



Available online at <http://scik.org>

Commun. Math. Biol. Neurosci. 2026, 2026:101

<https://doi.org/10.28919/cmbn/10083>

ISSN: 2052-2541

A STATIONARY NEGATIVE BINOMIAL INAR(1) MODEL BASED ON THE BETA–BINOMIAL THINNING OPERATOR

SALEH WEMBO MOISE^{1,*}, BONIFACE MALENJE², MAKIMONA KIAKISOLAKO³

¹Department of Mathematics, Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI), JKUAT Campus, Juja, Kenya

²Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya

³Cavendish Laboratory Department of Physics, University of Cambridge, Cambridge, United Kingdom

Copyright © 2026 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. We propose a stationary negative binomial integer-valued autoregressive model of order one (NB-INAR(1)) built on a beta–binomial thinning operator. Allowing the thinning probability to vary according to a beta distribution induces a flexible survivor mechanism that preserves the binomial mean structure while inflating the conditional variance and capturing latent heterogeneity in persistence. Coupled with suitably specified negative binomial innovations that maintain the marginal distribution, the resulting process forms a Markov chain on the non-negative integers with tractable probabilistic properties, including geometrically decaying autocorrelation and an interpretable lag-one dependence parameter. Parameter estimation is investigated using conditional least squares, method of moments with feasibility diagnostics, and an optional maximum likelihood refinement based on convolution transition probabilities. A Monte Carlo study reveals that innovation parameters become increasingly difficult to identify under extreme persistence and pronounced thinning heterogeneity, underscoring the importance of reporting reliability measures such as feasibility and boundary-clamping rates. An application to weekly road-traffic collision counts (2009–2014) demonstrates the model’s practical performance in fitting and probabilistic forecasting.

*Corresponding author

E-mail address: wembo.saleh@students.jkuat.ac.ke

Received June 22, 2026

Keywords: INAR(1); beta-binomial thinning; count time series; negative binomial innovations; method of moments.

2020 AMS Subject Classification: 62M10, 62F10, 60G10.

1. INTRODUCTION

Count time series arise naturally in epidemiology, insurance, public safety, neuroscience, and many other fields where observations are nonnegative integers recorded sequentially [1]. A central difficulty is to represent serial dependence while respecting discreteness. Integer-valued autoregressive (INAR) models address this by replacing the linear autoregressive (AR) coefficient with a thinning operator that randomly “selects survivors” from the previous count, ensuring that the AR component remains integer-valued [2–6].

In the classical INAR(1) model defined by $X_t = \alpha \circ X_{t-1} + \varepsilon_t$, the parameter $\alpha \in (0, 1)$ has a direct interpretation as the expected survival rate and, under standard constructions, coincides with the lag-one autocorrelation [3, 4].

Broader discussions of count time series modeling, including observation-driven specifications, are given in [7, 8]. Forecasting aspects for discrete-valued series are treated in [9, 10].

However, many real count time series are overdispersed, exhibiting variances substantially larger than their means, together with heterogeneous persistence over time. Overdispersion and its testing within thinning-based models have also been studied, for example, in compound Poisson INAR constructions [11]. Fixed-probability binomial thinning is often too restrictive in such settings, because it imposes a constant survival rate and does not capture extra-binomial variability in the AR component [6]. Several extensions have therefore been proposed, including random-coefficient thinning and compound/iterated thinning mechanisms, which aim to increase flexibility in both dispersion and dependence [12–15]. Related developments for integer-valued (IN) moving average models and their estimation include [16]. In particular, randomization of the probability of thinning provides an interpretable way to model latent heterogeneity in survival rate and can induce additional variability beyond the binomial remark [17, 18].

Table 1 summarizes the review literature of INAR(1) models within the evolution of the count time series methodology based on thinning. Early work established the binomial-thinning INAR(1) principle and its extension to INAR(p), providing an integer-coherent analogue of linear AR dependence. Subsequent developments addressed a core empirical challenge—overdispersion

(Var $\gg$ Mean)—either by adopting negative binomial (NB) marginals or innovations, or by enriching the dependence mechanism through more flexible operators (compound, extended, or random thinning). The reviews in the table clarify two parallel modeling traditions: (i) thinning-based Markov constructions, where dependence is interpreted as a survivor mechanism, and (ii) observation-driven NB models (e.g., NB-INGARCH), where dependence is encoded through an evolving conditional intensity.

Our contribution differs from prior works by addressing both needs within a single stationary INAR(1) framework. Specifically, we introduce a beta-binomial (BB) thinning survivor mechanism in which the retention probability is randomized, and couple it with NB innovations selected to preserve a stationary NB marginal while maintaining the familiar geometric ACF decay, governed by the dependence parameter (τ) and an explicit heterogeneity/precision parameter (κ).

Beyond establishing the model properties, we provide a unified inference workflow using CLS and MME (with feasibility diagnostics), complemented by an optional MLE refinement. We also document how estimator accuracy and reliability vary across different persistence and heterogeneity regimes.

Table 1: **Literature review on thinning-based INAR models, overdispersion, and inference.**

Author(s)	Contribution	Inference focus	Importance
Steutel and van Harn [2]	Discrete analogues of self-decomposability/stability for count laws	Foundations / distribution theory	Provides a theoretical backbone for constructing stationary count processes with integer-coherent dependence.
McKenzie [3]	Early discrete autoregressive modeling for counts via thinning ideas	Model formulation / applications	Motivates integer-coherent dependence modeling for count data.
Al-Osh and Alzaid [4]	Classical INAR(1) with binomial thinning ($\alpha \circ X$)	Properties + basic inference	Baseline INAR(1): interpretable survivor mechanism and dependence parameter.
Du and Li [5]	Extension to INAR(p) via thinning and higher-order dependence	Model properties / estimation setup	Context for extending thinning-based models beyond order one.

Continued on next page

Table 1: **Literature review on thinning-based INAR models, overdispersion, and inference** (continued).

Author(s)	Contribution	Inference focus	Importance
Al-Osh and Aly [19]	First-order autoregressive count models with NB / geometric marginals	Overdispersion model- ing + estimation discussion	Key motivation for NB marginals/innovations when $\text{Var} \gg \text{Mean}$.
McKenzie [17]	ARMA-type count processes with NB / geometric marginals (non-thinning route)	Properties + likelihood-type ideas	Shows an alternative NB dependence construction, helping differentiate thinning-based vs marginal-driven approaches.
Weiss [6], Scotto et al. [15]	Surveys/reviews of thinning operators and thinning-based count TS	Review (operators taxonomy)	+ Positions flexible thinning (incl. randomized thinning ideas) within the broader literature.
Schweer and Weiss [11]	Compound Poisson INAR(1) + testing for overdispersion in thinning models	Properties + testing	Shows how overdispersion arises under thinning and how it can be diagnosed.
Liu and Zhu [20]	Modern extensions of thinning-based INAR models for added flexibility	Model extension + estimator comparisons	Supports reporting finite-sample behavior across CLS/MME/MLE-type estimators.
Wei [18]	Overview of <i>stationary</i> count time-series models (thinning-based and beyond)	Review (stationarity + inference toolbox)	+ Provides a unifying frame and standard inference/diagnostics used in stationary settings.
Karlis et al. [21]	Identifiability/separability of thinning operators under NB-INAR(1) simulations	Model comparison	+ Highlights that same NB marginals can hide different transitions (operator matters: BB vs Bin).
Gouriroux and Lu [22]	NB autoregressive process with stochastic intensity (observation-driven comparator)	Properties + estimation + applications	Benchmark alternative to thinning: dependence via latent/conditional intensity rather than survivor thinning.
Zhu [14]	NB-INGARCH (observation-driven model; likelihood-based inference)	MLE / comparisons (OD models)	Contrasts thinning-based INAR: dependence via conditional intensity dynamics, not survivor thinning.

In this paper, we develop and study a stationary NB INAR(1) model driven by a BB thinning operator. Specifically, we let survival probability vary over time according to a beta distribution, so that conditional on $Y_{t-1} = n$ the survivor count follows a BBeta law, and combine this with

NB innovations chosen to preserve a stationary NB marginal distribution [17, 18]. The resulting construction generalizes the fixed-coefficient NB-INAR(1) model, as the Beta precision increases (i.e., the Beta distribution concentrates at its mean), the random thinning degenerates to ordinary binomial thinning and the model reduces to the classical NB-INAR(1) setting [14, 19, 22].

Hence, our contributions are threefold.

- First, we formalize the model and summarize its key distributional properties, including stationarity of the mean and variance, NB marginals, and a geometrically decaying autocorrelation function of the form $\rho(h) = \tau^h$, where τ denotes the mean thinning probability and equals the lag-one autocorrelation under stationarity [14, 18, 19].
- Second, we investigate parameter estimation using conditional least squares (CLS), method of moments (MME), and maximum likelihood estimation (MLE), highlighting the trade-offs between computational simplicity and statistical efficiency in the presence of random thinning [20].
- Finally, through a Monte Carlo study, we assess finite-sample bias and root mean squared error (RMSE) across persistence and overdispersion regimes, and complement the simulation evidence with an application to a real count dataset to illustrate practical fitting and interpretation [23, 24].

The remainder of the paper is organized as follows. Section 2 presents the model formulation and its main probabilistic properties. Section 3 describes the three suggested estimation approaches. Section 4 reports the Monte Carlo design and finite-sample performance results. Section 5 presents a real count data application. Section 6 concludes the paper, highlighting potential avenues for future research.

2. MODEL FORMULATION

2.1. Thinning operators and the BB mechanism. Let X be a nonnegative integer-valued random variable. The classical binomial thinning operator $\alpha \circ X$, $0 < \alpha < 1$, is defined by

$$\alpha \circ X = \sum_{i=1}^X B_i, \quad B_i \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\alpha).$$

So that, conditionally on $X = n$, $\alpha \circ X \sim \text{Binomial}(n, \alpha)$.

This operator provides the standard route for constructing INAR processes while preserving discreteness and nonnegativity [2–6, 25].

To introduce extra-binomial variability and hence additional flexibility for overdispersed counts, we adopt the BB thinning operator [18, 25]. Let Θ be a random retention probability with Beta distribution,

$$\Theta \sim \text{Beta}(\alpha_1, \alpha_2), \quad \alpha_1 > 0, \alpha_2 > 0.$$

Given $\Theta = \theta$ and $X = n$, define

$$\Theta \circ X \mid (\Theta = \theta, X = n) \sim \text{Binomial}(n, \theta).$$

Marginally (i.e., after integrating out Θ), the thinned count has a BBeta distribution:

$$\Pr(\Theta \circ X = k \mid X = n) = \binom{n}{k} \frac{B(\alpha_1 + k, \alpha_2 + n - k)}{B(\alpha_1, \alpha_2)}, \quad k = 0, 1, \dots, n,$$

where $B(\cdot, \cdot)$ denotes the beta function [18].

It is convenient to re-parameterize the beta distribution by its mean and precision:

$$\tau = \frac{\alpha_1}{\alpha_1 + \alpha_2} \in (0, 1), \quad \kappa = \alpha_1 + \alpha_2 > 0,$$

so that $\alpha_1 = \tau\kappa$ and $\alpha_2 = (1 - \tau)\kappa$. Then, for $n \in \mathbb{N}$,

$$\mathbb{E}[\Theta \circ X \mid X = n] = \tau n,$$

and

$$\text{Var}(\Theta \circ X \mid X = n) = n \tau (1 - \tau) \frac{n + \kappa}{\kappa + 1},$$

showing that the BB operator preserves the binomial mean structure but inflates the conditional variance through the factor $(n + \kappa)/(\kappa + 1)$ [25].

Figure 1. highlights the main mechanism that differentiates our model from classical binomial-thinning INAR(1) (Table 1). By randomizing the survival probability via a Beta law, the beta-binomial (BB) operator preserves the usual conditional mean structure while inflating conditional variance through the precision parameter κ , providing a direct route to accommodate extra-binomial variability in dependent counts.

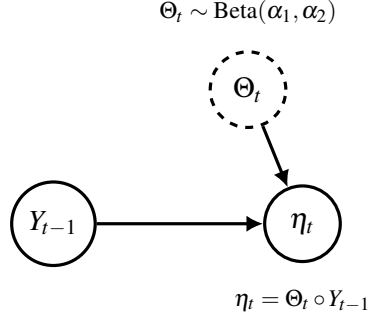


FIGURE 1. BB thinning mechanism: a random survival probability $\Theta_t \sim \text{beta}(\alpha_1, \alpha_2)$ drives the thinned count $\eta_t = \Theta_t \circ Y_{t-1}$.

2.2. NB-INAR(1) model with BB thinning. Let $\{Y_t : t = 0, 1, 2, \dots\}$ be a count time series. The proposed stationary NB-INAR(1) model is defined by the recursion

$$(1) \quad Y_t = \eta_t + \xi_t, \quad t \geq 1,$$

where

$$\eta_t = \Theta_t \circ Y_{t-1} \quad \text{and} \quad \{\Theta_t\}_{t \geq 1} \text{ are i.i.d. with } \Theta_t \sim \text{Beta}(\tau\kappa, (1-\tau)\kappa),$$

independent of the past $\{Y_{t-1}, Y_{t-2}, \dots\}$. Conditionally on $Y_{t-1} = n$, the thinning component follows

$$\eta_t \mid (Y_{t-1} = n) \sim \text{BB}(n; \tau\kappa, (1-\tau)\kappa).$$

The innovation sequence $\{\xi_t\}_{t \geq 1}$ is assumed i.i.d. and independent of $\{\Theta_t\}$ and $\{Y_{t-1}, \eta_t\}$. We take ξ_t to be Negative Binomial with parameters (r, p) under the convention

$$(2) \quad \Pr(\xi_t = k) = \binom{r+k-1}{k} p^r (1-p)^k, \quad k = 0, 1, 2, \dots,$$

so that

$$\mathbb{E}[\xi_t] = r \frac{1-p}{p}, \quad \text{Var}(\xi_t) = r \frac{1-p}{p^2}.$$

We set

$$\xi_t \sim \text{NB}((1-\tau)\mu, p),$$

which yields a stationary marginal distribution $Y_t \sim \text{NB}(\mu, p)$ when the process is initialized in stationarity (e.g., $Y_0 \sim \text{NB}(\mu, p)$) and $0 < \tau < 1$ [14, 19, 22].

2.3. Hierarchical representation and parameter interpretation. The model (1) admits the hierarchical form

$$\eta_t \mid (Y_{t-1} = n) \sim \text{BB}(n; \tau\kappa, (1-\tau)\kappa), \quad \xi_t \sim \text{NB}((1-\tau)\mu, p), \quad Y_t = \eta_t + \xi_t.$$

The parameter $\tau \in (0, 1)$ governs serial dependence and, under stationarity, equals the lag-one autocorrelation; (μ, p) specify the stationary NB marginal distribution with mean $\mathbb{E}[Y_t] = \mu(1-p)/p$ and variance $\text{Var}(Y_t) = \mu(1-p)/p^2$; and κ controls the strength of random-thinning heterogeneity (larger κ concentrates Θ_t around τ) [14, 25]. In the limit $\kappa \rightarrow \infty$, Θ_t degenerates at τ and the model reduces to the classical fixed-coefficient NB-INAR(1) of [14, 19].

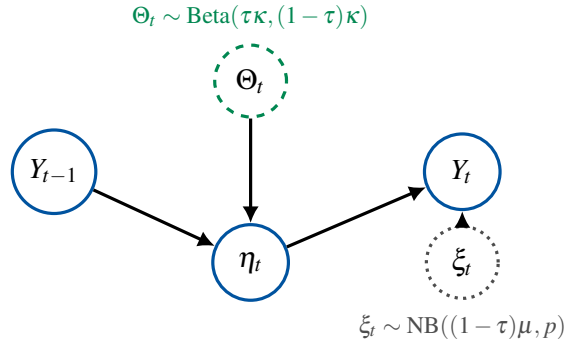


FIGURE 2. Graphical representation of the proposed NB-INAR(1) model with BB thinning: $\eta_t = \Theta_t \circ Y_{t-1}$ and $Y_t = \eta_t + \xi_t$.

Figure 2. clarifies the model decomposition $Y_t = \eta_t + \xi_t$ and the roles of parameters: (τ, κ) control the dependence and heterogeneity in the BB thinning, whereas (μ, p) control the NB innovation component. This separation is important for interpretation and also helps explain later finite-sample patterns (Tables 2–4), where dependence can be well-estimated even when innovation parameters are harder to identify under strong persistence and heterogeneity.

2.4. Stationarity and distributional properties.

2.4.1. One-step conditional distribution. From (1), given $Y_{t-1} = n$ we have the decomposition

$$Y_t \mid (Y_{t-1} = n) \stackrel{d}{=} \eta_t + \xi_t, \quad \eta_t \mid (Y_{t-1} = n) \sim \text{Beta-B}(n; \tau\kappa, (1-\tau)\kappa),$$

with $\xi_t \sim \text{NB}((1 - \tau)\mu, p)$ independent of η_t . Therefore, the one-step transition probabilities are given by the convolution

$$(3) \quad \Pr(Y_t = m \mid Y_{t-1} = n) = \sum_{k=0}^{\min(n, m)} \Pr(\eta_t = k \mid Y_{t-1} = n) \Pr(\xi_t = m - k), \quad m = 0, 1, 2, \dots,$$

where

$$\Pr(\eta_t = k \mid Y_{t-1} = n) = \binom{n}{k} \frac{B(\tau\kappa + k, (1 - \tau)\kappa + n - k)}{B(\tau\kappa, (1 - \tau)\kappa)}, \quad k = 0, 1, \dots, n,$$

and $\Pr(\xi_t = \cdot)$ is the NB pmf under the convention stated in (2). Hence, $\{Y_t\}$ is a time-homogeneous Markov chain on $\mathbb{N}_0$ with transition kernel (3) [14, 19].

2.4.2. Stationary marginal distribution via probability generating functions. Let $G_Y(s) = \mathbb{E}[s^{Y_t}]$ denote the pgf of Y_t , and $G_\xi(s) = \mathbb{E}[s^{\xi_t}]$ the pgf of the innovations. Using conditional independence,

$$\mathbb{E}[s^{Y_t} \mid Y_{t-1}] = \mathbb{E}[s^{\eta_t} \mid Y_{t-1}] G_\xi(s).$$

For the thinning component, conditioning first on Θ_t gives

$$\mathbb{E}[s^{\eta_t} \mid (Y_{t-1} = n, \Theta_t = \theta)] = (1 - \theta + \theta s)^n,$$

and averaging over Θ_t implies

$$\mathbb{E}[s^{\eta_t} \mid Y_{t-1}] = \mathbb{E}[(1 - \Theta_t + \Theta_t s)^{Y_{t-1}} \mid Y_{t-1}].$$

A key simplification for BB thinning is that, at the pgf level, the Markov recursion reduces to the same functional form as fixed-coefficient thinning:

$$(4) \quad G_Y(s) = G_\xi(s) G_Y(1 - \tau + \tau s), \quad |s| \leq 1,$$

with $\tau = \mathbb{E}[\Theta_t]$ [19].

Under the NB parameterization in (2), the innovation pgf is

$$G_\xi(s) = \left(\frac{p}{1 - (1 - p)s} \right)^{(1 - \tau)\mu}, \quad |s| < \frac{1}{1 - p}.$$

Consider the candidate stationary marginal pgf

$$(5) \quad G_Y(s) = \left(\frac{p}{1 - (1 - p)s} \right)^\mu.$$

Substituting (5) into (4) yields an identity, so (5) is invariant under the recursion. Consequently, if the chain is initialized in stationarity (e.g., $Y_0 \sim \text{NB}(\mu, p)$), then

$$Y_t \sim \text{NB}(\mu, p) \quad \text{for all } t,$$

and the process is strictly stationary for $\boxed{0 < \tau < 1.}$ [14, 19, 22].

2.4.3. First and second moments. Using the conditional mean $\mathbb{E}[\eta_t | Y_{t-1}] = \tau Y_{t-1}$ and $\mathbb{E}[\xi_t] = (1 - \tau)\mu(1 - p)/p$ (under (2)),

$$(6) \quad \mathbb{E}[Y_t | Y_{t-1}] = \tau Y_{t-1} + (1 - \tau)\mu \frac{1 - p}{p}.$$

At stationarity, $\mathbb{E}[Y_t] = \mathbb{E}[Y_{t-1}]$, so

$$(7) \quad \mathbb{E}[Y_t] = \mu \frac{1 - p}{p}.$$

Moreover, from $Y_t \sim \text{NB}(\mu, p)$,

$$(8) \quad \text{Var}(Y_t) = \mu \frac{1 - p}{p^2}.$$

Note that κ does not alter the stationary marginal law (5); instead, it controls the *conditional* dispersion of the thinned component through $\text{Var}(\eta_t | Y_{t-1})$, with the fixed-coefficient model recovered as $\kappa \rightarrow \infty$ [19, 25].

2.4.4. Autocovariance and autocorrelation. Because ξ_t is independent of the past, the lag-one autocovariance satisfies

$$\text{Cov}(Y_t, Y_{t-1}) = \text{Cov}(\eta_t, Y_{t-1}).$$

Using $\mathbb{E}[\eta_t | Y_{t-1}] = \tau Y_{t-1}$,

$$\text{Cov}(\eta_t, Y_{t-1}) = \text{Cov}(\mathbb{E}[\eta_t | Y_{t-1}], Y_{t-1}) = \text{Cov}(\tau Y_{t-1}, Y_{t-1}) = \tau \text{Var}(Y_{t-1}).$$

At stationarity $\text{Var}(Y_{t-1}) = \text{Var}(Y_t)$, hence

$$(9) \quad \text{Cov}(Y_t, Y_{t-1}) = \tau \text{Var}(Y_t).$$

By iterating the same argument (induction on the lag), the autocovariance function is

$$(10) \quad \gamma(h) = \text{Cov}(Y_t, Y_{t-h}) = \tau^h \text{Var}(Y_t), \quad h \geq 0,$$

and the autocorrelation function becomes

$$(11) \quad \rho(h) = \frac{\gamma(h)}{\gamma(0)} = \tau^h, \quad h \geq 0,$$

so τ equals the lag-one autocorrelation and governs geometric decay of dependence [14, 19].

3. ESTIMATION METHODS

This section presents the three estimators used in the simulation study and the real-data application: CLS, MME, and MLE. We work with the stationary NB-INAR(1) model in Section 2 and parameter vector

$$\boldsymbol{\vartheta} = (\tau, \kappa, \mu, p)^\top, \quad \tau \in (0, 1), \kappa > 0, \mu > 0, p \in (0, 1).$$

Throughout, the NB distribution is parameterized as in (2) so that

$$\mathbb{E}[Y_t] = \mu \frac{1-p}{p}, \quad \text{Var}(Y_t) = \mu \frac{1-p}{p^2}.$$

3.1. Conditional moments. From the model definition, the thinning component satisfies

$$\mathbb{E}[\eta_t | Y_{t-1}] = \tau Y_{t-1}, \quad \text{Var}(\eta_t | Y_{t-1}) = Y_{t-1} \tau (1 - \tau) \frac{Y_{t-1} + \kappa}{\kappa + 1},$$

and the innovations satisfy

$$\xi_t \sim \text{NB}((1 - \tau)\mu, p), \quad \mathbb{E}[\xi_t] = (1 - \tau)\mu \frac{1-p}{p}, \quad \text{Var}(\xi_t) = (1 - \tau)\mu \frac{1-p}{p^2}.$$

Hence the conditional mean can be written in regression form

$$(12) \quad \mathbb{E}[Y_t | Y_{t-1}] = \tau Y_{t-1} + \lambda, \quad \lambda := (1 - \tau)\mu \frac{1-p}{p},$$

and the conditional variance is

$$(13) \quad \text{Var}(Y_t | Y_{t-1}) = \tau(1 - \tau) Y_{t-1} \frac{Y_{t-1} + \kappa}{\kappa + 1} + (1 - \tau)\mu \frac{1-p}{p^2}.$$

Under stationarity, the lag-one autocorrelation equals $\rho(1) = \tau$ and $\rho(h) = \tau^h$ [14, 19].

3.2. Conditional least squares (CLS). CLS estimates (τ, λ) by minimizing squared one-step prediction errors based on (12). Define

$$e_t(\tau, \lambda) = Y_t - \tau Y_{t-1} - \lambda, \quad t = 2, \dots, T,$$

and the CLS objective

$$Q_T(\tau, \lambda) = \sum_{t=2}^T e_t(\tau, \lambda)^2.$$

The CLS estimator is

$$(\hat{\tau}_{\text{CLS}}, \hat{\lambda}_{\text{CLS}}) = \arg \min_{\tau \in (0,1), \lambda > 0} Q_T(\tau, \lambda).$$

In practice, this is equivalent to an OLS regression of Y_t on Y_{t-1} with intercept, followed by truncation of $\hat{\tau}$ to $(0, 1)$ if necessary. CLS is simple and typically stable across regimes [20, 25].

3.3. MME. MME recovers (μ, p) from the stationary mean and variance together with the CLS estimate of τ . Let $\bar{Y}$ and S_Y^2 denote the sample mean and variance. Under stationarity,

$$m := \mathbb{E}[Y_t] = \mu \frac{1-p}{p}, \quad S := \text{Var}(Y_t) = \mu \frac{1-p}{p^2} = \frac{m}{p}.$$

Using (12), the stationary mean also satisfies

$$m = \frac{\lambda}{1-\tau}.$$

Therefore, we estimate

$$\hat{m} = \frac{\hat{\lambda}_{\text{CLS}}}{1 - \hat{\tau}_{\text{CLS}}}, \quad \hat{p}_{\text{MM}} = \frac{\hat{m}}{S_Y^2}, \quad \hat{\mu}_{\text{MM}} = \frac{\hat{m}^2}{S_Y^2 - \hat{m}},$$

provided $S_Y^2 > \hat{m}$ (overdispersion). When $S_Y^2 \leq \hat{m}$ the NB moment equations are infeasible; we record such cases as MME failures and report their frequency. The remaining parameter κ controls extra-binomial variability in thinning and is estimated by matching a conditional-variance identity implied by (13); the explicit estimating equation is given in the Appendix A.2.

3.4. MLE. Likelihood inference uses the one-step transition probabilities. Given $Y_{t-1} = i$,

$$\eta_t \mid (Y_{t-1} = i) \sim \text{BB}(i; \tau\kappa, (1-\tau)\kappa), \quad \xi_t \sim \text{NB}((1-\tau)\mu, p),$$

independent, and $Y_t = \eta_t + \xi_t$. Hence

$$(14) \quad \Pr(Y_t = j \mid Y_{t-1} = i) = \sum_{k=0}^{\min(i,j)} \Pr(\eta_t = k \mid Y_{t-1} = i) \Pr(\xi_t = j-k), \quad j \geq 0.$$

The conditional log-likelihood (given Y_1) is

$$\ell_T(\boldsymbol{\theta}) = \sum_{t=2}^T \log \Pr(Y_t = y_t \mid Y_{t-1} = y_{t-1}; \boldsymbol{\theta}),$$

with transition probabilities computed by (14). We compute the MLE by constrained numerical maximization over $\tau \in (0, 1)$, $p \in (0, 1)$, $\kappa > 0$, $\mu > 0$. Because the convolution in (14) must be evaluated at each time point, MLE can be computationally heavy and may fail to converge in extreme regimes; therefore, we report convergence rates and treat MLE as optional, using CLS/MME as the main estimators [20, 25].

3.5. Simulation reporting metrics. Across Monte Carlo replications, we summarize estimator performance using bias and root mean squared error (RMSE):

$$\text{Bias}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \theta), \quad \text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \theta)^2},$$

for each parameter $\theta \in \{\tau, \kappa, \mu, p\}$. For MME, we additionally report the *failure rate* (percentage of replications where the moment equations are infeasible), and for MLE we report the *convergence rate* (percentage of replications where numerical optimization terminates successfully). These reliability summaries are essential: a method with slightly lower RMSE but frequent failures can be less useful in practice than a stable estimator with marginally higher RMSE.

4. MONTE CARLO SIMULATION STUDY AND RESULTS

This section evaluates the finite-sample accuracy and numerical reliability of the proposed estimators CLS, MME, and (optionally) MLE for the stationary NB-INAR(1) model with Beta-Binomial thinning. Because asymptotic properties may not accurately reflect performance at moderate sample sizes and under strong overdispersion/persistence, we conduct a Monte Carlo

study and summarize performance using bias, RMSE, and estimator-reliability diagnostics [23–25].

4.1. Design: DGP, factors, sample sizes, and initialization.

Data-generating process. The DGP follows the hierarchical model in Section 2. For $t \geq 1$,

$$(15) \quad Y_t = \eta_t + \xi_t, \quad \eta_t \mid (Y_{t-1}, \Theta_t) \sim \text{Bin}(Y_{t-1}, \Theta_t), \quad \Theta_t \sim \text{Beta}(\tau\kappa, (1-\tau)\kappa),$$

with innovations

$$\xi_t \sim \text{NB}((1-\tau)\mu, p),$$

so that the process is stationary with marginal $Y_t \sim \text{NB}(\mu, p)$ when initialized in stationarity. Equivalently,

$$\eta_t \mid Y_{t-1} \sim \text{BetaBin}(Y_{t-1}; \tau\kappa, (1-\tau)\kappa),$$

and the parameter vector is $\boldsymbol{\vartheta} = (\tau, \kappa, \mu, p)^\top$.

Design factors. To reflect regimes commonly observed in overdispersed count series, we vary:

- **Persistence** via $\tau \in (0, 1)$ (weak, moderate, and strong dependence).
- **Random-thinning heterogeneity** via $\kappa > 0$ (smaller κ implies larger variability in the survival probability Θ_t ; as $\kappa \rightarrow \infty$, Θ_t concentrates at τ and the model approaches fixed-coefficient thinning).
- **Marginal level/dispersion** through (μ, p) in the Negative Binomial innovations.

We consider a *main grid* using representative values $\tau \in \{0.2, 0.5, 0.8\}$, $\kappa \in \{5, 20, 200\}$ (low to high precision), and $p \in \{0.3, 0.6\}$ (heavier-tail vs. milder-tail), together with two mean levels (“low” vs. “high”) via μ . We also include a *stress regime* with very high persistence (e.g. $\tau = 0.95$) and low precision κ , which is known to challenge moment-based procedures and constrained optimization.

Sample sizes and replications. We simulate series lengths $T \in \{100, 200, 500, 1000\}$ and run $R = 300$ replications per configuration for the main grid (see run log `Sim_results/logs/step2_runlog.txt`). The optional MLE add-on uses $R = 200$ on a representative subset (see `Sim_results/logs/step7_runlog.txt`).

Initialization. To approximate stationarity, we initialize the chain at its stationary marginal distribution,

$$Y_0 \sim \text{NB}(\mu, p),$$

and generate $\{Y_t\}_{t=1}^T$ according to (1) (equivalently (15)). (Alternatively, one may use a burn-in of length B and discard the first B observations; stationary initialization was sufficient for the present comparisons.)

4.2. Estimators and implementation. For each replication, we compute:

- **CLS:** estimate (τ, λ) from the conditional mean regression and then map to (τ, μ, p, κ) under the model constraints.
- **MME:** use stationary moment relations for (μ, p) and an additional conditional-variance identity for κ (see Appendix). MME may be infeasible when the finite-sample variance condition fails (e.g. $\hat{s}^2 \leq \hat{m}$); we record such cases as MME failures.
- **MLE (optional):** maximize the conditional likelihood based on the one-step transition probabilities (convolution form). Because optimization can be heavy or unstable in extreme regimes, we report convergence rates and treat MLE as an optional refinement [20, 25].

4.3. Performance and reliability metrics. For each parameter $\theta \in \{\tau, \kappa, \mu, p\}$ we compute bias and RMSE over R replications:

$$\text{Bias}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R \left(\hat{\theta}^{(r)} - \theta \right), \quad \text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{R} \sum_{r=1}^R \left(\hat{\theta}^{(r)} - \theta \right)^2}.$$

We also report reliability summaries: (i) MME feasibility rate, (ii) MLE convergence rate (when attempted), and (iii) boundary-clamping rates for constrained parameters (notably p). These diagnostics are essential because an estimator with low RMSE conditional on success can still be impractical if it fails often.

4.4. Results: main grid.

Overall accuracy. Table 2 reports overall averages (across the full main grid) for RMSE and bias. As expected, RMSE decreases monotonically with sample size. Averaged across all main-grid configurations, RMSE for τ decreases from 0.093 at $T = 100$ to 0.028 at $T = 1000$, while RMSE for μ decreases from 1.180 to 0.352. For the innovation parameter p , RMSE decreases from 0.177 at $T = 100$ to 0.063 at $T = 1000$.

Biases are generally negative for τ and p and positive for μ at small T , but diminish rapidly with T . For example, the grid-averaged bias for τ is -0.026 at $T = 100$ and -0.003 at $T = 1000$, while for p it is -0.061 at $T = 100$ and -0.010 at $T = 1000$.

TABLE 2. **Main grid (overall averages across configurations): RMSE and bias by sample size T .**

T	RMSE_τ	RMSE_μ	RMSE_p	Bias_τ	Bias_μ	Bias_p
100	0.093	1.180	0.177	-0.026	0.310	-0.061
200	0.064	0.801	0.134	-0.014	0.157	-0.036
500	0.040	0.499	0.089	-0.005	0.065	-0.019
1000	0.028	0.352	0.063	-0.003	0.031	-0.010

Persistence and innovation identifiability. Figures 3–4 show a robust trade-off: estimating τ becomes easier at higher persistence, while estimating the innovation parameter p becomes harder as τ increases. For example, at $T = 200$, average RMSE_p increases from 0.079 at $\tau = 0.2$ to 0.207 at $\tau = 0.8$, consistent with weaker identifiability of innovation-driven dispersion when the autoregressive component dominates.

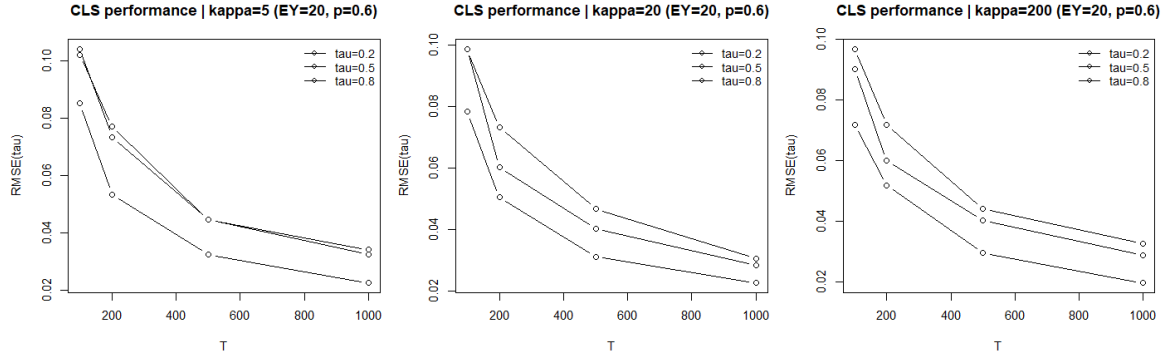


FIGURE 3. Main grid: RMSE for τ across persistence levels, faceted by T and other design factors.

Figure 3. visually supports the main-grid averages reported in Table 2: RMSE for τ decreases as T increases and remains comparatively stable across configurations. This reflects the fact that τ directly controls short-lag dependence, so the signal for τ is strong even when the thinning mechanism is randomized.

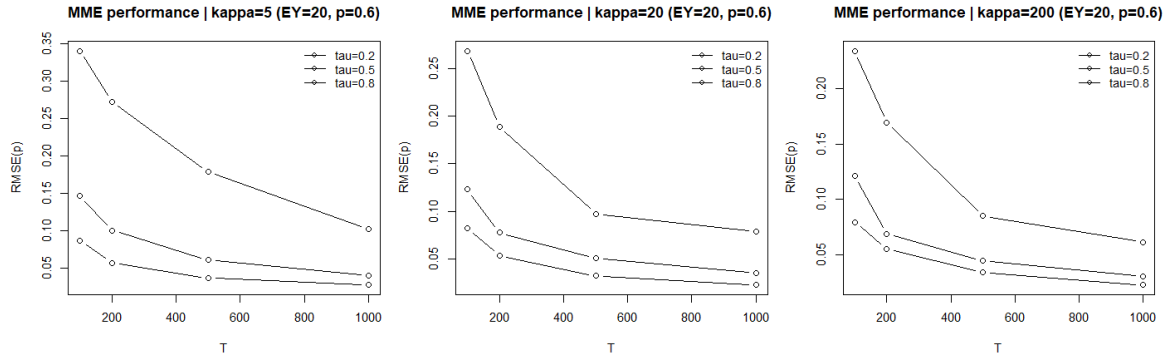


FIGURE 4. Main grid: RMSE for p across persistence levels. RMSE_p increases with τ , consistent with weaker identifiability of innovation parameters under strong persistence.

Figure 4. complements Table 2 by showing that p becomes increasingly difficult to estimate as persistence grows. When τ is large, more of the dynamics are explained by the survivor component, leaving less independent information to identify innovation dispersion; this motivates reporting reliability diagnostics (Table 3) in addition to RMSE/bias.

Effect of thinning heterogeneity (κ). Across the main grid, larger κ (less heterogeneity in the random thinning probability) yields modest but consistent gains in accuracy, particularly for

p . For example, at $T = 100$ the grid-averaged RMSE_p decreases from 0.193 at $\kappa = 5$ to 0.165 at $\kappa = 200$; at $T = 1000$ the corresponding decrease is from 0.073 to 0.055. This aligns with the intuition that strong random-thinning variability inflates conditional uncertainty, making innovation parameters harder to recover in short samples.

Reliability. Table 3 shows high overall MME feasibility on the main grid (99.4% at $T = 100$ and essentially 100% for $T \geq 500$), but the hardest corner (high persistence, low κ , small T) can degrade noticeably: the minimum MME success rate is 83.3% at $(\tau, \kappa, T) = (0.8, 5, 100)$ with $E[Y_t] = 20$ and $p = 0.3$. Boundary clamping is mainly a small-sample issue for p (average p -clamp 11.9% at $T = 100$ dropping to 1.0% at $T = 1000$).

TABLE 3. **Main grid (overall averages across configurations): estimator reliability by sample size T .**

T	MME success rate	p -clamp rate	τ -clamp rate
100	99.4%	11.9%	1.3%
200	99.9%	6.4%	0.2%
500	100.0%	2.6%	0.0%
1000	100.0%	1.0%	0.0%

4.5. Results: stress regime. The stress regime ($\tau = 0.95$ with low κ) amplifies innovation non-identifiability. Compared with the main grid, RMSE_p deteriorates substantially: it is 0.316 at $T = 100$ and remains 0.160 at $T = 1000$, with strong negative bias ($\text{Bias}_p = -0.174$ at $T = 100$; Table 4). By contrast, RMSE_τ stays small and decreases quickly ($0.075 \rightarrow 0.017$), reflecting the dominant autoregressive signal under strong persistence.

TABLE 4. **Stress design (overall averages across configurations): RMSE and bias by sample size T .**

T	RMSE_τ	RMSE_μ	RMSE_p	Bias_τ	Bias_μ	Bias_p
100	0.075	0.926	0.316	-0.041	0.493	-0.174
200	0.045	0.547	0.271	-0.021	0.246	-0.131
500	0.025	0.300	0.206	-0.009	0.100	-0.085
1000	0.017	0.202	0.160	-0.004	0.046	-0.053

Reliability also degrades markedly in the stress regime. Averaged across stress configurations, the MME success rate is 85.1% at $T = 100$ and rises to 99.7% at $T = 1000$, while the average p -clamp rate remains elevated (from 38.9% at $T = 100$ to 10.1% at $T = 1000$). The worst stress-case corner exhibits an MME success rate as low as 41.7% at $(\tau, \kappa, T) = (0.95, 5, 100)$ with $E[Y_t] = 20$ and $p = 0.3$, driven primarily by the finite-sample variance condition failing (see `final_reliability_worst20.csv`). (Table S1).

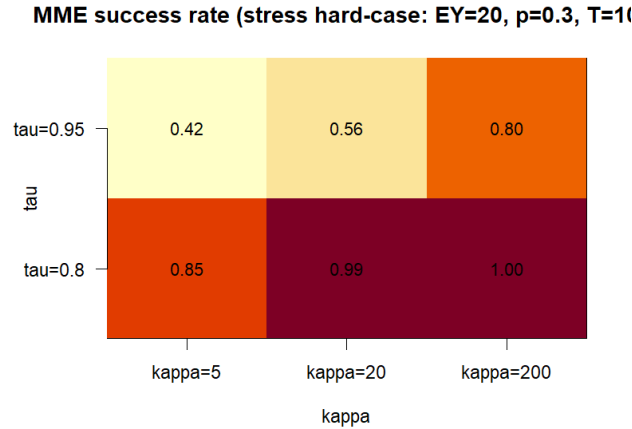


FIGURE 5. Stress design: MME success rate across the stress grid. The lowest success rates occur at small T , low κ , and $\tau = 0.95$.

Figure 5. complements the stress-design averages in Table 4 by displaying where MME feasibility deteriorates. The lowest success rates occur in the hard corner (small T , high τ , low κ), confirming that strong persistence combined with strong thinning heterogeneity can make moment conditions ill-conditioned in finite samples.

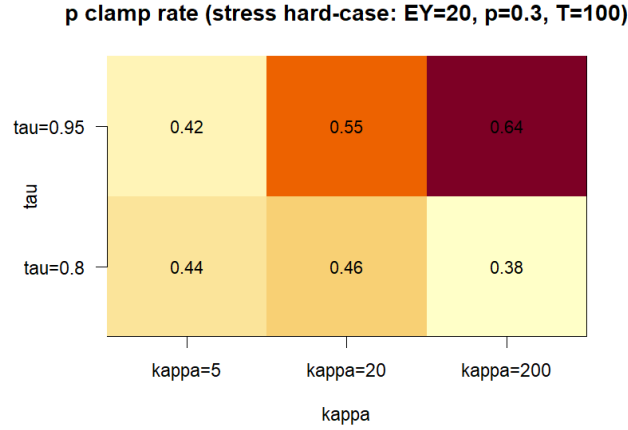


FIGURE 6. Stress design: p -clamp rate across the stress grid. Clamp events are concentrated in the high-persistence, low- κ corners where innovation parameters are weakly identified.

Figure 6. explains the boundary behavior summarized in Table S1. The concentration of clamp events at high τ and low κ indicates weak separation between the thinning-driven dependence and the innovation dispersion in extreme regimes. Tracking clamp rates therefore provides a necessary reliability complement to the RMSE/bias summaries in Table 4.

4.6. Optional MLE add-on. On a representative subset where optimization is stable, the MLE converges in 100% of runs (see `step8_mle_oneparagraph_numbers.csv`) and can improve efficiency, particularly at higher persistence. For configurations with $\tau = 0.8$, MLE reduces RMSE for τ by 0.0133 relative to CLS and reduces RMSE for p by 0.0348 relative to MME (Figure 7). These gains are meaningful, but MLE remains computationally heavier and is best viewed as an optional refinement rather than a default tool in hard regimes.

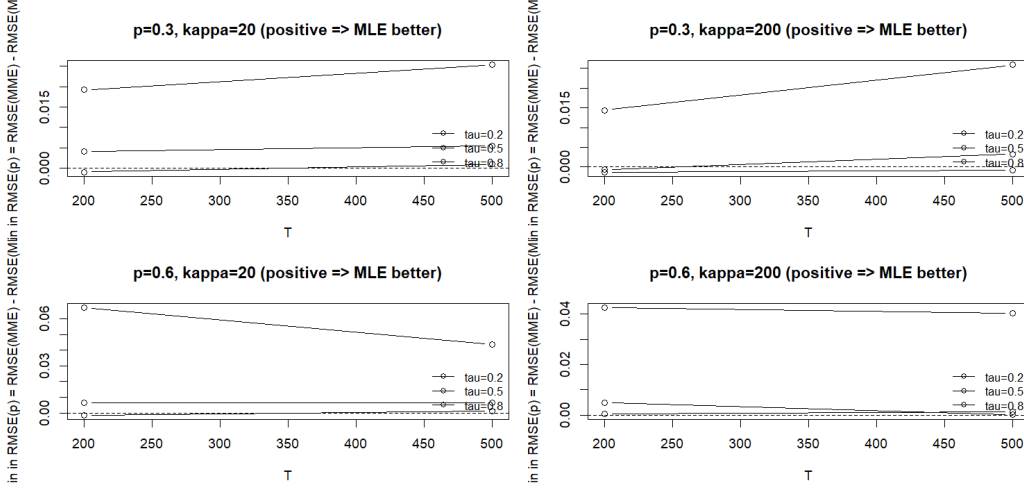


FIGURE 7. Optional MLE add-on (representative subset): RMSE gain for p versus τ (MME minus MLE). Positive values indicate that MLE improves over the moment-based estimate.

Figure 7. shows that an MLE refinement can improve estimation of p in representative settings, especially at higher persistence where MME is least informative about innovations. This figure should be read together with the baseline main-grid summaries in Table 2 and the reliability statistics in Table 3, because numerical stability (and feasibility) ultimately determines whether the MLE add-on is practically beneficial.

Practical takeaway. Across the main grid, all methods improve with T , with especially fast gains for τ . However, innovation parameters become difficult to recover under high persistence and strong thinning heterogeneity, making reliability reporting (feasibility and clamping) essential. When resources permit and optimization is stable, MLE can provide additional efficiency, especially for high persistence.

5. REAL-DATA APPLICATION

5.1. Data description and exploratory diagnostics. We illustrate the proposed BB thinning NB-INAR(1) model using a weekly count time series constructed from an accident-level dataset (`Accident.csv`) containing road-traffic collision records with attributes such as severity, weather, and lighting conditions, and local authority districts (e.g., Craven, Hambleton, Richmondshire, Ryedale, Scarborough, Selby, York). We aggregated the incident records into weekly totals, yielding the series `weekly_total_accidents_clean.csv` with $n = 262$

observations covering 2009–2014. Some summary statistics of these records are presented in Table 5.

Data source: The collision-level records were obtained from the UK Road Safety Data portal. The weekly accident-count series analyzed in this study was constructed from those records by aggregation to weekly totals.

The weekly series has a sample mean 31.90 and variance 56.33, giving a variance-to-mean ratio of 1.77, indicating clear overdispersion relative to a Poisson benchmark.

Dependence diagnostics in Table 6 show non-negligible short-lag persistence: the lag-one sample ACF is approximately 0.195, and the Ljung–Box test rejects the null of no serial correlation. These features motivate a model that jointly captures overdispersion and integer-valued dependence.

Figure 8 visualizes the weekly collision counts over 2009–2014 and complements the summary in Table 5 and the dependence checks in Table 6. The series fluctuates around a fairly stable level but exhibits clear bursts and short-run clustering, which is consistent with overdispersion ($\text{Var}/\text{Mean} > 1$ in Table 5) and motivates an NB-based marginal/innovation structure. The sharp drop at the very end of the plot suggests a potential edge-effect (e.g., an incomplete final aggregation window), which provides practical justification for reporting a trimmed variant (“Trim1”) alongside the full-sample fit in later comparison tables.

TABLE 5. **Weekly time series data summary statistics.**

n	Start	End	Mean	Variance	Var/Mean	Min	Max	Zeros
262	2009-01-05	2014-01-06	31.9008	56.3349	1.7659	3	55	0

TABLE 6. **Dependence and dispersion diagnostics.**

ACF(1)	Ljung–Box lag	Ljung–Box stat	p -value	Var/Mean
0.195	12	64.697	3.34×10^{-9}	1.766

Figure 9 (ACF/PACF) makes the dependence evidence in Table 6 visually explicit. The ACF shows a positive short-lag signal consistent with the reported $\text{ACF}(1) \approx 0.195$ (Table 6), while the PACF highlights that much of the dependence is concentrated at very small lags,

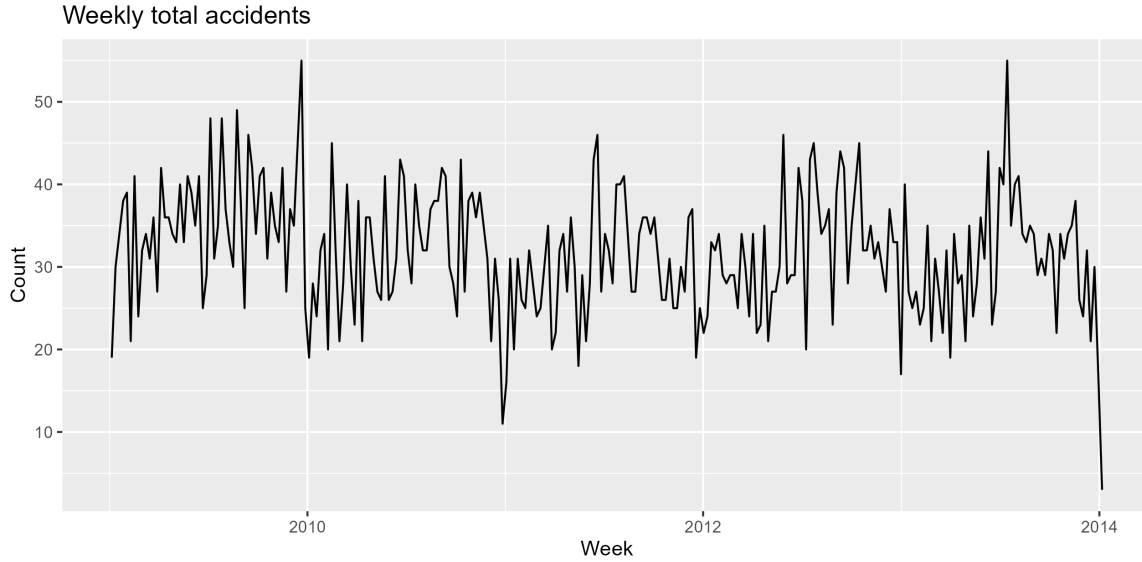


FIGURE 8. Observed weekly counts (2009–2014).

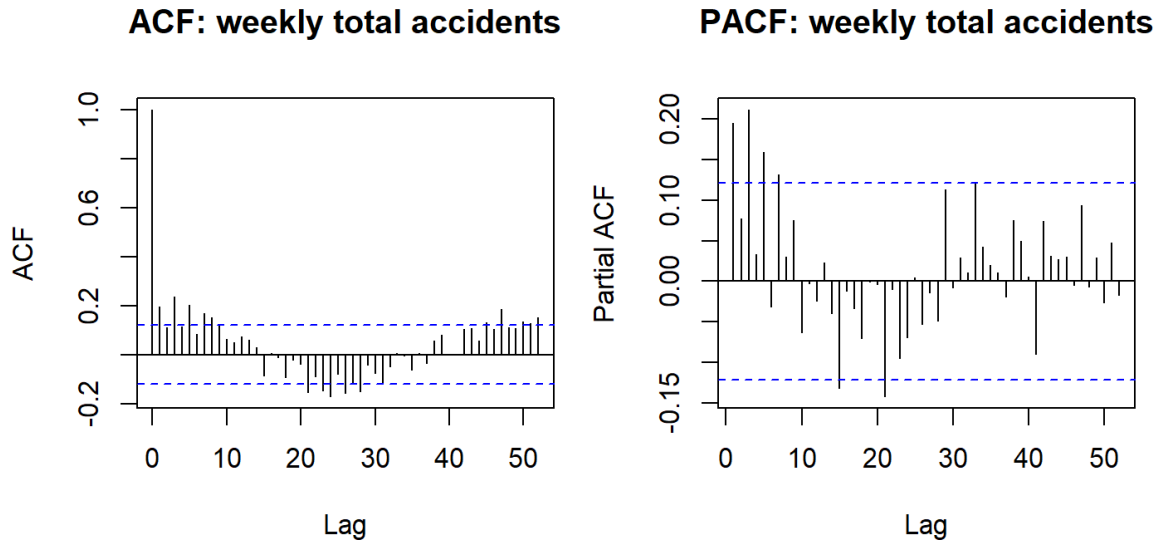


FIGURE 9. Empirical ACF and PACF of the weekly counts.

supporting the use of an order-1 integer-valued dependence mechanism. At the same time, noticeable structure at longer lags is suggestive of recurrent/seasonal effects in weekly data, which motivates the seasonal benchmark considered later (Tables 8, 9) and helps explain why an order-1 model may leave some residual dependence.

5.2. Baseline fits and motivation for thinning-based dependence. As simple baselines, we consider i.i.d. Poisson and i.i.d. Negative Binomial models. While the NB improves fit

over Poisson as shown in Table 7, both baselines are unable to account for the observed serial dependence, as reflected by strong Ljung–Box rejections. This motivates a discrete-valued time-series model.

TABLE 7. **Baseline model comparison (no seasonality).**

Model	logLik	AIC	BIC	Ljung–Box stat	p -value	Note
Poisson	-921.982	1845.964	1849.534	34.045	0.000659	Overdispersed
NB	-905.670	1815.341	1822.482	37.865	0.000136	Still dependent

Figure 10 illustrates why the i.i.d. Poisson and i.i.d. NB baselines in Table 7 are insufficient: their fitted conditional means are essentially level-based summaries and therefore cannot reproduce the observed run-to-run clustering seen in Figure 8. This limitation is also reflected in the Ljung–Box rejections reported in Table 7, confirming that improving marginal fit (NB vs Poisson) does not address serial dependence. These diagnostics collectively motivate moving from i.i.d. count models to a thinning-based time-series specification with an explicit dependence parameter.

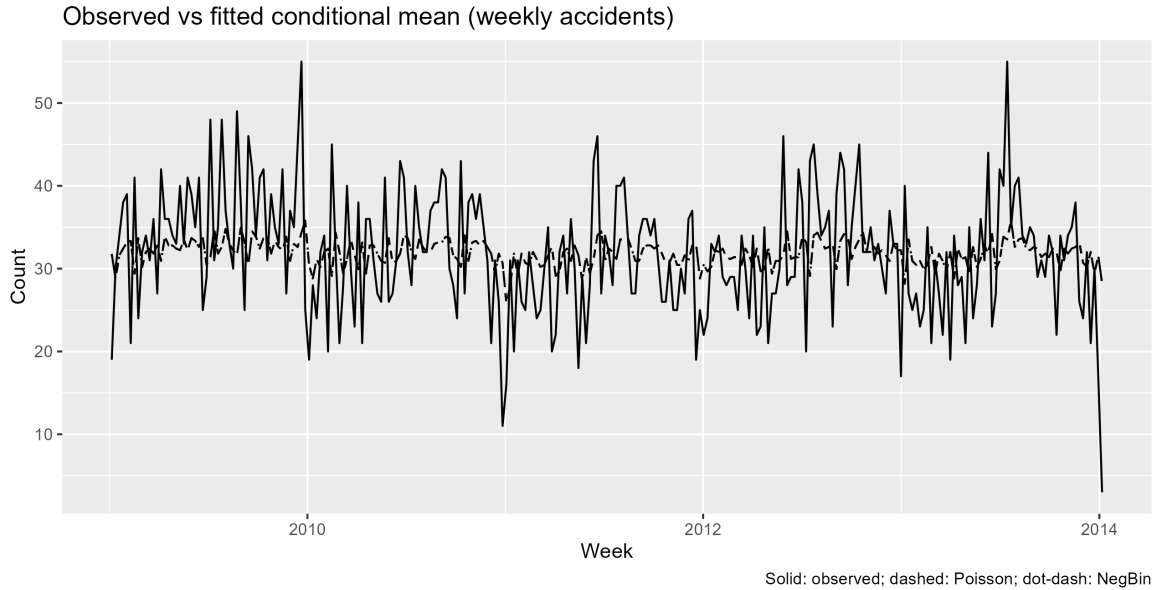


FIGURE 10. Observed series and fitted conditional means under i.i.d. Poisson and NB baselines.

5.3. Seasonal NB AR benchmark. Because weekly data often exhibit annual recurrence, we also consider a seasonal NB AR benchmark with lags 1 and 52 (hereafter “NB AR(1,52)”). This model improves information criteria relative to the i.i.d. baselines as presented in Table 8 and yields significant lag-1 and lag-52 effects as shown in Table 9, consistent with short-term persistence plus seasonal structure.

TABLE 8. **Model comparison including seasonality (benchmark).**

Model	logLik	AIC	BIC	Ljung–Box stat	<i>p</i> -value	Note
NB AR(1,52) (full)	-897.116	1802.232	1816.495	32.539	0.001545	Seasonal benchmark
NB AR(1,52) (trim1)	-880.419	1768.837	1783.080	34.406	0.000632	Trim1 benchmark

TABLE 9. **NB AR(1,52) parameter estimates.**

Fit	Intercept	Lag-1 coef	Lag-52 coef	Dispersion (size)
Full	18.630	0.196	0.074	61.245
Trim1	19.080	0.209	0.083	65.717

Figure 11 shows the observed series against the fitted conditional mean from the seasonal benchmark NB AR(1,52), complementing the information-criteria improvements in Table 8 and the significant lag-1 and lag-52 effects in Table 9. Visually, the benchmark smooths the series while incorporating both short-term persistence and annual recurrence, making it a strong comparator for any order-1 thinning model. This benchmark therefore sets a meaningful reference point for assessing whether BB–NB–INAR(1) can deliver competitive fit with fewer structural lags (Tables 10,11).

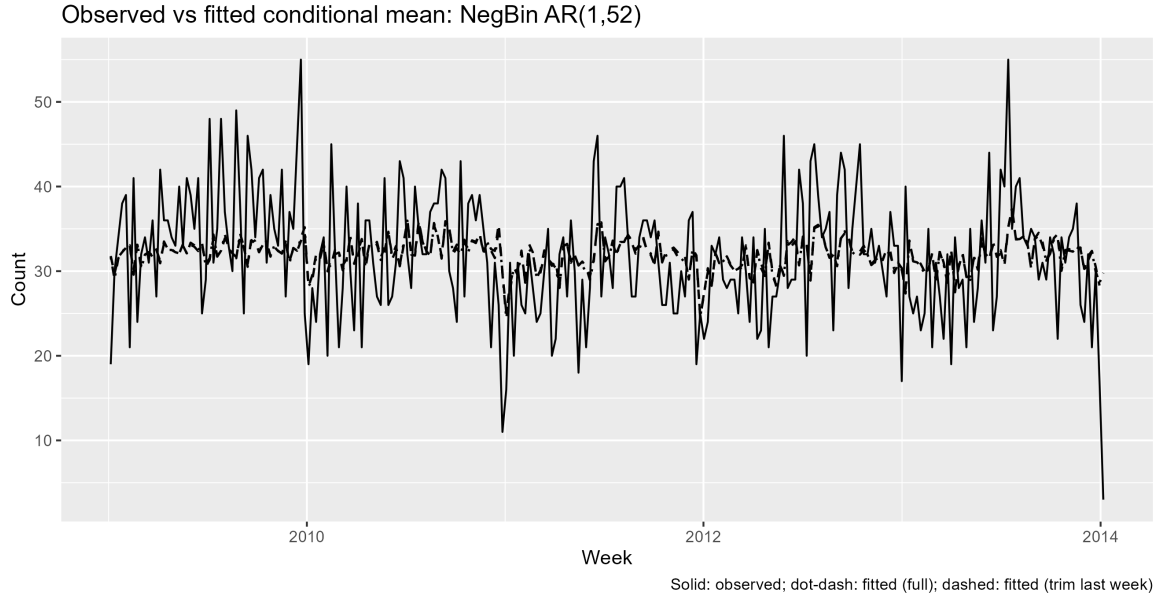


FIGURE 11. Observed series and fitted conditional mean under the seasonal benchmark NB AR(1,52).

5.4. Fit of the proposed BB–NB–INAR(1) model. We fit the proposed BB–NB–INAR(1) model by conditional maximum likelihood (CML) using constrained optimization. The optimizer converged successfully (BFGS with convergence code 0). Parameter estimates are reported in Table 10. The estimated persistence parameter $\hat{\tau}$ is close to the empirical ACF(1) (Table 6), supporting the interpretation of τ as the lag-one dependence strength. The estimated precision $\hat{\kappa}$ is moderate, indicating non-negligible heterogeneity in the thinning probability across time.

In terms of in-sample fit, BB–NB–INAR(1) attains slightly better information criteria than the seasonal benchmark in this dataset: for the full sample, AIC improves by about 3.18 (1799.05 vs. 1802.23), and for Trim1 by about 0.87 (1767.97 vs. 1768.84).

TABLE 10. BB–NB–INAR(1) parameter estimates (CML).

Fit	$\hat{\tau}$ (SE)	$\hat{\kappa}$ (SE)	$\hat{\rho}$ (SE)	$\hat{\phi}$ (SE)
Full	0.207 (0.075)	37.117 (30.733)	0.536 (0.123)	61.047 (138.433)
Trim1	0.180 (0.067)	55.378 (49.093)	0.633 (0.176)	28.557 (45.221)

TABLE 11. **BB-NB-INAR(1) fit metrics.**

Fit	logLik	AIC	BIC	Optimizer
Full	-895.525	1799.050	1813.308	BFGS
Trim1	-879.984	1767.968	1782.211	BFGS

Note on Trim1. Trim1 denotes the variant in which the final weekly observation is removed because the sharp terminal drop suggests a likely edge effect from an incomplete final aggregation window.

5.5. Calibration and residual diagnostics. To assess distributional calibration for the fitted BB-NB-INAR(1) model (Trim1 fit), we examine probability integral transform (PIT) diagnostics computed from one-step-ahead predictive distributions. The PIT summary statistics are close to the Uniform(0, 1) benchmark (mean ≈ 0.524 ; variance ≈ 0.083 , close to $1/12 \approx 0.0833$), suggesting reasonable marginal calibration as stated in Table 12.

However, the Ljung–Box test on randomized residuals still rejects at lag 12 ($p \approx 6.6 \times 10^{-4}$; Table 12), and the residual ACF indicates remaining serial structure. This is consistent with the fact that the proposed model is order-1 and does not explicitly include a seasonal lag (e.g., 52), so some seasonal dependence may persist in the residuals.

TABLE 12. **BB-NB-INAR(1) diagnostics (Trim1).**

Ljung–Box lag	Stat	p -value	PIT mean	PIT var	PIT min	PIT max
12	34.063	0.000659	0.523993	0.083378	0.000483	0.996235

Figure 12 provides a graphical companion to the PIT summary in Table 12. The near-uniform histogram shape is consistent with the PIT mean and variance being close to the Uniform(0, 1) benchmarks (Table 12), supporting the claim of reasonable one-step distributional calibration for the fitted BB-NB-INAR(1) model (Trim1). Importantly, this calibration evidence should be interpreted alongside the Ljung–Box result in Table 12: good marginal calibration can coexist with remaining serial structure in randomized residuals, which is plausible here because an order-1 dependence model does not explicitly encode seasonal lags.

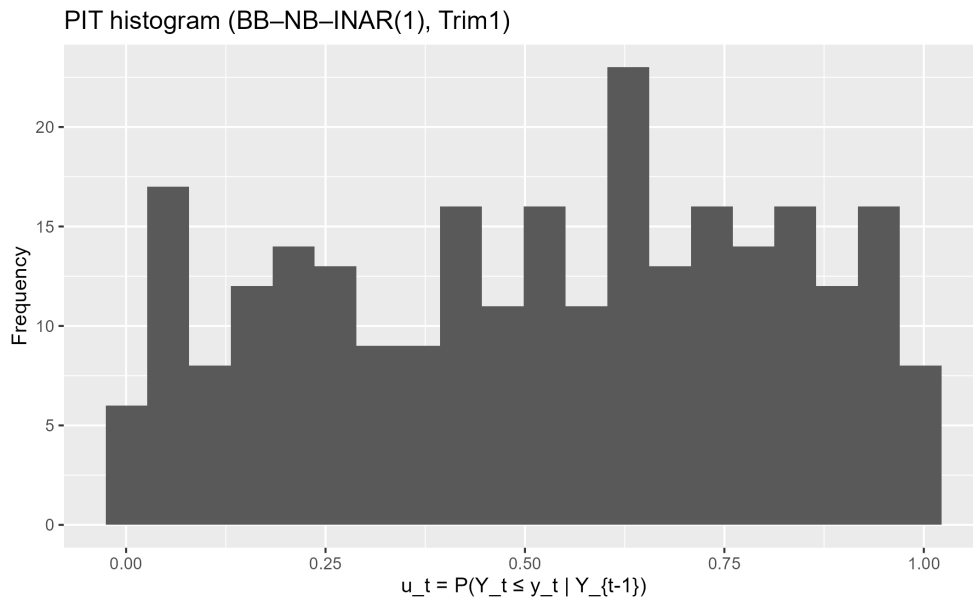


FIGURE 12. PIT histogram for BB–NB–INAR(1) (Trim1): near-uniform shape indicates reasonable calibration.

ACF of PIT residuals (BB–NB–INAR(1), Trim

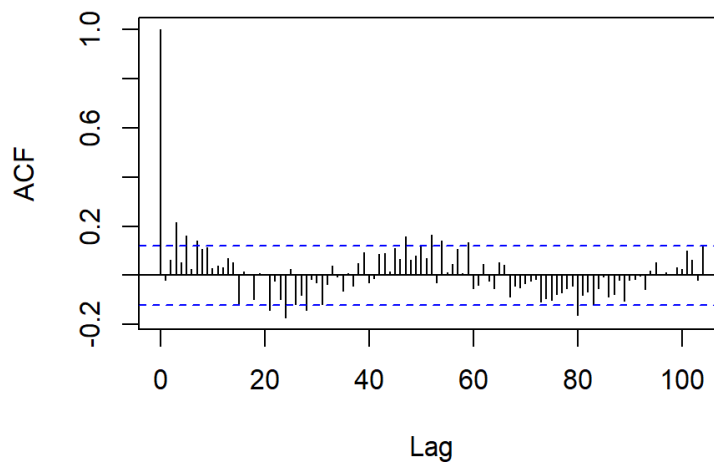


FIGURE 13. Residual ACF and PIT-based residual diagnostics for BB–NB–INAR(1) (Trim1). Remaining autocorrelation is consistent with unmodeled seasonal structure.

Figure 13. complements the diagnostic summaries in Table 12. It shows that good one-step calibration (PIT moments near Uniform(0,1)) can coexist with remaining autocorrelation in randomized/PIT-based residuals, which is plausible here because an order-1 dependence model does not explicitly encode seasonal lags.

5.6. Illustrative one-step probabilistic forecast. Figure 14 shows an illustrative one-step-ahead probabilistic forecast under the fitted BB–NB–INAR(1) model (Trim1). The predictive mean tracks the observed series reasonably well, while the predictive bands reflect uncertainty that is larger in high-count periods, consistent with the NB dispersion structure.

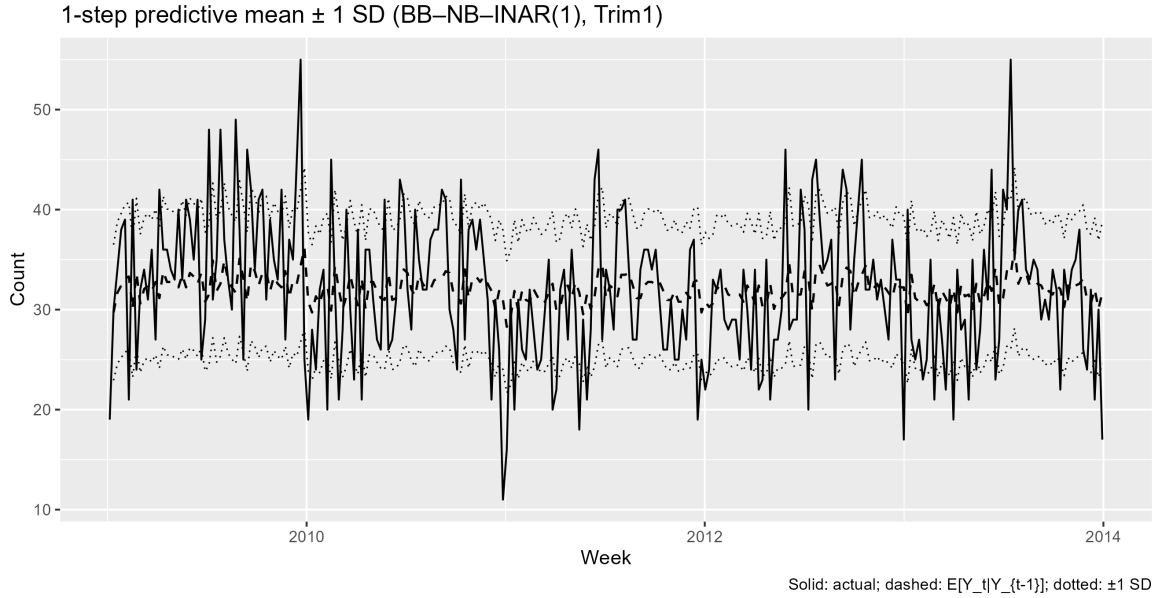


FIGURE 14. One-step-ahead probabilistic forecast under BB–NB–INAR(1) (Trim1): predictive mean and uncertainty bands.

Figure 14. translates the fitted BB–NB–INAR(1) parameters (Table 10) into an interpretable predictive output. The uncertainty bands widen during high-count episodes, reflecting NB dispersion, while the predictive mean tracks short-run persistence governed by $\hat{\tau}$; this aligns with the overall fit and diagnostic evidence reported in Tables 11–12.

6. CONCLUSION

This paper introduces a stationary negative binomial (NB) INAR(1) model built on a beta-binomial (BB) thinning operator. Randomizing the thinning probability through a Beta distribution adds an interpretable layer of heterogeneity to the autoregressive mechanism: τ governs dependence (and equals the lag-one autocorrelation under stationarity), while κ controls the strength of time-varying thinning (with the fixed-thinning NB-INAR(1) recovered as $\kappa \rightarrow \infty$). By pairing this thinning mechanism with NB innovations chosen to preserve a NB marginal distribution, the model provides a coherent framework for count series that are both serially dependent and overdispersed.

Finite-sample evidence from the Monte Carlo study supports three main messages. First, estimation accuracy improves consistently as the sample size increases across the main design grid. The RMSE declines monotonically for all parameters, while small-sample biases diminish rapidly with increasing T . Second, thinning heterogeneity matters for inference: larger κ (less heterogeneity) yields modest but consistent gains, especially for the innovation parameter p . Third, and most importantly for applied use, accuracy must be interpreted alongside reliability: in high-persistence settings with strong heterogeneity (small κ), innovation parameters become weakly identified in short samples, which is reflected both in higher RMSE_p and in increased boundary clamping and moment infeasibility. The stress regime ($\tau = 0.95$) makes this limitation explicit: while τ remains accurately estimable due to the dominant autoregressive signal, estimation of p deteriorates sharply and feasibility can drop in the hardest corner. When numerical optimization is stable, the optional MLE refinement can deliver efficiency gains, particularly at higher persistence, but it should be viewed as a complementary tool rather than a default estimator in extreme regimes.

The real-data application to weekly road-traffic collision counts illustrates the practical relevance of the proposed framework. The series is clearly overdispersed and exhibits statistically significant serial dependence, which i.i.d. Poisson and i.i.d. NB baselines cannot capture. Relative to a seasonal NB AR benchmark with lags 1 and 52, the BB–NB–INAR(1) achieves slightly improved information criteria and provides reasonably calibrated one-step-ahead predictive distributions based on PIT diagnostics. Remaining residual dependence is still detectable, consistent

with the fact that an order-1 thinning model does not explicitly incorporate seasonal lags; this suggests natural extensions.

Several directions follow directly from these results. A seasonal or higher-order extension (e.g., including lag 52 or combining multiple thinning components) is a principled way to address persistent seasonal autocorrelation in weekly data. More broadly, covariate-driven thinning or innovation components could incorporate external drivers when available. Finally, because near-unit-root persistence coupled with strong thinning heterogeneity challenges innovation identification, future work could investigate estimation strategies tailored to this regime, including regularized likelihood approaches and improved feasibility-aware moment procedures. All simulation outputs and real-data analysis artifacts are provided in the online supplement for reproducibility: <https://doi.org/10.5281/zenodo.18389818>.

APPENDIX A. SUPPLEMENTARY MATERIAL AND REDUCED APPENDIX

A.1. Online supplement (full tables and figures). Full simulation supplement. Complete scenario-by-scenario outputs (all bias/RMSE tables, full reliability tables, and all diagnostic figures) are available in the online supplement: <https://doi.org/10.5281/zenodo.18389818>

The supplement includes:

- Complete main-grid and stress-grid bias/RMSE tables for (τ, κ, μ, p) across all T .
- Complete reliability tables (MME feasibility, failure decomposition, clamp frequencies, and optional MLE convergence where applicable).
- All generated diagnostic plots (RMSE panels, heatmaps, trade-off plots, and MLE comparison plots).

To keep the manuscript concise, the appendix below reports only:

- (i) the hardest stress-case reliability configurations (Table S1) highlighting the regimes where MME feasibility deteriorates, and
- (ii) one representative main-grid RMSE slice (Table S2), illustrating estimation accuracy across sample sizes and parameter settings, together with
- (iii) one compact diagnostic figure summarizing the feasibility–clamping trade-off (Figure S1).

A.2. Worst-case reliability configurations (stress corner).

set	EY	p_true	kappa	tau_true	T	R	mme_success_rate	mme_failA_rate	mme_failVar_rate	p_clamp_rate	tau_clamp_rate	mu_clamp_rate
stress	20	0.3000	5	0.9500	100	300	0.4167	0.0000	0.5833	0.4160	0.0000	0.0000
stress	20	0.6000	5	0.9500	100	300	0.4933	0.0000	0.5067	0.3581	0.0000	0.0000
stress	20	0.3000	20	0.9500	100	300	0.5567	0.0000	0.4433	0.5509	0.0000	0.0000
stress	20	0.6000	20	0.9500	100	300	0.6400	0.0000	0.3600	0.4764	0.0000	0.0000
stress	20	0.3000	200	0.9500	100	300	0.7933	0.0000	0.2067	0.5140	0.0000	0.0000
stress	20	0.6000	200	0.9500	100	300	0.8467	0.0000	0.1533	0.4797	0.0000	0.0000
stress	20	0.3000	5	0.9500	200	300	0.7733	0.0000	0.2267	0.4091	0.0000	0.0000
stress	20	0.6000	5	0.9500	200	300	0.8433	0.0000	0.1567	0.3825	0.0000	0.0000
stress	20	0.3000	20	0.9500	200	300	0.8933	0.0000	0.1067	0.5409	0.0000	0.0000
stress	20	0.6000	20	0.9500	200	300	0.9500	0.0000	0.0500	0.4764	0.0000	0.0000
stress	20	0.3000	200	0.9500	200	300	0.9967	0.0000	0.0033	0.5011	0.0000	0.0000
stress	20	0.6000	200	0.9500	200	300	0.9967	0.0000	0.0033	0.4631	0.0000	0.0000
stress	20	0.3000	5	0.9500	500	300	0.9667	0.0000	0.0333	0.3075	0.0000	0.0000
stress	20	0.6000	5	0.9500	500	300	0.9867	0.0000	0.0133	0.2738	0.0000	0.0000
stress	20	0.3000	20	0.9500	500	300	0.9967	0.0000	0.0033	0.4258	0.0000	0.0000
stress	20	0.6000	20	0.9500	500	300	1.0000	0.0000	0.0000	0.3853	0.0000	0.0000
stress	20	0.3000	200	0.9500	500	300	1.0000	0.0000	0.0000	0.3867	0.0000	0.0000
stress	20	0.6000	200	0.9500	500	300	1.0000	0.0000	0.0000	0.3453	0.0000	0.0000
stress	20	0.3000	5	0.9500	1000	300	0.9933	0.0000	0.0067	0.1746	0.0000	0.0000
stress	20	0.6000	5	0.9500	1000	300	1.0000	0.0000	0.0000	0.1498	0.0000	0.0000

TABLE S1. Worst-case reliability configurations (top 20): MME success/failure decomposition and clamp rates across the hardest stress scenarios.

A.3. One representative scenario table (RMSE).

EY	p_true	kappa	tau_true	tau				mu				p			
				T=100	T=200	T=500	T=1000	T=100	T=200	T=500	T=1000	T=100	T=200	T=500	T=1000
20	0.6000	5	0.2000	0.1020	0.0770	0.0445	0.0340	2.1725	1.5227	0.9049	0.6333	0.0997	0.0719	0.0545	0.0325
20	0.6000	5	0.5000	0.1040	0.0733	0.0447	0.0323	1.9518	1.4272	0.8346	0.5820	0.1393	0.0980	0.0601	0.0441
20	0.6000	5	0.8000	0.0851	0.0533	0.0323	0.0224	1.6413	0.9856	0.5948	0.3920	0.2420	0.2107	0.1757	0.1248
20	0.6000	20	0.2000	0.1006	0.0746	0.0441	0.0312	2.1384	1.4996	0.8824	0.6215	0.0985	0.0707	0.0535	0.0320
20	0.6000	20	0.5000	0.1017	0.0727	0.0439	0.0314	1.8983	1.3889	0.8261	0.5786	0.1327	0.0922	0.0592	0.0435
20	0.6000	20	0.8000	0.0813	0.0502	0.0308	0.0218	1.5902	0.9467	0.5833	0.3872	0.2329	0.2057	0.1741	0.1241
20	0.6000	200	0.2000	0.1005	0.0746	0.0441	0.0312	2.1368	1.4992	0.8822	0.6215	0.0985	0.0707	0.0535	0.0320
20	0.6000	200	0.5000	0.1017	0.0727	0.0439	0.0314	1.8982	1.3888	0.8261	0.5786	0.1327	0.0922	0.0592	0.0435
20	0.6000	200	0.8000	0.0813	0.0502	0.0308	0.0218	1.5902	0.9467	0.5833	0.3872	0.2329	0.2057	0.1741	0.1241

TABLE S2. Representative main-grid slice (EY=20, $p = 0.6$): RMSE for (τ, μ, p) across T , κ , and τ .

A.4. One supporting supplementary figure.

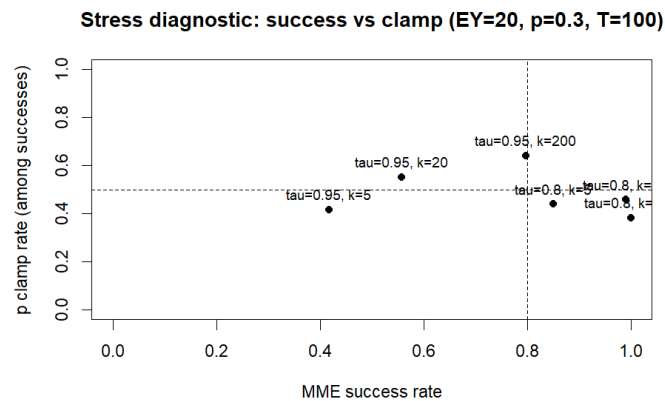


FIGURE S1. **Supplementary diagnostic: feasibility–clamping trade-off.** Summary visualization showing how MME success probability and boundary clamping co-move across difficult regimes.

Figure S1. summarizes the reliability message in Table S1: feasibility and boundary clamping tend to move together in the hardest regimes (high persistence and strong BB heterogeneity). This provides a compact practitioner-oriented summary of where finite-sample inference is expected to be numerically fragile, complementing the stress-design averages in Table 4.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author, S.W.M., upon reasonable request.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

- [1] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis: Forecasting and Control, 3rd ed., Prentice Hall, 1994.
- [2] F.W. Steutel, K. van Harn, Discrete Analogues of Self-Decomposability and Stability, Ann. Probab. 7 (1979), 893–899. <https://doi.org/10.1214/aop/1176994950>.
- [3] E. McKenzie, Some Simple Models for Discrete Variate Time Series, J. Am. Water Resour. Assoc. 21 (1985), 645–650. <https://doi.org/10.1111/j.1752-1688.1985.tb05379.x>.

- [4] M.A. Al-Osh, A.A. Alzaid, First-Order Integer-Valued Autoregressive (INAR(1)) Process, *J. Time Ser. Anal.* 8 (1987), 261–275. <https://doi.org/10.1111/j.1467-9892.1987.tb00438.x>.
- [5] J.-G. Du, Y. Li, The Integer-Valued Autoregressive (INAR(p)) Model, *J. Time Ser. Anal.* 12 (1991), 129–142. <https://doi.org/10.1111/j.1467-9892.1991.tb00073.x>.
- [6] C.H. Weiß, Thinning Operations for Modeling Time Series of Counts—A Survey, *AStA Adv. Stat. Anal.* 92 (2008), 319–341. <https://doi.org/10.1007/s10182-008-0072-3>.
- [7] K. Fokianos, Count Time Series Models, in: *Time Series Analysis: Methods and Applications, Handbook of Statistics*, Elsevier, 2012, pp. 315–347. <https://doi.org/10.1016/B978-0-444-53858-1.00012-0>.
- [8] R.A. Davis, H. Liu, Theory and Inference for a Class of Observation-Driven Models with Application to Time Series of Counts, *arXiv:1204.3915*, 2012. <https://doi.org/10.48550/arXiv.1204.3915>.
- [9] R.K. Freeland, B.P.M. McCabe, Forecasting Discrete Valued Low Count Time Series, *Int. J. Forecast.* 20 (2004), 427–434. [https://doi.org/10.1016/S0169-2070\(03\)00014-1](https://doi.org/10.1016/S0169-2070(03)00014-1).
- [10] R.C. Jung, A.R. Tremayne, Coherent Forecasting in Integer Time Series Models, *Int. J. Forecast.* 22 (2006), 223–238. <https://doi.org/10.1016/j.ijforecast.2005.07.001>.
- [11] S. Schweer, C.H. Weiß, Compound Poisson INAR(1) Processes: Stochastic Properties and Testing for Overdispersion, *Comput. Stat. Data Anal.* 77 (2014), 267–284. <https://doi.org/10.1016/j.csda.2014.03.005>.
- [12] H. Joe, Time Series Models with Univariate Margins in the Convolution-Closed Infinitely Divisible Class, *J. Appl. Probab.* 33 (1996), 664–677. <https://doi.org/10.2307/3215348>.
- [13] R. Zhu, H. Joe, A New Type of Discrete Self-Decomposability and Its Application to Continuous-Time Markov Processes for Modeling Count Data Time Series, *Stoch. Models* 19 (2003), 235–254. <https://doi.org/10.1081/STM-120020388>.
- [14] F. Zhu, A Negative Binomial Integer-Valued GARCH Model, *J. Time Ser. Anal.* 32 (2011), 54–67. <https://doi.org/10.1111/j.1467-9892.2010.00684.x>.
- [15] M.G. Scotto, C.H. Weiß, S. Gouveia, Thinning-Based Models in the Analysis of Integer-Valued Time Series: A Review, *Stat. Model.* 15 (2015), 590–618. <https://doi.org/10.1177/1471082X15584701>.
- [16] K. Brännäs, A. Hall, Estimation in Integer-Valued Moving Average Models, *Appl. Stoch. Model. Bus. Ind.* 17 (2001), 277–291. <https://doi.org/10.1002/asmb.445>.
- [17] E. McKenzie, Autoregressive Moving-Average Processes with Negative-Binomial and Geometric Marginal Distributions, *Adv. Appl. Probab.* 18 (1986), 679–705. <https://doi.org/10.2307/1427183>.
- [18] C.H. Weiß, Stationary Count Time Series Models, *WIREs Comput. Stat.* 13 (2021), e1502. <https://doi.org/10.1002/wics.1502>.
- [19] M.A. Al-Osh, E.-E.A.A. Aly, First Order Autoregressive Time Series with Negative Binomial and Geometric Marginals, *Commun. Stat. Theory Methods* 21 (1992), 2483–2492. <https://doi.org/10.1080/03610929208830925>.

- [20] Z. Liu, F. Zhu, A New Extension of Thinning-Based Integer-Valued Autoregressive Models for Count Data, *Entropy* 23 (2021), 62. <https://doi.org/10.3390/e23010062>.
- [21] D. Karlis, N. Mamode Khan, Y. Sunecher, The Negative Binomial INAR(1) Process under Different Thinning Processes: Can We Separate between the Different Models?, *Stats* 7 (2024), 793–807. <https://doi.org/10.3390/stats7030048>.
- [22] C. Gouriéroux, Y. Lu, Negative Binomial Autoregressive Process with Stochastic Intensity, *J. Time Ser. Anal.* 40 (2019), 225–247. <https://doi.org/10.1111/jtsa.12441>.
- [23] Averill M Law, W David Kelton, and W David Kelton, *Simulation modeling and analysis*, Mcgraw-hill New York, 2007.
- [24] B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall/CRC, 1993. <https://doi.org/10.1201/9780429246593>.
- [25] C.H. Weiß, *An Introduction to Discrete-Valued Time Series*, Wiley, 2018. <https://doi.org/10.1002/9781119097013>.