Revista Facultad de Ingeniería

UNIVERSIDAD DE ANTIOQUIA 1803

Title: Computational modeling of epidemiological count data using Non-Homogeneous Poisson Processes and functional data

Authors: Santiago Ortiz[1*] https://orcid.org/0000-0002-9121-3807, Juan Esteba Chavarría[2] and Henry Velasco[2] https://orcid.org/0000-0001-9595-7099

[1]Facultad de Ingeniería, Universidad San Buenaventura. Carrera 122 # 6 - 65. 760031. Cali, Colombia.

[2] Escuela de Ingeniería y Ciencias Aplicadas, Universidad EAFIT. Carrera 49 # 7Sur - 50. 050022. Medellín, Colombia.

Corresponding Author: Santiago Ortiz

E-mail: santy_ortiz@hotmail.com

DOI: **10.17533/udea.redin.20250367**

To appear in: *Revista Facultad de Ingeniería Universidad de Antioquia*

Received: August 20, 2024

Accepted: March 17, 2025

Available Online: March 17, 2025

Please cite this article as: S. Ortiz, J. E. Chavarría and H. Velasco. Computational modeling of epidemiological count data using Non-Homogeneous Poisson Processes and functional data, *Revista Facultad de Ingeniería Universidad de Antioquia*. [Online]. Available: https://www.doi.org/10.17533/udea.redin.20241247

redin Revista Facultad de Ingeniería, March 5, 2025

Computational Modeling of Epidemiological Count Data Using Non-Homogeneous Poisson Processes and Functional Data

Modelado Computacional de Datos Epidemiológicos Usando Procesos de Poisson No Homogéneos y Datos Funcionales

UNIVERSIDAD DE ANTIOQUIA
1 8 0 3

Authors: Double-blind review

KEYWORDS:
Intensity Function, Generalized Linear Model, Stochastic Trajectories, Depth Estimation, Confidence Envelopes

Función de Intensidad, Modelo Lineal Generalizado, Trayectorias Estocásticas, Estimación de Profundidad, Envolventes de Confianza

ABSTRACT: In this work, we introduce a novel methodology for modeling discrete count variables within the framework of stochastic processes. Our approach integrates two statistical areas: Non-Homogeneous Poisson Processes for the estimation and prediction of intensity functions based on explanatory variables and functional data estimation techniques. Through a comprehensive case study focusing on an infectious disease with viral characteristics, we demonstrate the potential of our methodology. We provide empirical evidence that our methodology offers a robust alternative for modeling count variables. Our findings support the utility of our approach in capturing the complex dynamics inherent in count data in infectious disease epidemiological phenomena.

RESUMEN: En este trabajo, presentamos una nueva metodología para modelar variables de conteo discretas dentro del marco de procesos estocásticos. Nuestro enfoque integra dos áreas estadísticas: los procesos de Poisson no homogéneos para la estimación y predicción de funciones de intensidad basadas en variables explicativas y las técnicas de estimación de datos funcionales. A través de un estudio de caso integral centrado en una enfermedad infecciosa de características virales, demostramos el potencial de nuestra metodología. Proporcionamos evidencia empírica de que nuestra metodología ofrece una alternativa robusta para modelar variables de conteo. Nuestros hallazgos apoyan la utilidad de nuestro enfoque para capturar la dinámica compleja inherente a los datos de conteo en los fenómenos epidemiológicos de enfermedades infecciosas.

1. Introduction

Modeling count data is essential in numerous real-world scenarios where understanding the frequency of discrete events is crucial for effective decision-making and policy development. For example, in environmental studies, modeling counts of wildlife sightings or occurrences of natural disasters supports conservation efforts and risk management strategies [1]. Similarly, in transportation engineering, analyzing vehicle counts at intersections or along roadways provides insights into traffic patterns, identifies congestion hotspots, and informs infrastructure planning [2]. In epidemiology, modeling the counts of infections over time enables researchers and public health officials to identify trends, assess intervention impacts, and forecast disease spread [3].

The study of count data dates back to the late 19th century, exemplified by the classic analysis of annual deaths caused by mule kicks in the Prussian Army [4], where the data were found to follow a Poisson distribution. Traditional models for count data often rely on assumptions about the underlying data distribution, such as Poisson or Gaussian distributions. These models can effectively explain count variables [5], but they may struggle to capture temporal patterns and seasonality, leading to suboptimal predictions and analyses [6].

Poisson regression and its variants are powerful tools for modeling count data but come with specific assumptions and limitations. Counts often exhibit heteroscedasticity, where variance increases with the mean, making ordinary least squares (OLS) regression unsuitable [7]. Poisson regression addresses issues of non-constant variance and non-normal error distributions but assumes equal mean and variance, which often leads to overdispersion in real-world scenarios. This results in inaccurate confidence intervals and hypothesis tests [8]. Negative binomial regression addresses overdispersion by allowing for a more flexible variance structure but still relies on specific assumptions [9]. Zero-Inflated Models tackle the prevalence of zeros by assuming two distinct processes generate the zero and non-zero counts [10, 11]. Furthermore, the assumption of independent counts, inherent in many models derived from Poisson regression, does not always hold, particularly in the presence of state dependence [12].

While traditional models do not explicitly account for time dependency, this can sometimes be addressed indirectly using time-related predictors. However, time series models, such as integer-valued autoregressive (INAR) models [13], are better suited for capturing temporal structures. INAR models account for event occurrences in one time period influencing subsequent periods through a conditional mean. Integer-valued Generalized Autoregressive Conditional Heteroscedasticity (INGARCH) models [14] extend this framework by introducing conditional heteroscedasticity, accommodating variance changes over time (volatility). Specific variants, such as INARCH [12], use a Poisson distribution with a conditional mean based on past observations. These are also referred to as autoregressive conditional Poisson (ACP) models [15] or linear-Poisson autoregressive models [16].

Regression-based or generalized linear models effectively incorporate explanatory variables into count data models. However, challenges like overdispersion and zero inflation persist. Count-based time series models excel in capturing dynamics, trends, and seasonality but often neglect the influence of explanatory variables on the counting process or its intensity function. Integrating these methodologies to model count phenomena as stochastic processes may address these limitations. Furthermore, incorporating functional data analysis (FDA) techniques into this framework could enhance both prediction and inference for count phenomena.

The main contribution of this work is the development of a novel methodology for modeling and predicting count data by combining Non-Homogeneous Poisson Processes (NHPP) with Functional Data (FDA) techniques. The key contributions are: (i) modeling and prediction of count data subject to explanatory variables and time evolution, and (ii) estimating the most significant trajectory and non-parametric confidence intervals (or envelopes) for cumulative counting processes. FDA complements NHPP by providing a flexible, non-parametric approach for estimating intensity functions and capturing functional relationships between variables. This integration enables robust modeling of the interactions between covariates and count data intensity, enhancing predictions and inference robustness.

The manuscript is organized as follows: Section 2 provides an overview of the materials and methods employed, delving into the theoretical underpinnings of the techniques utilized and offering a concise methodology outlining the approach to the problem and the proposed solution. In Section 3, the primary findings derived from the implementation of each technique are presented in detail. Finally, Section 4 offers a comprehensive discussion of the results alongside the concluding remarks of this research.

2. Methods

In this section, we explore discrete count data models crucial for analyzing phenomena characterized by random events over time. We begin by examining the dynamics of count data phenomena, starting with Poisson Regression, a foundational tool for modeling dependencies between count data and covariates. We then delve into NHPP counting processes, emphasizing the significance of incorporating time-varying intensity functions in count data modeling and elucidating their generation process. These discussions lay the groundwork for our proposed methodology, which integrates NHPP and FDA to effectively model and predict count data scenarios influenced by both temporal evolution and exogenous variables.

2.1 Homogeneous and Non-Homogeneous Poisson Processes

A count data phenomenon refers to the accumulation of the number of times that some event occurs during a fixed time-space interval [17], for example, the arrival of clients at a window to request some service or the moments in which certain machinery requires repair. As previously mentioned, the Poisson distribution is fundamental for modeling these events in time. Assuming events occur randomly at an average rate λ, and the occurrence of one event is independent of any other, then the number of events arising in a unit time interval has a Poisson distribution with parameter λ [18, Chapter 7]. This independence property makes it particularly useful for stochastic models based on probability functions [19, pp. 3].

Poisson Regression

In the conventional regression framework, a dependent variable N is a discrete non-negative random variable whose conditional mean depends on some vector of regressors X. If $N \sim Poisson(\lambda)$, it allows λ to depend on regressors, such that, for the i^{th} independent event

$E[N_i|X_i] = \lambda(X_i, \beta)$, an exact dependency without any other source of stochastic variation, defines the Poisson regression [20]. For a stochastic dependence on this intensity function, other distributions should be taken into consideration.

Poisson regression has a special property, the mean and variance are the same. This property is rarely satisfied in real data, so overdispersion, i.e. excess variability, and underdispersion arise as problems for this approach [21]. The first can be solved with a distribution that considers a dispersion parameter, such as a negative binomial [22], the second may be treated through a generalized Poisson model [23]. Zero counts are also a concern, whether they are excluded as a possibility from the model, which requires a zero-truncated model, or present in an excessive amount (zero-inflated model) since they can prevent the sum of probabilities to one from being fulfilled [7].

NHPP Counting Process

Formally, a counting process $\{N(t); t \geq 0\}$ is a stochastic process that represents the number of events that occur in a time interval $[0, t]$ that satisfies the following properties

- $N(0) = 0$
- $N(t) \in \{0, 1, 2, \dots\} \quad \forall t \geq 0$
- $N(t) - N(s)$ is the number of events that ocur in the interval $[s, t]$ or increments

The classical approach for modeling a counting process is to assume that the n events or counts during an interval $I = [s\,,\,t]\ ,\ s \neq 0$ follow a Poisson distribution, whose probability mass function is given by $f_{N_{[s,t]}}(n) = e^{\lambda(t-s)}(\lambda(t-s))^n/n!$, where, $\lambda(t-s)$ denotes the intensity parameter of the process which measures the average events per interval. This parameter characterizes the distribution satisfying $E(N_{[s,t]}) = V(N_{[s,t]}) = \lambda(t-s)$. Nonetheless, if it is assumed that λ varies as a function of time, i.e, $\lambda(t)$, a Non-Homogeneous Poisson Processes (NHPP) is reached, this kind of counting process has the following properties

- $N(0) = 0$
- $N(t)$ has independent increments
- $P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h)$
- $P(N(t+h) - N(t) \geq 2) = o(h)$

The intensity function $\lambda(t)$ plays a pivotal role in NHPP, distinguishing them from their Homogeneous counterparts (HPP). While HPP assumes a constant intensity rate over time, NHPP allows for the intensity rate to vary with time, making it a versatile tool for modeling diverse counting phenomena. The intensity function represents the rate at which events occur per unit of time and is essential for characterizing the temporal dynamics of the counting process. Unlike HPP, where the intensity function remains constant, NHPP accommodates changes in the intensity function, reflecting real-world scenarios where counting processes may deteriorate or improve over time. An NHPP can be seen as a ”Minimal Repair Process” because the counting system is deteriorating and/or improving over time [24]. This inherent flexibility in modeling the temporal dynamics of counting processes sets the importance of NHPP, as it provides a more realistic representation of phenomena where changes in count intensity occur naturally.

Intensity Function Estimation

Many estimation procedures exist for the intensity function of a NHPP process in the literature. In [25], a general methodology with the use of an exponential-polynomial trigonometric function for cyclic behavior and trends was proposed. In [26], a piecewise-polynomial intensity function was contributed, and [27, pp. 407] suggested a piecewise-constant function through a simple nonparametric procedure [28], but dependent on some arbitrary parameters. These various approaches highlight the complexity and intricacy of intensity function modeling, driving the need to choose the appropriate method based on the specific characteristics of the data and the objectives of the analysis.

Valuable insights into the estimation of $\lambda(t)$ for NHPPs, which are essential for statistical modeling based on point processes, are provided in [29]. In an NHPP, the intensity of the process $\lambda(t)$ is not constant over time but varies according to certain covariates, such as time trends, seasonal terms, or external factors. The time NHPP behavior is typically represented by modeling the intensity function $\lambda(t)$ as a function of these covariates. Ensuring that the intensity function $\lambda(t)$ remains positive, a common approach involves using a logarithm link function. Specifically, the logarithm of the intensity function is modeled as a linear combination of covariates and parameters, given by

$$\log(\lambda(t; \beta)) = \mathbf{X}^\top(t)\beta, \tag{1}$$

where $\mathbf{X}^\top(t)$ represents the row vector of covariates at time t, and β denotes the vector of parameters to be estimated. By employing this logarithmic link function, the estimation of the intensity function is facilitated while ensuring it is non-negative.

In an NHPP, $\lambda(t) \geq 0 \ \forall t \in (0, S]$, and is continuous for almost every $t \in (0, S]$, with S being a known constant that could be the upper limit for a time cycle. Having k realizations of NHPP, it is possible to obtain a cumulative intensity function or intensity measure of the process, on $(0, S]$ [30] and can be estimated with a structure $\lambda(\mathbf{X}, t)$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the matrix where each row denotes a temporal observation and each column an explanatory variable.

Random Generation of NHPP Trajectories

Available methods to generate an NHPP can be classified into three categories: inversion methods, order statistics methods, and acceptance-rejection methods. One of the earliest techniques is to take the cumulative intensity function inversion to compute the required event times. Another approach is to assume the event times as order statistics from a sample with a distribution function in terms of the cumulative intensity function [31, Section 23]. The last method is the most popular, as it uses thinning to determine a constant rate function and rejects a certain fraction of generated events to achieve the desired intensity [32, pp. 179].

The first and second methods can be computationally expensive when the cumulative intensity adjusts to a function that does not allow an efficient analytic inversion, which is common. Modern methods employ the first and second approaches, through the construction of a majorizing function in a favorable structure that benefits from the thinning process [33, pp. 2015-2018]. The Cinlar method [34] provides an efficient approach for generating points from an NHPP. Its key advantage lies in avoiding discretization of the intensity function, similar to the thinning approach [35]. This makes the Cinlar method faster, especially when the inversion of the cumulative intensity function can be easily computed. A brief overview of the Cinlar method is as follows:

1. Transformation to a Unit Rate Homogeneous Poisson Process (HPP): The method starts by transforming the NHPP into a unit-rate HPP using a time-scale transformation

   $$t_i^H = \int_0^{t_i^{NH}} \lambda(t) dt. \tag{2}$$

   This step involves generating points t_i^H in a unit rate HPP through independent exponential distances [36].

2. Transformation Back to NHPP: The points t_i^{NH} in the NHPP with intensity $\lambda(t)$ are obtained by transforming the points t_i^H using the inverse of the transformation function. Since our study assumes that $\lambda(t)$ is constant in $[t, t+1)$, the points t_i^{NH} are calculated iteratively such that the sum of the intensities up to t_i^{NH} equals t_i^H.

The Cinlar method is implemented in our methodology through the function `simNHP.fun` from the `NHPoisson` package for `R` [37]. This function enables the generation of points within a specified period $(0, T)$, where the intensity at each time point must be provided. The length of the intensity vector determines the value of T. Furthermore, to ensure reproducibility, a seed can be set in the generation process using the `fixed.seed` argument.

2.2 Functional Data Estimation Techniques

Given a set of values that follow a generated NHPP trajectory, it is possible to transform these observations into a function or curve, i.e. a functional observation, by an interpolating or smoothing procedure, for instance, using a linear combination of a known base function like polynomial, spline, wavelet, or Fourier [38].

A functional data $Y_i = Y_i(t) : t \in T \ (i = 1, \ldots, n)$ with $T \subset \mathbb{R}$, is the observation of n functional variables Y_i which takes values in an infinite dimensional space [39]. For this kind of data, there are multiple tools to perform exploratory data analysis in order, for example, to estimate the main, central, or most representative curve, estimate functional dispersion, dependence, confidence intervals, etc. [40, 41]. In [42], the Modified Band-Depth (MBD) estimator is proposed. A methodology to compute a ranking estimation of all the curves in the functional data sample is denoted by

$$\mathrm{MBD}(Y_r) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{\begin{array}{r}\lambda(\{t \in \mathcal{I} \mid \min(Y_i(t), Y_j(t)) \leq \\ Y_r(t) \leq \max(Y i_(t), Y_j(t))\})\end{array}}{\binom{n}{2}\lambda(\mathcal{I})}. \tag{3}$$

The $\mathrm{MBD}(Y_r)$ measures the average proportion of times that Yr is within the envelope generated by the curves Y_i and Y_j $(i \neq j)$. So, the higher the $\mathrm{MBD}(Y_r)$ the deeper the observation Yr. Therefore, an estimation of the functional median, based on depth rankings, is $Y_{[1]}(t) = \max_{r=1;\ldots;n} \mathrm{MBD}(Y_r)$.

Another exploratory data analysis tool, extended to the functional case, is the boxplot. In [43], it is proposed a functional version of the boxplot, which is constructed analogously in comparison to the common univariate boxplot. The functional boxplot is computed as follows: The functional median is estimated by the MBD, specifically by $Y_{[1]}(t) = \max_{r=1,\dots,n} \text{MBD}(Y_r)$ and the functional interquartile range ($C_{0.5}$) is determined as the region where 50% of the deepest curves are in, which is defined by

$$\begin{aligned} \mathrm{C}_{0.5} = \{&(t, Y(t)) \\ &: \min_{r=1,\dots,\frac{n}{2}} Y_{[r]}(t) \\ &\leq Y(t) \\ &\leq \max_{r=1,\dots,\frac{n}{2}} Y_{[r]}(t)\}. \end{aligned} \tag{4}$$

Additionally, the corresponding functional whiskers (F_W) are given by

$$\mathrm{F}_W = \text{Envelope}(\mathrm{C}_{0.5}) \pm 1.5\mathrm{C}_{0.5}, \tag{5}$$

where the notation Envelope($\mathrm{C}_{0.5}$) denotes the pair of functional observations that enclose the 50% deepest curves within their envelope.

It is worth considering the interpretation of the envelope of the functional boxplot as a confidence region. This perspective leads to a proposed adaptation of the envelope to create confidence regions of type $1 - \alpha$, where α represents a specified level of error or significance. Such an adaptation offers the potential to provide insights into the uncertainty surrounding the estimated functional median, enhancing the inferential capabilities of the proposed methodology, explained in the next subsection.

2.3 Proposed Methodology

We propose the combination of both NHPP and FDA for modeling and predicting count data problems, we denote it as NHPP-FD, following the next steps:

1. Model the intensity function $\lambda(t)$ as a function of p covariates $\mathbf{X} = (X_1, \dots, X_p)$, where $X_j = (x_{1j}, \dots, x_{tj})'$ $(j = 1, \dots, p)$, using the theoretical regression model in Equation (6)

   $$\lambda(\mathbf{X}, t) = \exp(\mathbf{X}\boldsymbol{\beta}), \tag{6}$$

   where $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ is a vector of the unknown regression coefficients, to obtain $\hat{\lambda}(\mathbf{X}, t)$.

2. Define a future time period $t + L$ and predict L future values for each X_j, denote it $\hat{X}_j^{[t+L]} = (\hat{x}_{(t+1)j}, \hat{x}_{(t+2)j}, \dots, \hat{x}_{(t+L)j})'$ for $j = 1, \dots, p$. Then, store all predictions in a new matrix $\hat{\mathbf{X}}_{[t+L]} = \left(\hat{X}_1^{[t+L]}, \dots, \hat{X}_p^{[t+L]}\right)$. To obtain $\hat{X}_j^{[t+L]}$ we propose a model of the form $\hat{X}_j^{[t+L]} = g(X_j)$, where $g(\cdot)$ is any time series model; we suggest the implementation of either Holt-Winters, LOESS or ARIMA [44]. In our empirical findings, the Holt-Winters smoother shows a proficient performance.

3. Predict $t + L$ periods the estimated intensity function $\hat{\lambda}(\mathbf{X}, t)$ with $\hat{\mathbf{X}}_{[t+L]}$ by the NHPP model in Equation (6) following the expression presented in Equation (7)

   $$\hat{\lambda}(t + L) = \exp\left(\hat{\mathbf{X}}_{[t+L]}\hat{\boldsymbol{\beta}}\right). \tag{7}$$

4. Once $\hat{\lambda}(t + L)$ is estimated from (7), generate random k NHPP trajectories, following the definition presented in Cinlar [34]

   $$\begin{aligned} &N_{k,\hat{\lambda}(t+L)} \\ &= \left[\left\{N_1\left(t, \hat{\lambda}(t{+}L)\right)\right\}, \dots, \left\{N_k\left(t, \hat{\lambda}(t{+}L)\right)\right\}\right]'. \end{aligned} \tag{8}$$

   Now, from the Independent Increments assumption, each $\left\{N_i\left(t, \hat{\lambda}(t + L)\right)\right\}$ $(i = 1, \dots, k)$ follows the original counting process $\{N(t)\}$, the observed cumulative counting variable. Moreover, note that all $\left\{N_i\left(t, \widehat{\lambda}(t)\right) : t \geq 0\right\}$ are both NHPP trajectories.

5. Compute the MBD for the sample $N_{k,\widehat{\lambda}(t)}$ and denote the most representative or deepest NHPP trajectory as $N_{k,\widehat{\lambda}(t+L)_{[1]}} = \max_{r=1,\dots,k} \text{MBD}\left(N_{k,\widehat{\lambda}(t+L)}\right)_{[r]}$.

6. Finally, compute a functional bootstrap procedure of B replicates to the sample $N_{k,\hat{\lambda}(t+L)}$ in order to obtain a set composed by MBD statistics of the resamplings [45]. Thus, we define $B_{\text{MBD}} = \left\{\widehat{\text{MBD}}_{[1]}\left(N^{Boot=1}_{k,\hat{\lambda}(t+L)}\right), \dots, \widehat{\text{MBD}}_{[1]}\left(N^{Boot=B}_{k,\hat{\lambda}(t+L)}\right)\right\}$ the set of deepest curves of the B bootstrap replicates. Define $A = \min_{r=1,\dots,(1-\alpha)B} \widehat{\text{MBD}}_{[1]}\left(N^{Boot}_{k,\hat{\lambda}(t+L)}\right)_{[r]}$ and $B = \max_{r=1,\dots,(1-\alpha)B} \widehat{\text{MBD}}_{[1]}\left(N^{Boot}_{k,\hat{\lambda}(t+L)}\right)_{[r]}$, then, given a confidence level $1 - \alpha$, then the confidence envelope of $N_{i,\widehat{\lambda}(t)[1]}(t)$ is denoted as

   $$\begin{aligned} \mathrm{C}_{1-\alpha} = \Big\{&\left(k, \widehat{\text{MBD}}_{[1]}\left(N^{Boot}_{k,\hat{\lambda}(t+L)}\right)\right) \\ &: A \\ &\leq \widehat{\text{MBD}}_{[1]}\left(N^{Boot}_{k,\hat{\lambda}(t+L)}\right) \\ &\leq B\Big\}. \end{aligned} \tag{9}$$

$\mathrm{C}_{1-\alpha}$ represents the envelope in which the $(1-\alpha)100\%$ of the estimated deepest trajectories will be in it. Therefore, the borders of $\mathrm{C}_{1-\alpha}$ can be seen as the confidence bands of the prediction $N_{k,\widehat{\lambda}(t+L)_{[1]}}$.

3. Results and Discussion

In this section, we present the outcomes of applying our proposed methodology to analyze count data from an infectious disease. While it would be interesting to design a case application that is less specific and relies on a controlled simulation environment to ensure more consistent and unbiased results, this is challenging. The complexity arises from the influence of covariates on the time series and other factors that are difficult to specify in such a complex scenario for our proposed counting model. We begin by detailing the data collection process and providing a thorough description of the dataset. Building upon the insights gained from the data analysis, we then compare our proposed methodology with existing approaches. Finally, we evaluate the performance of our methodology in predicting Dengue cases over time through a comparative analysis with competing methodologies.

Data and Information Collection

The data comprises weekly Dengue cases in the department of Antioquia, Colombia, during the years 2008 and 2012. Additionally, Google search trends for the word ”dengue” (X_1) and climatological data such as temperature (X_2), precipitation (X_3), and relative humidity (X_4) are employed as covariates to adjust the proposed model as described in step 4 of our methodology. X_1 was obtained from the Google Trends tool as an open resource, and the rest was obtained through a formal solicitation to the *Secretaría de Salud del Departamento de Antioquia*, and the *Instituto de Hidrología, Meteorología y Estudios Ambientales* (IDEAM, spanish acronym) [46].

To explore the relationship between Dengue cases and these covariates, Figure 1 visually illustrates the temporal trends of Dengue cases and related variables, offering insights into potential associations and patterns within the data.

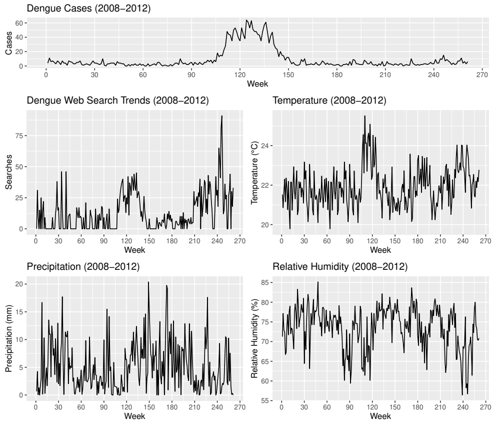

Figure 1 The plot shows the temporal evolution of Dengue cases alongside various covariates. Dengue cases reach their maximum peak of approximately 60 cases per week around week 120. During this period, there is also a graphical depiction of an increase in average temperature, suggesting a potential relationship between these variables. Moreover, the search trends for the word ”dengue” show a strong relationship with cases, indicating their potential as factors in the spread of the disease.

Following the exploration of the dataset, Figure 2 presents a comprehensive visualization of the covariates. The figure encapsulates density graphs, scatter plots, and Pearson correlations, offering a multifaceted insight into the characteristics of the dataset. Scatter plots elucidate relationships between pairs of variables, providing visual cues to potential associations. Moreover, Pearson correlations quantify these associations, highlighting notable relationships between temperature and precipitation with relative humidity.

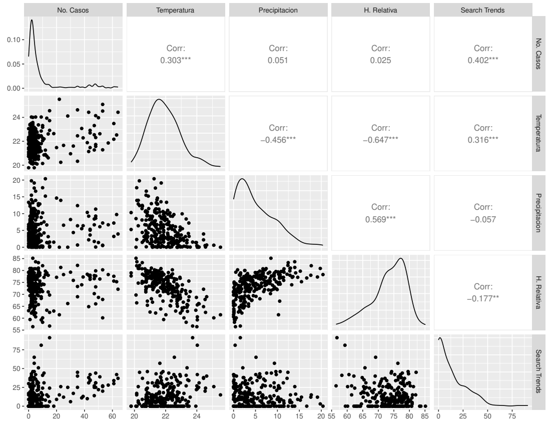

Figure 2 Comprehensive visualization of covariates: Density Graphs, Scatter Plots, and Pearson Correlations. Density graphs depict the distribution of each covariate, revealing right asymmetries for all variables except relative humidity, which exhibits left asymmetry. This figure also includes scatter plots and Pearson correlations for each pair of variables, highlighting notable relationships between temperature and precipitation with relative humidity.

Comparison with Other Methodologies and Evaluation

Considering one of the most common models for count data in the literature, we use INGARCH in the case of the time series approach and Poisson regression in the regression framework. A comparison of their performance will test how well these models fit the datasets against our methodology.

We conducted our research using the R software [47], a preferred tool for research in various fields due to its open-source nature and diverse range of packages for statistical methodologies and machine learning algorithms. Specifically, for fitting generalized linear models and conducting Poisson regression, we utilized the `glm()` function from the base R package `stats`. Similarly, to implement the INGARCH model with covariates, we utilized the `tsglm()` function, from the `tscount` package [48]. Within the `tsglm()` function, we specified various model parameters, such as the type of distribution, link function, and the inclusion of external covariates. For prediction purposes, we utilized the `predict()` function to generate forecasts based on the fitted models, using simulated covariate data for which predictions were desired. Specifically, we utilized 12 covariate data points to obtain predictions for the response variable over the subsequent 12 periods.

To illustrate the role of covariates in NHPP-FD, consider the intensity function, a core aspect of this methodology. Figure 3 depicts this function for the Dengue fever case study. The intensity function essentially captures how the expected number of cases changes over time.

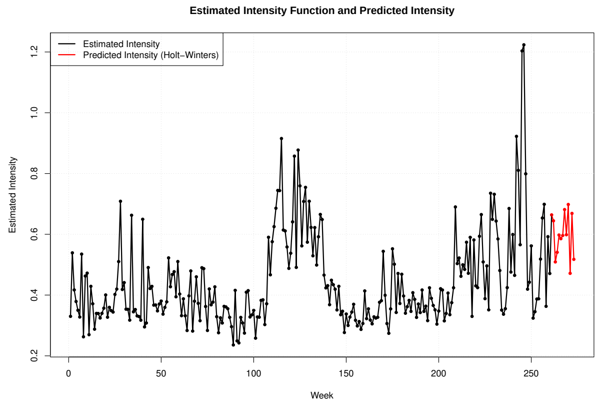

Figure 3 Estimated intensity function for the Dengue fever case study. The black line shows the estimated intensity function based on the observed Dengue fever data. In contrast, the red line represents the predicted intensity function based on forecasted covariate values using Holt-Winters. This visual comparison demonstrates how NHPP-FD incorporates covariates to dynamically adjust the expected number of cases over time.

In our evaluation of proposed methodologies, we conducted a comparative analysis of raw count predictions for the 12 forecasted periods with the simulated covariate values, starting with a simplistic Poisson regression model, which served as a foundational benchmark. Despite its simplicity, the Poisson regression model demonstrated a remarkable ability to stick closely to the midpoint of the confidence band, indicating a high degree of reliability in estimating overall uncertainty. However, further examination revealed that while the Poisson regression model stood out in maintaining proximity to the midpoint of the confidence band, it struggled to match the peaks of the time series data and lacked variability. Subsequent comparison with the INGARCH model reinforced these findings, showcasing improved performance in capturing dynamic structure but still falling short in peak modeling precision. Figure 4 presents a comparison of counts for INGARCH, Poisson Regression, and NHPP-FD, highlighting the aforementioned differences in performance, as well as a visual comparison of cumulative counts across competing methodologies.

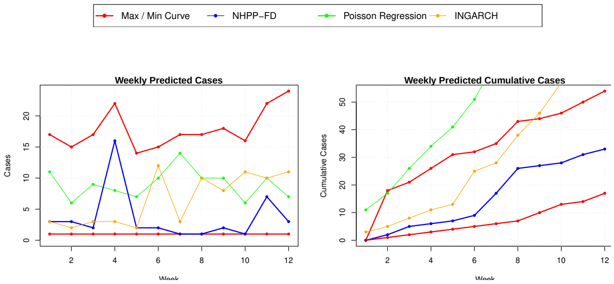

Figure 4 Comparison of predicted weekly cases and cumulative cases using INGARCH, Poisson Regression, NHPP-FD, and the proposed methodology. For Weekly Cases INGARCH and Poisson Regression tend to be closer to the midpoint of the confidence band, indicating reliability in overall uncertainty estimation, but struggle to capture the dynamic structure of the time series, particularly the peaks. In contrast, the proposed methodology shows peak modeling precision. For Cumulative Cases, competing methodologies fail to remain within the confidence band of the cumulative cases. The proposed approach consistently stays within this band throughout the predicted values. Additionally, NHPP-FD exhibits a varying behavior, showcasing its ability to capture dynamic phenomena effectively.

The performance in cumulative counts is a crucial metric in understanding the overall progression of the observed phenomena. Strikingly, we observed that the methodologies under scrutiny consistently failed to remain within the confidence band of the cumulative cases, indicating significant discrepancies in their predictions. However, our proposed approach demonstrated remarkable consistency, maintaining alignment within the confidence band throughout the predicted values. Additionally, while Poisson regression exhibited near-constant behavior with a more linear pattern, NHPP-FD displayed a varying behavior, showcasing its ability to capture dynamic phenomena effectively. This visual representation offers valuable insights into the comparative performance of methodologies in predicting cumulative counts and underscores the robustness of our proposed approach. In Figure 5, we show how these predicted trajectories compare to the defined time series of cumulative counts of the phenomena, and also present the non-cumulative counts, offering a detailed view of the weekly variations.

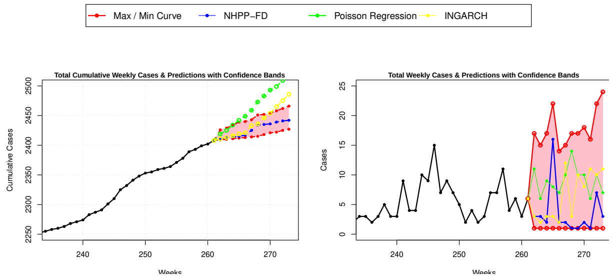

Figure 5 Comparison of Cumulative and Weekly Counts. The total known count trajectory of dengue cases is illustrated by the black dotted line time series, with forecasted predictions starting after the 261st period. The cumulative counts figure highlights the overall trend and the long-term accumulation of cases, while the weekly counts figure focuses on the short-term fluctuations, providing a more granular perspective on the model performance and the inherent variability in the data.

In Table 1, we present the proportion of cumulative predictions that fall within the confidence band for the various forecasting predictions we obtained. The analysis focuses on evaluating the consistency of each method in accurately capturing the dynamic behavior of the observed data while maintaining adherence to the predefined confidence band. Notably, NHPP-FD demonstrates remarkable consistency, regularly achieving a proportion of 1 for predictions within the confidence band. This proves the robustness of our approach in providing reliable forecasts that align closely with the observed data dynamics.

Method	**Proportion**
Poisson Regression	0.083
NHPP-FD	1.000
INGARCH	0.583

Table 1 Proportion of cumulative predictions within the confidence bands. It illustrates the proportion of times the cumulative predictions from different methods fall within the confidence band. NHPP-FD consistently maintains this criterion while capturing dynamic behavior.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are widely used statistical measures for model selection, providing a balance between goodness of fit and model complexity. Lower values of the AIC and BIC indicate better model fit with a preference for simpler models. In this analysis, we compare the AIC and BIC values obtained from different models, including Poisson regression, NHPP-FD model, and INGARCH model. Table 2 presents AIC and BIC of the models. The results shed light on the relative performance of these models in capturing the underlying data dynamics.

Model	AIC	BIC
Poisson Regression	3138.22	3156.05
NHPP-FD	55.05	72.88
INGARCH Model	1787.72	2661.03

Table 2 Comparison of AIC and BIC values for different models. The results suggest that our model demonstrates significantly lower AIC and BIC values compared to the other models, indicating superior goodness of fit and parsimony.

The inherent properties of the Poisson process underpinning NHPP-FD offer distinct advantages. By preserving these properties in our counts generation, we ensure a faithful representation of the underlying phenomenon, enhancing the explanatory power of the model. The non-parametric approach to modeling intensity further contributes to the versatility and accuracy of the methodology, allowing for a comprehensive understanding of the process dynamics. Furthermore, NHPP-FD enables quantification of the influence of explanatory variables through its structured intensity function. This feature facilitates a deeper insight into the factors driving the observed phenomena capabilities. In contrast, competing models lack this level of granularity. Overall, NHPP-FD stands out as a fantastic choice for modeling complex processes, offering advantages in accuracy, versatility, and interpretability over traditional techniques by providing a robust framework for analyzing dynamic phenomena and making informed predictions.

4. Concluding Remarks

This work introduces a methodology for analyzing count data by combining NHPP and FDA approaches, with an application to infectious disease cases. The NHPP-FD method preserves the inherent properties of the underlying stochastic process governing the occurrence of Dengue cases. By maintaining fidelity to these properties, we ensure the reliability and accuracy of our model predictions. NHPP-FD harnesses the concepts of NHPP Counting Processes and Functional Data Analysis, combining temporal evolution and exogenous variables to model and predict count data scenarios. This approach enables us to capture the dynamic nature of the data. Moreover, the incorporation of covariates given in our data set as search trends and climatological data further enhances the predictive capabilities of our model, allowing for a more nuanced analysis of count data phenomena dynamics. Comparative analysis against existing methodologies reaffirms the strengths of our approach. When contrasting our results with those obtained from traditional techniques such as Poisson regression and INGARCH, NHPP-FD performed well in modeling peaks and fluctuations over time.

Designing a case application based on specific conditions and established covariates presents challenges in planning a simulation scenario that allows for equal comparison of different models' competitiveness. Ensuring a fair and reliable competition among models is difficult to achieve due to the unique characteristics of each case. The results obtained in this study focus on dengue, a specific scenario within epidemiology. As future work, it would be interesting to design a simulation framework that allows for more general comparisons across various methodologies applicable to count data. This would ensure a more reliable and fair assessment of different approaches.

This framework can be extended to analyze diverse datasets beyond infectious disease counts. By exploring datasets that vary across different disciplines and even different dimensions such as space rather than time, we can leverage the inherent properties of NHPP-FD to gain deeper insights into complex phenomena. For instance, integrating our technique with spatial analysis methods like Geographic Information Systems (GIS) could facilitate the exploration of spatial patterns and correlations in disease spread. Additionally, interdisciplinary collaborations with fields such as ecology and environmental science could enable us to model the impact of ecological factors on disease dynamics and other phenomena. Moreover, integrating machine learning techniques such as neural networks or ensemble methods could enhance the predictive capabilities of our model, enabling more accurate and timely forecasting of disease outbreaks.

Declaration of Competing Interest

We declare that we have no significant competing interests including financial or non-financial, professional, or personal interests interfering with the full and objective presentation of the work described in this manuscript.

Acknowledgments

The authors thank Alexandra Catano-Lopez for providing her help in data acquisition. Finally, the authors thank the two anonymous reviewers and the editor for their constructive comments that improved the final presentation of this manuscript.

Funding

This research was partially supported by Universidad EAFIT, grant number 954-000002 (S. O.) and by Ministerio de Ciencia Tecnología e Innovación de Colombia, projects (I) Formación de Capital Humano de Alto Nivel–Universidad EAFIT—Corte 2 Nacional, BPIN 2020000100778 (H. V.), and (II) Convocatoria 909-2 2022 (S. O.). The funders had no role in study design, data collection and analysis, publication decisions, or manuscript preparation.

Author Contributions

Data curation S. O.; conceptualization: S. O.; Formal analysis: S. O. and J. E. C.; investigation: S. O., J. E. C. and H. V.; methodology: S. O., J. E. C. and H. V.; writing—original draft: S. O. and J. E. C.; writing—review and editing, S. O. All authors read and approved the final version of this manuscript.

Data Availability Statement

The authors confirm that the data supporting the findings of this study, including the codes as well, are available under request.

References

[1] Jay M. Ver Hoef and Peter L. Boveng. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.

[2] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Springer US, 1989.

[3] Scott L. Zeger and Bahjat Qaqish. A regression model for time series of counts. *Biometrika*, 75(4):621–629, 1988.

[4] M. P. Quine and E. Seneta. Bortkiewicz's data and the law of small numbers. *International Statistical Review / Revue Internationale de Statistique*, 55(2):173, 1987.

[5] Alexandra M. Schmidt and João Batista M. Pereira. Modelling time series of counts in epidemiology. *International Statistical Review*, 79(1):48–69, Apr 2011.

[6] S. Lundbye-Christensen, C. Dethlefsen, A. Gorst-Rasmussen, T. Fischer, H. C. Schønheyder, K. J. Rothman, and H. T. Sørensen. Examining secular trends and seasonality in count data using dynamic generalized linear modelling: a new methodological approach illustrated with hospital discharge data on myocardial infarction. *European Journal of Epidemiology*, 24(5):225–230, May 2009.

[7] Stefany Coxe, Stephen G. West, and Leona S. Aiken. The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of Personality Assessment*, 91(2):121–136, 2009.

[8] Adrian Colin Cameron and P. K. Trivedi. *Regression analysis of count data.* Econometric society monographs. Cambridge University Press, 1998.

[9] J. Scott Long. *Regression models for categorical and limited dependent variables.* Advanced quantitative techniques in the social sciences. Sage Publications, Thousand Oaks, 1997.

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

[11] Daniel B. Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039, Dec 2000.

[12] Christian Weiß. *An Introduction to Discrete-Valued Time Series.* Wiley, 1 edition, Jan 2018.

[13] A. A. Alzaid and M. Al-osh. An integer-valued pth–order autoregressive structure (inar(p)) process. *Journal of Applied Probability*, 27(2):314–324, Jun 1990.

[14] Tina Hviid Rydberg and Neil Shephard. Bin models for trade-by-trade data. modelling the number of trades in a fixed interval of time. *Econometric Society World Congress 2000 Contributed Papers*, Aug 2000.

[15] Andréas Heinen. Modelling time series count data: An autoregressive conditional poisson model. *SSRN Electronic Journal*, 2003.

[16] Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439, Dec 2009.

[17] Chester L. Britt, Michael Rocque, and Gregory M. Zimmerman. The analysis of bounded count data in criminology. *Journal of Quantitative Criminology*, 34(2):591–607, Jun 2018.

[18] Stuart Coles. *An introduction to statistical modeling of extreme values*, chapter 7, A Point Process Characterization of Extremes. Springer Series in Statistics. Springer-Verlag, London, 2001.

[19] Joseph M. Hilbe. *Chapter 1: Varieties of Count Data.* Cambridge University Press, 2014.

[20] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[21] Matthew J. Hayat and Melinda Higgins. Understanding poisson regression. *Journal of Nursing Education*, 53(4):207–215, 2014.

[22] M. Katherine Hutchinson and Matthew C. Holtman. Analysis of count data using poisson regression. *Research in Nursing & Health*, 28(5):408–418, 2005.

[23] P.C. Consul and Felix Famoye. Generalized poisson regression model. *Communications in Statistics - Theory and Methods*, 21(1):89–109, 1992.

[24] Steven E. Rigdon and Asit P. Basu. *Statistical methods for the reliability of repairable systems.* Wiley series in probability and statistics. Wiley, 2000.

[25] Sanghoon Lee, James R. Wilson, and Melba M. Crawford. Modeling and simulation of a nonhomogeneous poisson process having cyclic behavior. *Communications in Statistics - Simulation and Computation*, 20(2-3):777–809, 1991.

[26] Edward P. Kao and Sheng-Lin Chang. Modeling time-dependent arrivals to service systems: A case in using a piecewise-polynomial rate function in a nonhomogeneous poisson process. *Management Science*, 34(11):1367–1379, 1988.

[27] A.M. Law and W.D. Kelton. *Simulation Modelling and Analysis.* McGraw-Hill series in industrial engineering and management science. McGraw Hill, 1991.

[28] P. A. W. Lewis and G. S. Shedler. Statistical analysis of non-stationary series of events in a data base system. *IBM Journal of Research and Development*, 20(5):465–482, 1976.

[29] Ana Cebrian, Jesús Abaurrea, and Jesús Asin. Nhpoisson: An r package for fitting and validating nonhomogeneous poisson processes. *Journal of statistical software*, 64, 03

2015.

[30] Lawrence M Leemis. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous poisson process. *Management Science*, 37:886–900, 1991.

[31] D. R. Cox and P. Lewis. *The Statistical Analysis of Series of Events [by] D.R. Cox and P.A.W. Lewis.* Methuen's monographs on applied probability and statistics. Methuen, 1966.

[32] Paul Bratley, Bennett L. Fox, and Linus E. Schrage. *A Guide to Simulation*, chapter 5, Nonuniform Random Numbers. Springer New York, New York, NY, 1987.

[33] Raghu Pasupathy. *Generating Homogeneous Poisson Processes.* Wiley, 1st edition, January 2011.

[34] Erhard Cinlar. *Introduction to stochastic processes*, chapter 4, Poisson Processes. Prentice-Hall, 1975.

[35] P. A. W Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, September 1979.

[36] Daryl J. Daley and D. Vere-Jones. *An introduction to the theory of point processes.* Springer, New York, 2nd ed edition, 2003.

[37] Ana C. Cebrian. *NHPoisson: Modelling and Validation of Non Homogeneous Poisson Processes*, 2020. R package version 3.3.

[38] Martin Gaston, Teresa León, and Fermín Mallor. Functional data analysis for non homogeneous poisson processes. *2008 Winter Simulation Conference*, pages 337–343, 2008.

[39] Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics).* Springer-Verlag, Berlin, Heidelberg, 2006.

[40] James O. Ramsay and B. W. Silverman. *Functional data analysis.* Springer series in statistics. Springer, New York, 2nd ed edition, 2005.

[41] J. O. Ramsay, Giles Hooker, and Spencer Graves. *Functional data analysis with R and MATLAB.* Use R! Springer, Dordrecht; New York, 2009.

[42] Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, Jun 2009.

[43] Ying Sun and Marc G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, Jan 2011.

[44] Daniel Peña. *Análisis de Series Temporales.* Alianza Editorial, October 2010.

[45] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*, 51(2):1063–1074, November 2006.

[46] Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), 2022.

[47] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2024.

[48] Tobias Liboschik, Roland Fried, Konstantinos Fokianos, and Philipp Probst. *tscount: Analysis of Count Time Series*, 2020. R package version 1.4.3.