Causal Root-Cause Analysis for Multivariate Process Data

George Weale
University of California, Santa Barbara
Santa Barbara, CA, USA
gweale@ucsb.edu

***Abstract*—Industrial process monitoring detects departures from normal operation, but a high anomaly score does not identify the initiating disturbance. This paper develops a causal root-cause analysis framework for multivariate process data with feedback control, autocorrelation, missingness, and changing operating regimes. The framework combines a process-informed causal graph, regime-aware time-series preprocessing, lagged structural causal models, intervention-aware benchmarks, and conservative diagnosis reporting. It distinguishes statistical association, model-supported causal attribution, and verified physical cause, three evidence categories that are often conflated. Known interventions in dynamic benchmarks support quantitative evaluation, while identifiability conditions bound attribution from observational records. The output is a ranked, uncertainty-aware diagnostic report that exposes causal paths, assumptions, alternative candidates, and abstention conditions for engineering review.**

***Index Terms*—process monitoring, causal discovery, fault diagnosis, time series, root cause analysis, uncertainty quantification**

I. Diagnostic objective

The target setting is a continuous or batch process with measurements, manipulated variables, quality variables, alarms, and a partial process flowsheet. Under declared graph, temporal, and identifiability assumptions, the method ranks plausible initiating disturbances rather than merely listing correlated signals. It compares the causal, time-aware ranking with principal-component, correlation, and prediction-only baselines, while reserving physical verification and safety decisions for qualified engineering review.

II. Why correlation is not diagnosis

Let $X_1, \ldots, X_p$ be measured process tags. A conventional detector may score an observation using Hotelling's T^2, squared prediction error, a reconstruction residual, or a forecasting residual. These scores can identify that a normal-data model no longer fits. They cannot, by themselves, distinguish the following: a valve disturbance causing a composition change; a composition change causing a controller action; or a common feed disturbance causing both. In a feedback-controlled process, the manipulated variable can be strongly associated with a quality shift even when it is a compensatory response.

The framework uses three evidence labels:

1) *Associated variable*: statistically co-moving with the alarmed outcome after stated preprocessing.
2) *Model-supported candidate cause*: an upstream variable ranked by a causal model under declared graph and identifiability assumptions.
3) *Verified physical cause*: a candidate independently confirmed by a controlled intervention, maintenance record, or other external evidence.

Only the first two can generally be inferred from observational time series; the third requires evidence outside the model.

III. Related work and motivation

Industrial fault-detection research distinguishes model-based, qualitative, and process-history approaches, each of which is useful for detecting a departure but limited in the causal interpretation of correlated tags [1], [2], [3]. PCA residual monitoring remains a valuable baseline because it is simple, inspectable, and often effective for broad multivariate shifts [4]. Change-detection theory adds a precise language for alarm delay and false alarm control [10]; modern process-data analytics places those tools in a larger workflow that includes contextual variables, data quality, and operator decision support [11].

Causal-inference literature provides a different target: an intervention effect, not merely a predictive association [5], [7], [6]. In time series, lag structure can provide leverage, but it does not eliminate confounding, feedback, or regime dependence. Methods such as PCMCI are useful candidates for large nonlinear time series because they make conditioning choices explicit [8]; nonlinear additive-noise models provide another possible orientation mechanism under restrictive assumptions [9]. In a process setting, however, a learned graph should be constrained by known material flow, instrumentation location, and controller direction. Otherwise it can convert a compensatory actuator movement into an implausible cause.

The motivation for this paper is operational rather than philosophical. Alarm floods and correlation rankings can make a real disturbance harder to locate when a plant is under feedback control. A conservative graph-and-evidence report improves the structure of a human investigation by marking what is identifiable, what depends on model assumptions, and what remains unresolved. This focus is aligned with fault-diagnosis practice, in which process knowledge and verification remain essential [12].

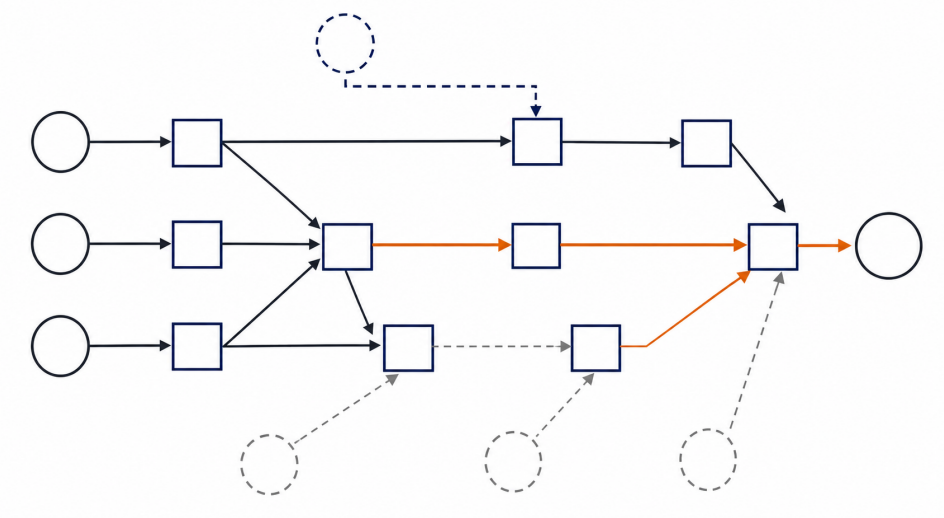

Fig. 1. Process-informed causal representation for diagnostic ranking. Candidate upstream paths, intervention links, and ambiguous latent pathways are carried into the evidence and uncertainty report.

IV. Evidence Model and Diagnostic Boundary

The method separates association, model-supported candidate cause, and externally verified cause in both its data schema and reporting language. Time-indexed diagnosis supports investigation without authorizing autonomous control action or replacing hazard analysis. Transparent dynamic benchmarks with known injected disturbances, including the Tennessee Eastman challenge process [13], establish the evaluation basis before permissioned industrial records are considered. Fixed graph constraints, time splits, intervention manifests, abstention rules, and root-cause metrics make failures of causal orientation as visible as successful rankings.

V. Process Representation

A. Dynamic structural causal model

For process tag X_i at discrete time t, the structural equation is

$$X_i(t) = f_i\big(\mathrm{Pa}_i^{(0)}(t), \mathrm{Pa}_i^{(1)}(t-1), \ldots, \mathrm{Pa}_i^{(L)}(t-L), U_i(t)\big), \tag{1}$$

where $\mathrm{Pa}_i^{(\ell)}$ denotes measured parents at lag ℓ, U_i represents exogenous disturbances, and L is a maximum physical or control-relevant lag. Directed edges must be constrained by a process diagram when available: mass and energy flow, known controller signal directions, and measurement location can rule out graph edges that pure data fitting would otherwise propose.

The graph includes process states S, measurements M, manipulated variables A, disturbance candidates D, and quality variables Y. Measurement bias and drift are represented separately from process disturbances where possible:

$$S(t+1) = F(S(t), A(t), D(t)) + \varepsilon_S(t), \tag{2}$$

$$M(t) = H(S(t)) + b(t) + \varepsilon_M(t), \tag{3}$$

$$A(t) = \pi(M(t), r(t)) + \varepsilon_A(t). \tag{4}$$

This separation prevents the model from automatically treating an actuator response as the initiating event.

B. Counterfactual diagnostic target

For an anomalous observation $\boldsymbol{x}_{t_0:t_1}$, candidate D_k is assessed through the contrast

$$\Delta_k = \mathbb{E}\left[Y_{t_1} \mid \mathrm{do}(D_k = d_k^{\mathrm{anom}}), \mathcal{H}_{t_0}\right] - \mathbb{E}\left[Y_{t_1} \mid \mathrm{do}(D_k = d_k^{\mathrm{base}}), \mathcal{H}_{t_0}\right], \tag{5}$$

where $\mathcal{H}_{t_0}$ is a permitted history. In deployment, the do-quantity is not directly observed; it is estimated only when the graph and adjustment conditions support identification. If these conditions fail, the system must report that causal ranking is not identified rather than produce a false-precision score.

VI. Data Protocol

A. Data provenance and segmentation

Every tag shall have a data dictionary including units, sensor type, sampling interval, physical location, control-loop role, calibration history if available, and access restrictions. The record is segmented before causal modeling into operating regimes defined by production grade, throughput band, equipment configuration, and major setpoint changes. Combining incompatible regimes can create spurious edges through Simpson's paradox or changing controller policies.

The preprocessing order is fixed in advance: timestamp alignment, unit validation, engineering range flags, missingness labeling, regime assignment, and only then model-specific normalization. Imputation must be causal in time: using future observations to fill a past gap is prohibited in an online early-warning analysis. Missingness itself may be informative; a sensor-dropout channel is retained when plausibly related to process state.

B. Avoiding leakage

Training, validation, and test sets are split in contiguous time blocks or by complete fault episodes. Random row-level splits are not allowed because autocorrelation would leak near-duplicate observations across sets. Feature construction must use only information available at the claimed decision time. Maintenance annotations or final laboratory assays may be used only in a post-hoc verification track, never as inputs to a real-time diagnosis model unless they would actually be available online.

VII. Causal diagnosis method

A. Regime-aware residualization

Known setpoints and planned operating variables are first modeled to remove predictable regime effects. For a tag X_i,

$$R_i(t) = X_i(t) - \widehat{m}_i\left(r(t), q(t), X_i(t-1:t-L)\right), \qquad (6)$$

where r represents setpoints and q represents known production context. Residualization is a diagnostic aid, not proof that confounding has been removed. The model must retain the original signals for physical interpretation.

B. Graph learning with hard and soft knowledge

Candidate graphs are generated by a time-series causal-discovery procedure over residuals, while a whitelist and blacklist encode physical and control constraints. Hard constraints include impossible upstream directions; soft constraints penalize but do not forbid uncertain relations. Stability selection resamples time blocks and reports an edge-selection frequency $\hat{p}_{ij}^{(\ell)}$. An edge is not interpreted as stable unless its direction, lag, and sign are consistent across specified resamples and regimes.

For each candidate graph, the study estimates either linear vector autoregressive structural equations or nonlinear additive-noise models. The latter are considered only when diagnostics demonstrate sufficient data support. Bootstrap intervals must reflect both parameter uncertainty and graph-selection uncertainty; reporting a confidence interval conditional on a single selected graph is insufficient.

C. Root-cause ranking

At an alarm time, each candidate source is scored by a transparent composite:

$$\mathrm{Score}_k = w_1\widetilde{\Delta}_k + w_2\widetilde{L}_k + w_3\widetilde{C}_k - w_4\widetilde{U}_k, \qquad (7)$$

where $\widetilde{\Delta}_k$ is the estimated counterfactual impact, $\widetilde{L}_k$ is temporal precedence consistency, $\widetilde{C}_k$ is consistency with allowed process paths, and $\widetilde{U}_k$ is uncertainty or instability. Weights are prespecified or calibrated only on training episodes. The interface must expose the path, lag, confidence or stability measure, alternative candidates, and the assumptions required for the ranking.

TABLE I
Evaluation targets and metrics for diagnostic performance.

Evaluation target	Metric
Was an anomaly detected?	Episode recall, false alarms per operating hour, detection delay.
Was the source ranked highly?	Top-k root-cause accuracy, mean reciprocal rank, path precision.
Was uncertainty useful?	Calibration error, coverage of intervals, selective-risk curve when abstaining.
Did the model survive a regime change?	Performance by held-out regime and degradation relative to in-regime test.
Could an operator audit it?	Fraction of reports with valid path, lag, and evidence fields; expert review protocol.

VIII. Benchmark and evaluation design

A. Synthetic and semi-synthetic testbed

The first testbed is a published process simulator or a transparent mass-energy-balance model with injected disturbances. Examples include feed-composition shifts, heat-transfer degradation, sensor bias, valve stiction, controller retuning, and unmeasured disturbances. The intervention time, location, magnitude, and downstream propagation are recorded as ground truth. A semi-synthetic track adds controlled disturbances to normal public or permissioned archived data while preserving the original observations.

B. Baselines and ablations

Baselines include PCA/T^2 and squared-prediction-error monitoring, dynamic PCA, forecasting-residual ranking, Granger-style lag association, and a process-graph heuristic without learned effects. Ablations remove (i) regime segmentation, (ii) controller nodes, (iii) physical graph constraints, (iv) uncertainty penalty, and (v) time lags. The causal method is not considered supported if it cannot outperform or at least explain its differences from these simpler approaches on known interventions.

IX. Human-auditable diagnostic report

The system output is not a single alarm label. It is a compact case report designed for an engineer who can inspect the process context. Every candidate-source ranking identifies the alarmed variable, relevant regime, candidate path, temporal ordering, uncertainty category, and evidence category. This structure distinguishes "the model found a stable predecessor" from "the model established a physical cause." It also creates a useful failure mode: if a required field cannot be populated, the system reports an insufficient-evidence condition instead of filling the report with a plausible narrative.

This reporting layer is a substantive part of the method because root-cause language can otherwise be overread. A high-ranked candidate can guide inspection, but it cannot validate a causal mechanism when key disturbances are unmeasured. The study should therefore evaluate not only ranking accuracy but also selective accuracy after the model abstains. A model that declines ambiguous episodes while producing better-supported

TABLE II
REQUIRED FIELDS IN THE HUMAN-AUDITABLE DIAGNOSIS REPORT.

Field	Function
Episode context	Time window, operating regime, data-quality flags, and alarm definition.
Candidate path	Ordered tags, allowed directions, lag range, and controller nodes on the path.
Evidence category	Association, model-supported candidate, externally verified, or abstention.
Uncertainty	Graph stability, effect interval where identified, and alternative candidates.
Verification prompt	Physical check, maintenance record, or safe test that could confirm or refute the ranking.

reports on the remainder may be operationally more useful than a model forced to name one cause for every alarm.

X. Assumptions, threats, and decision rules

Observational causal discovery relies on strong assumptions, including causal sufficiency or explicit latent-confounder handling, correct temporal resolution, adequate excitation, and a graph class compatible with the process. Fast unmeasured dynamics can reverse apparent lag order; feedback can make contemporaneous direction unidentifiable; sensor drift can masquerade as a disturbance. Therefore, the reporting rule is conservative:

- report "association" where graph orientation is unsupported;
- report "candidate source under model assumptions" where a stable, process-consistent path exists;
- reserve "root cause" for externally verified episodes.

The model must be allowed to abstain when top candidates are too close, graph stability is low, or the observation falls outside the training support.

XI. Analysis specification and verification

The main analytic choices are fixed before root-cause accuracy is inspected: tag inclusion rules, regime definition, maximum lag, graph whitelist and blacklist, data split, resampling scheme, and abstention threshold. A causal analysis is particularly vulnerable to unreported flexibility because a researcher can obtain a plausible graph by changing lags, conditioning variables, or exclusions after seeing a fault episode. A signed configuration file and changelog record every deviation and its reason.

Verification is separated from model fitting. For simulated episodes, the intervention manifest is hidden from the graph-learning process until scoring. For permissioned recorded episodes, an independent reviewer examines maintenance records, operator notes, or laboratory evidence after the model produces its ranked report. The reviewer can mark an episode as unresolved rather than forcing a binary true-or-false label. This distinction protects against circular verification in which the same alarms used to create a candidate path are later presented as confirmation.

The credible output is a set of episode-level reports with an audit trail, not a global causal diagram treated as permanently true. When the process configuration changes, the study should require a new regime assessment and potentially a new graph. This is a limitation but also an engineering benefit: the method makes its assumptions visible at the point where an operator needs to decide whether a recommendation can be trusted.

XII. Operator workflow and change management

The workflow begins with an alarm or quality deviation, but it does not begin causal discovery from scratch. A versioned process map, tag dictionary, controller inventory, and regime definition are assembled before deployment. At the episode boundary, the system freezes the available data window, records completeness and quality flags, and produces a ranked report. A qualified reviewer then chooses whether to inspect a physical pathway, consult a maintenance record, request a safe test, or mark the episode unresolved. The review action and its evidence are appended to the episode record rather than overwritten into the original model output.

Model updates require their own change-management process. A sensor replacement, control retuning, new feedstock, or equipment configuration can invalidate dependencies learned from earlier records. The workflow therefore uses a change trigger that pauses causal ranking or lowers its confidence until the affected regime has been requalified. This is more conservative than continuous self-updating, but it makes the model's operating support clear. A system that silently learns from an unverified upset can convert a one-time failure into a persistent diagnostic bias.

The method is decision support only. It should not open or close a valve, change a setpoint, or generate a safety conclusion without separate control, authorization, and hazard-review processes. The operational value is instead in shortening the path from a broad alarm pattern to a traceable set of inspection hypotheses. That value can be measured through report completeness, abstention quality, and agreement with later verification, rather than through an unsupported claim of autonomous root-cause resolution.

XIII. Reproducibility and governance

The reproducibility package should contain a de-identified tag map or synthetic substitute, data-split manifest, graph constraints, all preprocessing code, simulator intervention manifest, seeds, and calibration plots. If industrial data cannot be released, a synthetic counterpart with the same causal motifs should be provided. The protocol should not recommend autonomous control changes; diagnosis outputs are decision support that require qualified review.

XIV. Data stewardship and claim governance

Process records can contain commercially sensitive tags, operator notes, and maintenance information. The analysis uses a minimum-necessary data extract, role-based access, and a disclosure review before any artifact is shared. A releaseable synthetic counterpart is not a substitute for access controls, but it can preserve the graph motifs, feedback patterns, missingness mechanisms, and intervention timing needed to test

the method. The synthetic-data generator, parameter ranges, and differences from the protected record are documented so readers do not mistake it for a plant replica.

Each diagnosis report should receive a stable episode identifier and a model-version identifier. Subsequent verification must be appended as a new evidence record; it must not overwrite the model's original ranking, path, or abstention decision. This preserves a fair audit trail when an investigator later discovers a maintenance event, a previously unavailable laboratory result, or an unmeasured disturbance. It also makes it possible to evaluate whether the system's uncertainty statements were honest at the time of decision rather than merely consistent with information obtained later.

The claim register should enforce the evidence vocabulary defined in this paper. A report can be labeled associated, model-supported, verified, or unresolved, but not "root cause" merely because a graph has an arrow. Any publication-facing summary should link back to the supporting report and explicitly list assumptions that, if changed, would alter the conclusion. This governance layer is a technical requirement for cautious causal inference, not a substitute for it.

XV. Conclusion

This paper establishes a disciplined path from multivariate anomaly detection to evidence-calibrated causal diagnosis. Process-informed graph constraints, regime-aware time-series models, intervention benchmarks, uncertainty-aware ranking, and explicit abstention rules connect statistical detection to an auditable engineering report. The framework exposes when causal attribution is identifiable, when only association is supported, and when the system should defer to additional physical evidence.

References

[1] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis. Part I: Quantitative model-based methods," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 293–311, 2003.

[2] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis. Part II: Qualitative models and search strategies," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 313–326, 2003.

[3] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis. Part III: Process history based methods," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 327–346, 2003.

[4] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.

[5] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[6] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[7] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference*. Cambridge, MA, USA: MIT Press, 2017.

[8] J. Runge *et al.*, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Sci. Adv.*, vol. 5, no. 11, eaat4996, 2019.

[9] P. O. Hoyer, D. M. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Proc. NeurIPS*, 2009, pp. 689–696.

[10] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[11] S. J. Qin, "Process data analytics in the era of big data," *AIChE J.*, vol. 60, no. 9, pp. 3092–3100, 2014.

[12] L. H. Chiang, E. L. Russell, and R. D. Braatz, *Fault Detection and Diagnosis in Industrial Systems*. London, U.K.: Springer, 2001.

[13] N. L. Ricker, "Decentralized control of the Tennessee Eastman challenge process," *J. Process Control*, vol. 6, no. 4, pp. 205–221, 1996.