. Compute p_s and p_c. If $0 \leq p_s \leq 1$ and $0 \leq p_c \leq 1$, then take these values and stop.
2. If p_s or p_c are outside the acceptable ranges, then compute candidate values of p_s from the special case equations above giving $(p_s, p_c) = (p_{s,p_c=0}, 0)$ and giving $(p_s, p_c) = (p_{s,p_c=1}, 1)$. Insert both these pairs into the log likelihood expression and take whichever pair gives the greater value.

Algorithm

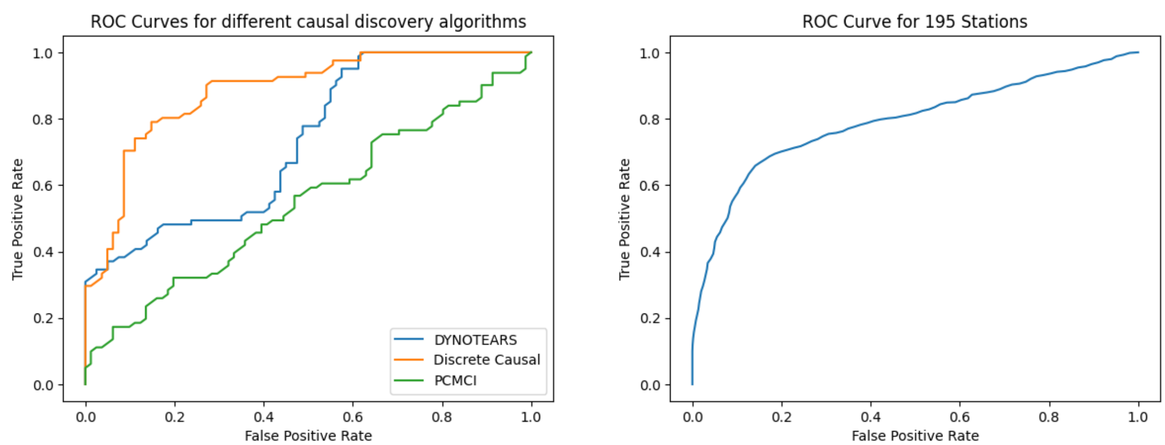
Although the above approach works for a single pair of stations and set lag, we provide a framework that scales our approach to a set of stations and outputs their results. First, we provide our framework with the set of N stations (nodes) that we want to consider, the maximum delay l_{max} (ie, a maximum delay of 8 corresponds to a maximum traffic propogation of $5 * 8 = 40$ minutes), and historical time series data for each station as input. For our experiments, we used 6 months of Caltrans PeMS data described in Section "Traffic Data." We then use the median week approach from Section "Event-Based Traffic Causality" to determine events representing the leading edges of traffic slowdowns.

We then consider all ${}^{N}P_2$ candidate permutations of station pairs. For simplicity, we do not evaluate a station against itself (i.e., no pair of stations (i, j) where $i = j$), but we do consider (i, j) and (j, i) to be distinct pairs. For each candidate pair of stations (i, j), we consider all the lags k where $0 \leq k \leq l_{max}$, and model the probability of a causal relationship as explained in Section "Event-Based Traffic Causality." When we model the probability of a causal relationship, we can adjust the error threshold α to increase or decrease the sensitivity of classifying a period as anomalous. If a pair of stations (i, j) at lag k produces a causal probability p_c greater than some threshold, we declare (i, j) to be a causal pair, i.e. traffic slowdowns at station i cause traffic slowdowns at station j with a temporal lag of k.

Results

We evaluated our internal data set and ground truth on the first six months of PeMS data from 2024 on 195 select traffic sensors with at least 90% complete data, as illustrated in Figure 4. All approaches use $l_{max} = 8$. In preliminary experiments, we tested our approach using data without the restriction of having a high threshold of data completeness. We found that our model struggled to properly identify anomalous periods and did not accurately classify our ground truth data. We attribute this to the event-based nature of our approach: imputed data is by nature non-anomalous and therefore hides events.

When evaluating our results, because the number of positive ground truth edges in our dataset is significantly less than the number of negative edges (since many station pairs and station lags are improbable/impossible), we reduce the number of negative edges to only consider the most extreme examples, i.e. edges with the highest driving times D_{ij}. This way, we retain a $1:1$ ratio between positive and negative ground truth edges in our dataset to ensure a more realistic evaluation of our results.



(a) ROC Curves of different algorithms. *(b) Our approach's ROC curve on 195 stations.*

Figure 9: ROC curve comparing various approaches. In (a), we show how our method outperforms existing baselines DYNOTEARS and PCMCI on a small subset of 30 stations. In (b), we show the performance of our model on 195 stations, with an AUC of 0.77.

*Table 1: Comparing performance between different models. Due to runtime limitations on the competing methods, we ran the experiments with 30 stations that are a subset of our 195 stations, sampled to ensure reasonable coverage. Time is in HH:MM:SS format. Our proposed method not only outperforms existing baselines, but is also more computationally efficient and scalable to larger datasets. * indicates an approximate time.*

Framework	AUC	Time
PCMCI	0.55	22:00:00*
DYNOTEARS	0.75	30:00:00*
Discrete Causal	**0.88**	**00:49:03**

Table 2: Grid search of different levels of anomaly threshold and the corresponding results using 195 stations.

Threshold	AUC
$\alpha = 0.15$	0.70
$\alpha = 0.25$	**0.77**
$\alpha = 0.40$	0.73

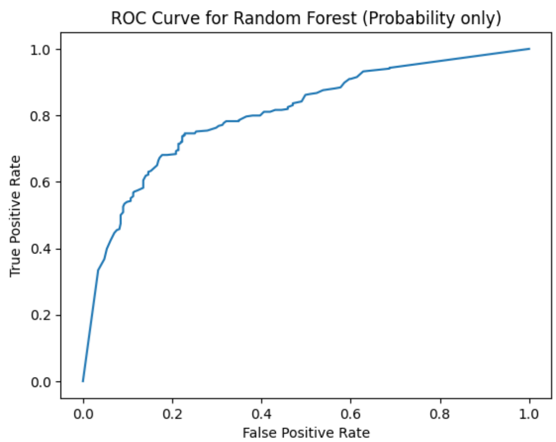
Comparison to State-of-the-Art Baselines

Table 1 shows the performance of our algorithm versus state-of-the-art benchmarks PCMCI and DYNOTEARS described in Section "Related Work." For efficiency and fair comparison, we conducted all of our experiments on an NVIDIA GeForce RTX 3090 and AMD Ryzen 9 5900 12-Core Processor. We found that it was infeasible to test the two baselines for our domain using $N = 195$ and $l_{max} = 8$ (since running these methods on our data would take on the order of weeks), so we selected a smaller subset of 30 stations from the 195 stations in our working set. This subset was selected strategically to detect if the models are capable of extracting clearly positive ground truth pairs, as opposed to selecting based on station completeness. Because our coverage and number of stations is smaller, we also set $l_{max} = 5$ for the baseline. When evaluated against our ground truth, our algorithm significantly outperforms other methods in both speed and accuracy. Thus existing methods are not well suited for the traffic causality problem due to performance and scalability.

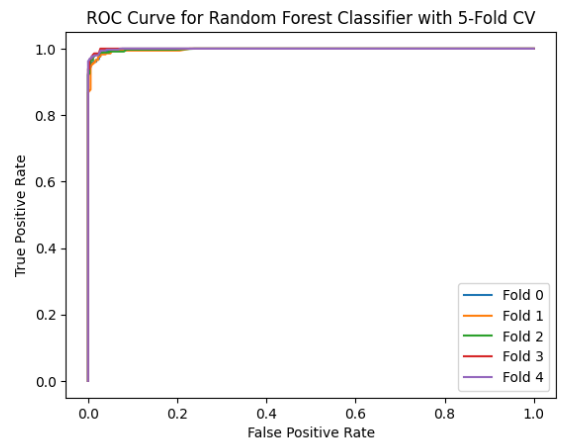
Figure 9a shows the ROC curves for all three baseline approaches on the set of 30 stations, and Figure 9b shows the results of our approach on the set of 195 stations, which took roughly one day to run. We plot our ROC curves based on a top-k threshold as opposed to a sampled threshold, since the probabilities p_c computed by our framework are not distributed in any meaningful way. Note that we will not be able to obtain 100% AUC for any of the approaches we consider. This is because, although certain pairs of stations are theoretically causal, it may not be the case that they are empirically observed to be causal. For example, in our six months of PeMS data, there may be minimal traffic slowdowns along certain sections of the highway system further from downtown Los Angeles, leading to few observed anomalies and a low p_c.

In Table 2, we complete a simple scan through different values of α, which is the threshold for declaring a traffic slowdown. By increasing the alpha threshold, we effectively tell our model to be more conservative about what regions of the time series are classified as anomalous, thus likely looking at a smaller set of anomalies. We show that $\alpha = 0.25$ is the most effective in correctly classifying our ground truth pairs.

Classification Accuracy



(a) Random Forest ROC (p_c only).



(b) Cross Validated Random Forest ROC (all features).

Figure 10: ROC curves for various random forest models. In (a), we show the ROC of only using the p_c value we computed, with an AUC of 0.80. In (b), we show the performance of our model with cross validation using all 5 features in a random forest, with an average AUC of 0.9985.

Instead of using a top-k boundary on p_c to construct an ROC curve (as seen in Figure 9), we can learn a more complex decision boundary using a random forest of decision trees. In addition to p_c, we can also use the correspondence counts A_{00}, A_{01}, A_{10}, and A_{11} (from the MLE equation in Section "Event-Based Traffic Causality") as inputs to the random forest.

Using various combinations of these features, we train different random forest models and extract a classification probability p_{forest} for each candidate tuple (i, j, k), determined from the classification vote in the ensemble. Figure 10 shows the effectiveness of our method at classifying between ground truth positive and negative pairs in dataset with 195 stations. When we use only p_c as a feature, our AUC is 0.80 (Figure 10a). When we add A_{00}, A_{01}, A_{10}, and A_{11} as features to the random forest, our AUC rises to 0.9985 on average. Because our downsampled data is randomly split $80/20$ for training and testing, respectively, we also conduct a simple 5-fold cross validation to show the effectiveness of our method, as pictured in Figure 10b.

Freight Routes

There are certain sections of the Los Angeles highway system that are especially subject to congestion that affects freight traffic. These areas are shown in Figure 12a from a technical report (Giuliano et al. 2017). We are interested in inferring causality at these locations.

Our first step in this analysis was to determine which of the PeMS measurement stations correspond to the freight congestion locations. We manually superimposed the station map

(Figure 12b) with the freight congestion map (Figure 12a) to find the corresponding stations shown in blue in Figure 12b. Figure 13 shows the data completeness for the 84 freight-related measurement stations. We find that four stations that meet the minimum data completeness requirement of 90% are also contained in our set of 195 stations.

Finally, we can train a similar classifier from Subsection 5.2 to determine pairs of stations (i, j) where j is a freight station. In the list of all tuples (i, j, k), we have 6984 candidates that include a relevant freight station in the effect. This is obtained from mapping the four freight stations in the effect to each of the other 194 stations in the set as a potential causal station. Out of these, only 40 candidate tuples are actually labeled as true by our ground truth dataset G. Thus, we downsample our dataset again to only include negatively labeled edges that contain the farthest driving distance between stations to obtain a similar 1: 1 ratio.

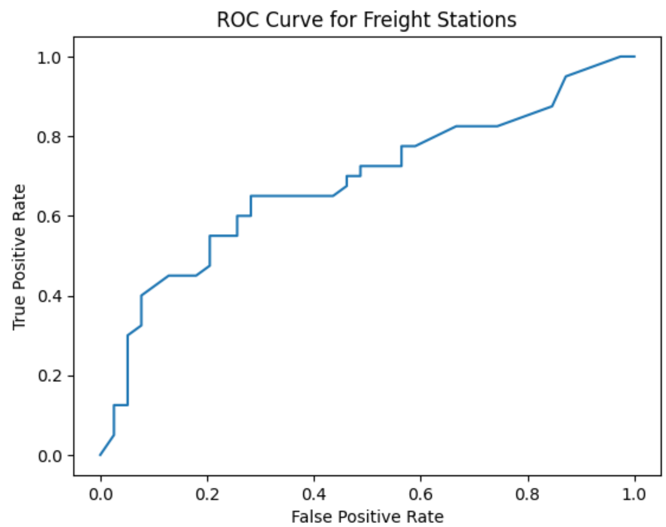
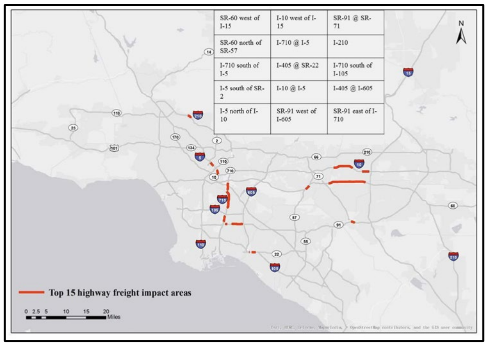
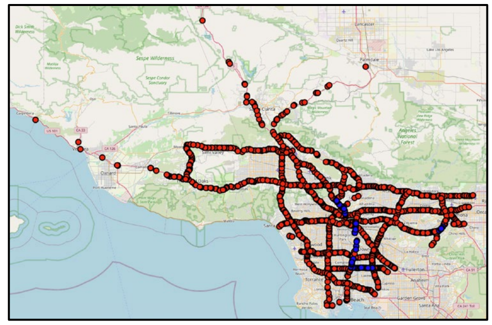


Figure 11: Freight random forest ROC (AUC = 0.68)

We then train a random forest classifier on pairs of stations (i, j) where the affected station j is not a freight station. Our forest uses the positive and negative labels from our ground truth dataset for each candidate pair. Then we test our data on the withheld pairs where j is a freight station. We use this method to classify candidate edges as positive or negative, and determine how well our framework can discover the causes of freight slowdowns. Figure 17 shows the ROC curve for the performance of our method with respect to the ground truth. One reason we see significantly lower AUC in this approach than with training the random forest previously is because we have significantly less training and testing data after downsampling and using a balanced training set. We provide a visualization of some positively labeled outputs obtained from a decision tree classifier on all five features (p_c, A_{00}, A_{01}, A_{10}, and A_{11}) in Figures 14a and 14b.



(a) Freight impact areas in the Los Angeles area (used with author's permission) (Giuliano et al. 2017)



(b) Measurement stations corresponding to freight routes as blue dots

Figure 12: The map in (a) shows impacted freight areas from another report. The corresponding mainline measurement stations from District 7 are shown as blue dots in (b).

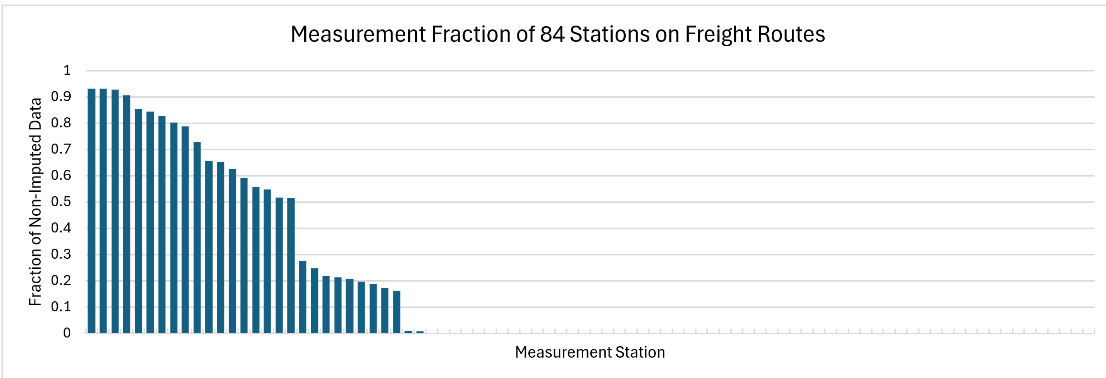
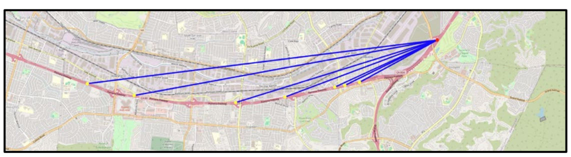


Figure 13: Speed data availability at measurement stations associated with freight congestion



(a) CA 60 Pomona Freeway Freight

(b) CA 91 Artesia Freeway Freight

Figure 14: Causality pairs along known freight routes. The yellow dots indicate the causal location, and the red dots indicate the affected location.

Conclusion

Our goal is to discover which locations in the road network tend to cause slowdowns at other locations. These especially causal locations are good candidates for further investigation and investment. Time series of speed data serve as a foundation for causal discovery. This problem stands out as one that offers somewhat certain ground truth data in that certain pairs of road locations clearly have a cause/effect relationship, and certain pairs do not. This means we can not only quantify the accuracy of our results, but we can employ a machine learning approach with ground truth data allocated to training and testing. We found that traditional Granger causality and two state-of-the-art causality discovery methods perform poorly. The state-of-the-art methods also run very slowly on our relatively large dataset.

Our new approach uses traffic slowdown events detected from traffic speed time series along with a maximum likelihood algorithm that computes the probability of one sequence of events causing another. Our tests show it is both quick to run and accurate, compared to the state-of-the-art approaches. Overall, we identify the following advantages our new approach:

- Our approach makes no assumption about the functional relationship between speed values on different parts of the road. For instance, Granger causality assumes a linear relationship. Instead, our approach looks only for corresponding slowdowns.
- Our approach implicitly eliminates the confounding effects of time. We know that traffic varies periodically over the day and week. These periodic fluctuations can be confused as causal connections. However, we eliminate the time dependence by extracting slowdowns that are *not* well-predicted by time, but are instead abnormal given the time of day and day of week.
- Related to the point above, our approach relies explicitly on natural experiments in the data in the form of unexpected traffic slowdowns. These events are useful for identifying how traffic slowdowns propagate from one part of the road system to other parts. By explicitly concentrating on events, we relieve the downstream causal inference system of having to discover the natural experiments.
- The time series signals are reduced from real (speed) values to binary events. This means the subsequent algorithm (Section 4) and processing are relatively fast, especially compared to the state-of-the-art baselines we tested against.
- Our approach is both faster and more accurate than the two state-of-the-art baselines.

For future work, we envision the following improvements:

- Expand to other regions of California.
- Expand to roads other than highways as both causes and effects.
- Experiment with other traffic slowdown detectors.
- Use actual traffic incident data to find spontaneous slowdowns.
- Use cell phone GPS data instead of road sensor data, because road sensors have a significant amount of imputed data.
- Expand our mathematical algorithm to account for a causal graph with multiple causal edges and multiple signal nodes.

References

California Department of Transportation. 2023. "Caltrans Performance Measurement System." https://pems.dot.ca.gov/.

Caltrans. 2020. *PeMS Introduction and User Guide*. https://pems.dot.ca.gov/Papers/PeMS_Intro_User_Guide_v6.pdf.

Fei, Wen-peng, Guo-hua Song, Fan Zhang, Yong Gao, and Lei Yu. 2017. "Practical Approach to Determining Traffic Congestion Propagation Boundary Due to Traffic Incidents." *Journal of Central South University* 24 (2): 413–22.

Giuliano, Genevieve, Catherine Showalter, Quan Yuan, and Rui Zhang. 2017. "Managing the Impacts of Freight in California." CA17-3133. California Department of Transportation. https://dot.ca.gov/-/media/dot-media/programs/research-innovation-system-information/documents/final-reports/ca17-3133-finalreport1-a11y.pdf.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37 (3): 424–38.

McCarthy, Niall. 2020. "Chart: Congestion Costs u.s. Cities Billions Every Year." https://www.statista.com/chart/21085/annual-economic-losses-from-traffic-congestion/.

Microsoft. 2024. "Bing Maps API." https://www.bingmapsportal.com/.

Pamfil, Roxana, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. 2020. "Dynotears: Structure Learning from Time-Series Data." In *International Conference on Artificial Intelligence and Statistics*, 1595–605. Pmlr.

Runge, Jakob. 2020. "Discovering Contemporaneous and Lagged Causal Relations in Autocorrelated Nonlinear Time Series Datasets." In *Conference on Uncertainty in Artificial Intelligence*, 1388–97. Pmlr.

Sager, Monica. 2024. "You Wasted $771 Sitting in Traffic Last Year." https://www.newsweek.com/traffic-study-wasted-money-economy-congestion-2012671.

Zheng, Xun, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. "DAGs with No Tears: Continuous Optimization for Structure Learning." *Advances in Neural Information Processing Systems* 31.

Data Management Plan

All our data came from the Caltrans Performance Measurement System (PeMS, https://pems.dot.ca.gov/) which is publicly available for free data downloads. We used data from California District 7 for the first six months of 2024.