Count Time Series Forecasting Applied to Crime Series in a Multi-Neighborhood Urban Region

Boanerges T. de Almeida[1], **Abraao H. S. Rosal**[1], **José E. B. Maia**[1]

[1]Universidade estadual do Ceará - UECE, 60714-903 - Fortaleza - CE - Brasil

boanerges.teixeira,abraao.rosal@aluno.uece.br

Abstract. *This article addresses count time series forecasting applied to crime series in Fortaleza City. It involves modeling and forecasting series whose observations are non-negative integer counts, generally with overdispersion and excess zeros. This regime violates the Gaussian assumptions of classical time series models (ARIMA, exponential smoothing), which motivates a specific family of methods. The series of criminal occurrences in Fortaleza (2022–2025) is divided into two contrasting targets: CVLI (Intentional Lethal Violent Crimes), a rare and sparse event; and CVP (Violent Crimes against Property), a frequent and strongly overdispersed event. Prediction experiments with various counting models, namely Poisson/Negative Binomial Regression, INGARCH, INAR, show that they outperform naive baselines. The marginal gain attributable to the autoregressive structure is also investigated.*

1. Introduction

Count time series are sequences of positive integer values generated by phenomena such as the number of daily hospital admissions, the number of monthly disease cases in an epidemic, or the number of weekly crimes in a region or city. Models such as AR and ARMA do not directly apply to count time series because they were designed for continuous variables and assume a normal (Gaussian) distribution of the data. For example, common linear models can predict negative values, which is biologically or physically impossible for a count.

This article discusses count time series forecasting applied to crime series in Fortaleza City. According to contemporary literature reviews [Butt et al. 2020, Yu et al. 2024], crime analytics can be categorized based on its data dimensions. While crime prediction often focuses on classification models to infer types of crimes or offender profiles, crime hotspot prediction relies heavily on spatial algorithms to discover geographical clusters of high-risk density. Distinctly, crime time series prediction isolates the temporal component to predict future count trends and seasonal fluctuations over fixed intervals [Yu et al. 2024].

Additionally, traditional continuous evaluation metrics, such as RMSE and MAPE, are mathematically suboptimal for count time series forecasting. As demonstrated by [Hyndman and Koehler 2006], percentage errors yield infinite or undefined values when zero-count events occur. Furthermore, evaluating discrete data under quadratic loss assumptions can bias predictive performance; optimizing for Mean Absolute Error (MAE) in low-count sparse series inherently drives forecasts toward the median (often zero), rendering the evaluation misleading [Kolassa 2016]. Therefore, predictive assessment for count-type data requires specialized scoring models or scaled metrics that respect the underlying discrete distribution [Czado et al. 2009].

In this paper, the central research question is: how do the properties of time series of Violent Crimes against Property (CVP) and Intentional Lethal Violent Crimes (CVLI) (dispersion, autocorrelation, seasonality, proportion of zeros) vary with the temporal (daily, weekly, monthly) aggregation scale, and how does this variation condition the appropriate model class for forecasting short-term CVP and CVLI crime event counts?

This paper divides the series of criminal occurrences in Fortaleza (2022–2025) into two contrasting regimes: CVLI (Intentional Lethal Violent Crimes), a rare and sparse event; and CVP (Violent Crimes against Property), a frequent and strongly overdispersed event. Prediction experiments with various counting models, namely Poisson/Negative Binomial Regression, INGARCH, INAR, show that they outperform naive baselines. The marginal gain attributable to the autoregressive structure is also investigated.

The remainder of this article is organized as follows: Section 2 reviews some works strongly related to this one. Section 3 provides an exploratory analysis of the data. Section 4 is a brief review of the fundamentals of the methods, and Section 5 presents the processing results accompanied by a discussion. Section 6 concludes the article.

2. Related Works

This section presents a brief review of some works directly related to our research, aiming to establish the state of the art.

[Fokianos et al. 2009] introduced the Poisson autoregression model, in which the conditional intensity λ_t follows a multiplicative dynamic analogous to GARCH: $\lambda_t = exp(d + a.log(y_{t-1} + 1) + b.log(\lambda_{t-1}))$. The authors established ergodic stationarity conditions and proved the consistency and asymptotic normality of maximum likelihood estimators, providing the rigorous statistical basis for the INGARCH (Integer-valued GARCH) class. The extension to Negative Binomial likelihood accommodates residual overdispersion. **Data and results**.The model was validated on real count series — monthly credit transaction rates and network traffic series — with well-characterized serial autocorrelation. The results demonstrate that the autoregressive structure captures temporal dependence more parsimoniously than linear models, producing more calibrated predictions when the series exhibits memory. INGARCH has established itself as a canonical reference for count series with temporal dynamics. **Link to the present work**. INGARCH(1,1) is the central temporal component of the proposed methodology. In the CVLI data from Fortaleza - a series with weak autocorrelation (ACF = 0.04) and dispersion close to 1.45 - the INGARCH Poisson reduces wMAPE by 8.1 p.p. compared to GLM with trend and seasonality (44.6% vs. 52.7%), confirming the gain of the autoregressive structure even in regimes of weak memory. The NB-INGARCH version provides the best log-score in the overdispersed regime (CVP), connecting regime diagnosis and model selection in a methodologically clear way.

[Mohler et al. 2011] modeled crimes as Hawkes self-exciting point process (SEPP), in which each event temporarily raises the intensity of future events in the space-time vicinity. The intensity decomposes into a background component (long-term level) and a decaying core excitation in time and space, estimated via likelihood maximization over the complete sequence of point events. The model explicitly captures the contagion or opportunity structure that characterizes many types of crime. **Data and results**. Applied to Los Angeles data (residential burglary and homicide), SEPP outperformed

static kernel density (KDE) estimation in hotspot accuracy, producing the first controlled demonstration that the spatiotemporal structure of the process enhances the anticipation of where future events will occur. The model revealed that crimes tend to cluster in space-time on scales of hours to days—a pattern interpreted as contagion or opportunity exploitation. **Link to the present work**. SEPP motivates two elements of the proposed methodology: (i) the process-derived sparsity attributes (time since the last event, distance to the recent event, recent counts in radii of 3/5/10 km), used as covariates in the Top-K ranking; and (ii) the finding that the Fortaleza CVLI — a rare event with a persistent concentration pattern — shares the sparsity and clustering regime studied by Mohler et al., making these attributes directly applicable to spatial ranking.

[Adepeju et al. 2016] identified that conventional error metrics (RMSE, R2) are inadequate for evaluating hotspot predictions in sparse point processes, as they are dominated by the many zero-count cells. As a solution, they proposed the **PAI** (Predictive Accuracy Index) — the ratio between the capture rate (fraction of the events in the test period contained in the predicted hotspot) and the prioritized area fraction: $PAI = Recall@K/(TopK_Area/Total_Area)$ — along with complementary metrics (PEI: relative predictive efficiency; ALS: accuracy by location). The protocol uses an out-of-sample sliding window. **Data and results**. Applied to UK theft data, the study showed that methods ranked as superior by RMSE were inferior by PAI, reversing the performance ranking. PAI revealed that some models obtained high recalls simply by covering large fractions of the territory — behavior masked by RMSE but exposed by the separate numerator and denominator of PAI. The metric has become a reference for evaluating sparse hotspots in the crime prediction literature. **Link to the present work**. PAI is the main metric of this article. The Pareto frontier produced by sweeping λ is organized exactly in terms of PAI: when increasing λ (area penalty), the GA solution migrates from PAI = 1.27 ($\lambda = 0$, area 177 km2) to PAI = 3.58 ($\lambda = 2$, area 17.7 km2), systematically covering the entire curve. The budget penalty mechanism in the objective function is, in the language of Adepeju et al., the strategy that prevents the denominator of the PAI from growing freely — making GA an optimizer of PAI, not just of crude recall.

[Villegas et al. 2022] systematically investigated the effect of spatiotemporal sparsity on homicide prediction models — an analogous target for CVLI. The authors compared statistical (Poisson, Negative Binomial, hurdle) and machine learning (Random Forest, XGBoost) models at different spatial (census tracts, neighborhoods, zones) and temporal (week, month, quarter) granularities, measuring the systematic impact of sparsity on the performance of each model class across multiple Colombian cities. **Data and results**. The main results are: (i) no model dominates in all scenarios; (ii) sparsity increases monotonically with spatial and temporal resolution, degrading models differentially — count models are more robust at fine granularity than Gaussian regression models; and (iii) the zero ratio is the strongest predictor of relative performance between model classes, more so than autocorrelation or overdispersion alone. Overdispersion requires the Negative Binomial; in low autocorrelation regimes, the gain from the temporal component is limited. **Link to the present work**. The regime studied by Villegas et al. — sparse homicide at high resolution — is the direct analogue of the CVLI in Fortaleza. The findings corroborate and anticipate the results of this dissertation: the scale dependence found in Fortaleza (daily series with 13.6% zeros changes regime; weekly is almost equidispersed) replicates the pattern reported for Colombian cities. The con-

clusion that no model is universally superior reinforces the multiple models and multiple metrics protocol to be adopted.

[Andrade et al. 2025] proposed STEval, a modular framework for standardized evaluation of spatiotemporal crime prediction models. The framework divides the territory into regular grids of parameterized size (100 m to 1000 m), trains and evaluates four models — STKDE (Spatiotemporal Density Kernel), SARIMA, STHSL (Hypergraph Self-Supervised Learning) and Extra Trees — in multiple spatiotemporal granularity scenarios, and compares the results by a unified set of metrics (HR, PAI, PEI, ALS), with significance by the paired Wilcoxon test. **Data and results**. Validated on 40 million theft records from Minas Gerais, STEval demonstrated that: (i) no model is universally superior — STKDE dominates in fine granularities and short term; SARIMA is efficient in long time intervals; STHSL excels in horizons far from training; (ii) relative performance changes systematically with granularity; and (iii) the set of metrics (HR, PAI, PEI, ALS) is necessary to capture complementary dimensions of predictive performance. **Link to the present work**. STEval is the closest national precedent. The differences are in nature and regime. Nature: STEval is an evaluation framework — it compares models under standardized conditions; this dissertation is a predictive method with mesh co-optimization and ranking. Spatial representation: in STEval, the grid is a fixed parameter; here, geometry, side, displacement, and rotation are decision variables optimized by GA. Regime: STEval validates on theft (dense event); the focus here is CVLI (rare, sparse, 3031 events in 4 years), a regime that requires sparsity metrics and optimization regularization. As a point of convergence, the STEval set of metrics is adopted in its entirety to ensure comparability with national literature.

By cross-referencing the five studies, the gap is precisely identified. The literature on count series [Fokianos et al. 2009] models temporal dynamics over given spatial partitions; the self-exciting process [Mohler et al. 2011] motivates sparsity attributes but does not optimize the grid; the metrics of [Adepeju et al. 2016] evaluate hotspots but do not optimize representation; [Villegas et al. 2022] confirm scale dependence without optimizing it; and [Andrade et al. 2025] standardizes evaluation using fixed grids. What is not established is the integrated treatment: co-optimizing spatial representation with the ranking function of rare event count series, under explicit budget constraints, evaluating the result with metrics appropriate to sparsity (PAI, Recall@K). The contribution is positioned precisely at this intersection.

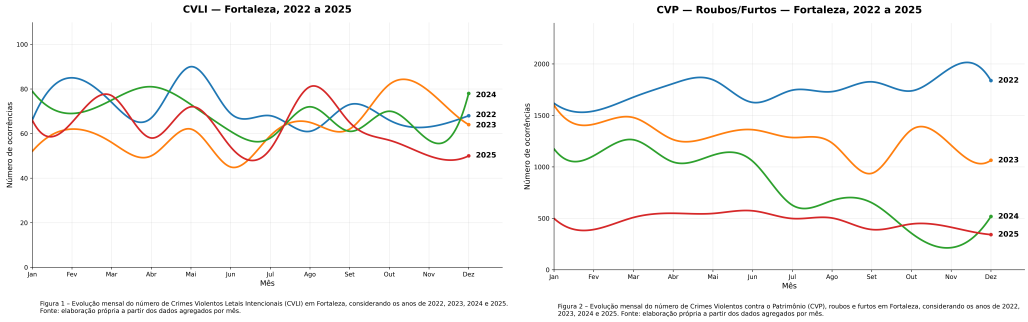


Figure 1. Annual overview of the CVLI and CVP time series.

3. Data

Table 1 and Figure 2 and the Figure 4 summarize the statistical properties important for choosing prediction models. The total number of events in each series in the 2022-2025 interval was CVLI = 3,031 and CVP = 109,343. Figure 1 provides more details about the spatial profile of crime series in Fortaleza City.

Figure 1 shows the annual (smoothed) view of the CVLI and CVP time series. Two features should be noted. The first is that CVLI varies between 40 and 90 monthly events, while CVP varies between 400 and 2000 monthly events. Another property is that CVP falls systematically year after year, while CVLI fluctuates but maintains approximately the same monthly average over the years.

Table 1. Additional data from the Crime Time Series in Fortaleza City.

	CVLI (rare events)	CVP (frequent events)
Total 2022–2025	∼3,031	∼109,343
Units (neighborhoods)	118	149
50% of events	∼5 neighborhoods	25 neighborhoods
80% of events	∼18 neighborhoods	61 neighborhoods
Top 1 neighborhood	Dom Lustosa / Bom Jardim	Centro (3,308)

Source: Values based on CVLI (violent crimes with lethal outcome) and CVP (violent crimes against property) data from the SJC (State Justice Council) and SESED (State Secretariat for Public Security and Social Defense) between January 2022 and June 2025. Data is provisional and subject to change. CVLI = Crimes Violentos Letais Intencionais (Violent Intentional Lethal Crimes); CVP = Crimes Violentos Contra o Patrimônio (Violent Crimes Against Property).

Note from Table 2 the significant differences in the properties of the ACF (Autocorrelation Function), Dispersion Index, and number of zeros, both between the CVP and CVLI series, and within each type of crime, for the different temporal aggregations.

On the other hand, note from the last column of Table 2 how these properties condition the choice of the prediction model to be employed, according to the properties and purposes of the models and the recommendations of the best literature in the field.

Thus, for daily CVLI, the high rate of zero values recommends a hurdle/ZIP model, while for CVP on the same scale, it recommends negative binomial GLM regression.

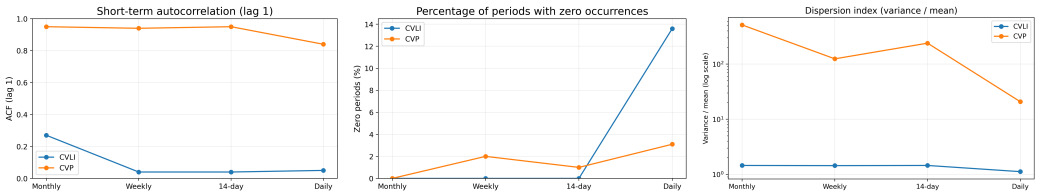


Figure 2. Visualization of the most important data from Table 1.

At the other extreme of the scale, the monthly scale, the low dispersion and weak autocorrelation of CVLI recommend a Poisson GLM regression model, while for CVP, the strong autocorrelation (strong memory) along with overdispersion recommend a seasonal NB+AR model.

Table 2. Descriptive Analysis and Models Indicated by Series and Scale.

Series	Scale	Mean	Disp.	ACF(1)	Regime / Indicated model
CVLI	Monthly	63.1	1.45	0.27	Near-equidispersed, weak autocorr. → Poisson acceptable
CVLI	Weekly	14.5	1.44	0.04	Near-equidispersed, no memory → limited temporal gain
CVLI	14 days	29.1	1.45	0.04	Moderate 14-day seasonality (ACF seasonal ∼0.33)
CVLI	Daily	2.07	1.12	0.05	Low count, 13.6% zeros → evaluate hurdle/ZIP
CVP	Monthly	2162.8	513.2	0.95	Highly overdispersed, strong memory → NB + AR/seasonal
CVP	Weekly	496.7	124.6	0.94	Highly overdispersed, strong memory → NB/quasi-Poisson
CVP	14 days	997.7	239.6	0.95	Idem; strong seasonality
CVP	Daily	71.1	20.8	0.84	Highly overdispersed → Negative Binomial (NB)

4. Methods

In this section we briefly review the theoretical foundations of the methods used. Readers wishing to delve deeper should consult the references listed in each paragraph.

Baseline:

The seasonal naïve predictor sets the forecast for period t equal to the observation at lag s (s = 52 for weekly, s = 12 for monthly), i.e., $\hat{y}_t = y_{t-s}$. It requires no estimation and serves as the lower performance bound that any model must surpass to justify its complexity.

Poisson GLM Regression:

The Poisson Generalized Linear Model [MacCullagh and Nelder 1989] models the conditional mean as $log(\lambda_t) = x_t^T \beta$, where x_t contains trend and Fourier seasonal terms. The count $y_t|x_t$ follows a Poisson distribution with mean λ_t:

$$log(\lambda_t) = x_t^T \beta, y_t|x_t \sim Poisson(\lambda_t) \quad (1)$$

By construction, Poisson GLM imposes equidispersion $Var[y_t] = E[y_t]$, which is often violated in practice.

Negative Binomial GLM Regression:

The NB2 extension [Cameron and Trivedi 1998] introduces a dispersion parameter $\alpha > 0$ so that $Var[y_t] = \lambda_t + \alpha\lambda_t^2$, accommodating overdispersed counts. The conditional mean retains the same log-linear form as Poisson GLM; the additional parameter is estimated by maximum likelihood alongside β:

$$Var[y_t] = \lambda_t + \alpha\lambda_t^2, \alpha > 0. \tag{2}$$

INGARCH:

The INGARCH(p,q) model [Fokianos et al. 2009] specifies a multiplicative autoregressive dynamic for the conditional intensity analogous to GARCH. The log-linear specification - numerically stable and guaranteed positive - is:

$$log(\lambda_t) = d + a.log(y_{t-1} + 1) + b.log(\lambda_{t-1}) \tag{3}$$

where $y_t|F_{t-1} \sim Poisson(\lambda_t)$ or $NB(\lambda_t, \alpha)$; $d > 0, a \geq 0, b \geq 0$; stationarity: $a+b < 1$. Parameters are estimated by conditional maximum likelihood. The NB variant replaces the Poisson likelihood to accommodate residual overdispersion (CVP regime).

INAR:

The INAR(1) model [Al-Osh and Alzaid 1987] represents the count at time t via binomial thinning of the previous count plus an independent Poisson innovation:

$$y_t = \alpha.y_{t-1} + \epsilon_t, \epsilon_t \sim Poisson(\mu) \tag{4}$$

The thinning operator $\alpha.y_{t-1} = \sum_{k=1}^{y_{t-1}} B_k$, $B_k \sim Bernoulli(\alpha) i.i.d.$; each past event independently survives to t with probability $\alpha \in [0, 1)$.

Hurdle / Zero-Inflated:

For series with excess zeros, two-component models are appropriate. The Hurdle model [Mullahy 1986] separates the binary event of a non-zero count (logistic component) from its positive value (truncated Poisson). The Zero-Inflated Poisson (ZIP) [Lambert 1992] mixes a point mass at zero with a standard Poisson, allowing zeros to arise from either the Poisson component or a latent always-zero state:

$$P(Y = 0) = \pi + (1 - \pi)\epsilon^{-\lambda}; P(Y = y) = (1 - \pi)\frac{\lambda^y \epsilon^{-\lambda}}{y!}, y > 1 \tag{5}$$

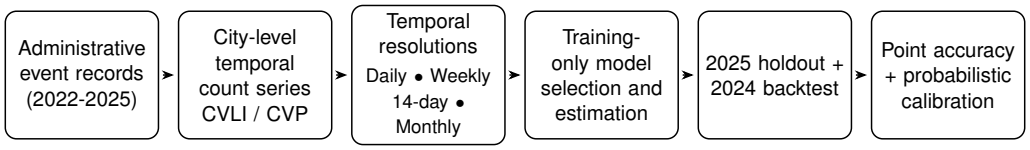
$\pi = P(structuralzero)$; as $\pi \to 0$ ZIP reduces to Poisson. Both are identified by maximum likelihood; the choice depends on whether zeros are structural or low-count tail.

5. Results and Discussion

5.1. Experimental Workflow and Metrics

Figure 3 shows the experimental workflow adopted to ensure that there is no information leakage between training and testing. The test set (the year 2025) is kept out of any training phase procedure, and all metrics presented are about this subset of data.

Figure 3. Experimental design and evaluation workflow



Primary forecast horizon: one step ahead • Moving forecast origin • No test information used for model selection

Metrics: Point forecasts are summarized using MAE, RMSE, percentage bias, and weighted MAPE (wMAPE):

$$\mathrm{wMAPE} = 100\,\frac{\sum_t |y_t - \hat{y}_t|}{\sum_t y_t}. \tag{6}$$

Because the denominator is the aggregate observed count, the measure remains defined when individual observations are zero.

For discrete predictive distributions, the logarithmic score and ranked probability score are defined as

$$\mathrm{LogS}_t = -\log p_t(y_t), \tag{7}$$

$$\mathrm{RPS}_t = \sum_k \left[F_t(k) - 1\{y_t \le k\}\right]^2. \tag{8}$$

Lower values indicate better probabilistic forecasts. Central 80% and 95% predictive interval coverage and mean interval width are also reported to distinguish calibration from sharpness.

Uncertainty in point-forecast comparisons is assessed by paired moving-block bootstrap confidence intervals for differences in wMAPE, using block lengths of 7, 4, 3, and 2 observations for daily, weekly, 14-day, and monthly series, respectively, and a fixed resampling seed of 20260810. A paired Diebold–Mariano-type comparison of absolute errors uses HAC variance estimation [Diebold and Mariano 1995]. Differences in log-score and RPS between Poisson and Negative Binomial versions are evaluated analogously. Ljung–Box tests [Ljung and Box 1978] on training residuals are used to diagnose remaining serial dependence.

5.2. Results

All forecasts in this paper are for a one-step-ahead horizon, using a model trained only once and evaluated via a moving-origin approach.

Table 3 reports the out-of-sample wMAPE of GLM count models against a seasonal naive baseline across scales and series. Three patterns emerge directly from the data. First, for CVLI, the gains of both Poisson and Negative Binomial GLM over the baseline are modest - at most 1.2 p.p. at the 14-day scale and 0.8 p.p. at the weekly scale - consistent with the weak autocorrelation ($ACF_1 \approx 0.04$) diagnosed in Table 1: when the series exhibits little temporal memory, the information extractable by a static regression

structure is limited. Second, for CVP, the NB GLM consistently outperforms Poisson GLM, and in the weekly scale the Poisson model actually worsens against the baseline (20.9% vs. 16.5%), while NB corrects this bias (15.5%). This is a direct consequence of the extreme overdispersion of CVP (dispersion index $\approx$ 124.6 at the weekly scale): forcing Poisson's equidispersion constraint inflates forecast variance and produces systematically biased point estimates. Third, the NB advantage grows with scale resolution - at the daily CVP level (dispersion $\approx$ 20.8), the gap between Poisson (33.9%) and NB (29.2%) narrows relative to coarser scales, suggesting that aggregation partially absorbs the excess variance.

Table 3. Comparison of Models by Scale and Reading Type.

Series	Scale	Baseline (wMAPE)	Poisson (wMAPE)	Neg. Bin. (wMAPE)	Reading
CVLI	14 days	19.4%	18.2%	18.3%	Small gain (weak memory)
CVLI	Weekly	27.0%	26.2%	26.2%	Poisson $\approx$ NB
CVP	14 days	12.6%	10.3%	9.0%	NB wins (overdispersion)
CVP	Weekly	16.5%	20.9%	15.5%	NB corrects Poisson bias
CVP	Daily	28.7%	33.9%	29.2%	NB necessary

Table 4 extends the comparison to autoregressive count models (INGARCH, INAR) and to the full probabilistic scoring triplet (wMAPE, log-score, RPS). For CVLI, INGARCH(1,1) Poisson achieves the best wMAPE (44.6%), an 8.1 p.p. gain over the GLM Poisson (52.7%) and a 12.5 p.p. gain over the naive baseline (57.1%). This result is notable precisely because the CVLI series exhibits weak autocorrelation: even a modest serial structure, when explicitly modeled by the INGARCH mechanism, yields a non-trivial improvement in point forecast accuracy. On probabilistic scores, INGARCH(1,1) Poisson also attains the lowest log-score (3.63) and RPS (3.47) for CVLI, confirming better distributional calibration over GLM alternatives. The NB version of INGARCH, however, performs below its Poisson counterpart in wMAPE for CVLI (51.7% vs. 44.6%), suggesting that the additional dispersion parameter is not warranted in a near-equidispersed regime and may introduce variance in estimation that degrades point forecasts. INAR(1) Poisson (52.3%) offers no improvement over the GLM, consistent with its more restricted thinning structure.

For CVP, INGARCH(1,1) Poisson achieves the best wMAPE (27.9%), outperforming both the naive baseline (85.3%) and the GLM Poisson (41.9%). The magnitude of this gain — 14 p.p. over GLM Poisson — reflects the strong autocorrelation of CVP ($ACF_1 \approx 0.94$–0.95): the INGARCH dynamic explicitly captures this temporal memory, which the GLM ignores. On probabilistic scoring, however, the NB-INGARCH(1,1) dominates with the lowest log-score (5.85) and lowest RPS (49.3) for CVP, while the INGARCH Poisson log-score (11.72) reveals distributional overconfidence — a Poisson's equidispersion assumption severely underestimates the predictive uncertainty in a series where variance exceeds the mean by two orders of magnitude. This split result — Poisson INGARCH winning on wMAPE while NB-INGARCH wins on calibration — illustrates

that point accuracy and distributional accuracy are complementary but not interchangeable criteria, and both are needed to fully characterize model performance on count data, as argued by Czado et al. (2009). INAR(1) Poisson (30.6% wMAPE, log-score 18.64) consistently underperforms INGARCH across all metrics for CVP.

Table 4. Comparison of Models for the CVLI and CVP Series.

Series	Model	wMAPE (%)	Log-score	RPS
CVLI	Seasonal naïve (lag 52)	57.1	–	–
CVLI	GLM Poisson (trend+season.)	52.7	4.21	4.25
CVLI	GLM Negative Binomial	52.7	3.82	4.10
CVLI	INGARCH(1,1) Poisson	44.6	3.63	3.47
CVLI	INGARCH(1,1) Negative Binomial	51.7	3.78	4.01
CVLI	INAR(1) Poisson	52.3	4.18	4.20
CVP	Seasonal naïve (lag 52)	85.3	–	–
CVP	GLM Poisson (trend+season.)	41.9	16.34	84.7
CVP	GLM Negative Binomial	34.9	6.06	59.4
CVP	INGARCH(1,1) Poisson	27.9	11.72	54.2
CVP	INGARCH(1,1) Negative Binomial	33.8	5.85	49.3
CVP	INAR(1) Poisson	30.6	18.64	63.1

Table 5 reports probabilistic scores across temporal scales for the Poisson and NB families, using both log-score and RPS. For CVP, the NB advantage in log-score is largest at the monthly scale (6.87 vs. 12.76), where overdispersion is most severe (dispersion $\approx$ 513.2), and decreases progressively at weekly (5.44 vs. 10.02) and daily (3.95 vs. 5.37) scales — in direct proportion to the decline in the dispersion index. This scale-monotone pattern confirms that the NB distributional correction is most critical where variance inflation is greatest, exactly as predicted by the statistical properties in Table 1. The RPS divergence follows the same direction for weekly CVP (NB: 27.1 vs. Poisson: 38.4) but is smaller at the monthly scale (NB: 106.7 vs. Poisson: 104.2), reflecting the RPS's lower sensitivity to tail behavior compared to the log-score. For CVLI, both models produce nearly identical log-scores and RPS at weekly (2.85 NB vs. 2.88 Poisson) and daily (1.704 vs. 1.704) scales, providing strong empirical confirmation that the near-equidispersion of CVLI (Table 1) renders the NB extension unnecessary at the series level — the Poisson already provides adequate calibration.

Taken together, the results from Tables 2–4 provide a coherent and consistent answer to the research question: the statistical properties of each series at each temporal scale are the primary determinants of the appropriate model class. For CVLI, the weak autocorrelation limits GLM gains but leaves room for INGARCH improvement; equidispersion renders the NB extension redundant at the series level. For CVP, strong overdispersion makes NB mandatory for calibration across all scales, and strong temporal memory makes the INGARCH structure indispensable for point accuracy. No single model dominates both series and both accuracy dimensions simultaneously — a finding consistent with Villegas et al. (2022) and Andrade et al. (2025).

Table 5. Comparison of Scales for the CVP and CVLI Series.

Series	Scale	Log-score Poisson	Log-score NB	RPS Poisson	RPS NB
CVP	Monthly	12.76	6.87	104.2	106.7
CVP	Weekly	10.02	5.44	38.4	27.1
CVP	Daily	5.37	3.95	8.05	6.38
CVLI	Weekly	2.88	2.85	2.47	2.44
CVLI	Daily	1.704	1.704	0.778	0.777

6. Conclusion

This paper investigated count time series forecasting applied to two contrasting crime series in Fortaleza: CVLI (rare, near-equidispersed, weak temporal memory) and CVP (frequent, strongly overdispersed, strong temporal memory). The central research question — how statistical properties vary across temporal aggregation scales and how this variation conditions the appropriate model class — is answered with empirical evidence across four temporal scales and six model families.

Three conclusions emerge from the experimental results. First, the statistical regime of a count series is the primary driver of model selection: the dispersion index and first-order autocorrelation measured at a given temporal scale reliably predict which model family will excel at that scale, validating the diagnostic approach of Table 3 as a prescriptive instrument. Second, the autoregressive structure of INGARCH provides consistent gains over static GLM in both series — even for CVLI, where autocorrelation is weak, the gain reaches 8.1 p.p. in wMAPE — while the INAR(1) structure, more restrictive in its thinning representation, fails to improve over GLM in either series. Third, point accuracy and distributional calibration are complementary dimensions that require separate assessment: for CVP, INGARCH Poisson achieves the best wMAPE while NB-INGARCH achieves the best log-score and RPS, underscoring that the Negative Binomial is essential for honest uncertainty quantification in overdispersed regimes even when it does not minimize point error.

These results confirm that classical Gaussian time series models are inadequate for crime count series: Poisson GLM actively worsens upon the naive baseline for weekly CVP, and its log-score for CVP is nearly twice that of NB at all scales. The use of proper scoring rules (log-score, RPS) alongside wMAPE is necessary to avoid misleading conclusions, as previously argued by Czado et al. (2009) and Kolassa (2016).

Future work will extend the evaluation to hurdle and zero-inflated models for daily CVLI — where the 13.6% zero rate shifts the distributional regime — and will incorporate spatial aggregation scales (neighborhood, AIS, hexagonal grid) to address the spatial dimension of the research question, connecting the temporal forecasting component with spatiotemporal ranking under sparsity.

References

Adepeju, M., Rosser, G., and Cheng, T. (2016). Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions-a crime case study. *International Journal of Geographical Information Science*, 30(11):2133–2154.

Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (inar (1)) process. *Journal of Time Series Analysis*, 8(3):261–275.

Andrade, Y., Teodoro, G., Ferreira, R., and Rocha, L. (2025). Um framework para avaliação de modelos de previsão espaço-temporal de crimes. *Revista Eletrônica de Iniciação Científica em Computação*, 23:136–142.

Butt, U. M., Letchmunan, S., Hassan, F. H., Ali, M., Baqir, A., and Sherazi, H. H. R. (2020). Spatio-temporal crime hotspot detection and prediction: a systematic literature review. *IEEE access*, 8:166553–166574.

Cameron, A. and Trivedi, P. (1998). Regression analysis of count data. *Econometric Society Monographs*.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, 32(3):788–803.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

MacCullagh, P. and Nelder, J. (1989). Generalized linear models. *Monographs on statistics and applied probability (37)*.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the american statistical association*, 106(493):100–108.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

Villegas, Á. J. R., Pabón, J. S. M., Rubio, M. D., Quintero, S., Vargas, J. G., and García, H. (2022). Spatio temporal sparsity in homicide prediction models. *IEEE Access*, 10:14359–14367.

Yu, G., Zou, J., Hu, X., Aviles-Rivero, A. I., Qin, J., and Wang, S. (2024). Revitalizing multivariate time series forecasting: Learnable decomposition with inter-series dependencies and intra-series variations modeling. *arXiv preprint arXiv:2402.12694*.