CGPA WORKING PAPER
2026-19

Statistical Measurement of Bilateral Trade Restrictions: A Monthly Index Based on Rolling-Baseline SPE Decomposition

Cheng Zhenghao



DREXEL UNIVERSITY
Center for
Global Policy Analysis
LeBow College of Business

The CGPA working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the CGPA Board. The CGPA does not hold any responsibility for the content and correctness of the papers in this series.

Statistical Measurement of Bilateral Trade Restrictions: A Monthly Index Based on Rolling-Baseline SPE Decomposition

Abstract: Export controls and economic sanctions have become increasingly frequent and covert, yet existing measurements either record de jure measures without observing enforcement, or presume known events at annual frequency. This paper transplants the monitoring paradigm of statistical process control (SPC) onto bilateral trade data and constructs the Rolling-Baseline Trade Suppression (RBTS) index, which relies solely on monthly customs records. A rolling baseline with selective updating identifies the in-control regime; idiosyncratic residuals from factor decomposition are calibrated and smoothed, yielding restriction readings in σ units that are comparable across commodity-partner pairs and over time. Two classified case studies and a blind test covering 120 sanction pairs on the full U.S. export matrix show that explicitly restricted pairs respond at the event or announcement month, pairs listed on paper but not actually restricted stay flat, and covert restrictions without formal texts are quantified; in the blind test, readings rise strictly monotonically with the ex-ante verified severity of measures (median shares of above-threshold cells 0.58%, 3.26% and 5.47% across three severity tiers). Simulations delineate the effective working zone of the index, with a minimum detectable effect of about 50% sustained suppression. The index complements list-based measures and provides a new statistical tool for monitoring trade restrictions.

Keywords: trade restriction index; statistical process control; principal component residuals; export controls; mirror trade data

1. Introduction

Export controls and economic sanctions have become routine instruments of international economic relations. Since 2018 alone, the United States has reimposed sanctions on Iran and repeatedly tightened controls on advanced computing and semiconductors destined for China; Japan has placed key materials bound for Korea under licensing administration; and Saudi Arabia has mounted a boycott of Turkey without any formal text. Restrictive measures have grown increasingly diverse, covert, and window-guidance-based. Relative to this rapid policy evolution, the statistical measurement of trade restrictions lags behind: list- and text-based approaches record “de jure controls” but cannot observe enforcement; structural-gravity and quasi-experimental studies presume known event timing and treated objects, mostly at annual frequency; price-based approaches and trade restrictiveness indices

rely on structural parameters such as demand elasticities and remain at the "country × sector × year" granularity. Researchers lack a descriptive statistical tool that can answer: how strongly is country A's export of commodity k to country B actually restricted in month t?

This paper transplants the SPC paradigm of "estimating parameters in an in-control period and monitoring deviations period by period" onto bilateral trade matrices, and constructs the Rolling-Baseline Trade Suppression (RBTS) index. The index takes monthly records from UN Comtrade as its only input and requires no policy-text prior: for each month, the preceding 48 months form a candidate window; each row drops the months in which that row breached its limits before parameters are estimated (row-wise selective updating); the factor subspace is estimated only on low-contamination months within the window; the current vector is then decomposed into a common-factor projection and an idiosyncratic residual. The row-wise components of the residual, after empirical-dispersion calibration, one-sidedness, and exponentially weighted moving average (EWMA) smoothing, yield a restriction reading in σ units that measures how rare the deviation is relative to the row's own historical fluctuation.

Validation proceeds on two levels. At the case level, controls are designed on the four combinations of "formal text × actual restriction": in the overt case, hydrogen fluoride and photoresist—both on Japan's 2019 control list—exhibit opposite reading patterns, the former crossing the reference line in the month after the event with a peak of 9.24σ, the latter staying flat throughout (peak 1.21σ), showing that the index measures actual restriction rather than paper listing; in the covert case, Saudi Arabia's textless boycott of Turkey is quantified, with peak readings of 9.65–9.98σ across three chapters that do not collapse after the COVID window is removed. At the system level, a blind test on the full U.S. export matrix (about 1.26 million monthly records) covering all partners and all chapters takes 120 sanction pairs mechanically generated from the Global Sanctions Data Base (GSDB), with criteria frozen before computation; readings rise strictly monotonically with the ex-ante case-by-case verified severity of measures, and the index further uncovers verifiable event footprints absent from the database. Simulations and property analysis delineate the effective working zone: the minimum detectable effect is about 50% sustained suppression, and the nominal probability of the reference line is diluted on strongly autocorrelated series—these boundaries are given together with the reading rules.

The marginal contributions are threefold. Methodologically, the rolling baseline with row-wise selective updating removes the core obstacle in transplanting SPC—the in-control phase is unknown and the timing of controls unpredictable; ablation shows that this design strengthens the readings of persistent restrictions by a factor of 3.5–5.3 (chapter-by-chapter ratios, throughout). Retaining the row-wise components of the SPE rather than compressing them into a single scalar localizes monitoring to specific "commodity × partner" cells. In terms of data, the index relies solely on publicly available monthly trade flows, with frequency and granularity ("commodity × partner × month") superior to existing indices, and works in both value and quantity terms. In application, the index provides an independent cross-validation channel for list-based measures, and its ability to identify covert

restrictions and to preserve the ex-ante severity ranking offers a quantitative screening tool for policy research.

2. Literature Review

The quantitative measurement of export controls and sanctions spans international trade, political economy, and statistical monitoring. We review five strands and identify the gap they leave.

2.1 List- and text-based measures

The most systematic efforts begin with structured compilations of sanction and control lists. Hufbauer et al. (2007) built a database of economic sanction episodes covering the twentieth century; Felbermayr et al.'s (2020) Global Sanctions Data Base (GSDB) extends coverage to bilateral and multilateral sanctions since 1950, graded by type, objective, and severity. For export controls proper, Stinnett et al. (2011) construct a multi-indicator composite index of national nonproliferation export-control regimes. In the Chinese-language literature, Wei and Zhang (2018) use the G-L and Aquino indices to confirm the presence of U.S. high-tech export controls against China and their effect on bilateral imbalance; Liu and Li (2023) compile Entity List and Commerce Control List data to study the dual effect of U.S. export controls on Chinese firm innovation. List-based measures provide comparable records of "paper controls," but suffer three structural gaps: coverage depends on public texts, so unannounced covert restrictions are systematically missed; lists lag policy changes; and paper listing does not equal actual enforcement—the gap between de jure measures and de facto restriction is well known.

2.2 Causal estimates of trade effects

A second strand estimates the trade effects of sanctions and controls. With structural gravity, Larch et al. (2024) find that economic sanctions significantly reduce agricultural trade with sizable trade diversion; Jena et al. (2024) use the GSDB to estimate bilateral suppression and deflection toward third parties. For specific episodes, Makioka and Zhang (2024) find that Japan's export controls in the 2019 Japan–Korea dispute sharply reduced related exports to Korea; Hayakawa et al. (2023), using firm-level data, find that dual-use controls significantly shrank the number of exporting firms. This literature answers "how large the average effect of a control is"; identification requires prior knowledge of event timing and the treated set, and frequency is mostly annual—causal inference conditional on controls being "known," rather than discovering and measuring restriction from the data itself.

2.3 Price-based measures and trade restrictiveness indices

A third strand converts non-tariff measures into price equivalents. Cadot and Gourdon (2016) estimate ad valorem equivalents of non-tariff measures from price gaps; Nordås (2016) measures the trade effects of regulatory differences based on the OECD Services Trade Restrictiveness Index; Chen

et al. (2014) construct a generalized measure of China's trade restrictiveness accounting for processing trade. Price-based measures are economically interpretable and addable across instruments, but depend on price data and structural parameters such as demand elasticities, and their granularity typically stays at "country × sector × year," making it hard to trace restriction intensity at "exporter × importer × commodity × month."

The division of labor with existing trade restrictiveness indices should be clarified: the OTRI of Kee et al. (2009) and the generalized measure of Chen et al. (2014) are structural measures that convert tariffs and non-tariff measures into ad valorem equivalents using import demand elasticities, at annual frequency and "country × sector" granularity; Nordås (2016) is a de jure coding of regulatory texts. RBTS complements rather than substitutes for these indices: they measure paper measures or institutional differences, whereas this index measures de facto restriction in the flow, at "exporter × importer × commodity × month" granularity—the gap between paper listing and actual suppression is itself a measure of enforcement intensity.

2.4 Statistical process control: the methodological basis

Since Shewhart (1931), SPC has formed a mature monitoring paradigm: estimate parameters from in-control data, set control limits, and compute monitoring statistics period by period (Page, 1954; Roberts, 1959; Lucas and Saccucci, 1990). In the multivariate direction, MacGregor and Kourti (1995) established the PCA-based monitoring framework in which out-of-subspace variation is monitored by residual statistics; the squared prediction error (SPE) of Jackson and Mudholkar (1979), which measures an observation's deviation from the principal-component subspace, is the direct ancestor of our main statistic; Qin (2003) surveys such data-driven monitoring methods. Transplants of SPC into economics also exist: Leisch et al. (2000) apply CUSUM-type statistics to online monitoring of regression models; Groen et al. (2013) remove cross-sectional correlation before multivariate fluctuation monitoring, close to our "factor first, idiosyncratic residual second" design; Aue and Kirch (2024) review cumulative-sum sequential changepoint tests. Common-factor extraction rests on the large-dimensional factor-model literature (Stock and Watson, 2002; Bai and Ng, 2002); Efron's (2004) empirical null provides the basis for our calibration step. Note that standard SPC targets controlled industrial processes, and bilateral trade matrices are ill-suited in three ways: rows are driven by common cyclical factors; once a control shock enters the baseline sample it contaminates parameter estimates; and the null distribution of monthly flows need not be theoretically normal. These three issues are what the methodology section must handle.

2.5 Mirror trade data and this paper's position

Some cases below rely on mirror data reported by partner countries, a practice with established applications in trade-data quality and informal-trade research (Carrère and Grigoriou, 2015). In sum, the gaps of the five strands intersect: list-based measures cannot observe enforcement, causal estimates

presume known controls, price-based measures lack granularity, and the SPC paradigm has not been used to measure trade restriction. This paper combines the SPC monitoring paradigm with large-dimensional factor models on the "exporter × importer × commodity × month" trade matrix, constructing a descriptive trade-restriction index that relies solely on bilateral trade data and requires no policy-text prior, thereby filling the gap.

3. Index Construction

3.1 Unit of analysis and the rolling baseline

The unit of analysis is the bilateral trade flow at "reporter × partner × commodity × month." For a fixed reporter (the exporter or its importer's mirror), customs monthly records are organized into a matrix: rows are "commodity × partner" cells, columns are months, and entries are trade value or quantity—value terms are used throughout; quantity terms serve as a robustness check and are advisable for commodities with violent price or unit-value swings (Case 1 reports both; see Section 4). Commodity granularity is the HS six-digit level, aggregated to chapters where necessary. Preprocessing rules: stretches without records are treated as zero trade; series with reported months fewer than half of the sample span are dropped; sporadic internal gaps are linearly interpolated; no share normalization or demand-elasticity adjustment is applied, to keep construction transparent.

SPC presumes parameters estimated from an in-control phase, but the timing of controls cannot be known in advance. Our solution is a rolling baseline with selective updating: for each month t, the preceding 48 months form a candidate window, and each row retains only the months whose standardized residual is below 3 in absolute value (the standardized residual, defined in (6), is available recursively month by month), forming the row's estimation set:

$$C_{it} = \{j \in [t-48, t-1] : |z_{ij}| \leq 3\} \quad (1)$$

$z_{ij}$is the standardized residual of row $i$in month j; if the estimation set has fewer than 24 months, the full window is used (warm-up at the start of the sample). The column set of the common-factor subspace carries a further constraint: only months in which no more than 1% of rows breach limits are used, and if fewer than 24 such months exist, the 24 months with the lowest breach shares are taken. The row-wise rule improves on a matrix-level rule (disabling an entire month whenever any row breaches): the latter would exhaust clean months during broad-spectrum episodes such as trade wars and force estimation back into contaminated windows (Appendix B); row-wise removal keeps sufficient clean months for over 99% of rows. Moving-window PCA monitoring has prototypes in the process-monitoring literature (Wang et al., 2005; Li et al., 2000), but unconditional updating adaptively absorbs persistent shifts into the baseline; controls are precisely the persistent shifts we seek to detect, and selective updating closes this absorption channel—its necessity is shown by the ablation in Appendix B. The updating criterion of 3 and the reading reference line of 2 are not double standards: the former acts on unsmoothed single-month standardized residuals and is deliberately loose so as not to

mistakenly remove borderline normal months; the latter acts on the EWMA-smoothed statistic. The two are different statistics.

3.2 Index construction

On the estimation set, the index is computed month by month through the following chain:

$$x_{it} = ln(1 + y_{it}) \tag{2}$$

$$\tilde{x}_{it} = \frac{(x_{it} - \mu_i(t))}{\sigma_i(t)} \tag{3}$$

$$K_t = min\left\{max\left\{k : \frac{\Sigma_{j \le k} s_j^2}{\Sigma_j s_j^2} \ge 0.8, 1\right\}, 3\right\} \tag{4}$$

$$\hat{x}_t = U_K(t) U_K(t)' \tilde{x}_t, \quad e_t = \tilde{x}_t - \hat{x}_t \tag{5}$$

$$z_{it} = \frac{e_{it}}{s_i^e(t)} \tag{6}$$

$$\tilde{z}_{it} = max\{ -10, min[\frac{z_{it}}{\sigma_0}, 10] \} \tag{7}$$

$$D_{it} = max(0, -\tilde{z}_{it}) \tag{8}$$

$$D_{it} = \lambda D_{it} + (1 - \lambda) D_{i,t-1} \tag{9}$$

$y_{it}$is the raw flow of row i(a commodity × partner cell) in month t. Step (2) applies the log transform to compress magnitude differences and accommodate zeros; (3) row-standardizes with the row's estimation-set mean μ_i(t) and standard deviation σ_i(t), making cells of different scales comparable; (4) applies singular value decomposition to the standardized window matrix, where $s_j$is the j-th singular value, and the leading $K_t$directions reaching 80% cumulative variance form the common-factor subspace (Bai and Ng, 2002)—the cap of 3 keeps the residual subspace rich, the 80% threshold follows factor-model convention, and Appendix B3 perturbs it to 70% and 90% (i.e., perturbs the selection rule of K_t), with key findings rank-stable in that neighborhood; in (5), U_K(t) collects the first $K_t$left singular vectors, and the current standardized vector $\tilde{x}_t$is decomposed into the common-factor projection $\hat{x}_t$and the idiosyncratic residual e_t—this is the row-wise expansion of the squared prediction error (SPE) of Jackson and Mudholkar (1979): whereas SPE compresses deviation into a single alarm scalar, we retain the row components to localize to specific "commodity × partner" cells; (6) standardizes by the row standard deviation of estimation-set residuals s_i^e(t), and the resulting $z_{it}$is the selective-updating criterion in (1); (7) is empirical calibration and winsorization: the theoretical null N(0,1) does not hold on monthly trade data, and a right-shoulder half-sample estimate (Efron, 2004) on pooled $z$from training cases gives an empirical dispersion σ_0 = 1.22 (pool composition and leave-one-out sensitivity in Appendix B6), while the fitted degrees of freedom of a t-distribution null degenerate to the boundary, indicating that heavy tails are a property of the signal rather than the noise; winsorization at ±10 bounds single-month influence; (8) exploits the one-sidedness of restriction and accumulates only negative deviations; (9) applies EWMA smoothing (λ = 0.3; Roberts, 1959; Lucas and Saccucci, 1990) with recursion initialized at 0, suppressing single-month spikes and amplifying persistent suppression. Among these parameters only $\sigma_0$is calibrated from data; the rest take conventional design values in the

SPC literature, with sensitivity covered in Appendix B3.

The σ-unit $\bar{D}$measures how rare a deviation is relative to the row's own historical fluctuation: once cross-row volatility differences are removed, readings are directly comparable across cells, cases, and periods. If a percentage-magnitude reading is desired, the counterfactual gap $g = 1-\exp(e)$ can be converted from the log-scale deviation and read directly as a suppression percentage for large-flow cells (error smaller than the reciprocal of the counterfactual flow); the magnitude reading is used only for the dimension-fidelity reconciliation in Appendix B, and the main text reads in σ.

3.3 Reading rules, conditions of use, and boundary statements

The descriptive positioning of the index should be clear: RBTS gives a continuous restriction reading for each "commodity × partner × month" and does not output a binary verdict of "restricted or not"; interpretation is left to the user with external information. Solely for validation and cross-case comparison, we take 2σ as an empirical reference line: it is a built-in product specification of the index, not the result of ex post tuning—in null simulations the share of cells above the reference line is below 0.5% (Appendix A); in real matrices, no-event rows occasionally cross (1.40% of no-event cells in the U.S. matrix; the 18 no-event control rows in the Case-2 matrix have a median peak of 1.00σ and a maximum of 5.57σ), so crossings must be read jointly by persistence and intensity, and the readings of crossing cells can be traced one by one.

Reliable use of the index depends on the following conditions, of which (i)–(iv) follow from the two-phase logic of SPC (Chakraborti et al., 2008) and (v)–(vii) are calibrated by simulation and empirics:

(i) A warm-up period of no fewer than 24 months (48 recommended); series with insufficient estimation sets receive no reading.

(ii) The number of rows far exceeds the number of factors (with $K = 3$ the cumulative variance share should not fall below 70%; narrow matrices should widen commodity chapters or partner coverage).

(iii) Control rows are not exhausted: when most partners belong to a sanctioning coalition and are suppressed synchronously, suppression seeps into the common factor and the matrix composition should be changed.

(iv) A relatively stable baseline: coefficient of variation below 0.5.

(v) Sufficient flow scale: very small series lack the signal-to-noise ratio; the scale boundary calibrated in the blind test is about USD 100 million of annual trade.

(vi) Simultaneously suppressed rows should be a minority (sparsity): broad-spectrum shocks must instead be identified through the common-shock channel.

(vii) The reading does not distinguish "restriction" from partner-specific demand contraction: partner recessions and import-structure changes also fall into the idiosyncratic residual, so we recommend cross-checking with the partner's imports from the world.

Two structural assumptions: an approximate factor model in which macroeconomic and industry shocks are pervasive while control suppression is sparse (Stock and Watson, 2002); and one-sidedness—only negative suppression is measured, while positive changes from re-exports are outside

the scope.

Simulations give two quantitative boundaries (Appendix A). First, a depth boundary: the minimum detectable effect is about 50% sustained suppression—shallow suppression below 40% is essentially undetectable within a 24-month event window, while suppression deeper than 78% is stably detected; entity-list-level measures are undetectable precisely because the equivalent suppression depth of listed entities within chapter flows is extremely low—the granularity boundary and the depth boundary are two faces of the same boundary. Second, a correlation boundary: on strongly autocorrelated trade series (AR(1) coefficient above 0.6), the nominal probability of the 2σ reference line is diluted by about an order of magnitude, so interpretation should incorporate event-window length—a single-month crossing is not evidence; persistent crossing is.

3.4 Properties of the index

The construction implies six basic properties (sketch proofs in Appendix A). First, row-scale invariance: multiplying a flow by any positive constant does not change its reading (exact for large flows, asymptotic for small ones), so changing units or currencies has no effect. Second, bounded influence: the immediate contribution of any single month never exceeds 3.0σ (the product of winsorization and smoothing), so extreme values do not dominate. Third, geometric forgetting: a reading from *k*months ago decays at 0.7 to the power *k*, a half-life of about two months, so readings recede within months after a control is lifted. Fourth, bounded contamination: under selective updating an event month contributes zero to baseline estimation, and if it leaks in, its influence is bounded by the reciprocal of the estimation-set size (about 1/24 in the worst case), so the index does not adaptively dull during long episodes. Fifth, a decomposition identity: on the log scale the total deviation equals exactly the sum of the idiosyncratic residual and the common-factor component (an orthogonal-projection identity), the mathematical basis for separating “common shocks” from “bilateral suppression.” Sixth, factor consistency: as rows and window length grow together, the estimated factor space converges in probability to the true one (Bai and Ng, 2002), and our rolling implementation extends this result to the monitoring setting.

4. Case Studies: One Overt and One Covert

4.1 Case selection and data

The validation unit is the “commodity × partner” cell, classified on “formal text × actual restriction” into four types: overt-A (text, restricted) and overt-B (text, not restricted) test “responding when it should and staying quiet when it should not”; covert-A (no text, restricted) tests the identification of covert restrictions; covert-B (neither) tests that no-event cells stay silent. One typical overt and one typical covert case are taken, with type-B controls drawn from inside the same matrix—same pipeline, same period as the target cells, the strictest control. Table 1 summarizes the configuration.

Data are UN Comtrade monthly bilateral records: Case 1 uses Korea's mirror imports from Japan (value terms, with quantity terms as a check), and Case 2 uses Turkey's directly reported export values.

Table 1 Case configuration

Case	Event	Flow & unit	Type-A cells (listed & restricted)	Type-B cells (control)	Event date
1 (overt)	Japan tightens export controls on Korea	Korean imports from Japan (mirror, value)	281111 hydrogen fluoride (391190 fluorinated polyimide as further control)	370710 photoresist: same list, licenses granted as usual	2019-07
2 (covert)	Saudi informal boycott of Turkey	Turkish exports to Saudi Arabia (value)	Chapters 61, 62, 94 (apparel, furniture)	Ch.94 → Germany: same chapter, no event	2020-10

Note: cell types are defined on "formal text × actual restriction"; type-B controls share the matrix, pipeline, and period with type-A cells.

4.2 Overt case: Japan's export controls on Korea (July 2019)

In July 2019, Japan removed three key materials bound for Korea—hydrogen fluoride (HS281111), photoresist (HS370710), and fluorinated polyimide (HS391190)—from the fast-track licensing channel; the three stood side by side on the control list. Figure 1 shows that the reading of hydrogen fluoride → Korea crossed the 2σ reference line in the month after the event (August 2019), peaked at 9.24σ in November 2020, and stayed above the line for 22 consecutive months; from mid-2021 it receded as Korean domestic substitution advanced, consistent with the timeline of actual relaxation. Photoresist, on the same list, showed the opposite pattern: the control was executed through per-order licensing with approvals granted as usual, no persistent suppression appeared in the actual flow, and its reading stayed flat throughout with a peak of only 1.21σ. Fluorinated polyimide likewise stayed below the line in value terms (peak 1.64σ). Recomputing the same matrix in quantity terms as a robustness check: hydrogen fluoride peaks at 9.74σ (June 2020), photoresist at 1.08σ, and fluorinated polyimide at 3.18σ—the intermediate position of fluorinated polyimide in quantity terms matches public records that its enforcement was looser than hydrogen fluoride's; the "one crossing, one flat" conclusion is unchanged across the two units. One announcement, three materials, three readings—the index measures actual restriction rather than paper listing: listing is determined by texts, restriction by flows, and the gap between them is exactly where a descriptive indicator is useful.

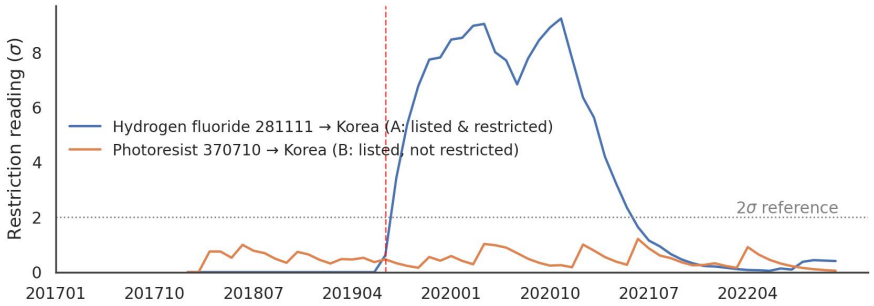


Figure 1 Case 1: restriction readings of hydrogen fluoride and photoresist → Korea (Korean imports from Japan, value terms)

4.3 Covert case: Saudi Arabia's informal boycott of Turkey (October 2020)

From October 2020, major Saudi retailers and chambers of commerce openly called for a boycott of Turkish goods and informal customs delays appeared, yet the Saudi government never issued a formal ban—list- and text-based measures have nothing to record here. Figure 2 shows that the readings of knitted apparel (Ch.61), woven apparel (Ch.62), and furniture (Ch.94) from Turkey to Saudi Arabia crossed the reference line successively between November and December 2020, peaking at 9.78σ, 9.65σ, and 9.98σ (close to the winsorization cap in (7)) and staying above the line for 24–27 consecutive months; same-chapter exports to Germany in the same matrix (no event background) peaked at only 0.33σ over the same period, flat against the floor. Among all no-event cells of the Case-2 matrix, the 18 control rows with at least 12 months of valid readings in the event window have a median peak of only 1.00σ and a maximum of 5.57σ (Section 3.3), and the target chapters' 9.65–9.98σ exceed their entire range. Readings receded below the line around late 2022 to early 2023, corresponding to the rapprochement period. The event window overlaps COVID-19; confounding is addressed in Section 5.4: after removing the pandemic window the three chapters still peak at 6.16–9.20σ—the readings do not collapse. Covert restrictions have no formal text and are precisely where flow-based measurement has its comparative advantage: no searchable control list exists, yet the index gives a quantitative restriction trajectory consistent with the event timeline.

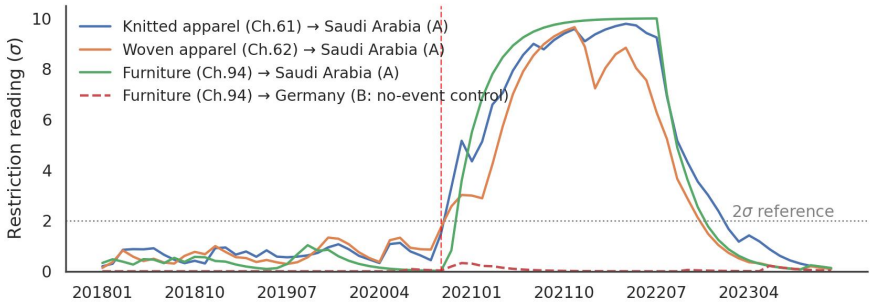


Figure 2 Case 2: restriction readings of three chapters under the Saudi boycott and a same-chapter no-event control (Turkish exports, value terms)

5. Systematic Validation: A Blind Test of 120 Sanction Pairs

The case-study cells were chosen by the researcher, leaving room for selection-bias concerns. This section reports a blind test with criteria frozen before computation: on the complete U.S. export matrix, we test whether readings rise monotonically with the ex-ante knowable severity of measures, and whether, within the same event, the readings of different commodities diverge according to actual suppression.

5.1 Design: data, labels, and criteria all fixed before computation

The data are the full U.S. export matrix: 120 months from January 2015 to December 2024, all partners × all HS chapters, about 1.26 million raw monthly records organized into 18,403 rows × 120 months. The reasons for choosing the U.S. matrix are unrelated to expected results: high reporting quality, full coverage, and a high density of sanction events. Labels come from the Global Sanctions

Data Base (GSDB version 5; Felbermayr et al., 2020; Kirilakha et al., 2021), mechanically filtered on the United States as sender, yielding 120 (target country × start-end) pairs—label generation does not depend on any reading. Criteria were frozen before computation: for each pair, the five chapters with the largest pre-event trade values, and an event window of 18 months after the event. The blind test aggregates to chapters for two reasons: six-digit monthly bilateral cells have too many zeros and insufficient signal-to-noise; and the GSDB records measure scope mostly by industry. Performance at six-digit granularity is shown at the case level (Case 1 is six-digit).

A crucial step is case-by-case verification: the GSDB records "policy events," many of which are individual-level designations (Global Magnitsky human-rights and anti-corruption designations, terrorism-financing network listings, etc.) that cannot possibly move chapter-level trade flows—counting such events in the evaluation denominator of a flow-based indicator is a caliber mismatch. We therefore coded the measure scope of every pair against policy texts and Treasury announcements before reading any readings; the GSDB's own trade-severity field cannot carry this ordering—it has almost no variation across the 120 pairs (only one full trade embargo among U.S.-initiated measures)—and the coding rules and the case-by-case coding table are published with the replication materials. Severity (how harsh the policy is) and measurability (how sensitive the instrument is) are two dimensions: insufficient scale (annual pre-event trade below USD 100 million), baselines already depressed by earlier rounds, and calibers finer than statistical chapters belong to the declared measurability boundaries (Section 3.3) and are reported separately as boundary evidence, outside the severity ordering.

5.2 Readings rise monotonically with ex-ante severity

All 93 measurable pairs (those with sufficient pre-event data) are assigned to three ex-ante severity tiers without any exclusion. Using the above-threshold cell share—the proportion of all chapters × all event-window months of the pair whose readings exceed 2σ, leaving no room for chapter selection—as the statistic, the three tiers give 0.58%, 3.26%, and 5.47%, strictly monotonic (Table 2); the Spearman rank correlation is 0.533 ($p = 3.8\times10^{-8}$), the Mann-Whitney test between the top and bottom tiers gives $p = 4.1\times10^{-6}$, and a cluster bootstrap over target countries (20,000 resamples) gives a one-sided $p \approx 5\times10^{-5}$, so the conclusion is robust to within-country correlation among multi-round pairs. The unbalanced tier sizes (74/7/12) faithfully reflect the distribution of measure types: rank correlation is based on pair-level ranks and the cluster bootstrap is not stratified by tier, so the imbalance does not alter the testing properties. The pair-peak caliber gives a consistent ordering (Spearman 0.453, $p = 5.1\times10^{-6}$). It is worth noting that tier 0's 0.58% is only slightly above the null-simulation baseline of 0.45% (Appendix A1): low-tier readings at the floor are not an index failure but a faithful record of the policies' actual intensity—the index measures not "whether a policy exists" but "how strongly the flow is actually restricted."

Table 2 Blind test: readings by ex-ante severity tier (93 measurable pairs)

Tier	Meaning	n	Median above-threshold cell share	Median pair peak
0	Individual/entity-level designations (Magnitsky, terrorism financing, sporadic Entity List additions)	74	0.58%	3.65σ
1	Round- or scale-limited (baseline depressed by earlier rounds, or small exposure)	7	3.26%	5.11σ
2	Sectoral and comprehensive measures (Iran program, Russia 2022, China-U.S. tariffs, etc.)	12	5.47%	6.82σ

Note: Spearman(tier, above-threshold cell share) = 0.533, $p = 3.8\times10^{-8}$; pair-peak caliber Spearman = 0.453, $p = 5.1\times10^{-6}$. The pair peak takes the maximum of roughly 1,700 cells (97 chapters × 18 months), so even tier 0's median peak exceeds 2σ—which is exactly why the above-threshold cell share (a proportion, not an extreme value) is the primary statistic.

As a supplementary reading against the 2σ line: among the 19 pairs in tiers 1 and 2, the Myanmar round of January 2023 is a follow-up round (its baseline had already been depressed by the previous round and is excluded from detection statistics by the frozen rule); of the remaining 18 sectoral-or-above pairs, 12 crossed within the event window (66.7%; 55.6% under the conservative caliber that removes COVID-window peaks), with a median peak of 4.21σ among hits; the six non-crossing pairs all have concrete records (short-lived executive orders, derivative designations, unverified cases conservatively retained, and so on). Boundary evidence is also as expected: the first Iran round in 2018 reads a strong 9.98σ, while post-2021 follow-up rounds read markedly lower—once the base is depressed, equally harsh measures no longer have a corresponding magnitude of measurable suppression in the flow. Detection per se is not the design goal; the magnitude of the reading is.

5.3 Within-event distribution: divergence across commodities in the same event

The second piece of blind-test evidence comes from inside an event: do the readings of different commodities diverge according to actual suppression? Taking the six-digit matrix of U.S. exports to China during the 2018 dispute (the 194 tariff lines with 2017 scale above USD 100 million, without any commodity selection), 30 of the 194 lines crossed within the event window (18 months from July 2018; reading trajectories of three representative groups in Figure 3): 22 are on China's retaliation lists (a 13.4% crossing rate within the lists), and commodities that suffered substantive suppression give significant readings with matching timing—soybeans 2.72σ (November 2018, consistent with public records of the fourth-quarter collapse of exports to China), sorghum 9.76σ, propane 7.36σ, finished vehicles 7.74σ. The lower soybean reading relative to sorghum and propane reflects the persistence of suppression rather than its presence: the soybean collapse was concentrated in Q4 2018 and recovered to about 65% of its 2017 scale in 2019, so the EWMA accumulated little from the short shock, whereas sorghum and propane were suppressed throughout the window. Of the 8 crossing lines outside the lists, 7 are waste paper and metal waste and scrap (2.8σ–10.0σ), whose crossing timings correspond one by one to the phased implementation of the concurrent solid-waste import ban—the index independently identified, without priors, another restrictive measure concurrent with the tariff war; the remaining one (memories 854232, peak 2.19σ) hugs the reference line, a weak signal combining the memory-chip

price downturn with a specific firm's sales ban. Among the 68 calm lines whose drawdown was below 20%, only one crossed (1.5%), the reading surface flat overall. Lines whose flows halved but did not cross are mainly energy and bulk commodities, whose declines were driven by the common price factor and removed by the factor decomposition—the index reports only bilateral idiosyncratic suppression, by design rather than by omission; occasional weak readings near the reference line are normal background noise, and the reading rules (Section 3.3) are designed precisely for this distinction. Within the same event, the same matrix, and the same pipeline, the divergence of readings corresponds exactly to the presence or absence of actual suppression, independent of any researcher selection.

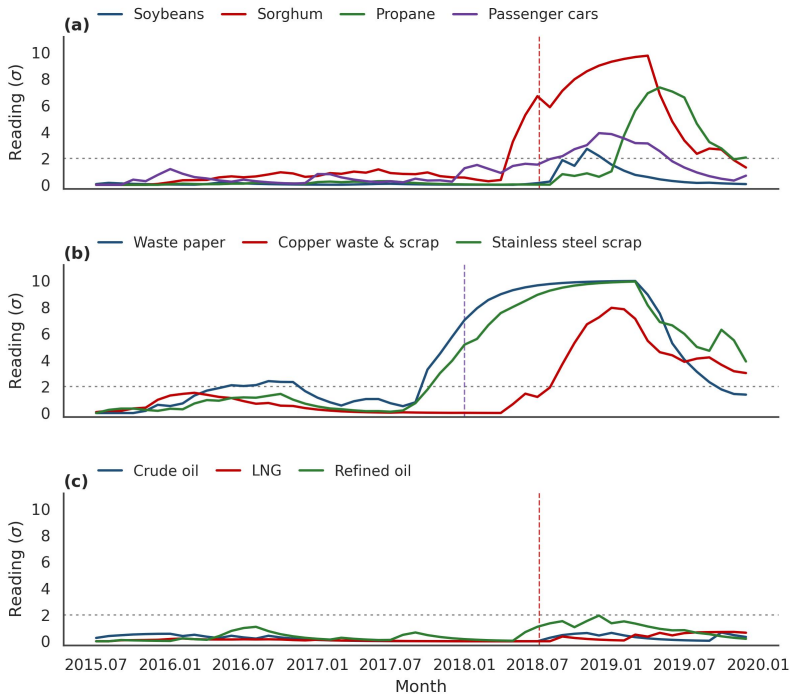


Figure 3 Within-event divergence of readings across commodities: targets of the retaliation lists (a), waste-and-scrap lines covered by the solid-waste import ban (b), and a price-collapse control group (c) (U.S. exports to China, six-digit level, value terms). The dashed line in (a) and (c) marks the effective date of the additional tariffs on China (July 2018), and in (b) the formal effective date of the solid-waste import ban (January 2018); the dashed lines serve only as timing references, and readings are capped by the winsorization parameter 10σ

5.4 Confounding, benchmarks, and parameter sensitivity

COVID confounding. The Case-2 event window overlaps the pandemic; recomputing after removing March 2020 to December 2021, the three chapters peak at 6.16–9.20σ and do not collapse; the same-chapter no-event control peaks at 0.33σ over the same period. The pandemic does not explain the readings.

Benchmark comparison. On an extended case pool covering five event types and 15 overtly restricted cells (Appendix B), RBTS is computed alongside three benchmarks: a single-row one-sided CUSUM (rolling-window self-standardization, no common-factor removal), Bai-Perron structural breaks (BIC-selected breakpoints), and a time-invariant-baseline SPE (estimation set fixed at all pre-event months). RBTS's advantage lies not in detection latency (1 month, the same as CUSUM) but in false-alarm control and reading shape: single-row methods without common-factor removal treat

industry cycles and macro shocks all as restrictions, with negative-cell false-alarm rates of 51.0% (CUSUM) and 77.7% (Bai-Perron pre-event breaks) against RBTS's 12.7%; moreover, both benchmarks answer only "whether a break exists" and give no continuous restriction reading, and both produce false alarms on the type-B cells "listed on paper but not actually restricted," while RBTS stays quiet.

Parameter sensitivity. Recomputing the extended case pool over 18 configurations of $\lambda \in \{0.1, 0.2, 0.3\}$ × variance threshold $\in \{70\%, 80\%, 90\%\}$ × window length $\in \{36, 48\}$: the median Spearman correlation of peak orderings across configurations is 0.78, and key findings are rank-stable in the parameter neighborhood; absolute reading levels respond to λ and window length as expected, and borderline cells whose peaks hug 2σ are unstable in crossing status under extreme parameters, so interpretation should focus on medium-to-high readings (Appendix B).

6. Conclusion and Outlook

This paper transplants the monitoring paradigm of statistical process control onto bilateral monthly trade data and constructs the descriptive trade-restriction index RBTS. The methodological contributions are twofold: first, the rolling baseline with row-wise selective updating resolves the "unknown in-control phase" premise of transplanting SPC, closing the channel through which persistent control shifts are adaptively absorbed into the baseline—ablation shows the design strengthens readings of persistent restrictions by a factor of 3.5–5.3; second, retaining the row-wise components of principal-component residuals, together with empirical calibration, one-sidedness, and EWMA smoothing, yields σ readings directly comparable across cells, cases, and periods, and six properties plus simulations delineate the index's effective working zone.

The empirical findings concentrate on three points. First, the index measures actual restriction rather than paper listing: hydrogen fluoride and photoresist on the same control list give opposite patterns, and in the blind test readings rise strictly monotonically with ex-ante verified severity. Second, covert restrictions without formal texts can be quantified: the Saudi boycott of Turkey yields readings consistent with the event timeline that do not collapse when the pandemic window is removed, filling a systematic blind spot of list-based measures. Third, the index's boundaries are transparent and declarable: a minimum detectable effect of about 50% sustained suppression, a scale boundary near USD 100 million of annual trade, and reference-line dilution on strongly autocorrelated series are all stated in the methodology with simulation support.

Future research can proceed in three directions: extending single-reporter matrices to multi-reporter joint panels; building a standing cross-validation mechanism with list-based products such as the GSDB, forming a "paper–enforcement" dual-track monitoring system; and further modeling the distributional properties of readings. Data come from UN Comtrade and the GSDB; the index code, blind-test criteria, and the case-by-case coding table are submitted as replication materials.

References

Aue, A., & Kirch, C. (2024). The state of cumulative sum sequential changepoint testing 70 years after Page. Biometrika, 111(2), 367–384.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica, 70(1), 191–221.

Cadot, O., & Gourdon, J. (2016). Non-tariff measures, preferential trade agreements, and prices: New evidence. Review of World Economics, 152(2), 227–249.

Carrère, C., & Grigoriou, C. (2015). Can mirror data help to capture informal international trade? (FERDI Working Paper). FERDI.

Chakraborti, S., Human, S. W., & Graham, M. A. (2008). Phase I statistical process control charts: An overview and some results. Quality Engineering, 21(1), 52–62.

Chen, B., Ma, H., & Xu, Y. (2014). Measuring China's trade liberalization: A generalized measure of trade restrictiveness index. Journal of Comparative Economics, 42(4), 994–1006.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association, 99(465), 96–104.

Felbermayr, G., Kirilakha, A., Syropoulos, C., Yalcin, E., & Yotov, Y. V. (2020). The Global Sanctions Data Base. European Economic Review, 129.

Groen, J. J. J., Kapetanios, G., & Price, S. (2013). Multivariate methods for monitoring structural change. Journal of Applied Econometrics, 28(7), 1028–1044.

Hayakawa, K., Ito, K., Fukao, K., et al. (2023). The impact of the strengthening of export controls on Japanese exports of dual-use goods. International Economics, 174, 160–179.

Hufbauer, G. C., Schott, J. J., Elliott, K. A., et al. (2007). Economic sanctions reconsidered (3rd ed.). Peterson Institute for International Economics.

Jackson, J. E., & Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. Technometrics, 21(3), 341–349.

Jena, D., Akash, C., & Gupta, P. (2024). Deflecting economic sanctions: Do trade and political alliances matter? International Economics and Economic Policy, 21(3), 543–567.

Kee, H. L., Nicita, A., & Olarreaga, M. (2009). Estimating trade restrictiveness indices. The Economic Journal, 119(534), 172–199.

Kirilakha, A., Felbermayr, G., Syropoulos, C., et al. (2021). The Global Sanctions Database: An update that includes the years of the Trump Presidency. In P. A. G. van Bergeijk (Ed.), Research handbook on economic sanctions. Edward Elgar.

Larch, M., Luckstead, J., & Yotov, Y. V. (2024). Economic sanctions and agricultural trade. American Journal of Agricultural Economics, 106(4), 1477–1517.

Leisch, F., Hornik, K., & Kuan, C. M. (2000). Monitoring structural changes with the generalized fluctuation test. Econometric Theory, 16(6), 835–854.

Li, W., Yue, H. H., Valle-Cervantes, S., et al. (2000). Recursive PCA for adaptive process monitoring. Journal of Process Control, 10(5), 471–486.

Liu, B., & Li, Q. (2023). U.S. export controls on China and Chinese firm innovation. Research on Finance and Economics, 49(12), 19–33. (in Chinese)

Lucas, J. M., & Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. Technometrics, 32(1), 1–12.

MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. Control Engineering Practice, 3(3), 403–414.

Makioka, R., & Zhang, H. (2024). The impact of export controls on international trade: Evidence from the Japan–Korea trade dispute in semiconductor industry. Journal of the Japanese and International Economies, 74,

101336.

Nordås, H. K. (2016). Services Trade Restrictiveness Index (STRI): The trade effect of regulatory differences (OECD Trade Policy Papers, No. 189). OECD Publishing.

Page, E. S. (1954). Continuous inspection schemes. Biometrika, 41(1/2), 100–115.

Qin, S. J. (2003). Statistical process monitoring: Basics and beyond. Journal of Chemometrics, 17(8/9), 480–502.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. Technometrics, 1(3), 239–250.

Shewhart, W. A. (1931). Economic control of quality of manufactured product. Van Nostrand.

Stinnett, D. M., Early, B. R., Horne, C., et al. (2011). Complying by denying: Explaining why states develop nonproliferation export controls. International Studies Perspectives, 12(3), 308–326.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association, 97(460), 1167–1179.

Wang, X., Kruger, U., & Irwin, G. W. (2005). Process monitoring approach using fast moving window PCA. Industrial & Engineering Chemistry Research, 44(15), 5691–5702.

Wei, P., & Zhang, Z. (2018). U.S. high-tech export controls on China and their effect on bilateral trade. Industrial Technology and Economy, 37(1), 76–85. (in Chinese)

Appendix A Simulation Support for the Index's Properties

A1 Null calibration (Monte Carlo). Pure factor-model simulation (20 rows × 180 months, no controls): after calibration, $z$has mean −0.04 and standard deviation 1.12, with P($|z|>2$) = 7.4% (theoretical 4.6%)—the mild overdispersion of rolling estimation is exactly why the empirical calibration $\sigma_0 = 1.22$ exists. Under the null, the share of $\bar{D}$ cells above 2σ is 0.45%; the corresponding empirical figure on the U.S. matrix is 1.40% (including footprints of real events not in the labels). The baseline scenario's crossing share is 0.08%, so the 2σ reference line is conservative relative to the nominal level.

A2 Minimum detectable effect (power curve). On a generic factor matrix (n = 20, T = 180, 20 replications per depth), a single row is subjected to 24 months of sustained suppression, and detection is counted if the event-window peak exceeds 2σ: detection rates are 0 at suppression depths of 10%–26%, 10% at 39%, 20% at 50%, 50% at 63%, 85% at 78%, and 100% at 86% (Figure A1). The minimum detectable effect is about 50% sustained suppression; for narrow-measure × large-exposure combinations the equivalent suppression depth equals the listed entities' share of chapter flows, unifying the granularity boundary and the depth boundary.

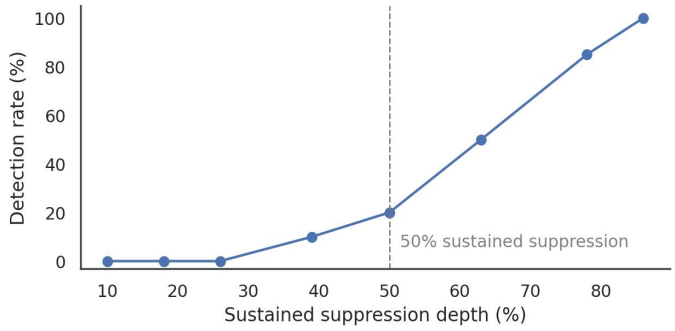


Figure A1 Detection rate against suppression depth (power curve)

A3 Stress tests (seasonality and autocorrelation). Null crossing shares above 2σ: baseline 0.08%; with deterministic seasonality 0.13%; AR(1) coefficient 0.3 gives 0.20%; AR(1) 0.6 gives 0.72% (maximum 1.12%); seasonality plus AR(1) 0.3 gives 0.16%. Seasonality and weak autocorrelation do

not change the nature of the reference line; on strongly autocorrelated series the nominal probability is diluted by about an order of magnitude, and interpretation should incorporate event-window length.

A4 Sketch proofs of the properties. P2 bounded influence: winsorization gives $D \leq 10$, and the EWMA recursion bounds a single month's contribution at $\lambda \times 10 = 3.0\sigma$; triggering winsorization requires $z \leq -12.2$. P3 geometric forgetting: weights 0.7 to the power k, half-life about 1.9 months, 90% of weight in the most recent 7 months. P4 bounded contamination: under selective updating an event month contributes zero to the baseline; if it leaks in, its influence on that month's in-window mean is at most 1/24. P5 decomposition identity: $\tilde{x} = \hat{x} + e$is an orthogonal projection, exact on the log scale after row scaling. log1p error bound: the gap reading $g = (C-A)/(1+C)$ differs from the true relative suppression $(C-A)/C$ by less than $1/C$, so for large-flow cells it reads directly as a percentage.

Appendix B Extended Robustness

B1 Ablation: selective versus unconditional updating. Recomputing the target cells of the two main-text cases with unconditional rolling updating: the event-window mean reading of persistent restrictions under selective updating is 3.5–5.3 times that under unconditional updating (chapter-by-chapter ratios: hydrogen fluoride 7.31σ versus 1.66σ; the three Saudi-boycott chapters 6.48–7.75σ versus 1.24–2.20σ). Unconditional updating adaptively absorbs persistent suppression into the baseline, so readings decay as suppression persists; selective updating closes this channel. The earlier matrix-level rule (disabling an entire month when any row breaches) exhausts clean months during broad-spectrum episodes and falls back to contaminated windows; the row-wise rule eliminates this fallback.

B2 Benchmark-comparison details. The extended case pool covers 15 overtly restricted cells across five event types—export controls, food bans, comprehensive sanctions, tariffs, and secondary sanctions—with 157 negative rows.

Table B1 Benchmark comparison (extended case pool)

Method	Event-window detection	Median latency	Negative false-alarm rate	Type-B controls
RBTS (this paper)	100%	1.0 month	12.7%	All quiet
Time-invariant SPE	93%	—	—	—
Single-row one-sided CUSUM	100%	1.0 month	51.0%	All false-alarm
Bai-Perron structural breaks	93%	2.5 months	77.7%	False-alarm or missed

Note: false-alarm rates are computed on event windows; time-invariant SPE is estimated on a fixed pre-event baseline and does not absorb shifts, so its window-mean readings are elevated, but it answers a different question from our ablation. Type-B controls are cells listed on paper but not actually restricted.

B3 Hyperparameter sensitivity. Recomputing the event-window peaks of the 15 target cells over 18 configurations of $\lambda \in \{0.1, 0.2, 0.3\}$ × variance threshold $\in$ {70%, 80%, 90%} (this dimension perturbs the selection rule of factor number K in (4): 70% corresponds to fewer factors, 90% often approaches the cap of 3) × window length $\in \{36, 48\}$: the Spearman rank correlation of peaks across configurations has minimum 0.557, median 0.779, and maximum 1.000; the all-configuration detection rate is 77.8%, with misses concentrated among borderline cells whose peaks hug 2σ. Stratified by λ: detection rates at $\lambda = 0.1, 0.2, 0.3$ are 46.7%, 90.0%, 96.7% (median event-window peaks 1.80σ, 2.65σ, 3.38σ)—$\lambda = 0.1$ over-smooths and dilutes persistent suppression, while $\lambda \geq 0.2$ is stable in both detection

and ordering; the main text's $\lambda = 0.3$ sits in the stable region.

B4 Deseasonalization check. Recomputing with month-of-year full-sample deseasonalization of the log series: the rank correlation of target-cell peaks is 0.711, but full-sample monthly means absorb part of persistent suppression into the "seasonal term" (event leakage), systematically depressing readings; the main line retains the non-deseasonalized caliber, and a variant with recursive deseasonalization inside estimation sets is left for future work.

B5 Dimension-fidelity reconciliation. Over 29 chapter-level observations in the extended case pool, the Spearman correlation between counterfactual-gap readings and actual flow drawdowns is 0.906: chapters that did not fall have gaps near zero, and a chapter that fell 80% reads a gap of 61%. The two numbers come from the same source (both computed from the same flow series), so this check calibrates the dimension and does not constitute independent evidence.

B6 Training pool and sensitivity of the empirical dispersion. The estimation pool of $\sigma_0$consists of standardized residuals from five case matrices (Japan–Korea 2019 in quantity terms, U.S.–China 2022, U.S.–Russia 2014, U.S.–Russia 2022, Turkey–Saudi 2020), n = 12,192 pooled estimation-set monthly observations; the right-shoulder half-sample estimate (Efron, 2004) gives mean −0.010 and dispersion 1.2125, taken as 1.22. Calibration–validation separation is guaranteed by construction: estimation uses only the right-shoulder half of pooled residuals, and suppression residuals in event periods fall in the left tail and do not enter the fit. Per-matrix dispersions range from 1.139 to 1.572; the Japan–Korea matrix is in quantity terms (prices fluctuated violently during that control episode; see the dual-caliber check in Section 4.2), and its value-terms per-matrix dispersion is 1.54, within the same range—the caliber choice does not change the calibration conclusion. Leave-one-out (re-estimating after dropping one matrix at a time) gives 1.182–1.403, with the U.S.–China 2022 matrix the most influential (1.403). The adopted value sits at the lower end of the leave-one-out interval, so the reference line errs on the conservative side.

B7 Severity-coding rules for the blind test. Coding was completed before reading any index readings. Rules: tier 2 = sectoral and comprehensive measures (full embargoes, sector-wide export restrictions, main-line trade-war tariffs); tier 1 = round- or scale-limited partial measures (baseline depressed by later rounds, or small exposure); tier 0 = individual- and entity-level designations (Global Magnitsky human-rights designations, terrorism-financing network listings, sporadic Entity List additions) and measures without actual trade measures (threats not implemented, non-trade designations); narrow measures finer than statistical chapters are handled under the declared measurability boundaries. Ambiguous cases are always coded one tier down (conservative direction). Of the 120 pairs, the 93 with sufficient pre-event data enter the test, with tier sizes 74/7/12. Coding was done by the author alone and inter-coder reliability was not computed; the case-by-case coding table is published with the replication materials for checking and re-coding.