Multivariate Ratio Edits Using Tolerance Intervals[1]

Daniel Tuyisenge[1], Dr. Derek Young[1], and Dr. Thomas Mathew[2]

1. University of Kentucky 2. University of Maryland

Department of Statistics

October 1, 2025

[1]Supported by MWSUG 2025 Student Travel Scholarship.

Introduction — What & Why

What are ratio data?

- Quotients of two linked items.
- Examples: Hours/Employee, Payroll/Employee, ...
- Ratios can be *current-cell* (same period) or *historic-cell* (across periods).

What are ratio edits?

- Compare observed ratios to program limits.
  - *Microdata:* record-level checks and follow-up.
  - *Macrodata:* tabulation/cell-level screening.
- Flagged data outside the limits for subject-matter review.

Goal of this project

- Build *multivariate* ratio checks via *rectangular* acceptance regions on vectors of ratios for continuous data.
- Evaluate detection with **Type I** (swamping) and **Type II** (masking) error rates to aid triage.

Background (1/3)-Definitions

- **Components and ratios.** Let $\mathbf{X} = (X_1, \ldots, X_q)^\top$ be original items and $\mathbf{R} = (R_1, \ldots, R_p)^\top$ be derived ratios with $R_j = X_{a(j)}/X_{b(j)}$ for linked elements of $\mathbf{X}$.
- **Ratio edit.** Check whether R_j lies within a program tolerance (possibly asymmetric). Edits operate on $\mathbf{R}$ as a *diagnostic surface*, but fixes are applied to $\mathbf{X}$.
- **Editing limits/regions.** Bounds chosen to control:
  - *Type I (false positive):* valid records flagged.
  - *Type II (false negative):* erroneous records not flagged.
- **Dependence & aliasing.** Shared numerators/denominators induce strong dependencies among ratios; two extreme components can yield a non-extreme ratio.
- **Masking & swamping.**
  - *Masking:* offsetting errors (e.g., both X_1 and X_2 inflated $\Rightarrow$ $R = X_1/X_2$ looks normal).
  - *Swamping:* heterogeneity/global limits over-flag valid records.
- **Ratio editing flavors.**
  - *Current-cell:* compare linked items within period t.
  - *Historic-cell:* compare the same item across t vs. $t-i$.

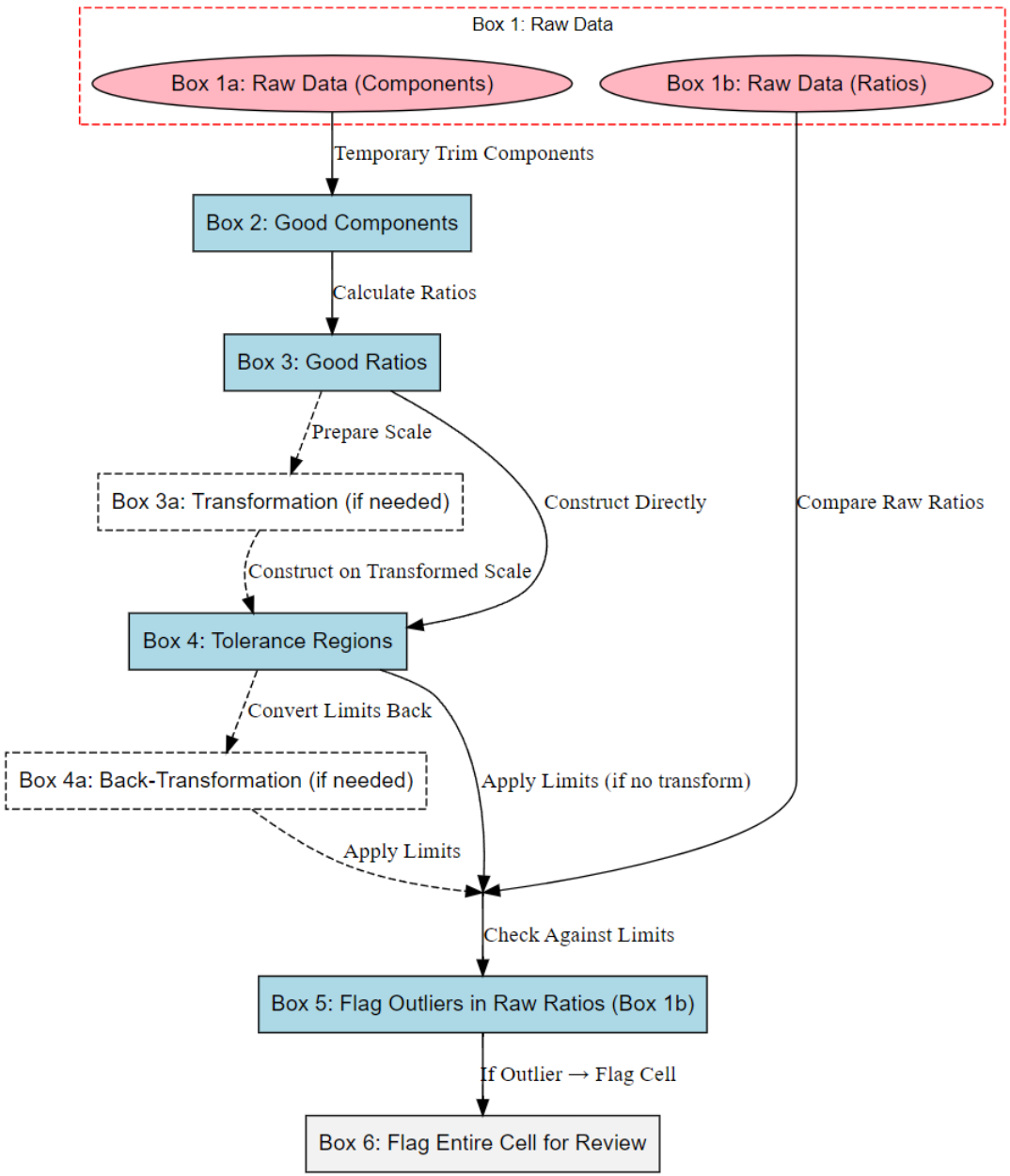
Background (2/3):Related Work

- **Previous work:**
  - Fellegi and Holt (1976) introduced the concept *editing and imputation* for continuous and categorical data using
    - logical/quantitative edits
    - minimal-change principle
  - The work of Hidiroglou and Berthelot (1986),Thompson and Sigman (1999),and Young and Mathew (2015) explored *Univariate ratio edits:*
    - Robust control limits
    - Resistant Fences
    - Hidiroglou-Berthelot method
    - D–MASO gap analysis,
    - Statistical Tolerance Limits-
  - The work of (Thompson and Ozcoskun, 2007) explored *Multivariate edits* using
    - Robust Mahalanobis on ratios vectors .

Background (3/3):Limitations and Solutins

- **Limitations to address.**
  - **Multivariate (Thompson 2007):** needs near-elliptical clouds and stable covariances; redundancy, masking/swamping, and low interpretability of distance-based flags.
  - **Univariate on multivariate data:** ignores dependence and inflates Type I error; offsetting errors slip through, yielding inconsistent triage across ratios.
  - **Skewed/heterogeneous cells:** symmetric limits and quadratic distances misfit right tails and size effects, causing domain-dependent over-flagging and unstable cutoffs.
- **Our contribution: Statistical Multivariate Tolerance Regions (TR).**
  1. Interpretable *rectangular* componentwise limits on **R**.
  2. Reduced swamping/masking via temporary trimming and transformations where needed.
  3. Distribution-free (nonparametric) TRs when transformation is not feasible.

This Talk in One Picture



Statistical Multivariate Tolerance Regions for Ratio Edits

- Let $F \in \mathcal{F}$ be an unknown p-variate distribution, $\mathcal{X} = \{X_1, \ldots, X_n\}$ i.i.d. random sample from F with $n > p$, and the random vector $\mathbf{X} \sim F$ independent of $\mathcal{X}$.
- For a random subset $T(\mathcal{X})$ construct using $\mathcal{X}$, define probability content of $T(\mathcal{X})$ under F:

$$C_F\big(T(\mathcal{X})\big) = \Pr_F\{\mathbf{X} \in T(\mathcal{X})\}.$$

- If

$$\inf_{F \in \mathcal{F}} \Pr\big\{\, C_F\big(T(\mathcal{X})\big) \geq P \,\big\} = \gamma,$$

  then $T(\mathcal{X})$ is a (P, γ) tolerance region of underlying distribution.
- The quantities (P, γ) are the desired probability content level and confidence level, respectively, of the tolerance set.
- In simple terms, a (P, γ) tolerance region, computed using a random sample, is meant to capture a specified proportion P or more of the underlying population distribution, with a confidence level of γ. See Krishnamoorthy and Mathew (2009).

Rectangular Central Tolerance Regions (RCTR)

- **Model/notation.**
  - Sample random vector $\mathbf{x}_{n\times q} = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iq})^T \in \mathbb{R}^{n\times q}$ with $\mathbf{x}_{n\times q} \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Sample mean vector $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_q)^\top$ and $S = (S_{ij})$ is the sample covariance matrix
  - Population parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^\top$, $\boldsymbol{\Sigma} = (\sigma_{ij})$ with σ_{ii} the ith marginal variance, and ρ the population correlation matrix of $\boldsymbol{\Sigma}$.
- According to Lucagbo and Mathew (2024), **Central rectangular content set.**

$$\mathcal{R}(P;\rho) = \left\{ \mu_i - c(\rho)\sqrt{\sigma_{ii}} \leq X_i \leq \mu_i + c(\rho)\sqrt{\sigma_{ii}},\ i = 1, \ldots, q \right\},$$

where $c(\rho)$ is chosen so that

$$\Pr(\mu_i - c(\rho)\sqrt{\sigma_{ii}} \leq X_i \leq \mu_i + c(\rho)\sqrt{\sigma_{ii}},\ i = 1, \ldots, q) = P.$$

- $c(\rho)$ is estimated using bootstrap as described in Algorithm 1(next slide).

Algorithm 1: Ratio Editing via Rectangular Central Tolerance Regions

1. Transform the ratio data $X = \{X_i\}_{i=1}^n$ to approximate $N_q(\mu, \Sigma)$ (see Section ??). Denote the working sample still by X.
2. Compute Mahalanobis distances d_i^2 from $(\bar{\mathbf{X}}, S)$; trim the largest $t\%$ to obtain X_t. Recompute $\bar{\mathbf{X}}, S$ and correlation $\hat{\rho}$ on X_t.
3. Numerically solve for $\hat{c} = c(\hat{\rho})$ such that the fitted $N_q(\bar{\mathbf{X}}, S)$ places probability P inside the symmetric rectangle $\bar{X}_i \pm \hat{c}\sqrt{S_{ii}}$.
4. Generate B bootstrap pairs

$$\bar{\mathbf{X}}_b^* \sim N_q\left(\mathbf{0}, \tfrac{1}{n_t}\hat{\rho}\right), \qquad (n_t - 1)S_b^* \sim W_q(\hat{\rho}, n_t - 1),$$

   and compute

$$W_b = \max_{1 \le i \le q} \frac{|\bar{X}_{b,i}^*| + \hat{c}}{\sqrt{S_{b,ii}^*}}, \qquad b = 1, \ldots, B.$$

5. Set $c_1(\hat{\rho})$ to the empirical γ-quantile of $\{W_b\}_{b=1}^B$.
6. Form the tolerance region

$$\mathcal{R}_{\mathrm{RC}}(P, \gamma) = \prod_{i=1}^q \left[\bar{X}_i - c_1(\hat{\rho})\sqrt{S_{ii}},\ \bar{X}_i + c_1(\hat{\rho})\sqrt{S_{ii}} \right].$$

7. Flag record $\mathbf{X}_i$ as an outlier if $\mathbf{X}_i \notin \mathcal{R}_{\mathrm{RC}}(P, \gamma)$.

Other Editing Methods

- **Mixed–Simultaneous TR (MSCTR):** combines two–sided bounds for some ratios with one–sided (upper/lower) bounds for others. When it is just two-sided it is becomes *Simultaneous TR (SCTR)*.
- **Statistically Equivalent Blocks (SEB):** partitions coordinates $\{1, \ldots, q\}$ into blocks of similar behavior. Liu et al. (2024)
- **Mahalanobis Depth TRs:** define depth via (robust) Mahalanobis distance and retain points whose depth exceeds a calibrated cutoff.Young and Mathew (2020)
- **Spatial Depth TRs:** use spatial depth based on mean direction vectors) and take the central set.Young and Mathew (2020)
- **Robust Mahalanobis Ellipsoids:** Use a robust center and scatter and use the ellipsoid to flag outliers. Thompson (2007))

Data Transformations for Ratios (1/2)

Setup. For observation i, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ with realized $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Each variable j has a transform parameter $\lambda_j \in \mathbb{R}$.

Motivation. Ratios are typically right-skewed and heteroscedastic; gentle power transforms stabilize scale/skew and help downstream TRs.

Box–Cox Box and Cox (1964) **(requires $x_{ij} > 0$):**

$$T_{\lambda_j}(x_{ij}) = \begin{cases} \dfrac{x_{ij}^{\lambda_j} - 1}{\lambda_j}, & \lambda_j \neq 0, \\ \log x_{ij}, & \lambda_j = 0. \end{cases}$$

Yeo–Johnson Yeo and Johnson (2000) **(allows $x_{ij} \in \mathbb{R}$):**

$$T_{\lambda_j}(x_{ij}) = \begin{cases} \dfrac{(x_{ij}+1)^{\lambda_j} - 1}{\lambda_j}, & x_{ij} \geq 0,\ \lambda_j \neq 0, \\ \log(x_{ij}+1), & x_{ij} \geq 0,\ \lambda_j = 0, \\ -\dfrac{(-x_{ij}+1)^{2-\lambda_j} - 1}{2-\lambda_j}, & x_{ij} < 0,\ \lambda_j \neq 2, \\ -\log(-x_{ij}+1), & x_{ij} < 0,\ \lambda_j = 2. \end{cases}$$

Data Transformations for Ratios (2/2)

Why transform? Stabilize scale/skew, improve approximate normality of $\mathbf{R}$, and make rectangular TRs more stable across domains.

Choose parameters: Marginal vs. Joint

- **Marginal (per ratio R_j):** estimate λ_j independently for each j.
- **Joint (all ratios):** estimate $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ jointly

Estimation targets (Gaussian working model).

- *Univariate likelihood (per j):* $\widehat{\lambda}_j = \arg\max_{\lambda} \sum_i \log f_{\mathcal{N}}\big(T_\lambda(R_{ij});\, \mu_j, \sigma_j^2\big)$
- *Joint likelihood (all j):* $\widehat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \sum_i \log f_{\mathcal{N}_p}\big(T_{\boldsymbol{\lambda}}(\mathbf{R}_{i\cdot});\, \boldsymbol{\mu}, \Sigma\big)$

Practical guardrails.

- Start marginal; move to joint only if cross-ratio dependence harms TR calibration.
- If using Box–Cox with zeros: add a small offset or switch to Yeo–Johnson.
- Refit $\boldsymbol{\lambda}$ on clear domain/period shifts; otherwise keep fixed for stability.

Jaccard Index for Agreement on Flags

Let $U = \{1, \ldots, N\}$ index the records. For method **a**, let $F_a \subset U$ be the set it flags. According to da F. Costa (2021), for two methods **a** and **b**, jaccard index is defined as

$$J(a, b) = \frac{|F_a \cap F_b|}{|F_a \cup F_b|} = \frac{n_{11}}{n_{11} + n_{10} + n_{01}} \in [0, 1],$$

where

- n_{11} is the number outliers flagged by both methods **a** and **b**
- n_{10} is the number outliers flagged by **a** only
- n_{01} s the number outliers flagged by **b** only
- n_{00} is the good data

Reporting:

- Emphasizes *agreement on outliers*, not on good data, n_{00}.
- Robust to class imbalance when the flagged proportion is small.

Data Generation & Contamination Regimes

Contaminated model

$$\boldsymbol{R}_i \;\sim\; (1-\nu)\, N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;+\; \nu\, N_q(\boldsymbol{\mu}^*,\; \boldsymbol{\Sigma}),$$

- ν: contamination proportion
- Mean shift parameterization: $\boldsymbol{\mu}^* = \boldsymbol{\mu} \pm k\,\boldsymbol{\sigma}$ with regimes $k \in \{8 \text{ (well)},\ 4 \text{ (moderate)},\ 2 \text{ (heavy)}\}$.

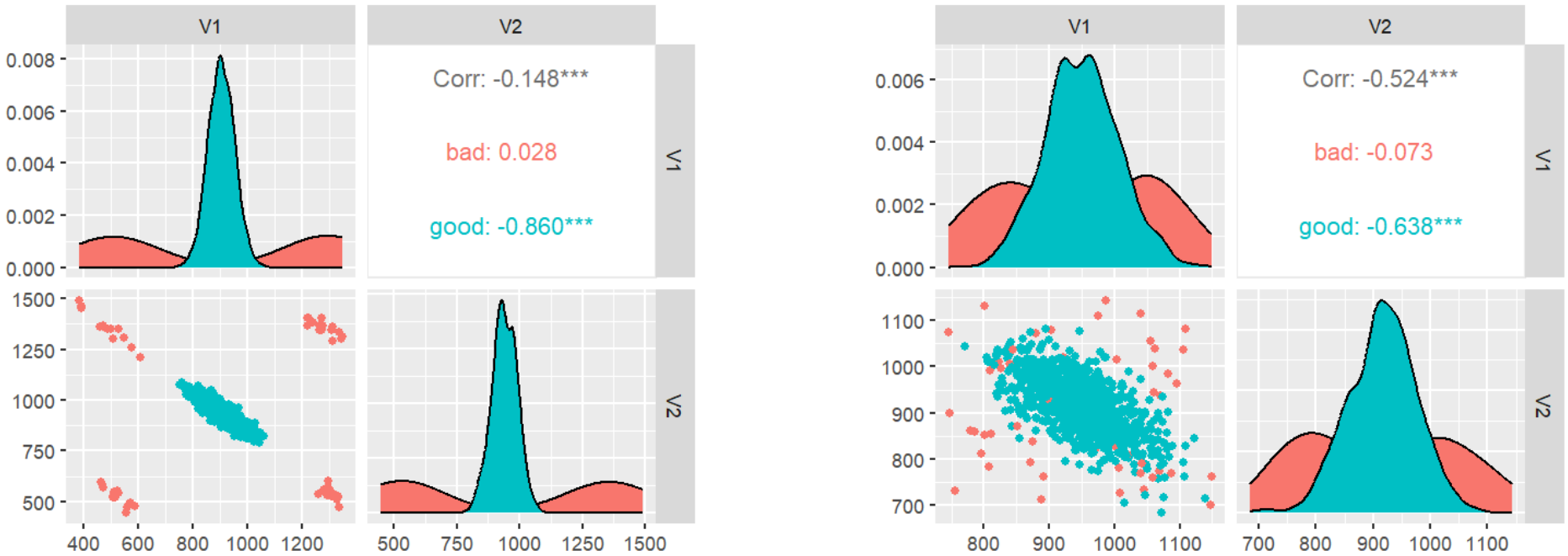


Figure 2: left:Well Separated vs right:Heavy overlapping on contamination

Simulation Setup

- Vary the following Variable: n,P,α,ν,regime, q, Editing Method, sides(for MSCTR)
- Boostrap:$B = 1000$, Number of Simulations: $N_{\text{sim}} = 1000$
- Metrics: **Type I Error Rate** , **Type II Error Rate**, **Volume** $|\mathcal{T}|$.
- One Example of the Simulations Setup
  - Method: RCTR, Content $P = 0.95$, Confidence $1 - \alpha = 0.95$.
  - $n = 1000$, dimensions $q = 2$.
  - Contamination $\nu = 0.05$
  - regimes: well and heavy.
  - Bootstrap replicates $B = 1000$.
  - Metrics: **Type I**, **Type II**, **Volume** $|\mathcal{T}|$.

Results (Tables 4 & 6): RCTR, Well vs Heavy, $P = .95$, $\alpha = .05$, $p = 2$

Table 1: Well Regime ($p = 2$)

Trim	Type I	Type II	Volume
0.01	0.0004	0.0000	1.86×10^5
0.05	**0.0643**	**0.0000**	$\mathbf{5.10 \times 10^4}$
0.10	0.1199	0.0000	3.81×10^4
0.15	0.1832	0.0000	2.95×10^4
0.20	0.2563	0.0000	2.29×10^4

Table 2: Heavy Regime ($p = 2$)

Trim	Type I	Type II	Volume
0.01	0.0271	0.4431	6.96×10^4
0.05	**0.0664**	**0.2691**	$\mathbf{5.03 \times 10^4}$
0.10	0.1176	0.1854	3.85×10^4
0.15	0.1758	0.1334	3.03×10^4
0.20	0.2432	0.0978	2.40×10^4

- **Observations:**
  - If our data is contaminated by 5% and we trim 5% of the data, it implies we have good data
  - in assessing the performance, we compare how close is Type I error rate to 1-P at the trim=0.05
  - In the table above(bold cell), the Type I error rate is very close to 0.05(i.e good performance)

Results (Tables 22 & 24): MSCTR, Well vs Heavy, $P = 0.95$, $\alpha = 0.05$, two-sided & upper, $p = 2$

Table 3: Well Regime ($p = 2$)

Trim	Type I	Type II	Volume
0.01	0.0002	0.0000	1.74×10^5
0.05	**0.0395**	**0.0000**	$\mathbf{6.96 \times 10^4}$
0.10	0.0768	0.0000	4.40×10^4
0.15	0.1022	0.0000	3.63×10^4
0.20	0.1283	0.0000	3.10×10^4

Table 4: Heavy Regime ($p = 2$)

Trim	Type I	Type II	Volume
0.01	0.0328	0.3691	7.35×10^4
0.05	**0.0552**	**0.2647**	$\mathbf{5.50 \times 10^4}$
0.10	0.0809	0.2067	4.36×10^4
0.15	0.1061	0.1715	3.68×10^4
0.20	0.1321	0.1458	3.18×10^4

- **Observations:**
  - For RCTR method, type I error is controlled to nominal level(contamination=0.05)
  - with heavy overlapping, we see an increase of Type II as expected

Methods Comparison — Heavy Regime ($P = 0.95$, $\alpha = 0.05$, trim $= 5\%$)

	$d = 3$			$d = 6$		
Method	Type I	Type II	Volume	Type I	Type II	Volume
Central Rect	0.0369	0.3052	1.91×10^7	0.0301	0.2472	7.78×10^{14}
Mixed Central	0.0910	0.1437	1.18×10^7	0.1571	0.0255	1.57×10^{14}
SEB	0.0460	0.2519	1.66×10^7	0.0407	0.1630	5.42×10^{14}
Spatial Depth	0.0388	0.2789	1.78×10^7	0.0269	0.2172	7.18×10^{14}
Mahal. Depth	0.0388	0.2795	1.79×10^7	0.0269	0.2175	7.18×10^{14}
Mahalanobis KT	0.0625	0.0875	5.60×10^6	0.0637	0.0030	2.08×10^{13}

- **Observations:**
  - Central Rectangular SEB methods perform better in lower and high dimension
  - Malanobis by Thompson and Ozcoskun (2007) performs good too in higher dimensions however it suffers the stated limitations earlier

Application (ASM): 2018 Back-Transformed TIs & Outlier Counts

- ASM data are the U.S. Census Bureau's Annual Survey of Manufactures tabulations and macrodata on manufacturing establishments covering shipments, employment, payroll, costs, inventories, and related measures.

Table 5: 2018 RCTR Tolerance Regions (95% content, 95% confidence) and outlier counts ($N = 648$).

Ratio	**Lower**	**Upper**	**Outliers**	**Non-outliers**
Payroll/Emp.	29.88	106.84	14	634
Wages/PW Hour	14.02	44.42	14	634
Revenue/Emp.	89.68	1999.03	15	633
VA/Emp.	45.48	782.09	12	636
VA/Payroll	1.03	9.62	15	633
joint	-	-	**32**	**616**

Notes: TIs are rectangular central tolerance intervals (back-transformed). Outlier counts are from RCTR; "Non-outliers" = N minus outliers.

Abbreviations Used

- Out: outlier / bad
- Non: non-outlier / good
- Cons. Non: consistently non-outlier (across years)
- Cons. Out: consistently outlier (across years)

Table 6: Per-ratio *and joint* counts by year (2018–2021). Each year $N = 648$.

	2018		2019		2020		2021	
Ratio / Group	Out	Non	Out	Non	Out	Non	Out	Non
Payroll/Emp.	14	634	11	637	15	633	18	630
Wages/PW Hour	14	634	11	637	13	635	13	635
Revenue/Emp.	15	633	15	633	13	635	13	635
VA/Emp.	12	636	15	633	9	639	13	635
VA/Payroll	15	633	17	631	11	637	13	635
joint	**32**	**616**	**33**	**615**	**36**	**612**	**30**	**618**

Table 7: Longitudinal transitions (2018→2021) under RCTR.

Ratio / Group	**N**	**Cons. Non**	**Cons. Out**	**Good→Out**	**Out→Good**
Payroll/Emp.	648	625	7	6	2
Wages/PW Hour	648	631	10	3	4
Revenue/Emp.	648	633	13	0	2
VA/Emp.	648	632	7	2	1
VA/Payroll	648	629	7	2	4
joint	**648**	**609**	**24**	**5**	**7**

Agreement Visualization (Figure 6)

- Jaccard Index agreement among the methods across 2018–2021 (higher = greater overlap in flagged records).

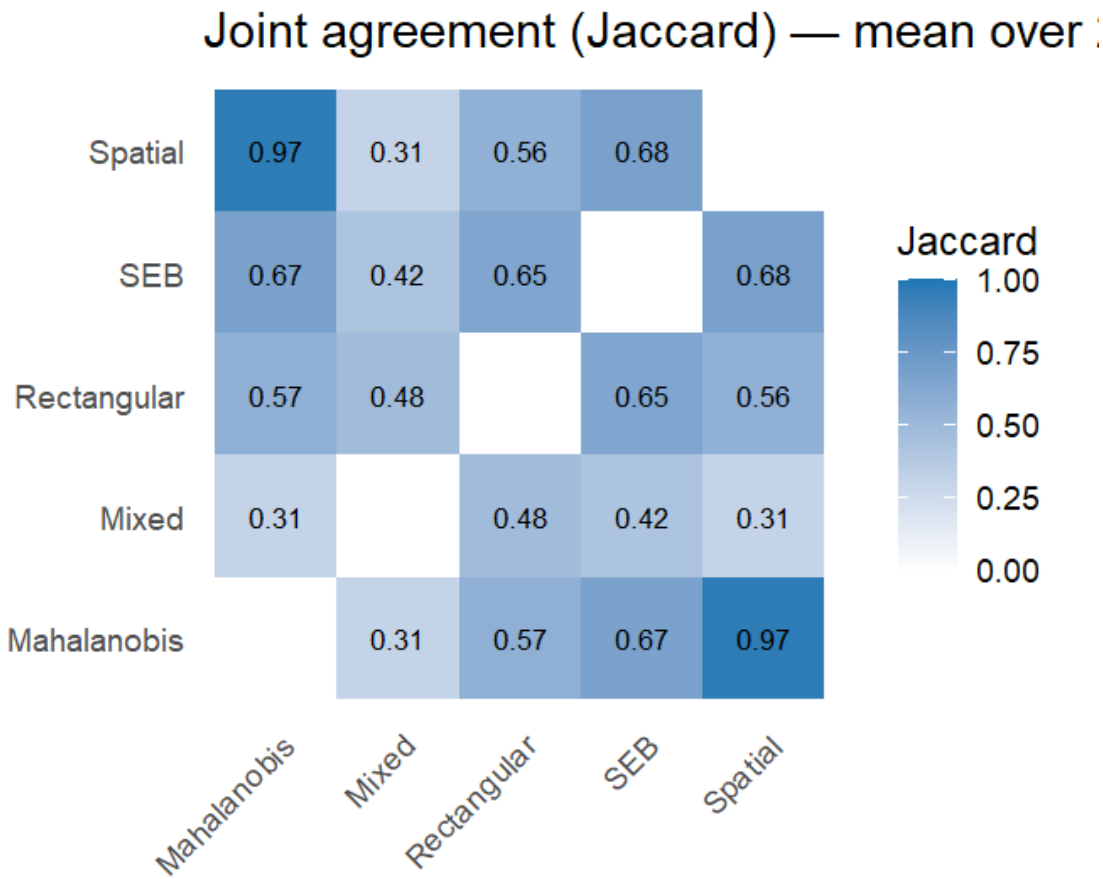


Figure 3: Agreement visualization across methods (Jaccard Index).

Conclusions — Simple Summary

- **What we did**
  - Built ratio-edit rules using multivariate tolerance regions.
  - Stabilized bounds via temporary trimming and optional transformations.
  - Evaluated editing methods using Type I/II, volume with simulations
  - Applied the editing methods using ASM data.
- **Key takeaways**
  - RCTR gives clear per-ratio limits and a good balance of flags vs. misses.
  - MSCTR when one-sided limits are required; SEB when assumptions are doubtful.
  - Depth methods are conservative, and useful with limited review capacity.
- **Future directions**
  - Unified edits for ratios *and* components; include categorical/discrete cases.
  - Use Bayesian pooling across industries and Machine Learning in editing

Acknowledgments

This presentation was supported by MWSUG 2025 Student Travel Scholarship.

Thank you!

Questions?

References I

George E. P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26(2): 211–252, 1964.

Luciano da F. Costa. Further generalizations of the jaccard index, 2021. URL https://arxiv.org/abs/2110.09619.

I. P. Fellegi and D. Holt. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71(353):17–35, 1976. URL http://www.jstor.com/stable/2285726.

Michel A. Hidiroglou and Jean-Marie Berthelot. Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12(1):73–83, 1986.

Kalimuthu Krishnamoorthy and Thomas Mathew. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 2009. ISBN 978-0-470-38026-0. doi: 10.1002/9780470473900.

Wei Liu, Frank Bretz, and Mario Cortina–Borja. Distribution–free hyperrectangular tolerance regions for setting multivariate reference regions in laboratory medicine. *Statistics in Medicine*, 43(8):1604–1614, 2024. doi: https://doi.org/10.1002/sim.10019. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.10019.

Michael Daniel Lucagbo and Thomas Mathew. Rectangular tolerance regions and multivariate normal reference regions in laboratory medicine. *Journal of Statistical Methodology*, 48(2):301–316, 2024. doi: 10.1080/03610920600683689.

Katherine Jenny Thompson. Investigation of macro editing techniques for outlier detection in survey data. In *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, pages 1186–1193, Montreal, Canada, 2007. American Statistical Association, Section on Survey Research Methods. URL https://ww2.amstat.org/meetings/ices/2007/proceedings/ices2007-000071.pdf.

Katherine Jenny Thompson and Laura Ozcoskun. An empirical investigation into macro editing. *Federal Committee on Statistical Methodology*, 2007. URL https://api.semanticscholar.org/CorpusID:216653461.

Katherine Jenny Thompson and Richard S Sigman. Statistical methods for developing ratio edit tolerances for economic data. *Journal of Official Statistics*, 15(4):517–535, 1999.

In-Kwon Yeo and Richard A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87 (4):954–959, 2000.

Derek S. Young and Thomas Mathew. Ratio edits based on statistical tolerance intervals. *Journal of Official Statistics*, 31(1): 77–100, 2015. doi: 10.1515/jos-2015-0004. URL https://doi.org/10.1515/jos-2015-0004.

Derek S. Young and Thomas Mathew. Nonparametric Hyperrectangular Tolerance and Prediction Regions for Setting Multivariate Reference Regions in Laboratory Medicine. *Statistical Methods in Medical Research*, 29(22):3569–3585, 2020. doi: 10.1177/0962280220933910.