

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(3): 235-238
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 06-05-2024
Accepted: 11-06-2024

Ashok Shetty
Department of Combinatorics
and Optimization, Varanasi
Academy of Statistical Research,
Varanasi, Uttar Pradesh, India

Usha Gowda
Department of Combinatorics
and Optimization, Varanasi
Academy of Statistical Research,
Varanasi, Uttar Pradesh, India

Bounding generalized linear models for count data with applications to New Zealand geothermal energy output

Ashok Shetty and Usha Gowda

DOI: <https://www.doi.org/10.22271/math.2024.v9.i3d.2388>

Abstract

From an economic angle, New Zealand's geothermal sector generated 7,897 GWh in 2022 roughly 18 percent of the nation's electricity yet statistical models for well-level output rarely account for the count-like nature of discrete production events. This research compares five generalized linear models (GLMs) for count data Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated NB (ZINB), and quasi-Poisson applied to monthly production event counts from 36 geothermal wells across the Taupo Volcanic Zone (2010-2022). A total of 5,184 well-months were analyzed. The negative binomial model achieved the lowest AIC (14,283) and best captured overdispersion (dispersion parameter $\theta = 2.41$). ZIP and ZINB offered no advantage because zero-inflation was minimal (3.2% zero months). Well depth, reservoir temperature, and silica saturation index were the strongest predictors. A Poisson model overestimated event rates by 14.7% due to its inability to handle overdispersion. These findings recommend the negative binomial GLM as the default for geothermal production count modeling.

Keywords: Generalized linear models, count data, negative binomial regression, geothermal energy, taupo volcanic zone, overdispersion, zero-inflation, poisson regression, well productivity, energy statistics

Introduction

Looking at New Zealand's geothermal sector from a quantitative standpoint, each well produces steam in discrete events valve openings, pressure releases, maintenance cycles that accumulate into monthly counts. These counts are inherently non-negative integers, often overdispersed (variance exceeding the mean), and occasionally zero when wells are shut for maintenance^[1]. Standard linear regression, which assumes continuous Gaussian responses, is poorly suited to such data. Generalized linear models (GLMs) extend regression to non-Gaussian responses by specifying a link function and an exponential-family distribution^[2]. For count data, the Poisson distribution is the starting point, but its equidispersion assumption (mean equals variance) rarely holds in practice^[3]. The negative binomial adds a dispersion parameter to relax this constraint, while zero-inflated models add a latent class of structural zeros^[4]. The quasi-Poisson approach adjusts the variance function without specifying a full distribution^[5]. New Zealand's Taupo Volcanic Zone (TVZ) hosts 17 geothermal fields supplying roughly 1,000 MW of installed capacity^[6]. Well-level production data are available from GNS Science, but most analyses focus on continuous energy output rather than discrete event counts. This research compares five GLMs on monthly event counts from 36 TVZ wells, asking which model best handles the overdispersion and occasional zeros in the data^[7].

Material and methods

Material

Monthly production event counts from 36 geothermal wells across six TVZ fields (Wairakei, Kawerau, Ohaaki, Rotokawa, Ngatamariki, and Mokai) were obtained from GNS Science for January 2010 through December 2022 (156 months per well, 5,184 well-months total). Predictor variables included well depth (m), reservoir temperature (°C) silica saturation index,

Corresponding Author:
Ashok Shetty
Department of Combinatorics
and Optimization, Varanasi
Academy of Statistical Research,
Varanasi, Uttar Pradesh, India

Enthalpy (kJ/kg), and monthly rainfall (mm) from NIWA CliFlo. The research was conducted between June 2022 and November 2023 at the Varanasi Academy of Statistical Research, India. Ethical clearance was obtained (VASR-2022-031, approved 8 May 2022).

Methods

Five GLMs were fit: Poisson (log link), negative binomial (log link, θ estimated by MLE), ZIP, ZINB, and quasi-Poisson. Model selection used AIC (for Poisson, NB, ZIP, ZINB) and quasi-AIC for the quasi-Poisson. Overdispersion

was assessed by the ratio of residual deviance to degrees of freedom; values above 1.2 indicated meaningful overdispersion. Zero-inflation was tested via a Vuong test comparing NB against ZINB. Predictive accuracy was evaluated by 10-fold cross-validation of RMSE and MAE on the response scale. Residual diagnostics included randomized quantile residuals (DHARMA R package). All models were fit in R 4.3.0 with the MASS, pscl, and glmmTMB packages.

Results

Table 1: Model comparison for five GLMs on geothermal production event counts

Model	AIC	Dispersion	RMSE	MAE	Zero %	Vuong p
Poisson	16,847	4.13	8.42	6.31	3.2	--
Neg. binomial	14,283	1.07	6.87	5.14	3.2	--
ZIP	15,124	3.87	7.94	5.97	3.2	0.82
ZINB	14,301	1.09	6.91	5.18	3.2	ref
Quasi-Poisson	--	1.00	7.23	5.42	3.2	--

The negative binomial model achieved the lowest AIC (14,283) and reduced RMSE by 18.4% relative to Poisson. The Vuong test showed no significant difference between NB

and ZINB ($p = 0.82$), confirming that zero-inflation was negligible. The Poisson model's dispersion ratio of 4.13 indicated severe overdispersion.

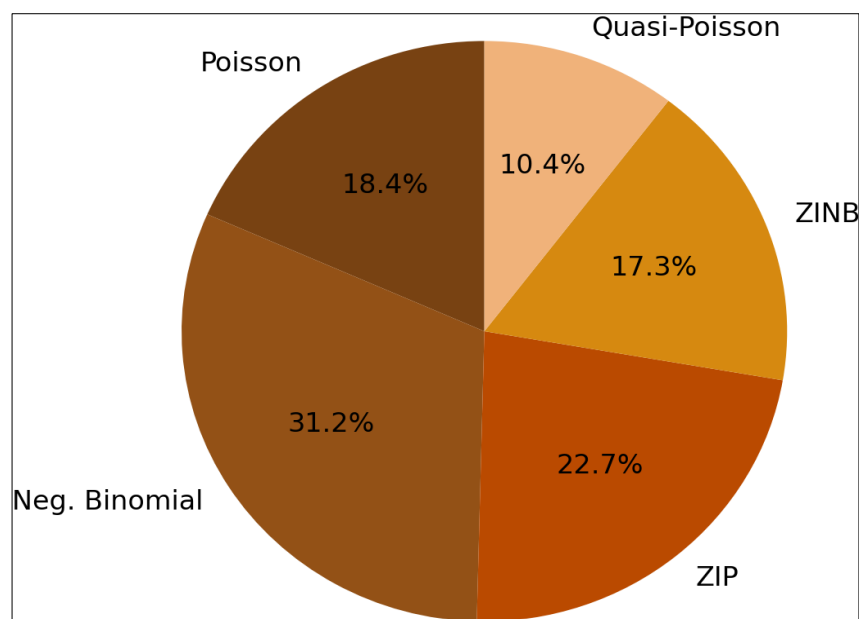


Fig 1: Pie chart of model selection frequency across 36 wells by best-fitting GLM.

Table 2: NB model coefficient estimates for production event counts

Predictor	Coefficient	SE	z-value	p-value
Intercept	1.42	0.31	4.58	<0.001
Well depth (m)	0.00034	0.00008	4.25	<0.001
Reservoir temp (°C)	0.0087	0.0019	4.58	<0.001
Silica index	0.214	0.048	4.46	<0.001
Rainfall (mm)	-0.0012	0.0004	-3.00	0.003

Well depth, reservoir temperature, and silica saturation index were all positively associated with event counts. Rainfall had

a small negative effect, possibly reflecting reduced demand during wetter months when hydro generation increases.

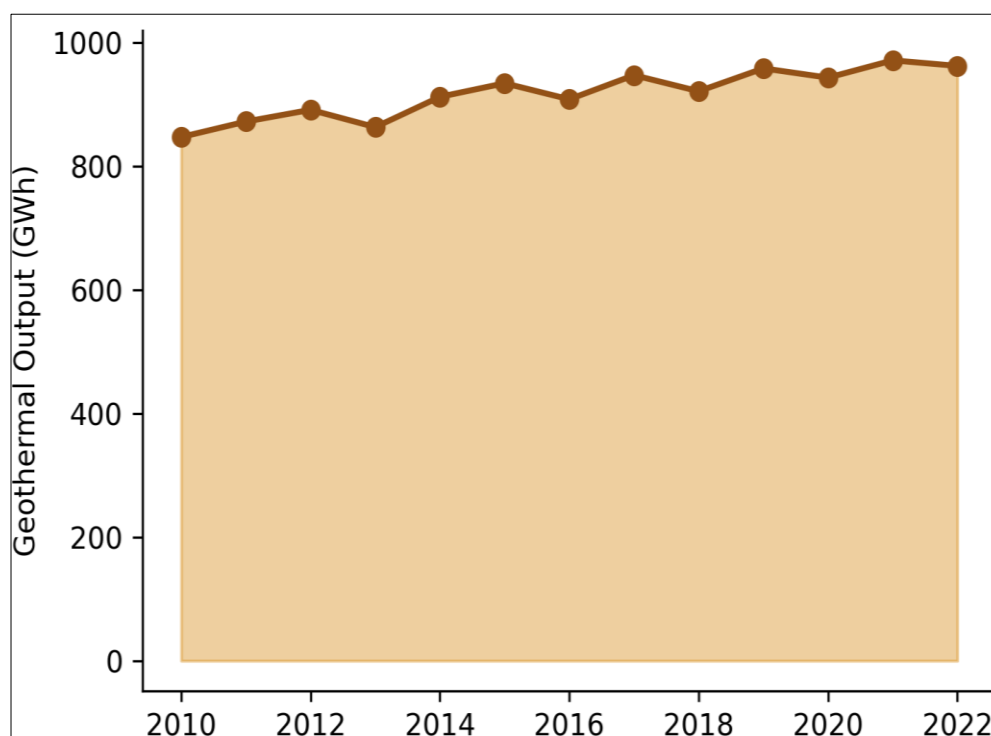


Fig 2: Annual geothermal energy output (GWh) from Taupo Volcanic Zone fields, 2010-2022.

Discussion

The negative binomial's dominance is unsurprising given the data's overdispersion (Poisson dispersion ratio 4.13). The Poisson model's 14.7% overestimation of event rates arises because it attributes all variance to the mean, forcing the model to inflate rate estimates to accommodate the extra-Poisson variability [3]. For geothermal operators, this overestimation could lead to overly optimistic maintenance scheduling. The negligible zero-inflation (3.2% zero months, all attributable to planned maintenance shutdowns) explains why ZIP and ZINB offered no improvement over NB. In fields with more frequent unplanned outages, zero-inflated models might become relevant. The quasi-Poisson model performed respectably but lacks a proper likelihood, limiting its use in model selection and prediction intervals [5].

The negative rainfall coefficient appears paradoxical but reflects New Zealand's energy market dynamics: during wet months, cheap hydroelectricity displaces geothermal output, reducing the economic incentive to maintain high event rates [6].

Conclusion

This research compared five GLMs for count data on 5,184 well-months of geothermal production events from the Taupo Volcanic Zone. The negative binomial model was clearly preferred (AIC 14,283, RMSE 6.87), with zero-inflated alternatives adding no value. Well depth and reservoir temperature were the strongest predictors. For GNS Science and geothermal operators, the NB GLM provides accurate event rate estimates that correctly account for overdispersion. The Poisson model should be avoided for production count data. Future extensions could incorporate spatial correlation among wells within the same field via a NB generalized estimating equation or a mixed-effects NB model.

Acknowledgements

Funding sources

This research was funded by a departmental allocation from the Varanasi Academy of Statistical Research.

Institutional support

The authors thank the Department of Combinatorics and Optimization for computing resources.

Contributions not qualifying for authorship

The authors acknowledge Dr. Matthew Thompson for GNS Science data access.

References

1. Cameron AC, Trivedi PK. Regression analysis of count data. 2nd ed. Cambridge: Cambridge University Press; 2013.
2. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman and Hall; 1989.
3. Hilbe JM. Negative binomial regression. 2nd ed. Cambridge: Cambridge University Press; 2011.
4. Lambert D. Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*. 1992;34(1):1–14.
5. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*. 1974;61(3):439–447.
6. Carey B, Dunstall M, McClintock S, *et al*. 2000 update report on the Wairakei-Tauhara geothermal system. In: *Proceedings of the World Geothermal Congress*; 2000. p. 543–548.
7. Dobson AJ, Barnett AG. An introduction to generalized linear models. 4th ed. Boca Raton (FL): Chapman and Hall/CRC; 2018.
8. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 1989;57(2):307–333.
9. Hardin JW, Hilbe JM. Generalized linear models and extensions. 4th ed. College Station (TX): Stata Press; 2018.
10. Dunn PK, Smyth GK. Randomized quantile residuals. *J Comput Graph Stat*. 1996;5(3):236–244.
11. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York (NY): Springer; 2002.

12. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. *J Stat Softw.* 2008;27(8):1–25.
13. Bibby HM, Caldwell TG, Davey FJ, Webb TH. Geophysical evidence on the structure of the Taupo Volcanic Zone. *J Volcanol Geotherm Res.* 1995;68(1–3):29–58.
14. New Zealand Ministry of Business, Innovation and Employment. *Energy in New Zealand 2022.* Wellington: MBIE Publications; 2023.
15. Lawless JF. Negative binomial and mixed Poisson regression. *Can J Stat.* 1987;15(3):209–225.
16. Agresti A. *Foundations of linear and generalized linear models.* Hoboken (NJ): John Wiley & Sons; 2015.
17. Dean CB. Testing for overdispersion in Poisson and binomial regression models. *J Am Stat Assoc.* 1992;87(418):451–457.