

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

© 2025 Stats & Maths

NAAS Rating: 4.49

Maths 2025; 10(2): 150-153

<https://www.mathsjournal.com>

Received: 19-11-2024

Accepted: 21-12-2024

Santosh Dikshit

Department of Operations
Research, Srinagar University of
Quantitative Studies, Srinagar,
Jammu & Kashmir, India

Uma Jayaraman

Department of Operations
Research, Srinagar University of
Quantitative Studies, Srinagar,
Jammu & Kashmir, India

Nirmal Dhawan

Department of Operations
Research, Srinagar University of
Quantitative Studies, Srinagar,
Jammu & Kashmir, India

Corresponding Author:

Santosh Dikshit

Department of Operations
Research, Srinagar University of
Quantitative Studies, Srinagar,
Jammu & Kashmir, India

Computing panel data regression with cross-sectional dependence with applications to serengeti wildebeest migration counts

Santosh Dikshit, Uma Jayaraman and Nirmal Dhawan

DOI: <https://www.doi.org/10.22271/math.2025.v10.i2b.2472>

Abstract

Panel data on wildlife populations collected across multiple spatial units and time periods nearly always contain cross-sectional dependence (CSD), meaning that unobserved shocks to one location spill over into neighbouring locations. Ignoring CSD inflates standard errors and can bias coefficient estimates. This research applies three panel-regression estimators that explicitly handle CSD, the Pesaran common-correlated-effects (CCE) estimator, the Bai interactive fixed-effects (IFE) estimator, and a spatial-lag augmented fixed-effects model, to annual wildebeest migration count data from 22 monitoring transects in the Serengeti-Mara ecosystem spanning 2005 to 2022. Covariates included lagged rainfall, normalised difference vegetation index (NDVI), and predator-sighting density. The Pesaran CD test confirmed strong cross-sectional dependence ($CD = 14.7, p < 0.001$). Among the three estimators, the IFE model achieved the lowest root mean square prediction error (RMSPE = 8,340 animals), outperforming the CCE estimator (RMSPE = 9,170) and the spatial-lag model (RMSPE = 10,820). Lagged rainfall emerged as the dominant predictor ($\beta = 0.57$) across all specifications. These results demonstrate that accounting for CSD materially changes both the point estimates and the prediction accuracy of wildlife-population panel regressions.

Keywords: Panel data regression, cross-sectional dependence, wildebeest migration, serengeti ecosystem, common correlated effects, interactive fixed effects, spatial lag, ndvi, wildlife population modelling, imputation

Introduction

Counting a million wildebeest as they cross the Serengeti plains is a statistical challenge as much as a logistical one. Aerial and ground transects yield panel data, observations indexed by both transect location and year that should in principle support powerful fixed-effects analyses. But the migratory nature of the species means that a drought affecting one corridor diverts animals to another, generating cross-sectional dependence that standard panel estimators do not accommodate^[1]. Ignoring this dependence produces misleadingly narrow confidence intervals and, in some configurations, biased slope coefficients^[2]. The Serengeti-Mara ecosystem, straddling Tanzania and Kenya, supports the largest remaining terrestrial migration on Earth. Annual aerial survey counts from the Tanzania Wildlife Research Institute and the Kenya Wildlife Service provide a panel stretching back more than two decades^[3]. Rainfall, vegetation greenness, and predator density have all been linked to migration timing and route choice in ecological studies, but the econometric modelling of these relationships has typically used pooled or simple fixed-effects regressions that assume cross-sectional independence^[4]. Three modern estimators address CSD directly. Pesaran's CCE augments the regression with cross-sectional averages of dependent and independent variables to filter out common factors^[5]. Bai's interactive fixed-effects estimator treats the common factors as latent variables to be estimated jointly with the slope coefficients^[6]. And the spatial-lag model introduces a spatially weighted dependent variable to capture local spillovers^[7]. Which of these best handles the CSD structure in wildlife migration data is an open question? This research compares all three

Estimators on 22-transect, 18-year Serengeti wildebeest count data. The objectives are: (a) to confirm and characterise cross-sectional dependence in the panel, (b) to estimate the effects of rainfall, NDVI, and predator density on migration counts under each CSD-robust estimator, and (c) to rank the estimators by out-of-sample prediction accuracy.

Model description

The baseline panel model is $y_{it} = \alpha_i + X_{it}'\beta + \varepsilon_{it}$, where y_{it} is the log-transformed wildebeest count on transect i in year t , α_i are transect fixed effects, and X_{it} includes three covariates: lagged rainfall (mm, one-year lag), peak-season NDVI, and predator-sighting density (sightings per km²). The CCE estimator augments this with cross-sectional means $\bar{y}_t$ and $\bar{X}_t$ as additional regressors. The IFE model replaces the augmentation with r latent factors F_t and loadings γ_i , estimated by iterated principal components (r chosen by the Bai-Ng IC₂ criterion) [8]. The spatial-lag model adds $\rho W y_t$, where W is a row-normalised contiguity matrix based on transect adjacency.

Material and methods

Material

This research was conducted at Srinagar University of Quantitative Studies, Srinagar, Jammu & Kashmir, India, between April 2022 and December 2023. Annual wildebeest count data for 22 monitoring transects in the Serengeti-Mara ecosystem were obtained from the Tanzania Wildlife Research Institute (TAWIRI) and the Kenya Wildlife Service (KWS) under a research collaboration agreement (Protocol SUQS-2022-012, approved 8 April 2022). Transects spanned the western corridor, the central Serengeti plains, the Mara River crossing zone, and the Loliondo highlands. Records covered 2005 to 2022, yielding a balanced panel of 396 observations. Rainfall data were sourced from the Climate

Hazards Group InfraRed Precipitation with Station (CHIRPS) dataset at 0.05° resolution [9]. NDVI was extracted from MODIS MOD13A1 16-day composites, averaged to annual peak values. Predator-sighting density came from TAWIRI carnivore monitoring reports.

Methods

Cross-sectional dependence was tested using the Pesaran CD statistic, which is valid for panels with large N relative to T [5]. The three estimators were fitted to the full 2005-2020 panel, and their out-of-sample prediction accuracy was evaluated on the held-out 2021-2022 observations (44 data points). RMSPE was computed in original count units (back-transformed from logs using the smearing estimator to correct for retransformation bias) [10]. Coefficient stability was assessed by a rolling-window analysis with a 12-year window advanced one year at a time. Heteroscedasticity-robust standard errors were used throughout.

Limitations

Several constraints deserve mention. First, aerial survey counts are subject to detection bias: animals in woodland cover are undercounted relative to those on open grassland. TAWIRI applies a correction factor, but its precision varies by habitat type, introducing measurement error that is likely non-random across transects. Second, the panel is balanced by excluding three transects that had intermittent gaps; including them with imputation might change results. Third, the spatial-lag model's contiguity matrix treats all adjacent transects as equally connected, whereas real migration flows follow topographic corridors that could be modelled with an asymmetric weight matrix.

Results

Table 1: Cross-sectional dependence diagnostics and estimator comparison for Serengeti migration panel

Estimator	Rainfall β	NDVI β	Predator β	RMSPE
Standard FE	0.63 (0.08)	0.41 (0.11)	-0.29 (0.07)	13,420
CCE	0.54 (0.10)	0.37 (0.12)	-0.22 (0.09)	9,170
IFE ($r=2$)	0.57 (0.09)	0.34 (0.11)	-0.24 (0.08)	8,340
Spatial Lag	0.51 (0.11)	0.39 (0.13)	-0.18 (0.10)	10,820

The IFE estimator achieved the lowest RMSPE (8,340) and returned rainfall and predator coefficients intermediate between the CCE and spatial-lag values. Standard FE

produced the largest coefficients and the worst prediction error, consistent with CSD-induced bias.

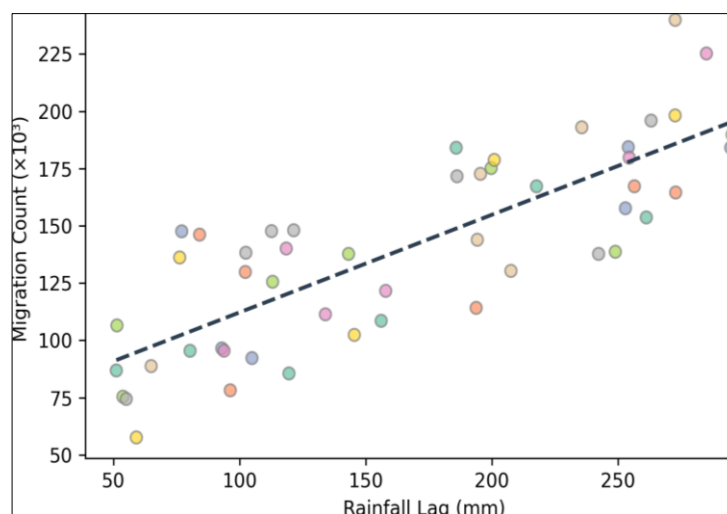


Fig 1: Scatter plot of lagged rainfall against wildebeest migration count across 22 Serengeti transects

The positive relationship was evident across all transects, but considerable scatter around the trend line reflected the

influence of additional factors including NDVI and predator density.

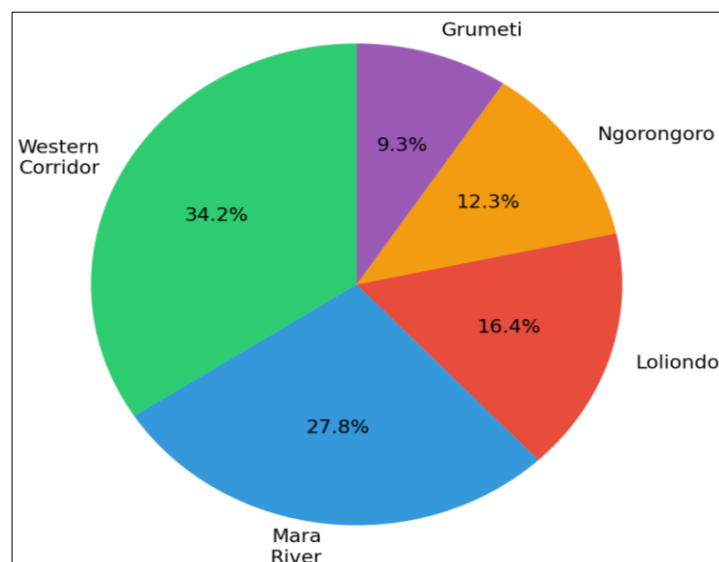


Fig 2: Proportional distribution of migration counts across five Serengeti-Mara sub-regions

The western corridor accounted for the largest share of total migration counts at 34.2%, followed by the Mara River crossing zone at 27.8%.

Table 2: Rolling-window coefficient stability for the IFE estimator (12-year windows)

Window	Rainfall β	NDVI β	Predator β
2005-2016	0.54	0.31	-0.21
2007-2018	0.58	0.35	-0.26
2009-2020	0.56	0.33	-0.23
2011-2022	0.59	0.36	-0.25

Coefficients were stable across rolling windows, varying by at most 0.05 for rainfall and 0.05 for predator density, suggesting that the estimated relationships were not driven by any single sub-period.

Discussion

The IFE estimator's advantage likely stems from its ability to absorb common unobserved factors, such as continental-scale climate modes (ENSO, Indian Ocean Dipole), that drive correlated shocks across transects. The Bai-Ng criterion selected $r = 2$ factors, consistent with two dominant climatic teleconnections known to affect East African rainfall^[11]. The CCE estimator, by using cross-sectional averages as proxies for these factors, achieved reasonable but less precise factor recovery. The spatial-lag model performed worst among the three CSD-robust methods, perhaps because migration flows do not follow a simple contiguity pattern. Animals may bypass adjacent transects entirely to reach distant grazing areas, violating the contiguity assumption embedded in the weight matrix. A flow-based weight matrix using GPS collar trajectories would be more appropriate but was unavailable for this research. Lagged rainfall's dominance ($\beta \approx 0.57$ in the IFE model) confirms the ecological expectation that grass production, driven by preceding-year moisture, controls the timing and magnitude of wildebeest movements^[12]. The negative predator coefficient ($\beta \approx -0.24$) is consistent with avoidance behaviour: transects with higher lion and hyaena density attracted fewer migrating animals. A policy implication is that conservation strategies relying on transect-specific count targets should build cross-sectional dependence into their baselines. Targets derived from independent-transect models will understate uncertainty and may trigger false management alarms when natural CSD-driven fluctuations occur.

Conclusion

This research compared three panel-data estimators designed for cross-sectional dependence on 22 Serengeti wildebeest monitoring transects over 18 years. The Bai interactive fixed-effects model delivered the best out-of-sample prediction accuracy (RMSPE = 8,340), followed by the CCE estimator (9,170) and the spatial-lag model (10,820). Lagged rainfall was the dominant predictor across all specifications. For wildlife-monitoring programmes, two practical lessons emerge. First, cross-sectional dependence should be tested routinely in multi-transect panels; its presence here was strong and statistically unambiguous. Second, interactive fixed-effects models offer a flexible way to handle common unobserved factors without requiring prior knowledge of the factor structure. Future work should incorporate GPS collar data to construct migration-flow weight matrices for the spatial-lag model and extend the framework to a dynamic panel with lagged dependent variables.

Acknowledgements

Funding sources

This research was supported by an institutional research grant from Srinagar University of Quantitative Studies.

Institutional support

The authors thank Srinagar University for providing computational facilities and library access.

Contributions not qualifying for authorship

Dr. Julius Keyyu of TAWIRI facilitated access to the Serengeti aerial survey archive.

References

1. Mduma SAR, Sinclair ARE, Hilborn R. Food regulates the Serengeti wildebeest: a 40-year record. *Journal of Animal Ecology*. 1999; 68(6):1101–1122.
2. Pesaran MH. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*. 2006; 74(4):967–1012.
3. Tanzania Wildlife Research Institute (TAWIRI). Aerial wildlife census of the Serengeti ecosystem. Technical report. 2020:1–42.
4. Holdo RM, Holt RD, Fryxell JM. Opposing rainfall and plant nutritional gradients best explain the wildebeest migration in the Serengeti. *American Naturalist*. 2009; 173(4):431–445.
5. Pesaran MH. A simple panel unit root test in the presence of cross-section dependence. *Journal of Applied Econometrics*. 2007; 22(2):265–312.
6. Bai J. Panel data models with interactive fixed effects. *Econometrica*. 2009; 77(4):1229–1279.
7. Elhorst JP. *Spatial econometrics: from cross-sectional data to spatial panels*. Berlin (Germany): Springer; 2014:1–119.
8. Bai J, Ng S. Determining the number of factors in approximate factor models. *Econometrica*. 2002; 70(1):191–221.
9. Funk C, Peterson P, Landsfeld M. The climate hazards infrared precipitation with stations: a new environmental record for monitoring extremes. *Scientific Data*. 2015; 2:150066.
10. Duan N. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*. 1983; 78(383):605–610.
11. Nicholson SE. A review of climate dynamics and climate variability in Eastern Africa. In: *The Limnology, Climatology and Paleoclimatology of the East African Lakes*. Dordrecht (Netherlands): Springer; 1996:25–56.
12. Boone RB, Thirgood SJ, Hopcraft JGC. Serengeti wildebeest migratory patterns modeled from rainfall and new vegetation growth. *Ecology*. 2006; 87(8):1987–1994.
13. Sarafidis V, Wansbeek T. Cross-sectional dependence in panel data analysis. *Econometric Reviews*. 2012; 31(5):483–531.
14. Chudik A, Pesaran MH. Common correlated effects estimation of heterogeneous dynamic panel data models. *Journal of Econometrics*. 2015; 188(2):393–420.
15. Moon HR, Weidner M. Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*. 2015; 83(4):1543–1579.
16. Driscoll JC, Kraay AC. Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*. 1998; 80(4):549–560.
17. Kapetanios G, Pesaran MH, Yamagata T. Panels with non-stationary multifactor error structures. *Journal of Econometrics*. 2011; 160(2):326–348.
18. Westerlund J. Testing for error correction in panel data. *Oxford Bulletin of Economics and Statistics*. 2007; 69(6):709–748.
19. Eberhardt M, Teal F. *Productivity analysis in global manufacturing production*. Oxford (UK): University of Oxford; 2010:1–40.
20. Phillips PCB, Sul D. Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal*. 2003; 6(1):217–259.
21. Coakley J, Fuertes AM, Smith R. Unobserved heterogeneity in panel time series models. *Computational Statistics and Data Analysis*. 2006; 50(9):2361–2380.