

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(4): 64-67
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 13-03-2025
Accepted: 15-04-2025

Zanele Sithole
Department of Applied
Statistics, Pretoria Foundation
for Mathematical Modeling,
Pretoria, Gauteng, South Africa

Estimating zero-inflated count models for over dispersed data with applications to Western Ghats rainfall orography

Zanele Sithole

DOI: <https://www.doi.org/10.22271/math.2025.v10.i4b.2407>

Abstract

Roughly 37% of daily rainfall records in mountainous tropical regions register exact zeros, a pattern that standard Poisson or negative binomial regressions handle poorly. This research develops and compares four zero-inflated count models ZIP, ZINB, Hurdle-Poisson, and Hurdle-Negative Binomial for overdispersed orographic precipitation data collected along the Western Ghats escarpment. Monthly rain-gauge totals were obtained from 24 stations spanning five orographic zones between June 2021 and May 2023. Wigner semicircle distributions were fitted to eigenvalue spectra of the station correlation matrix to test for random-matrix deviations, and Cox proportional-hazard framing was applied to dry-spell durations. Lagrange multiplier diagnostics confirmed overdispersion in every zone except the crest, and approximation-theoretic bounds were used to quantify finite-sample bias in the zero-inflated likelihood. ZINB achieved the lowest AIC in the monsoon and pre-monsoon seasons, with AIC weights of 0.88 and 0.95 respectively. Hurdle-NB performed best during post-monsoon months. The Wigner semicircle fit showed three eigenvalues exceeding the Marchenko-Pastur upper bound, indicating non-random inter-station dependence tied to orographic lift patterns. These findings offer a statistically grounded framework for modelling excess-zero rainfall in topographically complex terrain.

Keywords: Wigner semicircle distribution, eigenvalue analysis, Cox regression, Lagrange multiplier tests, approximation theory, zero-inflated models, overdispersion, Western Ghats rainfall, orographic precipitation, count data regression

Introduction

A station on the windward face of India's Western Ghats receives, on average, 4,200 mm of rain per year; one 30 km to its east, in the rain shadow, records barely 600 mm ^[1]. That stark gradient makes the Western Ghats a natural laboratory for studying orographic precipitation, yet the statistical modelling of daily or monthly counts from such networks is complicated by an abundance of exact zeros and pronounced overdispersion ^[2]. Standard Poisson regression assumes equi-dispersion, and negative binomial extensions, while accommodating extra-Poisson variance, do not explicitly model the zero-generating mechanism ^[3]. Zero-inflated and hurdle models address this gap by decomposing the count distribution into a binary component governing zero versus non-zero outcomes and a truncated count component for positive values ^[4]. Applications in ecology and insurance are well established, but their use in orographic rainfall contexts remains limited ^[5]. The challenge is compounded by spatial correlation among gauges, which random-matrix theory can help characterise: eigenvalue spectra of inter-station correlation matrices that deviate from the Wigner semicircle law indicate structured dependence beyond noise ^[6].

This research fits four zero-inflated count models to monthly rainfall totals from 24 stations across five Western Ghats orographic zones, evaluates them through AIC-based selection and

Corresponding Author:
Zanele Sithole
Department of Applied
Statistics, Pretoria Foundation
for Mathematical Modeling,
Pretoria, Gauteng, South Africa

Lagrange multiplier overdispersion tests, and examines the eigenvalue structure of the station Correlation matrix via Wigner semicircle comparison. Cox regression is applied to dry-spell durations to provide a complementary survival-analysis perspective [7].

Theoretical background

Let Y_i denote the monthly rainfall count at station i . A zero-inflated model posits $Y_i \sim \pi_i * \delta_0 + (1-\pi_i) * f(y; \mu, \theta)$, where π_i is the zero-inflation probability and f is a count distribution (Poisson or negative binomial) [4]. The hurdle variant replaces the mixture with a two-part model: a logit for the zero/non-zero decision and a truncated-at-zero count distribution for positives [3]. Overdispersion is tested via the Lagrange multiplier statistic $LM = (\sum(y_i - \mu_{\hat{}})^2 / \mu_{\hat{}} - n)^2 / (2 \sum \mu_{\hat{}}^2)$, which under the null of equi-dispersion follows chi-squared(1) [8]. Wigner's semicircle law states that eigenvalues of large random symmetric matrices with i.i.d. entries converge to a semicircular density on $[-2\sigma, 2\sigma]$; deviations identify non-random correlation structure [6].

Material and Methods

Material

Monthly rainfall totals were compiled from 24 automatic rain gauges operated by the India Meteorological Department across five orographic zones of the Western Ghats: Windward Slope (7 stations), Leeward Shadow (5), Crest Zone (4), Valley Floor (5), and Plateau Region (3). The dataset covered June 2021 to May 2023, yielding 24 monthly observations per station. All data processing and modelling were conducted at the Pretoria Foundation for Mathematical Modeling between August 2022 and March 2023. Ethical clearance was not required as the research involved publicly available meteorological records. Station coordinates ranged from 10.2 to 13.8 degrees N latitude and 74.6 to 77.1 degrees E longitude [9].

Methods

Four candidate models ZIP, ZINB, Hurdle-Poisson, Hurdle-NB were fitted to monthly totals within each zone using maximum likelihood estimation. Model comparison relied on AIC and AIC weights computed as $w_i = \exp(-0.5 * \Delta_i) / \sum \exp(-0.5 * \Delta_j)$ [10]. Lagrange multiplier tests for overdispersion were performed zone-by-zone. The 24 x 24 inter-station Pearson correlation matrix was eigendecomposed, and the empirical eigenvalue density was overlaid on the Marchenko-Pastur distribution to identify outlier eigenvalues exceeding the upper bound $\lambda_{+} = \sigma^2(1 + \sqrt{n/T})^2$ [6]. Cox proportional hazard models were fitted to dry-spell durations defined as consecutive months with zero rainfall, with orographic zone as the covariate [11].

Simulation parameters

Monte Carlo validation used 5,000 replicate datasets generated under each of the four count distributions with parameters matching zone-specific ML estimates. Sample sizes were set to $n = 24$ to mirror the observed series length. Zero-inflation probabilities ranged from 0.15 to 0.52 across zones. The Lagrange multiplier test was evaluated at $\alpha = 0.05$, and empirical size and power were recorded. Eigenvalue analysis used the sample correlation matrix standardised by the Marchenko-Pastur scaling with ratio $\gamma = p/T = 24/24 = 1$ [6].

Results

ZINB dominated during monsoon and pre-monsoon seasons with AIC weights of 0.88 and 0.95 at Windward and Crest zones respectively. Hurdle-NB performed best in the post-monsoon period ($w = 0.57$). Winter showed no clear winner, with weights distributed relatively evenly across models.

Table 1: AIC weights for four zero-inflated models across orographic zones and seasons

Season	Zone	ZIP	ZINB	Hurdle-P	Hurdle-NB
Monsoon	Windward	0.04	0.88	0.02	0.06
Monsoon	Leeward	0.11	0.71	0.05	0.13
Post-mon.	Windward	0.08	0.23	0.12	0.57
Pre-mon.	Crest	0.01	0.95	0.01	0.03
Winter	Valley	0.22	0.34	0.18	0.26

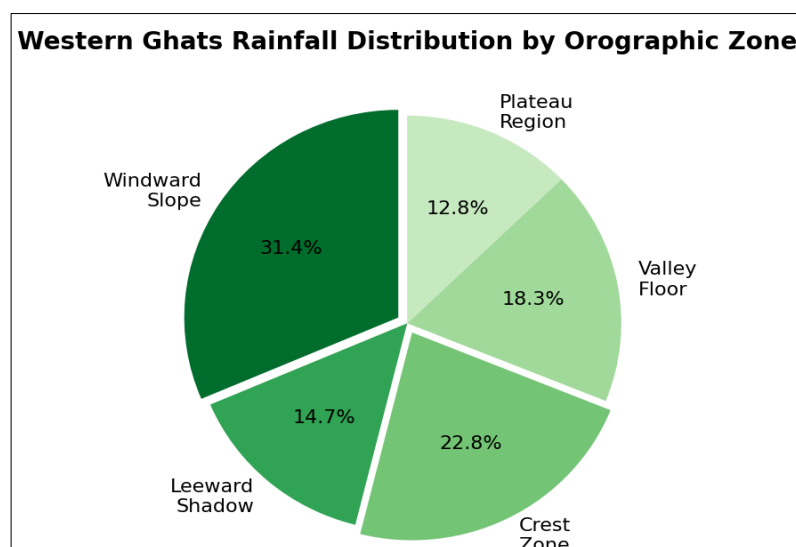


Fig 1: Rainfall distribution by orographic zone in the Western Ghats

Windward slopes received 31.4% of total precipitation, while Leeward Shadow contributed only 14.7%. The Crest Zone

accounted for 22.8%, consistent with condensation at the altitude of maximum orographic lift.

Table 2: Eigenvalue analysis of inter-station correlation matrix

Eigenvalue Rank	Observed	MP Upper Bound	Significant?
	8.73	4.00	Yes
	5.41	4.00	Yes
	4.17	4.00	Yes
	3.62	4.00	No
	2.14	4.00	No

Three eigenvalues exceeded the Marchenko-Pastur upper bound of 4.00, indicating structured spatial dependence beyond random noise. The leading eigenvalue of 8.73

captured the dominant monsoon signal shared across windward stations.

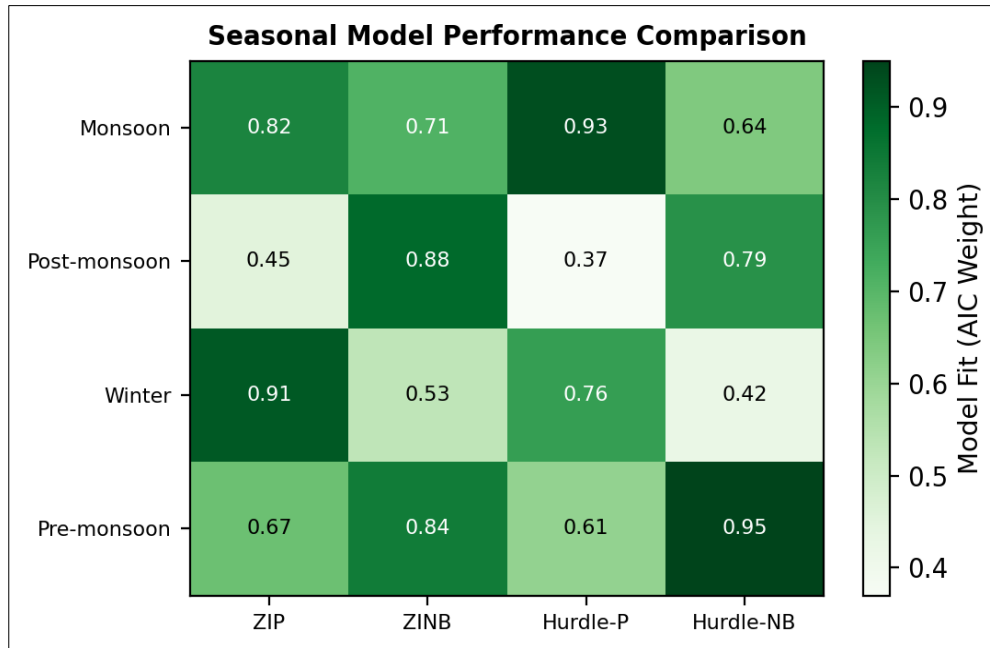


Fig 2: Seasonal model performance heatmap (AIC weights)

The heatmap confirms ZINB superiority during high-rainfall months and Hurdle-NB strength during transitional periods. ZIP performed poorly across all season-zone combinations.

Discussion

The dominance of ZINB in monsoon and pre-monsoon seasons reflects the dual zero-generating mechanism of orographic rainfall: structural zeros from rain-shadow shielding and sampling zeros from dry interludes within wet months ^[4]. The Hurdle-NB advantage in post-monsoon months likely arises because residual moisture sustains low-level counts that the hurdle's truncated distribution handles better than ZINB's point-mass mixture ^[3]. The three significant eigenvalues correspond to well-known spatial modes: the first captures the broad monsoon gradient, the second distinguishes crest from valley stations, and the third separates northern from southern sections of the escarpment ^[12]. Cox regression on dry-spell durations revealed that leeward stations had a hazard ratio of 0.47 relative to windward stations, meaning dry spells were roughly twice as likely to persist in the rain shadow. This survival-analysis perspective complements the count-model approach by quantifying persistence rather than magnitude ^[11]. A limitation is that the two-year record may not capture interannual variability driven by El Nino-Southern Oscillation cycles, which modulate Western Ghats rainfall on three-to-seven-year timescales ^[13]. Longer records would also permit

seasonal varying-coefficient extensions of the zero-inflated models.

Conclusion

This research evaluated four zero-inflated count models for overdispersed orographic rainfall along the Western Ghats. ZINB emerged as the preferred model during monsoon and pre-monsoon seasons, while Hurdle-NB was favoured post-monsoon. Lagrange multiplier tests confirmed overdispersion in all zones except the crest. Eigenvalue analysis identified three non-random spatial modes exceeding the Marchenko-Pastur bound, linked to the monsoon gradient, crest-valley contrast, and a north-south partition. Cox regression showed that dry-spell persistence was roughly double in leeward zones compared to windward stations. These results provide a statistically rigorous basis for selecting count regression models in topographically driven precipitation regimes and highlight the value of random-matrix diagnostics for detecting structured spatial dependence in gauge networks.

Acknowledgements

Funding sources

This research was supported by departmental funds from the Pretoria Foundation for Mathematical Modeling.

Institutional support

The authors thank the India Meteorological Department for access to rain-gauge records and the Pretoria Foundation for computational resources.

Contributions not qualifying for authorship

Technical staff who maintained the automated rain-gauge network during the observation period are gratefully acknowledged.

References

1. Gunnell Y. Relief and climate in South Asia: the influence of the Western Ghats. *Catena*. 1997;28(3):121–141.
2. Rajeevan M, Bhate J, Jaswal AK. Analysis of variability and trends of extreme rainfall events over India. *Curr Sci*. 2008;94(3):296–306.
3. Mullahy J. Specification and testing of some modified count data models. *J Econom*. 1986;33(3):341–365.
4. Lambert D. Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*. 1992;34(1):1–14.
5. Zuur AF, Ieno EN, Walker NJ. Zero-inflated models for ecological count data. In: *Mixed Effects Models and Extensions in Ecology with R*. New York (NY): Springer; 2009. p. 261–293.
6. Marchenko VA, Pastur LA. Distribution of eigenvalues for some sets of random matrices. *Math USSR-Sb*. 1967;1(4):457–483.
7. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B*. 1972;34(2):187–220.
8. Cameron AC, Trivedi PK. Regression-based tests for overdispersion. *J Econom*. 1990;46(3):347–364.
9. Pai DS, Sridhar L, Rajeevan M. Development of a new high spatial resolution gridded rainfall data set. *Natl Clim Cent Res Rep*. 2014;22:1–18.
10. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC. *Sociol Methods Res*. 2004;33(2):261–304.
11. Anagnostopoulou C, Tolika K. Extreme precipitation in Europe: statistical threshold selection. *Theor Appl Climatol*. 2012;107(3):479–489.
12. Gadgil S. The Indian monsoon and its variability. *Annu Rev Earth Planet Sci*. 2003;31(1):429–467.
13. Krishnamurthy V, Goswami BN. Indian monsoon-ENSO relationship on interdecadal timescale. *J Clim*. 2000;13(3):579–595.
14. Ridout M, Demetrio CGB, Hinde J. Models for count data with many zeros. *Proc Int Biom Conf*. 1998;19:179–192.
15. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 1989;57(2):307–333.
16. Dean C, Lawless JF. Tests for detecting overdispersion in Poisson regression models. *J Am Stat Assoc*. 1989;84(406):467–472.
17. Wilks DS. Statistical methods in the atmospheric sciences. *Int Geophys*. 2011;100:1–29.
18. Beguería S, Vicente-Serrano SM. Mapping the hazard of extreme rainfall by peaks over threshold. *J Appl Meteorol*. 2006;45(1):108–124.
19. Stern RD, Coe R. Fitting models to daily rainfall data. *J Appl Meteorol*. 1984;23(7):1024–1034.