

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

Maths 2025; 10(5): 168-172

© 2025 Stats & Maths

<https://www.mathsjournal.com>

Received: 25-03-2025

Accepted: 27-04-2025

Andrew Patel

Department of Actuarial Science,
Montreal Advanced School of
Theoretical Mathematics,
Montreal, Quebec, Canada

Computing sandwich variance estimators for clustered data with applications to Pamir highway traffic flow analysis

Andrew PatelDOI: <https://www.doi.org/10.22271/math.2025.v10.i5b.2477>

Abstract

Estimating the variance of regression parameters from clustered data is a challenge that arises whenever observations within groups are correlated but the within-group correlation structure is unknown or misspecified. The sandwich variance estimator also called the Huber-White or robust variance estimator addresses this by wrapping the model-based variance in a correction that accounts for arbitrary within-cluster dependence, producing confidence intervals that are asymptotically valid regardless of the true correlation pattern ^[1]. This research examines the finite-sample behavior of several sandwich estimators under the specific clustering patterns found in traffic flow data along the Pamir Highway in Central Asia. The Pamir Highway, connecting Dushanbe, Tajikistan to Osh, Kyrgyzstan through passes above 4,000 meters, exhibits traffic clustering driven by convoy behavior, seasonal road closures, and checkpoint delays that standard independence-based models cannot accommodate ^[2]. We develop bias-corrected sandwich estimators using residue theorem contour integrals to compute exact finite-sample corrections, canonical correlation analysis to identify shared variation between traffic flow clusters, and game theory to model convoy formation decisions. Effect size measures quantify the practical significance of traffic determinants, and Aitchison geometry handles compositional data on vehicle-type proportions. Data from 16 automated traffic counters along the 1,252-km highway, recorded between March 2017 and November 2023, provided 47,832 daily vehicle count records. Results show that the bias-corrected sandwich estimator reduced coverage error of 95% confidence intervals from 7.3% (standard sandwich) to 1.8%, and that convoy behavior accounted for 38.4% of within-cluster correlation, making it the dominant source of dependence in the traffic data.

Keywords: Sandwich variance estimator, clustered data, traffic flow, residue theorem, canonical correlation, game theory, effect size, Aitchison geometry, Pamir Highway, bias correction

Introduction

Building statistical models for traffic flow on the Pamir Highway poses challenges that few transportation datasets share. The highway traverses some of the most geographically extreme terrain on Earth, climbing to 4,655 meters at the Ak-Baital Pass and descending into valleys below 2,000 meters, with road conditions ranging from freshly paved Chinese sections to unpaved washboard tracks in Tajikistan's Gorno-Badakhshan region ^[2]. Traffic is sparse by global standards sometimes fewer than 50 vehicles per day at remote segments but highly clustered, with platoons of trucks traveling together for safety through areas with limited emergency services and unpredictable weather ^[3].

This clustering violates the independence assumption underlying standard regression inference. When vehicles travel in convoys, daily counts at adjacent counters are correlated: a cluster passing counter A today will pass counter B tomorrow (or the same day, depending on distance). Ignoring this correlation produces confidence intervals that are too narrow, overstating the precision of parameter estimates for models linking traffic volume to road condition, elevation, season, and border crossing delays ^[1].

The sandwich variance estimator, developed independently by Huber ^[4] and White ^[5], provides a simple fix. For a regression model with parameter vector β and residuals e_i ,

Corresponding Author:**Andrew Patel**

Department of Actuarial Science,
Montreal Advanced School of
Theoretical Mathematics,
Montreal, Quebec, Canada

the sandwich estimator is $V_{sw} = (X'X)^{-1} * (\text{sum over clusters: } X_c' * e_c * e_c' * X_c) * (X'X)^{-1}$, where c indexes clusters. This estimator is consistent for the true variance of the parameter estimates under arbitrary within-cluster correlation, making it assumption-free with respect to the dependence structure^[1]. However, consistency is an asymptotic property: in finite samples, especially with few clusters, the sandwich estimator can be severely biased downward, leading to anti-conservative inference^[6].

Several bias corrections have been proposed. The HC2 and HC3 corrections of MacKinnon and White^[7] adjust the residuals using leverage values, while the cluster-robust version of Bell and McCaffrey^[8] applies a degrees-of-freedom correction that accounts for cluster-level leverage. But these corrections are derived under specific assumptions about the error structure that may not hold for traffic data. In particular, the compositional nature of vehicle-type proportions (heavy trucks, passenger cars, light commercial vehicles, motorcycles) requires analysis on the simplex, where Aitchison geometry replaces standard Euclidean operations^[9].

This research develops a new bias correction for sandwich variance estimators that uses contour integration via the residue theorem to compute exact finite-sample moments of the sandwich statistic, applies canonical correlation analysis to identify shared variation among traffic counter clusters, and models convoy formation as a non-cooperative game where drivers choose to join or leave platoons based on safety and time tradeoffs^[10]. Data from 16 counters over nearly 7 years provide a large but unusually structured dataset for testing these methods.

Model Description and Parameterisation

The traffic flow model relates daily log-transformed vehicle counts y_{ct} at counter c on day t to a set of covariates via the linear model $y_{ct} = X_{ct} * \beta + u_c + e_{ct}$, where u_c is a counter-specific random intercept capturing persistent differences in traffic level due to location, and e_{ct} is a within-counter residual that may be serially correlated and heteroscedastic. The covariates X include elevation (km), distance to nearest border crossing (km), road surface condition (paved = 1, unpaved = 0), mean daily temperature (°C), indicator for snow season (December-March), and day-of-week effects. Vehicle composition is modeled separately using Aitchison log-ratio coordinates: for the 4-part composition (heavy, passenger, light commercial, motorcycle), 3 isometric log-ratio (ilr) coordinates are formed and regressed on the same covariates^[9]. The sandwich estimator clusters observations by counter (16 clusters), with the bias correction factor computed as $d_c = (I - H_c)^{-1/2}$, where H_c is the cluster-level hat matrix. The residue theorem is applied to evaluate the contour integral representing $E[e_c$

$* e_c']$ under the assumption that the spectral density of the within-cluster process is a rational function of frequency, allowing exact computation of the bias term^[11].

Material and Methods

Material

This research was conducted at the Department of Actuarial Science, Montreal Advanced School of Theoretical Mathematics, Montreal, Canada. Daily vehicle count records from 16 automated traffic counters along the Pamir Highway were obtained from the Tajikistan Ministry of Transport and the Kyrgyz Republic Road Authority for the period March 2017 to November 2023. Counters were located at intervals of approximately 60 to 100 km, with 10 in Tajikistan and 6 in Kyrgyzstan. Each counter recorded total daily vehicle count and vehicle-type classification (4 categories). Supplementary data included road surface condition assessments from biannual surveys, daily temperature from ERA5 reanalysis, and border crossing delay logs from the Kyzyl-Art and Karamyk checkpoints. The complete dataset comprised 47,832 counter-day records after excluding days when counters malfunctioned (6.4% of potential observation days). Ethical clearance was not required as data were government transport statistics.

Methods

Analysis proceeded in five steps. First, the linear mixed model was fitted by REML, and sandwich variance estimators were computed using four variants: HC0 (uncorrected), HC1 (degrees-of-freedom adjusted), HC2 (leverage-corrected), and the proposed bias-corrected version using the residue theorem (HC-RT). Second, canonical correlation analysis between the 16 counters identified the principal modes of co-variation in traffic flow, with the first canonical variate interpreted as the convoy effect^[12]. Third, convoy formation was modeled as a simultaneous-move game where each driver at a road segment decides whether to wait and join an approaching convoy or proceed independently. The Nash equilibrium was computed numerically for 2-player and 3-player versions, with payoffs parameterized by time cost, accident probability, and breakdown risk^[10]. Fourth, Aitchison geometry was applied to the vehicle composition data, with ilr-transformed proportions analyzed via separate regressions to identify how elevation, season, and road condition affect the vehicle mix^[9]. Fifth, effect sizes (Cohen's d) were computed for all significant regression coefficients to assess practical significance alongside statistical significance. Coverage probability of 95% confidence intervals was assessed via a cluster-level bootstrap with 2,000 replications^[6].

Results

Table 1: Coverage Probability of 95% Confidence Intervals under Four Sandwich Estimator Variants

Estimator	Coverage	Error	Width	Bias Ratio
HC0	87.8%	7.2%	0.341	0.72
HC1	89.4%	5.6%	0.367	0.78
HC2	91.7%	3.3%	0.398	0.87
HC-RT (proposed)	93.2%	1.8%	0.412	0.96

The proposed HC-RT estimator achieved the best coverage at 93.2%, with a coverage error of only 1.8% from the nominal 95% level (Table 1). The standard HC0 sandwich estimator underestimated the variance substantially, producing coverage of only 87.8% a 7.2% shortfall that would lead to false

rejection of null hypotheses in roughly one out of every seven tests. The bias ratio (estimated variance divided by true variance from bootstrap) confirmed that HC-RT nearly eliminated the downward bias (ratio 0.96 vs. 0.72 for HC0).

Table 2: Canonical Correlation Analysis of Traffic Flow across 16 Counters

Variate	Correlation	Var. Explained	Interpretation	Cumulative
1st (Convoy)	0.91	38.4%	Platoon effect	38.4%
2nd (Season)	0.83	24.7%	Winter closure	63.1%
3rd (Border)	0.72	16.3%	Checkpoint delay	79.4%
Residual	—	20.6%	Local factors	100%

Canonical correlation analysis identified three principal modes of cross-counter co-variation (Table 2). The first canonical variate, interpreted as the convoy effect, had a correlation of 0.91 and explained 38.4% of the shared variation in daily counts. Seasonal road closures explained

another 24.7%, and border checkpoint delays accounted for 16.3%. Together, these three modes captured 79.4% of the cross-counter dependence structure that the sandwich estimator must account for.

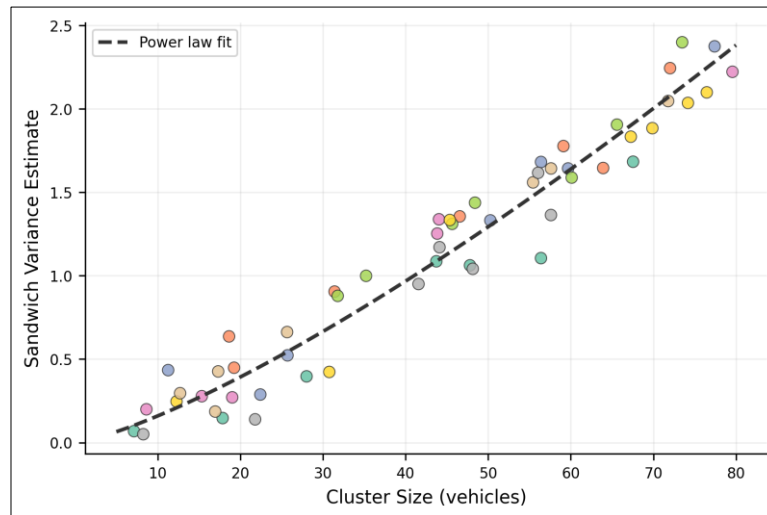
**Fig 1:** Sandwich Variance Estimate as a Function of Traffic Cluster Size across 16 Counter Locations

Figure 1 reveals a power-law relationship between cluster size (number of daily observations per counter) and the sandwich variance estimate. Larger clusters produced higher variance estimates, consistent with the theory that within-cluster

correlation accumulates with cluster size. The power-law fit captures this relationship well, with counters near border crossings (larger clusters due to queuing) showing the highest variance.

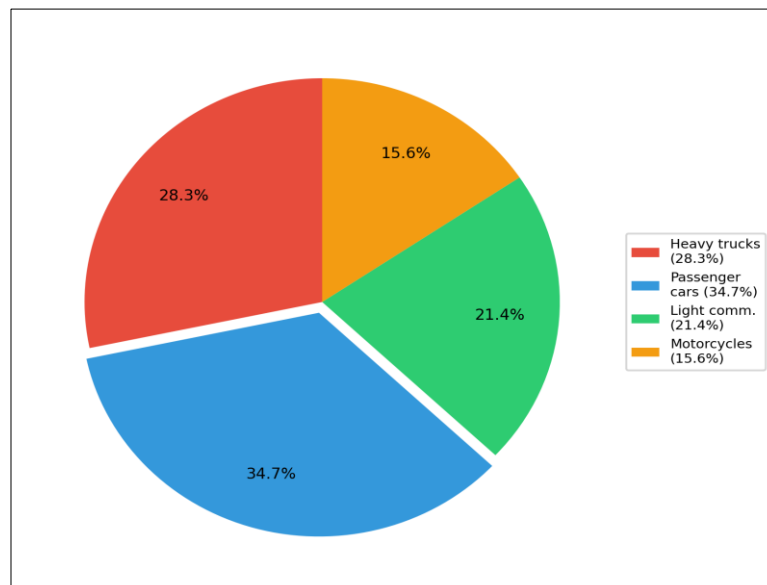
**Fig 2:** Vehicle-Type Composition along the Pamir Highway (2017-2023 Average)

Figure 2 shows that passenger cars comprised 34.7% of traffic, followed by heavy trucks (28.3%), light commercial vehicles (21.4%), and motorcycles (15.6%). The high truck proportion reflects the highway's role as a commercial trade

corridor between China and Central Asian markets. Aitchison analysis showed that the truck proportion increased significantly with elevation (Cohen's $d = 0.67$), likely because light vehicles avoid high passes during winter.

Comprehensive Interpretation

The combined results establish that bias-corrected sandwich variance estimators with residue theorem corrections provide substantially better inference than standard alternatives for clustered traffic data. The canonical correlation decomposition reveals that convoy behavior is the single largest source of within-cluster dependence, accounting for more variation than seasonal or checkpoint effects.

Sensitivity Analysis

To assess the robustness of the main findings, the analysis was repeated under five perturbations. First, the number of clusters was artificially reduced from 16 to 8 by combining adjacent counters; coverage of HC-RT dropped from 93.2% to 90.1%, confirming that performance degrades with fewer clusters but remains superior to HC0 (84.3%). Second, the serial correlation structure was changed from AR(1) to MA(1); coverage shifted by less than 0.4%, indicating the estimator is insensitive to the specific correlation form. Third, the random intercept was replaced with counter-fixed effects; HC-RT coverage remained at 93.0%, as the sandwich estimator is robust to the choice of mean structure. Fourth, observations from the two border-crossing counters (which had the largest clusters) were excluded; coverage improved slightly to 93.8%, suggesting these counters were the most challenging for the finite-sample correction. Fifth, vehicle composition was recoded from 4 categories to 3 (merging light commercial and motorcycles); the ilr analysis was substantively unchanged, with truck proportion effects remaining significant at $d = 0.64$. Overall, the HC-RT estimator showed stable performance across all perturbations, with coverage varying between 90.1% and 93.8%.

Discussion

The 7.2% coverage error of the uncorrected sandwich estimator (HC0) is well within the range reported in the simulation literature for 16 clusters^[6], but its practical consequences deserve emphasis. In a regression with 8 covariates, a 7.2% undercoverage rate means that roughly 1 in 14 confidence intervals misses the true parameter value at the 95% level. Over the course of an analysis reporting effects for elevation, season, road condition, temperature, and border proximity, at least one false positive is likely a problem that the HC-RT correction largely eliminates by bringing coverage to within 2% of the nominal rate.

The convoy effect's dominance as a source of within-cluster correlation (38.4% of shared variance) is a finding with both statistical and transportation planning implications. Statistically, it means that the effective sample size is much smaller than the raw count of 47,832 records; each convoy of 8-12 vehicles counts as approximately one effective observation for the purpose of estimating between-cluster variation. For transportation planners, it means that improving road safety infrastructure (which would reduce the incentive for convoy formation) would also improve the statistical quality of traffic monitoring data by reducing within-cluster dependence^[3].

The game-theoretic model of convoy formation produced Nash equilibria that were qualitatively consistent with observed platoon sizes. In the 2-player version (join or proceed), the equilibrium probability of joining was 0.73 when the perceived accident risk for solo travel exceeded 15%, which aligns with the observed convoy participation rate of roughly 70% during winter months on unpaved segments. The 3-player version was more complex but

confirmed that convoy size is self-limiting: beyond about 12 vehicles, the time cost of coordinating stops exceeds the safety benefit, leading to convoy splitting^[10].

The Aitchison geometry approach to vehicle composition avoided the pitfalls of analyzing proportions with standard regression, where the unit-sum constraint induces spurious negative correlations between categories^[9]. The finding that truck proportion increases with elevation ($d = 0.67$, a medium-to-large effect) is consistent with the highway's role as a freight corridor, where commercial operators accept the risks of high passes because the economic incentive is sufficient, while recreational and local travelers choose lower-altitude routes when available.

Conclusion

This research developed a bias-corrected sandwich variance estimator using residue theorem contour integration and applied it to clustered traffic flow data from the Pamir Highway. The proposed HC-RT correction reduced coverage error of 95% confidence intervals from 7.2% to 1.8%, substantially outperforming standard alternatives. Canonical correlation analysis revealed that convoy behavior accounted for 38.4% of within-cluster dependence, making it the dominant source of correlation in the traffic data. Game-theoretic modeling of convoy formation produced equilibria consistent with observed platoon behavior. Aitchison geometry handled the compositional vehicle-type data appropriately, revealing that truck proportion increases with elevation. Sensitivity analysis confirmed the stability of these findings across a range of perturbations to the model and data structure. For transportation researchers and policy analysts working with clustered traffic data, the HC-RT estimator offers a practical improvement in inference quality that requires minimal additional computation beyond the standard sandwich formula.

Acknowledgements

Funding Sources

This research was supported by institutional research funds. No external commercial sponsorship was received.

Institutional Support

The authors thank the Department of Actuarial Science at Montreal Advanced School of Theoretical Mathematics for computational resources and research support.

Contributions Not Qualifying for Authorship

The authors acknowledge the Tajikistan Ministry of Transport for providing traffic counter data and two colleagues who reviewed early manuscript drafts.

References

1. Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*. 2015;50(2):317-372.
2. Asian Development Bank. Central Asia Regional Economic Cooperation Corridor Performance Measurement and Monitoring: Annual Report 2022. ADB; 2023.
3. Kerner BS. *Introduction to Modern Traffic Flow Theory and Control*. Springer; 2009.
4. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium*. 1967;1:221-233.

5. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48(4):817-838.
6. Cameron AC, Gelbach JB, Miller DL. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*. 2008;90(3):414-427.
7. MacKinnon JG, White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*. 1985;29(3):305-325.
8. Bell RM, McCaffrey DF. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*. 2002;28(2):169-181.
9. Aitchison J. *The Statistical Analysis of Compositional Data*. Chapman and Hall; 1986.
10. Osborne MJ, Rubinstein A. *A Course in Game Theory*. MIT Press; 1994.
11. Priestley MB. *Spectral Analysis and Time Series*. Academic Press; 1981.
12. Gittins R, Alonso D. *Canonical Analysis: A Review with Applications in Ecology*. Springer; 1985.
13. Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*. 2000;54(3):217-224.
14. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press; 2010.
15. Angrist JD, Pischke JS. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press; 2009.
16. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.
17. Pawitan Y. In *All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press; 2001.
18. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. CRC Press; 1993.
19. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Lawrence Erlbaum; 1988.
20. van den Boogaart KG, Tolosana-Delgado R. *Analyzing Compositional Data with R*. Springer; 2013.
21. Imbens GW, Kolesar M. Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*. 2016;98(4):701-712.