A MULTIVARIATE POISSON INTEGER-VALUED AUTOREGRESSIVE MODEL FOR PUBLIC HEALTH SURVEILLANCE

By

Zielesa Zulu

A dissertation submitted to the University of Zambia in partial fulfillment of the requirements for the degree of Master of Science in Statistics

THE UNIVERSITY OF ZAMBIA
LUSAKA

2025

Declaration

I, **Zielesa Zulu**, hereby declare that this dissertation represents my own work (except where otherwise indicated), and that it has not previously been submitted for a degree, diploma or other qualification at this or another university.

Signature: ..... Date: .....

Certificate of Approval

This dissertation of Zielesa Zulu has been approved as fulfilling the requirements or partial fulfilment of the requirements for the award of the degree of Master of Science in Statistics by the University of Zambia.

.....	.....	.....
Examiner I	Signature	Date
.....	.....	.....
Examiner II	Signature	Date
.....	.....	.....
Examiner III	Signature	Date
.....	.....	.....
Chairperson Board of Examiners	Signature	Date
.....	.....	.....
Supervisor	Signature	Date

Abstract

This dissertation develops a first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) model to improve disease outbreak detection in public health surveillance systems. The MPINAR(1) model is proposed as a computationally tractable alternative to the general multivariate integer-valued autoregressive (MINAR(1)) model. While the MINAR(1) framework allows for correlated innovations, it introduces substantial estimation challenges and requires strong distributional assumptions that may not hold in practice. To address these limitations, the MPINAR(1) model simplifies the innovation structure by assuming that the innovation process consists of independent Poisson-distributed components. This assumption retains the discrete nature of the data while reducing computational complexity, allowing for efficient implementation using conditional maximum likelihood estimation.

A simulation study is conducted to evaluate the performance of the MPINAR(1) model in detecting simulated outbreaks within trivariate count time series. The model is assessed against independent univariate INAR(1) models using metrics such as average run length (ARL), detection rate, and false alarm rate (FAR). Results show that the MPINAR(1) model outperforms the univariate alternatives, particularly in scenarios involving moderate to large outbreak sizes.

The model is further applied to syndromic surveillance data comprising daily counts of fever, cough, and dyspnea among hospitalised COVID-19 patients. Covariate adjustments for day-of-week and seasonal effects are incorporated into the innovation process. The MPINAR(1) model demonstrates superior fit compared to univariate models, producing wider and more reliable prediction intervals and detecting additional potential outbreak events. Residual analyses confirm that the MPINAR(1) model provides a closer approximation to white noise and better captures temporal and cross-sectional dependencies.

Overall, the MPINAR(1) model provides a theoretically sound and computationally efficient framework for multivariate count time series modelling in surveillance settings. Its improved detection capabilities and robustness to overdispersion make it a valuable tool for enhancing early warning systems in public health. Future extensions could include more flexible innovation structures and refined alarm decision rules.

To my family, especially in honor of the treasured legacy of my late father, Levias Zulu.

Acknowledgements

I would like to thank my supervisor, Dr. J. Musonda, for his guidance and tireless efforts in bringing this work to fruition. I would also like to thank the Department of Mathematics and Statistics, School of Natural Sciences, University of Zambia for providing a conducive environment for study and research. I also thank all the lecturers that taught me during my postgraduate studies for the knowledge they imparted in me.

I am also grateful to my fellow postgraduate students in the Department of Mathematics and Statistics at the University of Zambia, both past and present, for their encouragement and support throughout my academic journey. Special thanks go to Ms. P. Mwewa, Mr. K. Tambuzi, Mr. R. Sichone, Mr. K. Mazimba, Mr. T. Kabunda, Mr. E. Mwansa, Ms. M. Manda and Mrs. P. Lungu, who never missed an opportunity to offer encouragement during my postgraduate studies at the University of Zambia.

My heartfelt thanks go to my mother, Loveness, for her lifelong support, love, and trust in my choices. I am also grateful to my brothers, Clinton and Levy, and my sisters, Ngawa, Elita, and Annet, for their unwavering support, prayers, and belief in me and my dreams throughout my studies.

Finally, I am indebted to many others, too numerous to mention, who played a vital role in making the writing of this dissertation a success. May God richly bless you all.

Lusaka, January, 2025

Zielesa Zulu

Table of Contents

List of Symbols and Terms

Below is a list of some symbols, abbreviations, distributions, and statistical models used in this dissertation, along with their corresponding meanings.

Symbols

θ	Vector of unknown parameters.
$l(\theta)$	Log-likelihood function.
$\circ$	Binomial thinning operator.

Abbreviations

MLE	Maximum Likelihood Estimation.
CLS	Conditional Least Squares.
YW	Yule–Walker.
ARL	Average Run Length.
EWMA	Exponentially Weighted Moving Average.
CUSUM	Cumulative Sum.

Distributions

POI(λ)	Poisson distribution with parameter $\lambda > 0$.
BIN$(1, \alpha)$	Bernoulli distribution with success probability α, where $0 \leq \alpha \leq 1$.

Statistical Models

AR(p)	Autoregressive model of order p, where p is a positive integer.
INAR(p)	Integer-valued autoregressive model of order p.
MINAR(p)	Multivariate integer-valued autoregressive model of order p.
MPINAR(p)	Multivariate Poisson integer-valued autoregressive model of order p.

Miscellaneous

$\boldsymbol{A}^T$	Transpose of the matrix $\boldsymbol{A}$
$\mathbb{N}$	The set of natural numbers: $\{1, 2, 3, \dots\}$
$\mathbb{N}_0$	The set of non-negative integers: $\{0, 1, 2, 3, \dots\}$
$\mathbb{Z}$	The set of integers: $\{\dots, -2, -1, 0, 1, 2, \dots\}$

Chapter 1

Introduction

The aim of public health surveillance systems is to detect disease outbreaks effectively and timely so that control measures for the elimination of disease transmission are taken rapidly. To achieve this, various statistical techniques, such as statistical process control methods and statistical modelling techniques, have been developed (see [29, 30, 31]). These methods are designed to identify unusual patterns in data series that may indicate disease outbreaks. However, health surveillance data possess unique characteristics that introduce statistical challenges in this research area.

Health data typically consist of non-negative counts representing the number of events observed at specific time points. For example, daily counts of diagnosed cases or emergency department visits are common metrics. Such data are often characterized by serial correlation (dependence over time), overdispersion (variance exceeding the mean), and the integer-valued nature of observations. Failure to properly account for these characteristics can lead to misspecified models, undermining their reliability for outbreak detection. Time series techniques, particularly autoregressive integrated moving average (ARIMA) models, have been widely used in public health surveillance to model these dependencies and variations over time (see [3, 16, 33]).

While these approaches have been useful, they are generally limited to monitoring single data series, which has restricted their effectiveness in complex real-world scenarios. For the early detection of large-scale epidemics or bioterrorism outbreaks, it is essential to consider multivariate data. Such data can represent a specific variable measured across multiple regions, different variables measured in a single region, or a combination of multiple variables across multiple regions. These scenarios often involve correlations between data series (cross correlation), as the same underlying process may influence several indicators.

These challenges highlight the need for a robust modelling framework capable of addressing the complexities of multivariate health surveillance data, particularly the integer-valued nature of health data, serial and cross correlations, and overdispersion. To address this gap, this study introduces a first-order multivariate Poisson integer-valued autoregressive model (MPINAR(1)), a robust framework designed to improve the effectiveness and timeliness of disease outbreak detection. The proposed model aims to contribute to better public health decision-making and resource allocation. The persistent challenges in health surveillance systems underline the importance of developing advanced modelling frameworks. These issues form the foundation of the problem explored in this study.

1.1 Statement of the Problem

Public health surveillance systems collect multivariate data to track health indicators and disease events. However, current statistical models often fail to account for complex interdependencies, such as serial and cross-correlations between disease indicators over time, or the integer-valued nature and overdispersion of the data. This limits timely and accurate outbreak detection, underscoring the need for a robust modelling framework that handles these challenges effectively.

1.2 Aim of the Study

The aim of this study is to develop a first-order multivariate Poisson integer-valued autoregressive model (MPINAR(1)) to enhance the effectiveness and timeliness of disease outbreak detection in public health surveillance.

1.3 Research Objectives

(i) To formulate a practical and theoretically grounded MPINAR(1) model for detecting disease outbreaks in public health surveillance.

(ii) To develop a parameter estimation approach and application strategy for the model.

(iii) To evaluate the model's performance through a simulation study, comparing it with independent univariate INAR(1) models under controlled conditions.

(iv) To assess the model's practical utility by applying it to real-world surveillance data.

1.4 Research Questions

(i) How can an MPINAR(1) model be formulated to balance theoretical rigour with practical utility in public health surveillance?

(ii) What parameter estimation method and application strategy can ensure the model's practical implementation and reliability?

(iii) How does the MPINAR(1) model perform in comparison to independent univariate INAR(1) models under controlled simulation settings?

(iv) How effective is the MPINAR(1) model in detecting outbreaks from real data?

1.5 Significance of the Study

This study contributes to the fields of public health surveillance and time series analysis by developing a multivariate integer-valued autoregressive model that effectively captures the integer-valued nature of health data, serial and cross-correlations, and overdispersion. The proposed model promises earlier disease outbreak detection, more informed decision-making, and improved resource allocation in public health responses.

1.6 Literature Review

Research in public health surveillance has grown rapidly in recent years. However, only a few approaches have simultaneously addressed the integer-valued nature of health surveillance data alongside its complex correlation structure. Among the most notable contributions are the works of Held et al. [12, 14, 15], who used multivariate branching processes to model infectious disease surveillance data. While effective, these methods can be computationally demanding and may not adequately capture the integer-valued property of the data.

This study proposes an alternative approach based on integer-valued autoregressive (INAR) processes. INAR models were introduced in the literature as discrete-valued counterparts of the standard Gaussian autoregressive process by Al-Osh and Alzaid [1] and McKenzie [21]. These models have been used in public health surveillance by several researchers. Cardinal et al. [2] demonstrated that INAR models outperform autoregressive integrated moving average (ARIMA) models in forecasting infectious disease incidence by yielding smaller relative forecast errors. Within a statistical process control framework, Weiß [35, 38] and Weiß and Testik [36] extended INAR models to develop cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) control charts for Poisson first-order INAR processes.

Some extensions of the INAR models to the multivariate case have been proposed by Franke and Rao [8], Latour [18], McKenzie [21], Quoeshi [28], Heinen and Rengifo [13] and Pedeli and Karlis [25]. The extension of the first-order integer-valued autoregressive (INAR(1)) process to a multi-dimensional framework is highly intriguing, yet it remains underdeveloped in the literature to date. Pedeli and Karlis [27] introduced a multivariate INAR(1) (MINAR(1)) process with correlated innovations. This model was designed to capture dependencies both within and across time series, providing a flexible framework for analyzing multivariate count data. By incorporating correlations between the innovations, the model is well-suited for applications where interactions between different health indicators are expected, such as during the simultaneous monitoring of multiple symptoms or diseases. However, the inclusion of correlated innovations increases the computational complexity of parameter estimation and requires strong distributional assumptions about the nature of the correlations, which may not always align with real-world data. To avoid such complications, Pedeli and Karlis [26] considered a constrained MINAR(1) model by assuming a single source of dependence between the univariate series that comprise the MINAR(1) process.

This study develops a simplified MINAR(1) model based on the work of Pedeli and Karlis [27], assuming independent innovations that follow a Poisson distribution. The choice of the Poisson distribution is motivated by its ability to reflect the integer-valued nature of health data while being the simplest discrete distribution to work with. By focusing on prediction thresholds for multivariate health data, this approach addresses key gaps in outbreak detection related to overdispersion, correlations, and computational feasibility, enhancing the effectiveness and timeliness of disease detection in public health surveillance.

1.7 Methodology

This section outlines the methodology used to develop the first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) model for public health surveillance. It begins with the model formulation, followed by an explanation of the parameter estimation approach and a strategy for applying the model to disease outbreak detection. The section then describes the design of a simulation study to assess the model's performance in controlled settings and concludes with its application to real-world surveillance data.

1.7.1 Model Formulation

The MPINAR(1) model is formulated as a simplification of the first-order multivariate integer-valued autoregressive (MINAR(1)) model introduced by Pedeli and Karlis [27]. The MINAR(1) process $\{\boldsymbol{X}_t\}$ is defined as:

$$\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{1.1}$$

where:

- $\boldsymbol{X}$ is a non-negative integer-valued n-dimensional random vector,
- $\boldsymbol{A}$ is an $n \times n$ matrix with entries α_{ij} satisfying $0 \leq \alpha_{ij} \leq 1$ for all $i, j = 1, \ldots, n$,
- $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})^T$ is an innovation vector whose components ε_{it} are **correlated** non-negative integer-valued random variables. The vector $\boldsymbol{\varepsilon}_t$ is independent of $\boldsymbol{A} \circ \boldsymbol{X}_{t-1}$, and has mean $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$,
- $\boldsymbol{A} \circ \boldsymbol{X}$ is an n-dimensional random vector with i-th component given by

$$(\boldsymbol{A} \circ \boldsymbol{X})_i = \sum_{j=1}^{n} \alpha_{ij} \circ X_j, \quad i = 1, \ldots, n, \tag{1.2}$$

  where $\alpha_{ij} \circ X_j$ are mutually independent binomial thinning operations, defined by

$$\alpha_{ij} \circ X_j = \sum_{k=1}^{X_j} Y_k^{(ij)}, \tag{1.3}$$

  where $Y_k^{(ij)}$ are independent and identically distributed Bernoulli random variables with

$$P\left(Y_k^{(ij)} = 1\right) = \alpha_{ij} \quad \text{and} \quad P\left(Y_k^{(ij)} = 0\right) = 1 - \alpha_{ij}. \tag{1.4}$$

  It follows from equation 1.2 that $\boldsymbol{A}\circ$ acts as the usual matrix multiplication while simultaneously adhering to the definition of the binomial thinning operation.

Traditionally, statistical models for public health surveillance data aim to effectively capture the endemic and epidemic dynamics of disease risk. In principle, the endemic component explains a baseline rate of cases with a stable temporal pattern and is identified with the innovations part of the model. More specifically, it describes the risk of new events as a

function of external factors that are independent of the history of the epidemic process. Examples of such external factors include seasonality, socio-demographic characteristics, population density, and vaccination coverage. The epidemic component, on the other hand, introduces infectiousness, that is, an explicit dependence between events. Consequently, the epidemic component is driven by past observations and corresponds to the autoregressive part of the model.

This additive decomposition of disease risk is well-embodied in the MINAR(1) model (1.1). However, the inclusion of correlated innovations increases the computational complexity of parameter estimation and requires strong distributional assumptions about the nature of the correlations, which may not always align with real-world data. To avoid such complications, we consider in this dissertation a simplification of the MINAR(1) model 1.1. In particular, we formulate the MPINAR(1) model by relaxing the degree of complexity of the model by assuming that the innovation series $\boldsymbol{\varepsilon}_t$ (the endemic components) are independent and follow a Poisson distribution (see Chapter 3 for the actual formulation). The choice of the Poisson distribution is motivated by its ability to reflect the integer-valued nature of health data while being the simplest discrete distribution to work with. The resulting model admits a realistic epidemiological interpretation and is extremely advantageous in terms of practical implementation since the distribution of the innovations becomes a product of univariate probability mass functions:

$$P\left(\varepsilon_{1t} = k_1, \ldots, \varepsilon_{nt} = k_n\right) = \prod_{i=1}^{n} P\left(\varepsilon_{it} = k_i\right). \tag{1.5}$$

Moreover, overdispersion, a typical characteristic of health surveillance data, can be easily accommodated even under our simplest parametric assumption of Poisson innovations.

1.7.2 Parameter Estimation

The MPINAR(1) process is defined as:

$$\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{1.6}$$

where $\boldsymbol{A}$ is a thinning matrix, $\boldsymbol{X}_t$ is the observed multivariate count process, and $\boldsymbol{\varepsilon}_t$ is a vector of innovations. The innovations ε_{it} are assumed to follow independent Poisson distributions with means λ_i and variances $v_i\lambda_i$, where $v_i > 1$ accounts for overdispersion.

The unknown parameter vector $\boldsymbol{\theta}$, which includes the elements of $\boldsymbol{A}$ and the innovation means λ_i are estimated using conditional maximum likelihood estimation (CMLE). The dispersion parameters v_i are then estimated using the residuals from the fitted model.

The likelihood function for the MPINAR(1) process is derived from the conditional distribution of $\boldsymbol{X}_t$ given $\boldsymbol{X}_{t-1}$. For a time series of length T, the likelihood function is given by:

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{t=2}^{T} f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right), \tag{1.7}$$

where $f(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta})$ is the conditional probability mass function (PMF) of $\boldsymbol{X}_t$ given $\boldsymbol{X}_{t-1}$. The estimation of $\boldsymbol{\theta}$ involves maximizing the log-likelihood function:

$$l(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{t=2}^{T} \log f(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}). \tag{1.8}$$

This is typically done using numerical optimization techniques. In this dissertation, this is implemented in R using the `optim` function from the `tscount` package. See Appendix C.

After obtaining the estimates for $\boldsymbol{A}$ and λ_i, the dispersion parameters v_i are estimated using the residuals from the fitted model. The residuals r_{it} are computed as the difference between the observed values x_{it} and the predicted values $\hat{x}_{it}$ based on the estimated model:

$$r_{it} = x_{it} - \hat{x}_{it}. \tag{1.9}$$

The dispersion parameters v_i are estimated by calculating the average of the squared residuals divided by the estimated mean $\hat{\lambda}_i$ for each component i over the time series, as follows:

$$\hat{v}_i = \frac{1}{T-1} \sum_{t=2}^{T} \frac{r_{it}^2}{\hat{\lambda}_i}. \tag{1.10}$$

We conclude this subsection by remarking that the quasi-likelihood framework provides an alternative approach for estimating the parameters of the MPINAR(1) model, including the dispersion parameters $v_1, \ldots, v_n$. This method is particularly useful when the innovations $\boldsymbol{\varepsilon}_t$ exhibit overdispersion, i.e., when $\text{Var}(\varepsilon_{it}) > \lambda_i$. Unlike the standard likelihood approach, which requires a fully specified probability distribution, the quasi-likelihood framework relies on specifying only the mean-variance relationship of the data. Specifically, the variance of the innovations is modelled as:

$$\text{Var}(\varepsilon_{it}) = v_i \lambda_i, \quad i = 1, \ldots, n, \tag{1.11}$$

where $v_i > 1$ are the dispersion parameters. The quasi-likelihood function is then constructed based on this relationship, allowing for efficient estimation of the parameters without requiring a full distributional assumption. This approach is computationally simpler and more flexible, making it well-suited for applications where overdispersion is present.

1.7.3 Strategy for Applying the Model to Outbreak Detection

The MPINAR(1) process can be used for modelling clean historical data and generating one-step-ahead forecasts for prediction-based monitoring. We should emphasize here that the set-up phase is assumed to be free of, or cleaned of, outbreaks. The proposed outbreak detection statistical methodology comprises of two steps: In the first step, the available series of data in the set-up phase (historical data) is modelled through an MPINAR(1) process and a parameter vector of maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ is obtained. The second step is dedicated to the successive monitoring of incoming observations in the operational

phase (surveillance data) using the model obtained from the set-up phase. In particular, each new observation $\boldsymbol{x}_{t+1}$ is assessed against a multivariate prediction threshold derived from the model fitted in the first step in order to determine whether an alarm should be triggered. Specifically, for each multivariate observation $\boldsymbol{x}_{t+1}$ in the operational phase, we estimate the one-step-ahead predictive distribution

$$\hat{P}(\boldsymbol{X}_{t+1} = \boldsymbol{x}_{t+1} | \boldsymbol{x}_t, \hat{\theta}), \quad \boldsymbol{x} \in \mathbb{N}_0^n \tag{1.12}$$

and obtain the marginal predictive probabilities

$$\hat{P}(X_{i,t+1} = x_{i,t+1} | \boldsymbol{x}_t, \hat{\theta}), \quad i = 1, \ldots, n. \tag{1.13}$$

For each observation $x_{i,t+1}$, we construct a $(1-\alpha)\%$ prediction interval with upper bound $x_{i,t+1}^{UB}$ equal to the $(1-\alpha)$-quantile of the corresponding marginal predictive distribution, where α is a pre-specified level of significance. The lower bound of the prediction interval is set to 0 as we are only interested in detecting positive deviations from the in-control model. Each series flags an alarm at time $t+1$ if the corresponding observation lies outside the prediction interval, i.e., if

$$x_{i,t+1} > x_{i,t+1}^{UB}. \tag{1.14}$$

Finally, for the overall alarm, a majority rule can be defined, i.e. flagging an alarm if a certain percentage of the series signals an alarm at the same point in time (see Vial [32]).

1.7.4 Simulation Study

A simulation study was conducted to evaluate the performance of the proposed MPINAR(1) model in the context of disease outbreak detection. The study involved generating synthetic trivariate time series data under controlled conditions, allowing comparison between the MPINAR(1) model and independent univariate INAR(1) models.

Time series data of length $n = 200$ were simulated from a trivariate INAR(1) model with independent Poisson innovations with parameters λ_i $(i = 1, 2, 3)$. It was assumed that the first 150 observations consist the set-up phase (that is a clean process without outbreaks) and the last 50 observations consist the monitoring phase. Subsequently, for each series i, an outbreak of expected size κ_i at time $t = 170$ was simulated from a Poisson distribution with mean equal to κ_i . Therefore this trivariate case of the MPINAR(1) model has the form

$$\begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \circ \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \\ X_{3,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{pmatrix}, \tag{1.15}$$

where ε_{it} are independent Poisson random variables with mean $E(\varepsilon_{it}) = \lambda_i + \kappa_i I(t = 170)$ and $I(\cdot)$ is an indicator function. The true parameter values were assumed to be

$$\boldsymbol{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} = \begin{bmatrix} 0.4 & 0.1 & 0.3 \\ 0.3 & 0.4 & 0.2 \\ 0.3 & 0.2 & 0.1 \end{bmatrix}, \tag{1.16}$$

and $\lambda_1 = \lambda_2 = \lambda_3 = 1$. It was also assumed that $\kappa_1 = \kappa_2 = \kappa_3 = \kappa$ and the values of κ were taken to be 5, 8 or 10 . The aim of choosing these specific values for κ is to study both cases where the outbreak is manifest, as well as cases where the outbreak cannot be easily distinguished from the typical range of values in the in-control state.

For each value of $\kappa \in \{5, 8, 10\}$, a total of 1000 replicates were generated. In each replicate, the MPINAR(1) model was fitted to the set-up phase using maximum likelihood estimation, and one-step-ahead upper prediction bounds were computed for the monitoring phase at the 90%, 95%, and 99% significance levels. For comparison, three independent univariate INAR(1) models with Poisson innovations were also fitted separately to each series.

The detection performance was evaluated using the following metrics:

- **Detection Rate (DR):** The proportion of replicates in which an alarm was raised at the true outbreak time $t = 170$.
- **False Alarm Rate (FAR):** The proportion of total alarms triggered at times $t \neq 170$, computed as the number of false alarms divided by 1000×49 (i.e., the number of monitoring points across all replicates).
- **Average Run Length (ARL):** For each series i, ARL_i was defined as the average number of time points before the first false alarm during the monitoring phase. A conservative overall ARL was defined as $\text{ARL} = \min_i \text{ARL}_i$.

An outbreak alarm was triggered if at least two out of the three series exceeded their corresponding prediction intervals at the same time point, following a 2/3 majority rule.

The simulation study was implemented in the R statistical software. The code used for data generation, model fitting, and evaluation is provided in Appendix B.

1.7.5 Application to Real-World Surveillance Data

The real-world applicability of the proposed MPINAR(1) model is assessed using syndromic surveillance data provided by the Brazilian Ministry of Health. The dataset, which is publicly available through the OpenDataSUS SRAG, contains daily records of symptoms commonly observed in COVID-19 patients. In this study, we focus on three distinct COVID-19 symptoms fever, cough, and dyspnea (shortness of breath) that are statistically significantly correlated, with cross-correlation values ranging from 0.20 to 0.45. This correlation structure makes them suitable for multivariate time series analysis.

The dataset is divided into two phases: a set-up phase (March 11, 2020 to August 7, 2020) for model training, and a monitoring phase (August 8, 2020 to September 26, 2020) for evaluation. Symptoms were recorded daily, resulting in $t_0 = 150$ observations for the set-up phase and $t_1 = 50$ observations for the monitoring phase.

To assess stationarity, time series plots of the three symptom series were examined. The

absence of visible trends or seasonal patterns supports the assumption of weak stationarity. The appropriate order of the trivariate model was determined by analysing autocorrelation and partial autocorrelation functions during the set-up phase. The exponentially decaying autocorrelation functions suggest the suitability of an autoregressive (AR) model, while the partial autocorrelation functions showing significant lags at 1 indicate that a first-order model is appropriate. Residual analysis following model fitting, showing minimal significant autocorrelations, suggests a good model fit.

We then apply the strategy outlined in Subsection 1.7.3 by fitting a trivariate INAR(1) model with independent Poisson innovations to the historical syndromic surveillance data. To account for covariates commonly used in infectious disease modelling, we express the expectation of the innovation series as a function of the available covariate information:

$$E\left(\varepsilon_{it}\right) = \exp\left(\boldsymbol{z}_t^T \boldsymbol{\beta}_i\right), \quad i = 1, 2, 3, \tag{1.17}$$

where $\boldsymbol{z}_t$ is a vector of covariates and $\boldsymbol{\beta}_i$ is the associated parameter vector. The candidate covariates include a binary indicator for the day of the week (weekdays vs. weekends) and trigonometric terms to capture potential seasonality. Since the series appear stationary, we do not include a time trend. Each marginal series is thus modelled as

$$X_{it} = \sum_{j=1}^{3} \alpha_{ij} \circ X_{j,t-1} + \varepsilon_{it}, \quad i = 1, 2, 3, \tag{1.18}$$

where ε_{it} are independent Poisson random variables with mean

$$E\left(\varepsilon_{it}\right) = \exp\left\{\beta_{i0} + \beta_{i1}\ \text{Weekday}\ + \beta_{i2}\cos\left(\frac{2\pi t}{365}\right) + \beta_{i3}\sin\left(\frac{2\pi t}{365}\right)\right\}. \tag{1.19}$$

For comparison, we also employ a univariate surveillance approach by fitting three independent INAR(1) regression models with Poisson innovations. Covariate information is incorporated in the same way as in the multivariate model. For both approaches, one-step-ahead predictive distributions are used to construct $(1-\alpha)\%$ prediction intervals. The upper bounds of these intervals serve as thresholds for alarm detection. We assume a component-wise type I error rate of $\alpha = 0.01$ and apply a majority rule for overall alarm detection: an alarm is triggered if at least two out of the three series flag an alarm at the same time point.

All analyses were conducted in the R statistical software. The code used for model fitting, prediction, and alarm detection procedures is provided in Appendix C.

1.8 Organisation of the dissertation

The rest of the dissertation is organized as follows:

Chapter 2. In this chapter, we present the preliminaries needed to understand the content of this dissertation. Section 2.1 provides an overview of the Poisson distribution, including its mathematical formulation, properties, and applications. Section 2.2 introduces the binomial thinning operator, highlighting its properties and significance in integer-valued time

series models. Section 2.3 discusses three commonly used parameter estimation methods: maximum likelihood estimation (MLE), conditional least squares (CLS), and Yule–Walker (YW) estimation. Finally, Section 2.4 outlines the steps involved in conducting simulation studies and emphasises their importance in evaluating statistical models.

Chapter 3. In this chapter, we introduce the first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) process as a simplification of the first-order multivariate integer-valued autoregressive (MINAR(1)) process. Section 3.1 presents the MINAR(1) process as a multi- variate extension of the first-order integer-valued autoregressive (INAR(1)) process, followed by the formulation of the MPINAR(1) process in Section 3.2. Section 3.3 details the parameter estimation approach, and Section 3.4 outlines a two-step methodology for applying the MPINAR(1) model in public health surveillance.

Chapter 4. In this chapter, we evaluate the performance of the first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) model through a simulation study and a real-world data application. The goal is to demonstrate the model's theoretical robustness and practical utility in public health surveillance. Section 4.1 presents a simulation study designed to assess the performance of the MPINAR(1) model in comparison to independent univariate INAR(1) models, using synthetic data generated under controlled conditions. Section 4.2 applies the MPINAR(1) model to COVID-19 syndromic surveillance data, focusing on three statistically significantly correlated symptoms: fever, cough, and dyspnea.

Chapter 5. In this chapter, we discuss the results of the first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) model as a practical alternative to the more general multivariate integer-valued autoregressive (MINAR(1)) process for modelling multivariate count time series data in public health surveillance. Section 5.1 covers the formulation and key statistical properties of the MPINAR(1) model. Section 5.2 discusses a simulation study conducted to assess the model's performance under various outbreak scenarios, while Section 5.3 discusses the application of the model to real syndromic surveillance data.

Chapter 6. In this final chapter, we summarise the main findings of the dissertation and highlight their implications for both theoretical and practical applications. Based on these insights, we offer recommendations to guide future research and potential improvements. Section 6.1 presents the summary, and Section 6.2 outlines the recommendations.

Appendices. These appendices provide supplementary materials supporting the main content of the dissertation. Appendix A presents the data set utilized for the real data example discussed in Chapter 4, offering a detailed overview of the variables and data collection methods. Appendix B contains the R code employed for the simulation study outlined in Chapter 4, facilitating a deeper understanding of the computational processes used in the analysis. Finally, Appendix C includes the R code relevant to the real data example in Chapter 4, further illustrating the practical implementation of the methods discussed.

Chapter 2

Preliminaries

In this chapter, we present the preliminaries necessary for understanding the content of this dissertation. Section 2.1 provides an overview of the Poisson distribution, including its mathematical formulation, properties, and applications. Section 2.2 introduces the binomial thinning operator, highlighting its properties and significance in integer-valued time series models. Section 2.3 discusses three commonly used parameter estimation methods: maximum likelihood estimation (MLE), conditional least squares (CLS), and Yule–Walker (YW) estimation. Finally, Section 2.4 outlines the steps involved in conducting simulation studies and emphasises their importance in evaluating statistical models.

2.1 Poisson Distribution

The Poisson distribution is one of the most prominent discrete probability distributions in probability theory and statistics, widely used in modelling events that occur independently over a fixed interval of time or space. Its relevance spans diverse fields such as queuing theory, epidemiology, finance, and insurance. This section provides a detailed overview of the Poisson distribution, including its mathematical formulation, properties, and applications.

2.1.1 Definition and Some Examples

A discrete random variable X is said to follow a Poisson distribution with parameter $\lambda > 0$ if its probability mass function is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots. \tag{2.1}$$

Indeed equation (2.1) defines a probability mass function, since

$$\begin{aligned}
\sum_{x=0}^{\infty} P(X = x) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\
&= e^{-\lambda} e^{\lambda}, \quad \text{since } \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \text{ is the Taylor series expansion of } e^{\lambda} \\
&= 1.
\end{aligned}$$

Some examples of random variables that generally follow the Poisson probability distribution (that is, they obey Equation (2.1)) are as follows:

(i) The number of misprints on a page (or a group of pages) of a book.

(ii) The number of people in a community who survive to age 100.

(iii) The number of wrong telephone numbers that are dialed in a day.

(iv) The number of packages of dog biscuits sold in a particular store each day.

(v) The number of customers entering a post office on a given day.

(vi) The number of vacancies occurring during a year in the federal judicial system.

(vii) The number of α-particles discharged in a fixed period of time from some radioactive material.

2.1.2 Moments and Properties

We write $X \sim \text{POI}(\lambda)$ to indicate that the random variable X follows a Poisson distribution with parameter $\lambda > 0$. A key property of the Poisson distribution is that its expected value and variance are both equal to its rate parameter. This property is formally stated in the following proposition.

Proposition 2.1.1. *If $X \sim \text{POI}(\lambda)$, then the expected value and variance of X are given by*

$$E(X) = \lambda \quad \textit{and} \quad \text{Var}(X) = \lambda, \tag{2.2}$$

respectively. Thus, both the expected value and the variance of a Poisson random variable are equal to its parameter λ.

Proof. We first prove the expected value of X as follows:

$$\begin{aligned}
E(X) &= \sum_{x=0}^{\infty} xP(X = x) \\
&= \sum_{x=0}^{\infty} \frac{xe^{-\lambda}\lambda^x}{x!}, \quad \text{by equation (2.1)} \\
&= \sum_{x=1}^{\infty} \frac{xe^{-\lambda}\lambda^x}{x(x-1)!} \\
&= \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}, \quad \text{where } y = x - 1 \\
&= \lambda e^{-\lambda} e^{\lambda}, \quad \text{since } \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \text{ is the Taylor series expansion of } e^{\lambda} \\
&= \lambda,
\end{aligned}$$

as required.

Next, we prove the variance of X by first computing $E(X^2)$ as follows:

$$\begin{aligned}
E\big(X^2\big) &= \sum_{x=0}^{\infty} x^2 P(X=x) \\
&= \sum_{x=0}^{\infty} \frac{x^2 e^{-\lambda}\lambda^x}{x!}, \quad \text{by equation (2.1)} \\
&= \lambda \sum_{x=1}^{\infty} \frac{x e^{-\lambda}\lambda^{x-1}}{(x-1)!} \\
&= \lambda \sum_{y=0}^{\infty} \frac{(y+1) e^{-\lambda}\lambda^{y}}{y!}, \quad \text{where } y = x-1 \\
&= \lambda \left(\sum_{y=0}^{\infty} \frac{y e^{-\lambda}\lambda^{y}}{y!} + \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^{y}}{y!} \right) \\
&= \lambda(\lambda+1),
\end{aligned}$$

where the last equality follows since in the preceding equality the first sum equals $E(X) = \lambda$, and the second equals 1 (sum of all probabilities). Finally, the variance is

$$\begin{aligned}
\mathrm{Var}(X) &= E\left(X^2\right) - (E(X))^2 \\
&= \lambda(\lambda+1) - \lambda^2 \\
&= \lambda,
\end{aligned}$$

which is the required expression for the variance of X. □

Another important characteristic of a Poisson random variable X is its moment generating function, $M_X(t)$, defined for all real values of t by

$$M_X(t) = E\big(e^{tX}\big). \tag{2.3}$$

Proposition 2.1.2. *If $X \sim \mathrm{POI}(\lambda)$, then the moment generating function of X is given by*

$$M_X(t) = e^{\lambda(e^t - 1)}. \tag{2.4}$$

Proof. By the definition of the moment generating function,

$$\begin{aligned}
M_X(t) = E(e^{tX}) &= \sum_{x=0}^{\infty} e^{tx} P(X=x) \\
&= \sum_{x=0}^{\infty} \frac{e^{tx}\lambda^x e^{-\lambda}}{x!}, \quad \text{by equation (2.1)} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\
&= e^{-\lambda} e^{\lambda e^t}, \quad \text{since } \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \text{ is the Taylor series expansion of } e^{\lambda e^t} \\
&= e^{\lambda(e^t - 1)},
\end{aligned}$$

which is the required expression for the moment generating function of X. □

Remark 2.1.3. *The expected value and variance of a Poisson random variable can also be derived from its moment generating function.*

Proposition 2.1.4. *If* $X_1 \sim \text{POI}(\lambda_1)$ *and* $X_2 \sim \text{POI}(\lambda_2)$ *are independent, then*

$$X_1 + X_2 \sim \text{POI}(\lambda_1 + \lambda_2). \tag{2.5}$$

Proof. By the definition of the moment generating function,

$$\begin{aligned} M_{X_1+X_2}(t) &= E(e^{tX_1+tX_2}) \\ &= E(e^{tX_1}e^{tX_2}) \\ &= E(e^{tX_1})E(e^{tX_2}), \quad \text{since } X_1 \text{ and } X_2 \text{ are independent} \\ &= M_{X_1}(t)M_{X_2}(t) \\ &= e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} \\ &= e^{(\lambda_1+\lambda_2)(e^t-1)}. \end{aligned}$$

This is the moment generating function of a Poisson random variable with parameter $\lambda_1+\lambda_2$. Since the moment generating function uniquely determines the distribution of a random variable, it follows that $X_1 + X_2 \sim \text{POI}\,(\lambda_1 + \lambda_2)$. □

This result can be extended to n Independent Poisson Random Variables.

Proposition 2.1.5. *If* $X_1, X_2, \ldots, X_n$ *are independent random variables such that* $X_i \sim \text{POI}(\lambda_i)$ *for* $i = 1, 2, \ldots, n$*, then*

$$\sum_{i=1}^{n} X_i \sim \text{POI}\left(\sum_{i=1}^{n} \lambda_i\right). \tag{2.6}$$

Proof. Let $S_n = \sum_{i=1}^n X_i$. By the definition of the moment generating function,

$$\begin{aligned} M_{S_n}(t) = E\big(e^{tS_n}\big) &= E\Big(e^{\sum_{i=1}^n tX_i}\Big) \\ &= E\left(\prod_{i=1}^{n} e^{tX_i}\right) \\ &= \prod_{i=1}^{n} E\big(e^{tX_i}\big), \quad \text{since } X_1, X_2, \ldots, X_n \text{ are independent} \\ &= \prod_{i=1}^{n} M_{X_i}(t) \\ &= \prod_{i=1}^{n} e^{\lambda_i\left(e^t-1\right)} \\ &= e^{\sum_{i=1}^n \lambda_i\left(e^t-1\right)}. \end{aligned}$$

This is the moment generating function of a Poisson random variable with parameter $\sum_{i=1}^n \lambda_i$. Since the moment generating function uniquely determines the distribution of a random variable, it follows that $\sum_{i=1}^n X_i \sim \text{POI}\,(\sum_{i=1}^n \lambda_i)$.

□

Corollary 2.1.6. *If $X_1, X_2, \ldots, X_n$ are independent identically distributed (i.i.d.) random variables, each following a* $\mathrm{POI}(\lambda)$ *distribution, then*

$$\sum_{i=1}^{n} X_i \sim \mathrm{POI}(n\lambda). \tag{2.7}$$

Proof. It follows directly from the Proposition 2.1.5 that

$$\sum_{i=1}^{n} X_i \sim \mathrm{POI}\left(\sum_{i=1}^{n} \lambda\right) = \mathrm{POI}(n\lambda), \tag{2.8}$$

as required. This completes the proof of the corollary. □

2.1.3 Examples of Applications of the Poisson Distribution

The Poisson distribution is widely applied in various fields to model the occurrence of events that happen independently and at a constant average rate over a fixed interval of time, space, or other dimensions. Below are several examples illustrating its practical applications.

1. Customer Arrivals in a Queue

In a queueing system, the number of customers arriving at a service point, such as a bank or a call center, can be modelled using the Poisson distribution. For instance, if a bank observes an average arrival rate of 10 customers per hour, then the number of arrivals in a given hour, denoted by $N(1)$, follows a Poisson distribution with parameter $\lambda = 10$, that is, $N(1) \sim \mathrm{POI}(10)$. The probability of having exactly 5 arrivals in an hour is given by:

$$P(N(1) = 5) = \frac{e^{-10}(10)^5}{5!} \approx 0.0378, \text{ or approximately } 3.78\%.$$

2. Natural Disasters

The Poisson distribution is used to model the occurrence of rare natural disasters such as earthquakes in a specific region over time. For example, if a region experiences an average of 3 earthquakes per year and X is the number of earthquakes in a year, then $X \sim \mathrm{POI}(3)$. The probability of observing exactly 2 earthquakes in a year is:

$$P(X = 2) = \frac{3^2 e^{-3}}{2!} = \frac{9e^{-3}}{2} \approx 0.224, \text{ or } 22.4\%.$$

3. Traffic Accidents

The number of traffic accidents at a specific intersection over a week can be modelled using the Poisson distribution. Suppose the average number of accidents is 5 per week and X is the number of accidents in a week, then $X \sim \mathrm{POI}(5)$. The probability of exactly 3 accidents in a week is given by:

$$P(X = 3) = \frac{5^3 e^{-5}}{3!} = \frac{125e^{-5}}{6} \approx 0.140, \text{ or } 14\%.$$

4. Disease Outbreaks

The Poisson distribution is frequently used to model the occurrence of rare disease cases in a population. For instance, if the average number of cases in a small community is 1 per month and X is the number of outbreaks in a month, then $X \sim \text{POI}(1)$, The probability of no cases reported in a month is:

$$P(X = 0) = \frac{1^0 e^{-1}}{0!} = e^{-1} \approx 0.367, \text{ or } 36.7\%.$$

5. Defects in Manufacturing

In manufacturing, the Poisson distribution can model the number of defects found in a batch of items. For example, if a factory produces 100 items per hour and the average defect rate is 2 per hour and X is the number of defects per hour, then $X \sim \text{POI}(2)$. The probability of finding no defects in an hour is:

$$P(X = 0) = \frac{2^0 e^{-2}}{0!} = e^{-2} \approx 0.135, \text{ or } 13.5\%.$$

6. Radiation Emissions

In physics, the Poisson distribution is often used to model the count of radioactive particles emitted from a source within a fixed period. For example, if an average of 8 particles are emitted per second, then $N(1)$, the distribution of particle counts over any one-second interval, follows a Poisson distribution with parameter $\lambda = 8$, that is, $N(1) \sim \text{POI}(8)$.

2.2 Binomial Thinning Operator

Applying Autoregressive models for count time series is not as straightforward as in the continuous case. Suppose we have the first order autoregressive (AR(1)) recursion

$$X_t = \alpha \cdot X_{t-1} + \varepsilon_t \tag{2.9}$$

This recursion cannot be applied to the count process framework even if the innovations $\varepsilon_t \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$. The reason for this is due to the multiplication problem, that is, the multiplication $\alpha \cdot X_{t-1}$ does not guarantee to preserve the discrete range. To tackle the multiplication problem, McKenzie [21] proposed the use of different mechanisms for reducing X_{t-1}. One of these proposals is called binomial thinning (binomial subsampling).

Definition 2.2.1. *(Weiß [40]) Let X be a random variable with values in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ and $\alpha \in (0, 1)$. Furthermore, let $Y_1, \dots, Y_X$ be independent and identically distributed binary random variables that are independent of X such that $Y_1, \dots, Y_X \sim \text{Bernoulli}(\alpha) = \text{Bin}(1, \alpha)$. The random variable*

$$\alpha \circ X = \sum_{i=1}^{X} Y_i \tag{2.10}$$

is said to arise from X by binomial thinning. Here, "o" is referred to as the binomial thinning operator. At the boundaries, where $\alpha = 0$ or $\alpha = 1$, define $0 \circ X = 0$ and $1 \circ X = X$. It is then clear that $\alpha \circ X$ can only have integer values between 0 and X.

Note that if the value of X is known in Definition 2.2.1, then

$$\alpha \circ X \mid X = Y_1 + Y_2 + \ldots + Y_X \sim \text{Bin}(\underbrace{1 + 1 + \ldots + 1}_{X}, \alpha) = \text{Bin}(X, \alpha) \tag{2.11}$$

Moreover, we see that

$$\begin{aligned} E(\alpha \circ X) &= E\Big(E(\alpha \circ X \mid X)\Big), \quad \text{by the law of total expectation} \\ &= E(\alpha \cdot X), \quad \text{by equation (2.11)} \end{aligned} \tag{2.12}$$

This shows that the mean of $\alpha \cdot X$ and $\alpha \circ X$ are the same. This further motivates substituting $\alpha \cdot X_{t-1}$ with $\alpha \circ X_{t-1}$ in the AR(1) recursion, since the mean does not change based on this substitution. However, the variance of $\alpha \cdot X$ and $\alpha \circ X$ are not the same as shown below:

$$\begin{aligned} \text{Var}(\alpha \circ X) &= \text{Var}\Big(E(\alpha \circ X \mid X)\Big) + E\Big(\text{Var}(\alpha \circ X \mid X)\Big), \quad \text{by the law of total variance} \\ &= \text{Var}(\alpha \cdot X) + E\Big(\alpha(1-\alpha) \cdot X\Big) \\ &\neq \text{Var}(\alpha \cdot X), \quad \text{by equation (2.11)} \end{aligned} \tag{2.13}$$

In the following proposition we present some properties of the binomial thinning operator which will be useful in Chapter 3.

Proposition 2.2.2. *Let X, Y and Z be random variables with values in $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, and let $\alpha, \beta, \gamma \in [0, 1]$ be constants. Then the following binomial thinning operations hold:*

(1) $0 \circ X = 0$

(2) $1 \circ X = X$

(3) $\alpha \circ (\beta \circ X) \overset{\text{d}}{=} (\alpha\beta) \circ X$, *that is, the random variables $\alpha \circ (\beta \circ X)$ and $(\alpha\beta) \circ X$ are distributed identically.*

(4) $E(\alpha \circ X) = \alpha E(X)$

(5) $E(\alpha \circ X)^2 = \alpha^2 E\left(X^2\right) + \alpha(1-\alpha)E(X)$

(6) $E(\alpha \circ X)^3 = \alpha^3 E\left(X^3\right) + 3\alpha^2(1-\alpha)E\left(X^2\right) + \alpha(1-\alpha)(1-2\alpha)E(X)$

(7) $E((\alpha \circ X)Y) = \alpha E(XY)$

(8) $\text{Cov}(\alpha \circ X, X) = \alpha \,\text{Var}(X)$

(9) $E\left((\alpha \circ X)^2 Y\right) = \alpha^2 E\left(X^2 Y\right) + a(1-\alpha)E(XY)$

(10) $E[(\alpha \circ X)(\beta \circ Y)] = \alpha\beta E(X)E(Y)$, *if* X *and* Y *are independent.*

Proof. Parts (1) and (2) follows directly from Definition 2.2.1. Furthermore, we have

$$\begin{aligned}\alpha \circ (\beta \circ X) &\sim \text{Bin}(\beta \circ X, \alpha) \\ &= \text{Bin}(X, \alpha\beta), \quad \text{since } \beta \circ X \sim \text{Bin}(X, \beta) \text{ by Definition 2.2.1,}\end{aligned}$$

proving part (3). Next, by equation 2.12, we have

$$E(\alpha \circ X) = E(\alpha \dot{X}) = \alpha E(X),$$

proving part (4). For part (5), we have

$$\begin{aligned}E(\alpha \circ X)^2 &= \text{Var}(\alpha \circ X) + \Big(E(\alpha \circ X)\Big)^2 \\ &= \text{Var}(\alpha X) + E\Big(\alpha(1-\alpha)X\Big) + \Big(\alpha E(X)\Big)^2, \quad \text{by equation 2.13 and part (4)} \\ &= \alpha^2 \,\text{Var}(X) + \alpha(1-\alpha)E\Big(X\Big) + \alpha^2 E(X)^2 \\ &= \alpha^2\Big(E(X^2) - \big(E(X)\big)^2\Big) + \alpha(1-\alpha)E\Big(X\Big) + \alpha^2 E(X)^2 \\ &= \alpha^2 E(X^2) - \alpha^2 E(X)^2 + \alpha(1-\alpha)E\Big(X\Big) + \alpha^2 E(X)^2 \\ &= \alpha^2 E(X^2) + \alpha(1-\alpha)E(X),\end{aligned}$$

proving part (5). For part (6), we have

$$\begin{aligned}E(\alpha \circ X)^3 &= E\Big[E\big((\alpha \circ X)^3 \mid X\big)\Big], \quad \text{by the law of total expectation} \\ &= E\Big[\alpha X\Big(X^2\alpha^2 + 3X(X-1)\alpha(1-\alpha) + (1-\alpha)(1-2\alpha)\Big)\Big] \\ &= \alpha^3 E\left(X^3\right) + 3\alpha^2(1-\alpha)E\left(X^2\right) + \alpha(1-\alpha)(1-2\alpha)E(X),\end{aligned}$$

where the second equality follows from equation (2.11) that $\alpha \circ X \mid X \sim \text{Bin}(X, \alpha)$ and also from the third moment of a binomial random variable. This proves part (6). Next,

$$\begin{aligned}E\big((\alpha \circ X)Y\big) &= \sum_{i=1}^{X} E(Z_i Y), \quad \text{since } \alpha \circ X = \sum_{i=1}^{X} Z_i \\ &= \sum_{i=1}^{X} E(Z_i)E(Y), \quad \text{since } Z_i \text{ and } Y \text{ are independent} \\ &= \sum_{i=1}^{X} \alpha E(Y), \quad \text{since } Z_i \sim \text{Bernoulli}(\alpha) \\ &= \alpha E(XY),\end{aligned}$$

proving part (7). For part (8), we have

$$\begin{aligned}
\operatorname{Cov}(\alpha \circ X, X) &= E\big((\alpha \circ X)X\big) - E(\alpha \circ X)E(X), \quad \text{by the definition of covariance} \\
&= \alpha E(X^2) - \alpha E(X)E(X), \quad \text{by parts (3) and (7)} \\
&= \alpha E(X^2) - \alpha\big(E(X)\big)^2 \\
&= \alpha \operatorname{Var}(X),
\end{aligned}$$

proving part (8). For part (9), we have

$$\begin{aligned}
E\left((\alpha \circ X)^2 Y\right) &= E\Big(\big(\alpha^2 X^2 + a(1-\alpha)X\big)Y\Big), \quad \text{by parts (5) and (7)} \\
&= E\Big(\alpha^2 X^2 Y + a(1-\alpha)XY\Big) \\
&= \alpha^2 E\left(X^2 Y\right) + a(1-\alpha)E(XY)
\end{aligned}$$

proving part (9). Finally, we have

$$\begin{aligned}
E\big[(\alpha \circ X)(\beta \circ Y)\big] &= \alpha E\big[X(\beta \circ Y)\big], \quad \text{by parts (4) and (7)} \\
&= \alpha\beta E(XY), \quad \text{by parts (4) and (7) again} \\
&= \alpha\beta E(X)E(Y), \quad \text{since } X \text{ and } Y \text{ are independent,}
\end{aligned}$$

proving part (10). □

We now define the integer-valued AR(1) process using the binomial thinning operator "∘".

Definition 2.2.3. *The first order integer-valued autoregressive (INAR(1)) model is given by*

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t, \quad t = 1, 2, 3, \ldots, \tag{2.14}$$

where $\{\varepsilon_t\}$ is an innovation process consisting of uncorrelated non-negative integer-valued random variables with finite mean and variance.

We now interpret binomial thinning, and for this, we present the example given by Weiß [40]. Suppose first that X_{t-1} represents some population at time $t-1$. At the next time step t, the population may shrink due to individuals dying. If we can assume that each individual survives independently of each other with the probability α of a individual surviving from time $t-1$ to t, then the population at time t can be given as

$$X_t = \alpha \circ X_{t-1}, \tag{2.15}$$

where $\alpha \circ X_{t-1}$ is the number of survivors from $t-1$. This can be represented as follows:

$$\underbrace{X_t}_{\text{Population at time } t} = \underbrace{\alpha \circ X_{t-1}}_{\text{Survivors from time } t-1} + \underbrace{\varepsilon_t}_{\text{Immigration}} \tag{2.16}$$

The conditional mean and variance are given by:

$$\begin{aligned}
E\left(X_t \mid X_{t-1}\right) &= \alpha \cdot X_{t-1} + \mu_\varepsilon, \\
\operatorname{Var}\left(X_t \mid X_{t-1}\right) &= \alpha(1-\alpha) \cdot X_{t-1} + \sigma_\varepsilon^2,
\end{aligned} \tag{2.17}$$

which are both linear functions of X_{t-1}, and $\mu_\varepsilon = E(\varepsilon_t)$ and $\sigma_\varepsilon^2 = \operatorname{Var}(\varepsilon_t)$.

2.3 Parameter Estimation in Time Series Analysis

This section explores three widely used methods for parameter estimation in time series analysis: maximum likelihood (ML) estimation, conditional least squares (CLS) and Yule–Walker (YW) estimation, estimation. Each method offers unique advantages and is tailored to specific types of time series models.

2.3.1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a statistical method for estimating parameters by maximizing the likelihood function, which quantifies the probability of observing the given data under different parameter values. Let $X_1, \ldots, X_n$ be random variables with a joint density function $f(x_1, x_2, \ldots, x_n \mid \theta)$. Given observed values $X_i = x_i$ $(i = 1, \ldots, n)$, the likelihood function is defined as

$$L(\theta) = f(x_1, x_2, \ldots, x_n \mid \theta). \tag{2.18}$$

If the $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables, the joint density simplifies to the product of the marginal densities:

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta) \tag{2.19}$$

To simplify computation, the natural logarithm of the likelihood, known as the log-likelihood, is often maximized instead of the likelihood itself. For an i.i.d. sample, the log-likelihood is:

$$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta). \tag{2.20}$$

Now consider the case where $X_1, \ldots, X_n$ are i.i.d. and follow a Poisson distribution and λ is the parameter to be estimated. The maximum likelihood estimate (MLE) of λ is obtained by maximizing the log-likelihood function:

$$\begin{aligned} \ell(\lambda) &= \sum_{i=1}^{n} \Big(X_i \log \lambda - \lambda - \log(X_i!) \Big) \\ &= \log \lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log(X_i!) \end{aligned} \tag{2.21}$$

Next, we differentiate the log-likelihood with respect to λ and set it equal to zero:

$$\ell'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} X_i - n = 0. \tag{2.22}$$

Solving for λ, the MLE is:

$$\hat{\lambda} = \bar{X}, \quad \text{where } \bar{X} \text{ is the sample mean}$$

MLE is widely regarded as a highly efficient and consistent method for parameter estimation in statistical models. It offers key advantages such as asymptotic normality, flexibility, and compatibility with likelihood ratio tests for hypothesis testing. However, it can be sensitive to small sample sizes and model misspecification, which may lead to biased estimates.

2.3.2 Conditional least squares Estimation

Conditional least squares (CLS) estimation is a widely used method for parameter estimation in time series models, particularly in autoregressive (AR) models, autoregressive moving average (ARMA) models, and integer-valued time series models such as INAR models. The CLS approach is derived from basic linear regression principles, where X_{t+1} is treated as the dependent variable (Y) in the regression model, and $X_t, \ldots, X_{t-p}$ are treated as the independent variables (X_i). This method leverages the conditional expectation of the dependent variable given its past values, ensuring efficient parameter estimation for models that rely on autoregressive relationships.

Consider an AR(p) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$, where X_t is the value of the time series at time $t, \phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the model, and ε_t is white noise with zero mean and constant variance. The conditional least square estimates are

$$\left(\hat{\theta}_1, \ldots, \hat{\theta}_p\right) = \underset{\phi_1,\ldots,\phi_p}{\operatorname{argmin}} \sum_{t=p+1}^{n} \left(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}\right)^2 \tag{2.23}$$

and

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-p} \sum_{t=p+1}^{n} \left(X_t - \hat{\phi}_1 X_{t-1} - \cdots - \hat{\phi}_p X_{t-p}\right)^2 \tag{2.24}$$

Notice there is a common term in the above two expressions. We call the term

$$S_C\left(\phi_1, \ldots, \phi_p\right) = \sum_{t=p+1}^{n} \left(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}\right)^2 \tag{2.25}$$

the conditional sum of squares function. This term represents the total squared error between the observed values and the model's predicted values over the time series. By minimizing S_C, the CLS method identifies the parameter values $\hat{\phi}_1, \ldots, \hat{\phi}_p$ that lead to the best fit for the data, in terms of the least squared residuals as

$$\hat{\theta}_1, \ldots, \hat{\theta}_p = \underset{\phi_1,\ldots,\phi_p}{\operatorname{argmin}} S_C\left(\phi_1, \ldots, \phi_p\right) \tag{2.26}$$

Once the parameters $\hat{\phi}_1, \ldots, \hat{\phi}_p$ are estimated, we can calculate the error variance $\hat{\sigma}_\varepsilon^2$, which measures the variance of the residuals:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-p} \cdot S_C\left(\hat{\phi}_1, \ldots, \hat{\phi}_p\right) \tag{2.27}$$

In summary, Conditional least squares (CLS) is a method used for estimating parameters in statistical models, where the objective is to minimize the sum of squared residuals conditional on the given data, typically applied when there is conditional dependence between variables, allowing for estimation based on conditional expectations or given information. This method is commonly used in time series models and other situations where conditional relationships are important for accurate estimation.

2.3.3 Yule-Walker Estimation

The Yule-Walker estimation method is used to estimate the parameters of autoregressive (AR) models by leveraging the relationship between the autocorrelation function and the model parameters by the use of Yule-Walker equations. The Yule-Walker equations are a set of linear equations that are used to estimate the coefficients of an autoregressive (AR) model. An AR model is a type of time series model that expresses a variable of interest as a linear combination of its own past values. The Yule-Walker equations provide a method for estimating the parameters of the AR model from the autocorrelation function of the time series data. An autoregressive model of order p, denoted as AR(p), can be written as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t \tag{2.28}$$

where X_t is the value of the time series at time $t, \phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the model, and ε_t is white noise with zero mean and constant variance. The Yule–Walker equations are derived from the autocorrelation function of the time series. The autocorrelation function, $\rho(k)$, measures the correlation between values of the time series at different time points, at lag k. For an AR(p) model, the Yule-Walker equations are given by:

$$\begin{aligned}
\rho(1) &= \phi_1\rho(0) + \phi_2\rho(1) + \ldots + \phi_p\rho(p-1) \\
\rho(2) &= \phi_1\rho(1) + \phi_2\rho(0) + \ldots + \phi_p\rho(p-2) \\
&\vdots \\
\rho(p) &= \phi_1\rho(p-1) + \phi_2\rho(p-2) + \ldots + \phi_p\rho(0)
\end{aligned} \tag{2.29}$$

For an AR(p) process, the Yule-Walker equations can be written in matrix form $R\phi = \boldsymbol{r}$. Where:

Vector of autocorrelations:

$$\boldsymbol{r} = [\rho(1)\rho(2)\ldots\rho(p)] \tag{2.30}$$

Vector of AR coefficients:

$$\phi = \begin{bmatrix} \phi_1 & \phi_2 \ldots \phi_p \end{bmatrix} \tag{2.31}$$

The autocorrelation matrix R is a Toeplitz matrix:

$$R = \begin{bmatrix}
1 & \rho(1) & \rho(2) & \ldots & \rho(p-1) \\
\rho(1) & 1 & \rho(1) & \ldots & \rho(p-2) \\
\rho(2) & \rho(1) & 1 & \ldots & \rho(p-3) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho(p-1) & \rho(p-2) & \rho(p-3) & \ldots & 1
\end{bmatrix} \tag{2.32}$$

To estimate the AR coefficients, one can solve the Yule-Walker equations using methods for solving linear equations, such as the Levinson-Durbin recursion, which is computationally

efficient for Toeplitz matrices. The solution provides estimates of the AR coefficients that are consistent and asymptotically efficient, meaning that as the size of the time series data grows, the estimates converge to the true values with minimal variance.

The Yule-Walker equations are a fundamental component of time series analysis, providing a method to estimate the parameters of autoregressive models. They leverage the autocorrelation structure of the data to produce parameter estimates, enabling the modelling and forecasting of time series. While useful for estimating α in the context of AR(1) models, Yule-Walker equations do not directly estimate λ and rely on stationary assumptions, which may limit their applicability in certain cases (Dunsmuir & Scott, [7]).

2.4 Simulation Studies to Evaluate Model Performance

Simulation studies are essential tools in statistical research for assessing the properties and performance of proposed models and methods under controlled and replicable conditions. Unlike real-world datasets, which often present unknown complexities and confounding factors, simulations allow researchers to systematically generate data from known distributions and model structures. This controlled environment enables an objective comparison of model behaviour across different scenarios.

2.4.1 Purpose and Importance

The primary purpose of a simulation study is to evaluate the finite-sample performance of a statistical method or model. Specifically, simulations help to:

- Assess the bias, variance, and mean squared error (MSE) of parameter estimates;
- Evaluate the coverage probability of confidence or prediction intervals;
- Investigate the power and type I error rate of hypothesis tests or detection procedures;
- Examine the model's robustness under model misspecification or variations in distributional assumptions;
- Compare the performance of competing models or estimation strategies.

Simulation studies are particularly valuable when theoretical analysis is intractable or when empirical validation using real-world data is limited.

2.4.2 General Steps in a Simulation Study

A typical simulation study follows these steps:

1. **Model Specification**: Define the data-generating process (DGP), including the statistical model, distributional assumptions, and true parameter values. This step should reflect the characteristics of the real-world setting that the model is intended to address.

2. **Data Generation**: Simulate multiple datasets (typically a large number, such as 1000 or more) from the DGP using a programming environment such as `R`, `Python`, or `MATLAB`. Each dataset should have the same sample size and structure as the data to which the model will be applied in practice.

3. **Model Fitting**: Fit the proposed model (and possibly alternative models) to each simulated dataset. This involves applying the estimation procedure(s) of interest and recording the resulting parameter estimates, predictions, or test statistics.

4. **Performance Evaluation**: For each simulation run, compute evaluation metrics such as bias, variance, root mean squared error, or empirical coverage. These metrics are then averaged across all simulated datasets to summarise the model's performance.

5. **Result Interpretation**: Summarise the findings using tables, graphs, and descriptive statistics. Interpret the results with respect to the research questions or modelling goals, and discuss under what conditions the model performs well or poorly.

2.4.3 Design Considerations

A well-designed simulation study should be:

- **Transparent**: All assumptions, parameter values, and procedures should be clearly documented;
- **Reproducible**: The study should be implemented using code that allows others to reproduce the results;
- **Comprehensive**: Include a range of scenarios, such as varying sample sizes, parameter values, or noise levels, to test model performance under diverse conditions.

2.4.4 Limitations

While simulation studies provide valuable insights, they are not without limitations. Results are often specific to the chosen DGP and parameter settings and may not generalise to all real-world situations. Moreover, simulation studies rely on the quality and realism of the scenarios considered.

2.4.5 Conclusion

Simulation studies serve as a cornerstone in validating statistical models, offering a practical means to explore their properties under various settings. When designed carefully, they provide essential evidence of model reliability, guiding both methodological development and real-world application.

Chapter 3

From MINAR(1) to MPINAR(1): Structure, Estimation, and a Two-Step Approach to Public Health Surveillance

In this chapter, we introduce the first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) process as a simplification of the first-order multivariate integer-valued autoregressive (MINAR(1)) process. Section 3.1 presents the MINAR(1) process as a multivariate extension of the first-order integer-valued autoregressive (INAR(1)) process, followed by the formulation of the MPINAR(1) process in Section 3.2. Section 3.3 details the parameter estimation approach, and Section 3.4 outlines a two-step methodology for applying the MPINAR(1) model in public health surveillance.

3.1 MINAR(1): A Multivariate Extension of INAR(1)

This study proposes a public health surveillance approach based on integer-valued autoregressive (INAR) processes. INAR processes were first introduced as discrete-valued analogues of the standard Gaussian autoregressive process by Al-Osh and Alzaid [1] and McKenzie [21]. The simplest integer-valued autoregressive process is of order one, abbreviated as INAR(1), and is defined as the non-negative process $\{X_t\}$ satisfying

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t, \quad t \in \mathbb{N}, \tag{3.1}$$

where $\alpha \in [0, 1]$ and $\{\varepsilon_t\}$ is a sequence of uncorrelated non-negative integer-valued random variables with finite mean μ and finite variance σ^2 (hereafter referred to as the innovations). The $\circ$ symbol denotes the binomial thinning operator, defined by

$$\alpha \circ X = \sum_{j=1}^{X} Y_j \tag{3.2}$$

where $\{Y_j\}$ is a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables, independent of X, such that

$$P(Y_j = 1) = \alpha \quad \text{and} \quad P(Y_j = 0) = 1 - \alpha. \tag{3.3}$$

We call $\{Y_j\}$ the counting sequence of $\alpha \circ X$. The binomial thinning operator serves as an analogue to scalar multiplication in standard Gaussian (real-valued) time series models while ensuring that only integer values occur. This operator thus preserves the integer nature of the INAR(1) process and introduces serial dependence through conditioning on X_{t-1}.

INAR processes are useful for modelling count time series in a univariate setting. However, many real-world applications, including public health surveillance, require modelling multiple interdependent count processes. This need has led to several extensions of INAR processes to the multivariate case. We consider a multivariate extension of the INAR(1) process, as proposed by Pedeli and Karlis [26, 27]. Before presenting this multivariate INAR(1) model, we first extend the binomial thinning operator to a matrix formulation.

Definition 3.1.1. *Let $\boldsymbol{A}$ be an $n \times n$ matrix with entries α_{ij} satisfying $0 \leq \alpha_{ij} \leq 1$ for $i, j = 1, \ldots, n$, and let $\boldsymbol{X}$ be a non-negative integer-valued n-dimensional random vector. The action of $\boldsymbol{A}\circ$ on $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, denoted by $\boldsymbol{A} \circ \boldsymbol{X}$, is an n-dimensional random vector with i-th component given by*

$$(\boldsymbol{A} \circ \boldsymbol{X})_i = \sum_{j=1}^{n} \alpha_{ij} \circ X_j, \quad i = 1, \ldots, n. \tag{3.4}$$

More explicitly, the matricial binomial thinning operator $\circ$ is defined by

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{pmatrix} \circ \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} \alpha_{1j} \circ X_j \\ \sum_{j=1}^{n} \alpha_{2j} \circ X_j \\ \vdots \\ \sum_{j=1}^{n} \alpha_{nj} \circ X_j \end{pmatrix}. \tag{3.5}$$

The univariate thinning operations $\alpha_{ij} \circ X_j$ are based on counting sequences $\{Y_k^{(ij)}\}_{k \in \mathbb{N}}$, which consist of i.i.d. Bernoulli random variables. Consequently, these univariate operations are mutually independent. Furthermore, equation (3.5) demonstrates that $\boldsymbol{A}\circ$ functions similarly to matrix multiplication but applies binomial thinning element-wise instead of standard arithmetic operations.

The following properties of the matricial binomial thinning operator are useful for deriving the moments of the MINAR(1) process.

Lemma 3.1.2. *Let $\boldsymbol{A} = (\alpha_{ij})$ and $\boldsymbol{B} = (\beta_{ij})$ be two $n \times n$ matrices, where $0 \leq \alpha_{ij}, \beta_{ij} \leq 1$ for all $i, j = 1, \ldots, n$. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two non-negative integer-valued n-dimensional random vectors. The matricial binomial thinning operator $\circ$ satisfies the following properties:*

(i) ***Linearity of Expectation:***

$$E(\boldsymbol{A} \circ \boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}). \tag{3.6}$$

(ii) ***Expectation of the Cross Product:***

$$E\big[(\boldsymbol{A} \circ \boldsymbol{X})(\boldsymbol{B} \circ \boldsymbol{Y})^T\big] = \boldsymbol{A}E(\boldsymbol{X}\boldsymbol{Y}^T)\boldsymbol{B}^T, \tag{3.7}$$

provided that all the counting sequences of $\boldsymbol{A} \circ \boldsymbol{X}$ and $\boldsymbol{B} \circ \boldsymbol{Y}$ are independent.

(iii) ***Expectation of the Quadratic Form:***

$$E\big[(\boldsymbol{A} \circ \boldsymbol{X})(\boldsymbol{A} \circ \boldsymbol{X})^T\big] = \boldsymbol{A}E(\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{A}^T + \operatorname{diag}\big(\boldsymbol{C}E(\boldsymbol{X})\big), \tag{3.8}$$

where $\boldsymbol{C} = (c_{ij})$ *is the* $n \times n$ *variance matrix of Bernoulli random variables* $\alpha_{ij} \circ X_{j,t-1}$*, with elements for all* $i, j = 1, \ldots, n$ *given by*

$$c_{ij} = \alpha_{ij}(1 - \alpha_{ij}), \tag{3.9}$$

and $\operatorname{diag}(\boldsymbol{M})$ *denotes the diagonal matrix of* $\boldsymbol{M}$*.*

Proof. By Definition 3.1.1, $\boldsymbol{A} \circ \boldsymbol{X}$ and $\boldsymbol{B} \circ \boldsymbol{Y}$ are n-dimensional random vectors with i-th components given by $(\boldsymbol{A} \circ \boldsymbol{X})_i = \sum_{j=1}^{n} \alpha_{ij} \circ X_j$ and $(\boldsymbol{B} \circ \boldsymbol{Y})_i = \sum_{j=1}^{n} \beta_{ij} \circ Y_j$. Therefore, parts (i) and (ii) follow directly from the univariate relations given in Proposition 2.2.2 by

$$E(\alpha \circ X) = \alpha E(X) \quad \text{and} \quad E[(\alpha \circ X)(\beta \circ Y)] = \alpha\beta E(X)E(Y), \tag{3.10}$$

assuming the counting sequences of $\alpha \circ X$ and $\beta \circ Y$ are independent. Part (iii) is proved for each component separately, i.e., for $k, l = 1, \ldots, n$, we consider

$$\Big(E\big[(\boldsymbol{A} \circ \boldsymbol{X})(\boldsymbol{A} \circ \boldsymbol{X})^T\big]\Big)_{k,l} = \sum_{i,j=1}^{n} E\big[(a_{ki} \circ X_i)(a_{lj} \circ X_j)\big]. \tag{3.11}$$

If $k \neq l$ or $i \neq j$, then $a_{ki} \circ X_i$ and $a_{lj} \circ X_j$ are conditionally independent binomial random variables given X_i and X_j, and

$$\begin{aligned} E\big[(a_{ki} \circ X_i)(a_{lj} \circ X_j)\big] &= E\big[E\big((a_{ki} \circ X_i)(a_{lj} \circ X_j) \mid X_i, X_j\big)\big], \quad \text{law of total expectation} \\ &= E\big[(a_{ki}X_i)(a_{lj}X_j)\big] \\ &= a_{ki}a_{lj}E(X_iX_j). \end{aligned} \tag{3.12}$$

But if $k = l$ and $i = j$, we have

$$\begin{aligned} E(a_{ki} \circ X_i)^2 &= E\big[E\big((a_{ki} \circ X_i)^2 \mid X_i\big)\big], \quad \text{by the law of total expectation} \\ &= E\big[a_{ki}^2X_i^2 + a_{ki}(1 - a_{ki})X_i\big], \quad \text{since } a_{ki} \circ X_i \sim \operatorname{Bin}(X_i, a_{ki}) \\ &= a_{ki}^2E(X_i^2) + a_{ki}(1 - a_{ki})E(X_i) \end{aligned} \tag{3.13}$$

Using the Kronecker delta,

$$\delta_{kl} = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l, \end{cases} \tag{3.14}$$

equations (3.12) and (3.13) can be combined to yield the following relation:

$$\Big(E\big[(\boldsymbol{A} \circ \boldsymbol{X})(\boldsymbol{A} \circ \boldsymbol{X})^T\big]\Big)_{k,l} = \sum_{i,j=1}^{n} a_{ki}a_{lj}E(X_iX_j) + \delta_{kl}\sum_{i=1}^{n} a_{ki}(1 - a_{ki})E(X_i). \tag{3.15}$$

From this relation, part (iii) follows immediately. □

The matricial binomial thinning operator provides a strong foundation for modelling multivariate count time series. In particular, it enables the extension of the univariate INAR(1) process to a multivariate setting while preserving key probabilistic properties. We now define the multivariate integer-valued autoregressive process of order one (MINAR(1)), which generalises the INAR(1) process to accommodate multiple interdependent count processes.

Definition 3.1.3. *Let $\boldsymbol{A}$ be an $n \times n$ matrix with entries α_{ij} satisfying $0 \leq \alpha_{ij} \leq 1$ for $i, j = 1, \ldots, n$, and let $\boldsymbol{X}$ be a non-negative integer-valued n-dimensional random vector. The first-order multivariate integer-valued autoregressive (MINAR(1)) process $\{\boldsymbol{X}_t\}$ is defined as:*

$$\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{3.16}$$

*where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})^T$ is an innovation vector whose components ε_{it} are **correlated** non-negative integer-valued random variables. The vector $\boldsymbol{\varepsilon}_t$ is independent of $\boldsymbol{A} \circ \boldsymbol{X}_{t-1}$, and has mean $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. For greater clarity, the MINAR(1) model is expressed as:*

$$\begin{pmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{nt} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{pmatrix} \circ \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \\ \vdots \\ X_{n,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{nt} \end{pmatrix}, \quad t \in \mathbb{Z}. \tag{3.17}$$

Thus, each component of the MINAR(1) process satisfies

$$X_{it} = \sum_{j=1}^{n} \alpha_{ij} \circ X_{j,t-1} + \varepsilon_{it}, \quad i = 1, \ldots, n,$$

where $\alpha_{ij} \circ X_{j,t-1}$ are mutually independent univariate binomial thinning operations.

To ensure the existence of a strictly stationary solution to the MINAR(1) process, we impose conditions on the matrix $\boldsymbol{A}$ and the innovation process $\boldsymbol{\varepsilon}_t$. The following proposition provides a sufficient condition for strict stationarity, as established by Franke and Rao [8].

Proposition 3.1.4. *The non-negative integer-valued process $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ is the unique strictly stationary solution of* (3.16) *if the largest eigenvalue of $\boldsymbol{A}$ is less than 1 and $E\|\boldsymbol{\varepsilon}_t\| < \infty$.*

We now derive the mean vector and variance-covariance matrix of the MINAR(1) process, using the properties of the matricial binomial thinning operator established in Lemma 3.1.2.

Proposition 3.1.5. *Let $\boldsymbol{X}_t$ be a strictly stationary MINAR(1) process. Then, the mean vector and variance-covariance matrix of $\boldsymbol{X}_t$, defined respectively as*

$$\boldsymbol{\mu} = E(\boldsymbol{X}_t) \quad \text{and} \quad \boldsymbol{\gamma}(h) = E\left[(\boldsymbol{X}_{t+h} - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})^T\right],$$

satisfy the following equations:

$$\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{A})^{-1} \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}, \tag{3.18}$$

$$\boldsymbol{\gamma}(h) = \begin{cases} \boldsymbol{A}\boldsymbol{\gamma}(0)\boldsymbol{A}^T + \operatorname{diag}(\boldsymbol{B}\boldsymbol{\mu}) + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}, & h = 0, \\ \boldsymbol{A}^h \boldsymbol{\gamma}(0), & h \geq 1, \end{cases} \tag{3.19}$$

where $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{B} = (\beta_{ij})$ is the $n \times n$ variance-covariance matrix of Bernoulli random variables $\alpha_{ij} \circ X_{j,t-1}$, with elements given by

$$\beta_{ij} = \alpha_{ij}(1 - \alpha_{ij}), \tag{3.20}$$

for all $i, j = 1, \ldots, n$.

Proof. The MINAR(1) process $\boldsymbol{X}_t$ is defined by equation 3.16 as

$$\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}. \tag{3.21}$$

Taking expectations on both sides of this equation, we obtain

$$\begin{aligned}
E(\boldsymbol{X}_t) &= E\left(\boldsymbol{A} \circ \boldsymbol{X}_{t-1}\right) + E\left(\boldsymbol{\varepsilon}_t\right) \\
E(\boldsymbol{X}_t) &= \boldsymbol{A}E\left(\boldsymbol{X}_{t-1}\right) + E\left(\boldsymbol{\varepsilon}_t\right), \quad \text{by Lemma 3.1.2 (i)} \\
E(\boldsymbol{X}_t) &= \boldsymbol{A}E\left(\boldsymbol{X}_t\right) + E\left(\boldsymbol{\varepsilon}_t\right), \quad \text{by stationarity of the process} \\
\boldsymbol{\mu} &= \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}, \quad \text{by definition of } \boldsymbol{\mu} \\
\boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\mu} &= \boldsymbol{\mu}_{\boldsymbol{\varepsilon}} \\
(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{\mu} &= \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}
\end{aligned}$$

Since $\boldsymbol{I}$ is the identity matrix and all the eigenvalues of $\boldsymbol{A}$ lie inside the unit circle, the matrix $\boldsymbol{I} - \boldsymbol{A}$ is invertible, and thus, we can multiply through by its inverse to obtain

$$\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}.$$

To prove the variance-covariance matrix equation (3.19), we first observe that

$$\begin{aligned}
\boldsymbol{\gamma}(h) &= E\big[(\boldsymbol{X}_{t+h} - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})^T\big] \\
&= E\big[(\boldsymbol{A} \circ \boldsymbol{X}_{t+h-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\mu})(\boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\mu})^T\big], \quad \text{by equation (3.16)}
\end{aligned} \tag{3.22}$$

For the case $h = 0$, we obtain

$$\begin{aligned}
\boldsymbol{\gamma}(0) &= E\big[(\boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\mu})(\boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\mu})^T\big] \\
&= E\big[(\boldsymbol{A} \circ \boldsymbol{X}_{t-1})(\boldsymbol{A} \circ \boldsymbol{X}_{t-1})^T\big] + E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^T) + \text{cross terms} \\
&= \boldsymbol{A}\boldsymbol{\gamma}(0)\boldsymbol{A}^T + \operatorname{diag}(\boldsymbol{B}\boldsymbol{\mu}) + E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^T) + \text{cross terms}, \quad \text{by Lemma 3.1.2 (iii)} \\
&= \boldsymbol{A}\boldsymbol{\gamma}(0)\boldsymbol{A}^T + \operatorname{diag}(\boldsymbol{B}\boldsymbol{\mu}) + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}},
\end{aligned} \tag{3.23}$$

where the cross terms vanish since $\boldsymbol{\varepsilon}_t$ is independent of $\boldsymbol{A} \circ \boldsymbol{X}_{t-1}$. For $h \geq 1$, we have

$$\begin{aligned}
\boldsymbol{\gamma}(h) &= E\big[(\boldsymbol{X}_{t+h} - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})^T\big] \\
&= E\big[(\boldsymbol{A} \circ \boldsymbol{X}_{t+h-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})^T\big] \\
&= \boldsymbol{A}E[(\boldsymbol{X}_{t+h-1} - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})^T], \quad \text{by Lemma 3.1.2 (ii)} \\
&= \boldsymbol{A}\boldsymbol{\gamma}(h-1)
\end{aligned} \tag{3.24}$$

We now iterate this recurrence relation as follows:

$$\begin{aligned}
\boldsymbol{\gamma}(h) &= \boldsymbol{A}^1\boldsymbol{\gamma}(h-1) \\
&= \boldsymbol{A}^2\boldsymbol{\gamma}(h-2) \\
&= \boldsymbol{A}^3\boldsymbol{\gamma}(h-3) \\
&\vdots \\
&= \boldsymbol{A}^h\boldsymbol{\gamma}(0),
\end{aligned} \tag{3.25}$$

which is the required variance-covariance matrix relation for $h \geq 1$. □

We conclude this section with a discussion on parameter estimation for the first-order multivariate integer-valued autoregressive (MINAR(1)) process, defined as:

$$\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{3.26}$$

where $\boldsymbol{A}$ is a thinning matrix, $\boldsymbol{X}_t$ is the observed multivariate count process, and $\boldsymbol{\varepsilon}_t$ is an innovation vector whose components ε_{it} are **correlated** non-negative integer-valued random variables. The vector $\boldsymbol{\varepsilon}_t$ is independent of $\boldsymbol{A} \circ \boldsymbol{X}_{t-1}$, and has mean $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$.

The unknown parameter vector $\boldsymbol{\theta}$, which includes the elements of $\boldsymbol{A}$, $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, can be estimated using conditional maximum likelihood estimation (see Pedeli and Karlis [27]). The likelihood function for the MINAR(1) process is derived from the conditional distribution of $\boldsymbol{X}_t$ given $\boldsymbol{X}_{t-1}$. For a time series of length T, the likelihood function is given by:

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{t=2}^{T} f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right), \tag{3.27}$$

where $f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right)$ is the conditional probability mass function (PMF) of $\boldsymbol{X}_t$ given $\boldsymbol{X}_{t-1}$, and $\boldsymbol{\theta}$ is the vector of unknown parameters. The conditional PMF $f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right)$ involve convolutions of n sums of binomial distributions, given by

$$f_i\left(x_i \mid \boldsymbol{x}_{t-1}\right) = P\left(X_{it} = x_i \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}\right), \quad i = 1, \ldots, n, \tag{3.28}$$

and a distribution of the form

$$g\left(k_1, \ldots, k_n\right) = P\left(\varepsilon_{1t} = k_1, \ldots, \varepsilon_{nt} = k_n\right), \tag{3.29}$$

which corresponds to the joint distribution of the innovation process $\{\boldsymbol{\varepsilon}_t\}$. Consequently, the conditional PMF $f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right)$ can be expressed as the multiple sum

$$f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right) = \sum_{k_1=0}^{m_1} \cdots \sum_{k_n=0}^{m_n} f_1\left(x_{1t} - k_1 \mid \boldsymbol{x}_{t-1}\right) \cdots f_n\left(x_{nt} - k_n \mid \boldsymbol{x}_{t-1}\right) g\left(k_1, \ldots, k_n\right), \tag{3.30}$$

where $m_i = \min\left(x_{it}, x_{i,t-1}\right)$ for $i = 1, \ldots, n$.

The estimation of $\boldsymbol{\theta}$ involves maximizing the log-likelihood function:

$$l(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{t=2}^{T} \log f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right). \tag{3.31}$$

This is typically done using numerical optimization techniques, such as the Newton-Raphson method or gradient-based algorithms. However, when assuming a cross-correlated innovation process, the complexity of the MINAR(1) model increases significantly with dimensionality, making numerical maximization computationally demanding (see Pedeli and Karlis [27]). To mitigate this challenge, in the next section, we introduce a simplification of the MINAR(1) model by assuming that the innovations ε_{it} are uncorrelated and follow a Poisson distribution.

3.2 The MPINAR(1) Model: A Simplified MINAR(1) Model for Public Health Surveillance

Traditionally, statistical models for public health surveillance data aim to effectively capture the endemic and epidemic dynamics of disease risk. In principle, the endemic component explains a baseline rate of cases with a stable temporal pattern and is identified with the innovations part of the model. More specifically, it describes the risk of new events as a function of external factors that are independent of the history of the epidemic process. Examples of such external factors include seasonality, socio-demographic characteristics, population density, and vaccination coverage. The epidemic component, on the other hand, introduces infectiousness, that is, an explicit dependence between events. Consequently, the epidemic component is driven by past observations and corresponds to the autoregressive part of the model (see Meyer [20]).

This additive decomposition of disease risk is well embodied in the MINAR(1) model (3.16). However, the inclusion of correlated innovations increases the computational complexity of parameter estimation and requires strong distributional assumptions about the nature of the correlations, which may not always align with real-world data. To mitigate these computational challenges, we introduce a simplified version of the MINAR(1) model (3.16). Specifically, we propose the first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) model, which reduces model complexity by assuming that the innovations $\boldsymbol{\varepsilon}_t$ (the endemic components) are independent (uncorrelated) and follow a Poisson distribution. The choice of the Poisson distribution is motivated by its ability to reflect the integer-valued nature of public health data while being the simplest discrete distribution to work with. The resulting model admits a realistic epidemiological interpretation and is extremely advantageous in terms of practical implementation since the distribution of the innovations becomes a product of univariate mass functions:

$$P\left(\varepsilon_{1t} = k_1, \ldots, \varepsilon_{nt} = k_n\right) = \prod_{i=1}^{n} P\left(\varepsilon_{it} = k_i\right). \tag{3.32}$$

The mean vector and variance-covariance matrix of the simplified MPINAR(1) process $\boldsymbol{X}_t$ remain consistent with equations (3.18) and (3.19), where $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is now a diagonal matrix.

Definition 3.2.1. *Let $\boldsymbol{A} = (\alpha_{ij})$ be an $n \times n$ matrix where $0 \leq \alpha_{ij} \leq 1$ for all $i, j = 1, \ldots, n$, and let $\boldsymbol{X}$ be a non-negative integer-valued n-dimensional random vector. The first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) process $\{\boldsymbol{X}_t\}$ is defined as*

$$\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{3.33}$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})^T$ is an innovation vector whose components ε_{it} are independent (hence uncorrelated) Poisson-distributed random variables with parameters λ_i for $i = 1, \ldots, n$. The innovation vector $\boldsymbol{\varepsilon}_t$ is independent of $\boldsymbol{A} \circ \boldsymbol{X}_{t-1}$, and has mean $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}} = (\lambda_1, \ldots, \lambda_n)^T$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \operatorname{diag}(v_1\lambda_1, \ldots, v_n\lambda_n)$, where $v_i > 1$, allowing for overdispersion.

Proposition 3.2.2. *Let $\boldsymbol{X}_t$ be a strictly stationary MPINAR(1) process. Then, the mean vector and variance-covariance matrix of $\boldsymbol{X}_t$, defined respectively as*

$$\boldsymbol{\mu} = E\left(\boldsymbol{X}_t\right) \quad \textit{and} \quad \boldsymbol{\gamma}(h) = E\left[(\boldsymbol{X}_{t+h} - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})^T\right], \tag{3.34}$$

satisfy the following equations:

$$\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}, \tag{3.35}$$

$$\boldsymbol{\gamma}(h) = \begin{cases} \boldsymbol{A}\boldsymbol{\gamma}(0)\boldsymbol{A}^T + \operatorname{diag}(\boldsymbol{B}\boldsymbol{\mu}) + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}, & h = 0, \\ \boldsymbol{A}^h\boldsymbol{\gamma}(0), & h \geq 1, \end{cases} \tag{3.36}$$

where $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{B} = (\beta_{ij})$ is the $n \times n$ variance-covariance matrix of Bernoulli random variables $\alpha_{ij} \circ X_{j,t-1}$, with elements given by

$$\beta_{ij} = \alpha_{ij}(1 - \alpha_{ij}), \tag{3.37}$$

for all $i, j = 1, \ldots, n$, and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \operatorname{diag}(\lambda_1 \ldots, \lambda_n)$ is a diagonal matrix.

Proof. The result follows directly from Proposition 3.1.5, since setting $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ as a diagonal matrix accounts for the assumption of uncorrelated innovations. □

The independent innovations of the MPINAR(1) model imply that overdispersion, a typical characteristic of publlic health surveillance data, can be easily accommodated even under our simplest parametric assumption of Poisson innovations. More specifically, in line with Pedeli and Karlis [27], if $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})^T$ are independent Poisson random variables with parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$, then the joint distribution of $\boldsymbol{X}_t$ is given by the product of n generalized Poisson distributions (see Gurland [10], Kemp [17]) with parameters that are nonlinear combinations of $\boldsymbol{\lambda}$ and powers of $\boldsymbol{A}$.

In the bivariate case ($n = 2$), the expectation vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ has elements given by

$$\begin{aligned} \mu_1 &= \frac{(1 - \alpha_{22})\lambda_1 + \alpha_{12}\lambda_2}{(1 - \alpha_{11})(1 - \alpha_{22}) - \alpha_{12}\alpha_{21}}, \\ \mu_2 &= \frac{(1 - \alpha_{11})\lambda_2 + \alpha_{21}\lambda_1}{(1 - \alpha_{11})(1 - \alpha_{22}) - \alpha_{12}\alpha_{21}}, \end{aligned} \tag{3.38}$$

and the variance-covariance matrix

$$\boldsymbol{\gamma}(0) = \begin{pmatrix} \gamma_{11}(0) & \gamma_{12}(0) \\ \gamma_{12}(0) & \gamma_{22}(0) \end{pmatrix}$$

has elements given by

$$\begin{aligned} \gamma_{11}(0) &= \frac{\alpha_{12}^2\gamma_{22}(0) + 2\alpha_{11}\alpha_{12}\gamma_{12}(0) + \alpha_{11}(1 - \alpha_{11})\mu_1 + \alpha_{12}(1 - \alpha_{12})\mu_2 + v_1\lambda_1}{1 - \alpha_{11}^2}, \\ \gamma_{22}(0) &= \frac{\alpha_{21}^2\gamma_{11}(0) + 2\alpha_{21}\alpha_{22}\gamma_{12}(0) + \alpha_{21}(1 - \alpha_{21})\mu_1 + \alpha_{22}(1 - \alpha_{22})\mu_2 + v_2\lambda_2}{1 - \alpha_{22}^2}, \\ \gamma_{12}(0) &= \frac{\alpha_{11}\alpha_{21}\gamma_{11}(0) + \alpha_{12}\alpha_{22}\gamma_{22}(0)}{1 - \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}}. \end{aligned} \tag{3.39}$$

We formally present and prove these results in the following proposition.

Proposition 3.2.3. *Let $\boldsymbol{X}_t = (X_{1t}, X_{2t})^T$ be a strictly stationary bivariate process satisfying*

$$\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{3.40}$$

where $\boldsymbol{A}$ is a 2×2 thinning matrix given by

$$\boldsymbol{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}, \tag{3.41}$$

with $0 \leq \alpha_{ij} \leq 1$ for all $i, j = 1, 2$, and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t})^T$ is an innovation vector of i.i.d. random variables with mean vector $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}} = (\lambda_1, \lambda_2)^T$ and variance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \operatorname{diag}(v_1\lambda_1, v_2\lambda_2)$, where $v_i, \lambda_i > 0$ for all $i, j = 1, 2$. Then, the expectation vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ is given by

$$\begin{aligned} \mu_1 &= \frac{(1-\alpha_{22})\lambda_1 + \alpha_{12}\lambda_2}{(1-\alpha_{11})(1-\alpha_{22}) - \alpha_{12}\alpha_{21}}, \\ \mu_2 &= \frac{(1-\alpha_{11})\lambda_2 + \alpha_{21}\lambda_1}{(1-\alpha_{11})(1-\alpha_{22}) - \alpha_{12}\alpha_{21}}. \end{aligned} \tag{3.42}$$

Furthermore, the variance-covariance matrix $\boldsymbol{\gamma}(0)$, given by

$$\boldsymbol{\gamma}(0) = \begin{pmatrix} \gamma_{11}(0) & \gamma_{12}(0) \\ \gamma_{12}(0) & \gamma_{22}(0) \end{pmatrix},$$

has elements

$$\begin{aligned} \gamma_{11}(0) &= \frac{\alpha_{12}^2\gamma_{22}(0) + 2\alpha_{11}\alpha_{12}\gamma_{12}(0) + \alpha_{11}(1-\alpha_{11})\mu_1 + \alpha_{12}(1-\alpha_{12})\mu_2 + v_1\lambda_1}{1-\alpha_{11}^2}, \\ \gamma_{22}(0) &= \frac{\alpha_{21}^2\gamma_{11}(0) + 2\alpha_{21}\alpha_{22}\gamma_{12}(0) + \alpha_{21}(1-\alpha_{21})\mu_1 + \alpha_{22}(1-\alpha_{22})\mu_2 + v_2\lambda_2}{1-\alpha_{22}^2}, \\ \gamma_{12}(0) &= \frac{\alpha_{11}\alpha_{21}\gamma_{11}(0) + \alpha_{12}\alpha_{22}\gamma_{22}(0)}{1-\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21}}. \end{aligned} \tag{3.43}$$

Proof. For the expectation vector, Proposition 3.2.2 gives the expectation equation:

$$\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}. \tag{3.44}$$

Thus, to derive the elements of the expectation vector, we first compute $\boldsymbol{I} - \boldsymbol{A}$:

$$\boldsymbol{I} - \boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} = \begin{pmatrix} 1-\alpha_{11} & -\alpha_{12} \\ -\alpha_{21} & 1-\alpha_{22} \end{pmatrix}. \tag{3.45}$$

Since $\boldsymbol{X}_t$ strictly stationary, the matrix $\boldsymbol{I} - \boldsymbol{A}$ is invertible and its inverse given by

$$(\boldsymbol{I} - \boldsymbol{A})^{-1} = \frac{1}{(1-\alpha_{11})(1-\alpha_{22}) - \alpha_{12}\alpha_{21}} \begin{pmatrix} 1-\alpha_{22} & \alpha_{12} \\ \alpha_{21} & 1-\alpha_{11} \end{pmatrix}. \tag{3.46}$$

The elements of the expectation vector are then found as follows:

$$\begin{aligned} \boldsymbol{\mu} &= (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\mu}_{\boldsymbol{\varepsilon}} \\ &= \frac{1}{(1-\alpha_{11})(1-\alpha_{22}) - \alpha_{12}\alpha_{21}} \begin{pmatrix} 1-\alpha_{22} & \alpha_{12} \\ \alpha_{21} & 1-\alpha_{11} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \\ &= \frac{1}{(1-\alpha_{11})(1-\alpha_{22}) - \alpha_{12}\alpha_{21}} \begin{pmatrix} (1-\alpha_{22})\lambda_1 + \alpha_{12}\lambda_2 \\ \alpha_{21}\lambda_1 + (1-\alpha_{11})\lambda_2 \end{pmatrix} \end{aligned} \tag{3.47}$$

For the variance-covariance matrix, Proposition 3.2.2 gives the moment equation:

$$\boldsymbol{\gamma}(0) = \boldsymbol{A}\boldsymbol{\gamma}(0)\boldsymbol{A}^T + \text{diag}(\boldsymbol{B}\boldsymbol{\mu}) + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}, \tag{3.48}$$

where $\boldsymbol{B} = \big(\alpha_{ij}(1-\alpha_{ij})\big)$ is a 2×2 matrix. We expand this moment equation as follows:

$$\begin{aligned}
\begin{pmatrix} \gamma_{11}(0) & \gamma_{12}(0) \\ \gamma_{12}(0) & \gamma_{22}(0) \end{pmatrix} &= \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} \gamma_{11}(0) & \gamma_{12}(0) \\ \gamma_{12}(0) & \gamma_{22}(0) \end{pmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}^T \\
&\quad + \text{diag}\left[\begin{pmatrix} \alpha_{11}(1-\alpha_{11}) & \alpha_{12}(1-\alpha_{12}) \\ \alpha_{21}(1-\alpha_{21}) & \alpha_{22}(1-\alpha_{22}) \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right] + \begin{pmatrix} v_1\lambda_1 & 0 \\ 0 & v_2\lambda_2 \end{pmatrix} \\
&= \begin{pmatrix} \alpha_{11}\gamma_{11}(0) + \alpha_{12}\gamma_{12}(0) & \alpha_{11}\gamma_{12}(0) + \alpha_{12}\gamma_{22}(0) \\ \alpha_{21}\gamma_{11}(0) + \alpha_{22}\gamma_{12}(0) & \alpha_{21}\gamma_{12}(0) + \alpha_{22}\gamma_{22}(0) \end{pmatrix} \begin{pmatrix} \alpha_{11} & \alpha_{21} \\ \alpha_{12} & \alpha_{22} \end{pmatrix} \\
&\quad + \begin{pmatrix} \alpha_{11}(1-\alpha_{11})\mu_1 + \alpha_{12}(1-\alpha_{12})\mu_2 & 0 \\ 0 & \alpha_{21}(1-\alpha_{21})\mu_1 + \alpha_{22}(1-\alpha_{22})\mu_2 \end{pmatrix} + \begin{pmatrix} v_1\lambda_1 & 0 \\ 0 & v_2\lambda_2 \end{pmatrix} \\
&= \begin{pmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{pmatrix} + \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix},
\end{aligned}$$

where

$$\begin{aligned}
A_{11} &= \alpha_{11}^2\gamma_{11}(0) + 2\alpha_{11}\alpha_{12}\gamma_{12}(0) + \alpha_{12}^2\gamma_{22}(0), \\
A_{12} &= \alpha_{11}\alpha_{21}\gamma_{11}(0) + (\alpha_{11}\alpha_{22} + \alpha_{12}\alpha_{21})\gamma_{12}(0) + \alpha_{12}\alpha_{22}\gamma_{22}(0), \\
A_{22} &= \alpha_{21}^2\gamma_{11}(0) + 2\alpha_{21}\alpha_{22}\gamma_{12}(0) + \alpha_{22}^2\gamma_{22}(0), \\
B_{11} &= \alpha_{11}(1-\alpha_{11})\mu_1 + \alpha_{12}(1-\alpha_{12})\mu_2 + v_1\lambda_1, \\
B_{22} &= \alpha_{21}(1-\alpha_{21})\mu_1 + \alpha_{22}(1-\alpha_{22})\mu_2 + v_2\lambda_2.
\end{aligned}$$

The individual components are then obtained element-wise as follows:

$$\begin{aligned}
\gamma_{11}(0) &= \alpha_{11}^2\gamma_{11}(0) + 2\alpha_{11}\alpha_{12}\gamma_{12}(0) + \alpha_{12}^2\gamma_{22}(0) + \alpha_{11}(1-\alpha_{11})\mu_1 + \alpha_{12}(1-\alpha_{12})\mu_2 + v_1\lambda_1 \\
\big(1-\alpha_{11}^2\big)\gamma_{11}(0) &= \alpha_{12}^2\gamma_{22}(0) + 2\alpha_{11}\alpha_{12}\gamma_{12}(0) + \alpha_{11}(1-\alpha_{11})\mu_1 + \alpha_{12}(1-\alpha_{12})\mu_2 + v_1\lambda_1 \\
\gamma_{11}(0) &= \frac{\alpha_{12}^2\gamma_{22}(0) + 2\alpha_{11}\alpha_{12}\gamma_{12}(0) + \alpha_{11}(1-\alpha_{11})\mu_1 + \alpha_{12}(1-\alpha_{12})\mu_2 + v_1\lambda_1}{1-\alpha_{11}^2},
\end{aligned}$$

$$\begin{aligned}
\gamma_{22}(0) &= \alpha_{21}^2\gamma_{11}(0) + 2\alpha_{21}\alpha_{22}\gamma_{12}(0) + \alpha_{22}^2\gamma_{22}(0) + \alpha_{21}(1-\alpha_{21})\mu_1 + \alpha_{22}(1-\alpha_{22})\mu_2 + v_2\lambda_2 \\
\big(1-\alpha_{22}^2\big)\gamma_{22}(0) &= \alpha_{21}^2\gamma_{11}(0) + 2\alpha_{21}\alpha_{22}\gamma_{12}(0) + \alpha_{21}(1-\alpha_{21})\mu_1 + \alpha_{22}(1-\alpha_{22})\mu_2 + v_2\lambda_2 \\
\gamma_{22}(0) &= \frac{\alpha_{21}^2\gamma_{11}(0) + 2\alpha_{21}\alpha_{22}\gamma_{12}(0) + \alpha_{21}(1-\alpha_{21})\mu_1 + \alpha_{22}(1-\alpha_{22})\mu_2 + v_2\lambda_2}{1-\alpha_{22}^2},
\end{aligned}$$

$$\begin{aligned}
\gamma_{12}(0) &= \alpha_{11}\alpha_{21}\gamma_{11}(0) + (\alpha_{11}\alpha_{22} + \alpha_{12}\alpha_{21})\gamma_{12}(0) + \alpha_{12}\alpha_{22}\gamma_{22}(0) \\
\big(1-\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}\big)\gamma_{12}(0) &= \alpha_{11}\alpha_{21}\gamma_{11}(0) + \alpha_{12}\alpha_{22}\gamma_{22}(0) \\
\gamma_{12}(0) &= \frac{\alpha_{11}\alpha_{21}\gamma_{11}(0) + \alpha_{12}\alpha_{22}\gamma_{22}(0)}{1-\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}}.
\end{aligned}$$

This completes the derivation of the variance-covariance components. □

3.3 Conditional Maximum Likelihood Estimation for the MPINAR(1) Process

Conditional maximum likelihood estimation (CMLE) is a widely used method for estimating the parameters of time-series models (see Pedeli and Karlis [27]). The first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) process is defined as:

$$\boldsymbol{X}_t = \boldsymbol{A} \circ \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \quad t \in \mathbb{Z}, \tag{3.49}$$

where $\boldsymbol{A}$ is a thinning matrix, $\boldsymbol{X}_t$ is the observed multivariate count process, and $\boldsymbol{\varepsilon}_t$ is a vector of innovations. The innovations ε_{it} are assumed to follow independent Poisson distributions with means λ_i and variances $v_i\lambda_i$, where $v_i > 1$ accounts for overdispersion. The goal of CMLE is to estimate the unknown parameter vector $\boldsymbol{\theta}$, which includes the elements of $\boldsymbol{A}$ and the innovation means λ_i. The dispersion parameters v_i are then estimated using the residuals from the fitted model.

The likelihood function for the MPINAR(1) process is derived from the conditional distribution of $\boldsymbol{X}_t$ given $\boldsymbol{X}_{t-1}$. For a time series of length T, the likelihood function is given by:

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{t=2}^{T} f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right), \tag{3.50}$$

where $f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right)$ is the conditional probability mass function (PMF) of $\boldsymbol{X}_t$ given $\boldsymbol{X}_{t-1}$, and $\boldsymbol{\theta}$ is the vector of unknown parameters. The conditional PMF $f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right)$ depends on the distribution of the innovations $\boldsymbol{\varepsilon}_t$. For the bivariate case with Poisson innovations, the conditional PMF can be expressed as:

$$\begin{aligned} f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right) = \sum_{k_1=0}^{m_1} \sum_{k_2=0}^{m_2} \Bigg\{ & e^{-(\lambda_1+\lambda_2)} \sum_{m=0}^{\min(k_1,k_2)} \frac{\lambda_1^{x_{1t}-k_1-m} \lambda_2^{x_{2t}-k_2-m}}{(x_{1t}-k_1-m)!(x_{2t}-k_2-m)!} \\ & \times \sum_{j_1=0}^{k_1} \binom{x_{1,t-1}}{j_1} \binom{x_{2,t-1}}{k_1-j_1} \alpha_{11}^{j_1} (1-\alpha_{11})^{x_{1,t-1}-j_1} \alpha_{12}^{k_1-j_1} (1-\alpha_{12})^{x_{2,t-1}-k_1+j_1} \\ & \times \sum_{j_2=0}^{k_2} \binom{x_{2,t-1}}{j_2} \binom{x_{1,t-1}}{k_2-j_2} \alpha_{22}^{j_2} (1-\alpha_{22})^{x_{2,t-1}-j_2} \alpha_{21}^{k_2-j_2} (1-\alpha_{21})^{x_{1,t-1}-k_2+j_2} \Bigg\}, \end{aligned}$$

where:

- $m_i = \min(x_{it}, x_{1,t-1}, x_{2,t-1})$ for $i = 1, 2$,
- $\boldsymbol{\theta} = \{\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, \lambda_1, \lambda_2\}$ is the parameter vector,
- α_{ij} are the elements of the thinning matrix $\boldsymbol{A}$,
- λ_1 and λ_2 are the parameters of the Poisson innovations.

The estimation of $\boldsymbol{\theta}$ involves maximizing the log-likelihood function:

$$l(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{t=2}^{T} \log f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right). \tag{3.51}$$

This is typically done using numerical optimization techniques, such as the Newton-Raphson method or gradient-based algorithms. The steps are as follows:

1. Initialize the parameter vector $\boldsymbol{\theta}^{(0)}$.
2. Compute the log-likelihood $l(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{x})$ at iteration k.
3. Update the parameter estimates using the optimization algorithm.
4. Repeat until convergence is achieved (e.g., when the change in the log-likelihood or parameter values falls below a predefined threshold).

After obtaining the estimates for $\boldsymbol{A}$ and λ_i, the dispersion parameters v_i are estimated using the residuals from the fitted model. The residuals r_{it} are computed as the difference between the observed values x_{it} and the predicted values $\hat{x}_{it}$ based on the estimated model:

$$r_{it} = x_{it} - \hat{x}_{it}. \tag{3.52}$$

The dispersion parameters v_i are estimated by calculating the average of the squared residuals divided by the estimated mean $\hat{\lambda}_i$ for each component i over the time series, as follows:

$$\hat{v}_i = \frac{1}{T-1} \sum_{t=2}^{T} \frac{r_{it}^2}{\hat{\lambda}_i}. \tag{3.53}$$

We conclude this section by remarking that while the current formulation assumes Poisson innovations, the MPINAR(1) model can be extended to accommodate other distributions, such as the negative binomial or generalized Poisson. For instance, with negative binomial innovations, the conditional PMF $f\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}\right)$ would need to be modified to include the additional dispersion parameter. The likelihood function and estimation procedure would remain conceptually similar, but the mathematical expressions would become more complex.

3.4 Two-Step Approach to Public Health Surveillance

The proposed MPINAR(1) process can be used for modelling historical data and generating one-step-ahead forecasts for prediction-based monitoring. We should emphasize here that the set-up phase is assumed to be free of, or cleaned of, outbreaks. The proposed outbreak detection statistical methodology comprises of two steps: In the first step, the available series of data in the set-up phase (historical data) is modelled through an MPINAR(1) process and a parameter vector of maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ is obtained. The second step is dedicated to the successive monitoring of incoming observations in the operational phase (surveillance data) using the model obtained from the set-up phase. In particular, each new observation $\boldsymbol{x}_{t+1}$ is assessed against a multivariate prediction threshold derived

from the model fitted in the first step in order to determine whether an alarm should be triggered. Specifically, for each multivariate observation $\boldsymbol{x}_{t+1}$ in the operational phase, we estimate the one-step-ahead predictive distribution

$$\hat{P}(\boldsymbol{X}_{t+1} = \boldsymbol{x}_{t+1} | \boldsymbol{x}_t, \hat{\theta}), \quad \boldsymbol{x} \in \mathbb{N}_0^n$$

and obtain the marginal predictive probabilities

$$\hat{P}(X_{i,t+1} = x_{i,t+1} | \boldsymbol{x}_t, \hat{\theta}), \quad i = 1, \ldots, n.$$

For each observation $x_{i,t+1}$, we construct a $(1-\alpha)\%$ prediction interval with upper bound $x_{i,t+1}^{UB}$ equal to the $(1-\alpha)$-quantile of the corresponding marginal predictive distribution, where α is a pre-specified level of significance. The lower bound of the prediction interval is set to 0 as we are only interested in detecting positive deviations from the in-control model. Each series flags an alarm at time $t+1$ if the corresponding observation lies outside the prediction interval, i.e., if

$$x_{i,t+1} > x_{i,t+1}^{UB}.$$

Finally, for the overall alarm, a majority rule can be defined, i.e. flagging an alarm if a certain percentage of the series signals an alarm at the same point in time (see Vial [32]).

Chapter 4

Evaluating the MPINAR(1) Model: Simulation and Application to COVID-19 Syndromic Surveillance Data

In this chapter, we evaluate the performance of the first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) model through a simulation study and a real-world data application. The goal is to demonstrate the model's theoretical robustness and practical utility in public health surveillance. Section 4.1 presents a simulation study designed to assess the performance of the MPINAR(1) model in comparison to independent univariate INAR(1) models, using synthetic data generated under controlled conditions. Section 4.2 applies the MPINAR(1) model to COVID-19 syndromic surveillance data, focusing on three statistically significantly correlated symptoms: fever, cough, and dyspnea.

4.1 Simulation Study

A simulation study was conducted to evaluate the performance of the proposed MPINAR(1) model. Time series data of length $n = 200$ were simulated from a trivariate INAR(1) model with independent Poisson innovations with parameters λ_i $(i = 1, 2, 3)$. It was assumed that the first 150 observations consist the set-up phase (that is a clean process without outbreaks) and the last 50 observations consist the monitoring phase. Subsequently, for each series i, an outbreak of expected size κ_i at time $t = 170$ was simulated from a Poisson distribution with mean equal to κ_i . Therefore this trivariate case of the MPINAR(1) model has the form

$$\begin{pmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{pmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \circ \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \\ X_{3,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{pmatrix},$$

where ε_{it} are independent Poisson random variables with mean $E(\varepsilon_{it}) = \lambda_i + \kappa_i I(t = 170)$ and $I(A)$ is an indicator function. The true parameter values were assumed to be

$$\boldsymbol{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} = \begin{bmatrix} 0.4 & 0.1 & 0.3 \\ 0.3 & 0.4 & 0.2 \\ 0.3 & 0.2 & 0.1 \end{bmatrix}, \tag{4.1}$$

and $\lambda_1 = \lambda_2 = \lambda_3 = 1$. It was also assumed that $\kappa_1 = \kappa_2 = \kappa_3 = \kappa$ and the values of κ were taken to be 5, 8 or 10 . The aim of choosing these specific values for κ is to study

both cases where the outbreak is manifest, as well as cases where the outbreak cannot be easily distinguished from the typical range of values in the in-control state. Figure 4.1 shows the cumulative distribution of the maximum values of 10,000 trivariate INAR(1) series with independent Poisson innovations and $n = 200$ observations. All trivariate series are free of outbreaks ($\kappa = 0$) and have been simulated with parameter values as already described.

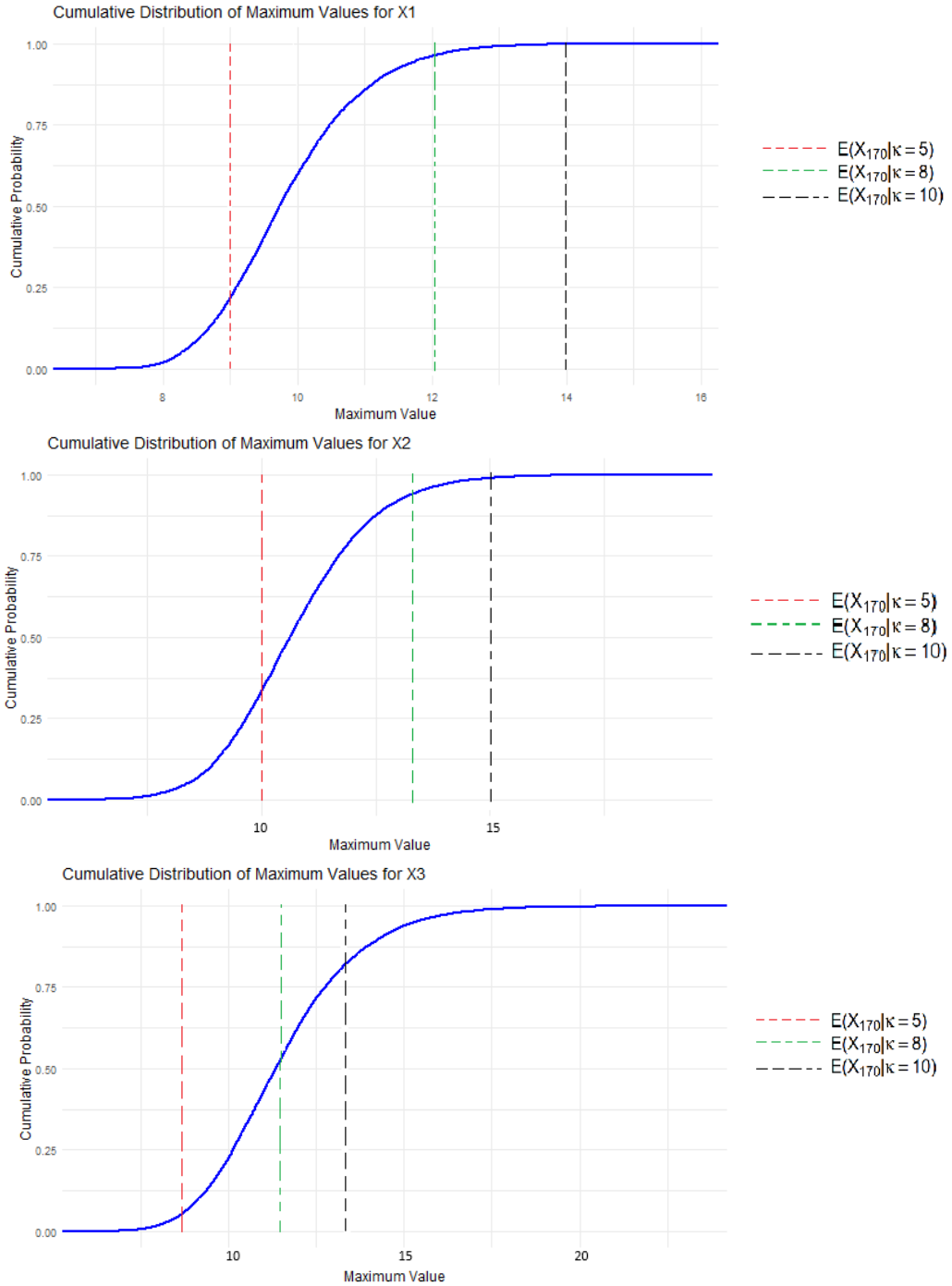

Figure 4.1: Cumulative distribution of the maximum values of 10,000 triavariate INAR(1) series with independent Poisson innovations.

The maximum values range between 8 and 25 for $\{X_1\}$ and $\{X_2\}$ and between 6 and 12 for $\{X_3\}$. Since the process is in contol until $t = 169$, the expected value of the trivariate series at

the time of the outbreak ($t = 170$) can be easily computed as $E(\boldsymbol{X}_{170}) = \boldsymbol{A}E(\boldsymbol{X}_{169}) + \boldsymbol{\lambda} + \kappa = \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\lambda} + \boldsymbol{\lambda} + \kappa$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$. The computed expectations for $\kappa = 5, 8$ and 10 are summarized in Table 4.1 below.

Table 4.1: Expected values of a triavariate INAR(1) series with independent Poisson innovations at the time of an outbreak of expected size κ equal to 5,8 and 10. Assumed parameter values are equal to $= (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \lambda_1, \lambda_2, \lambda_3) = (0.4, 0.1, 0.3, 0.3, 0.4, 0.2, 0.3, 0.2, 0.1, 1, 1, 1)$

κ	$E(X_{1t})$	$E(X_{2t})$	$E(X_{3t})$
5	9.36	10.08	8.69
8	12.35	13.08	11.69
10	14.36	15.08	13.69

Table 4.2 summarizes the empirical probabilities of the maximum value of each univariate series being greater than the corresponding expectation of the series at the time of an outbreak $(t = 170), P\left(\max_{t \neq 170}(x_{it}) > E(X_{i,170})\right), i = 1, 2, 3.$

Table 4.2: Empirical probabilities of the maximum value of each univariate series being greater than the corresponding expectation of the series at the time of an outbreak of expected size κ equal to 5, 8 and 10. Data have been simulated with $n = 200$ and $= (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33}, \lambda_1, \lambda_2, \lambda_3) = (0.4, 0.1, 0.3, 0.3, 0.4, 0.2, 0.3, 0.2, 0.1, 1, 1, 1)$. The outbreak is assumed to happened at time $t = 170$.

	$P\left(\max_{t \neq 170}(x_{it}) > E(X_{i,170})\right)$		
κ	$i = 1$	$i = 2$	$i = 3$
5	0.9993	0.4642	0.9626
8	0.7476	0.0349	0.2192
10	0.2996	0.0035	0.0315

From Figure 4.1 and Table 4.2 above we can conclude that $\kappa = 5$ corresponds to outbreaks that cannot be easily distinguished from the typical range of values in the in-control state, since the empirical probabilities $P\left(\max_{t \neq 170}(x_{it}) > E(X_{i,170})\right)$ are high for all univariate series. In contrast, $k = 8$ and $k = 10$ correspond to pronounced outbreaks with small empirical probabilities.

For each scenario ($\kappa = 5, 8$ or 10), we conducted 1000 simulation replicates. In each replicate, a trivariate INAR(1) model with independent Poisson innovations was fitted to the set-up phase and the parameter estimates were used to compute 90%, 95% and 99% upper prediction limits for the monitoring phase. For comparison purposes, we also fitted three independent INAR(1) models with Poisson innovations to the historical data and followed the same process for the computation of upper prediction limits. As evaluation measures we used the

detection rate and weekly false alarm rate based on a rule of 2/3 that is, assuming that an alarm is triggered if at least two out of the three series flagged an alarm at the same point in time. The detection rate was computed as the proportion of the 1000 replicates in which an alarm was triggered at time $t = 170$ while the weekly false alarm rate was defined as the number of cases in which an alarm was flagged at time $t \neq 170$ divided by 1000×49. Based on our simulations, we have also approximated the average run length (ARL) for different outbreak sizes κ and different significance levels α. In particular, for each univariate series we used the standard definition of ARL that is, we defined $\text{ARL}_i, i = 1, 2, 3$ as the average number of points in the monitoring phase that precede the very first indication of a false alarm. Then, to get an overall measure of the performance of the suggested multivariate surveillance approach, we followed a conservative approach defining $ARL = \min_i ARL_i$. However, it is important to note that this approximation and the related results should be treated with caution, first of all due to the limited number of simulations (see Weiß [39]), secondly because we are handling multivariate count data through a multivariate surveillance approach, and thirdly because our decision on the occurrence of an outbreak is based on a 2/3 rule rather than on modelling each series separately. The estimated ARL $_i$'s and overall ARLs are summarized in Table 4.3 below.

Table 4.3: Average run lengths for the three series (ARL $_i$'s) and overall (ARLs), for different outbreak sizes κ and different significance levels α.

		trivariate INAR(1)				**independent INAR(1)**			
Outbreak size	Sign. level	ARL_1	ARL_2	ARL_3	ARL	ARL_1	ARL_2	ARL_3	ARL
$\kappa = 5$	$a = 10\%$	16.0	15.5	15.3	15.3	17.0	13.0	14.0	13.0
	$a = 5\%$	20.5	20.1	19.5	19.5	21.5	17.5	18.2	17.5
	$a = 1\%$	25.8	27.1	24.7	24.7	26.0	25.0	24.5	24.5
$\kappa = 8$	$a = 10\%$	15.2	14.8	14.4	14.4	16.0	12.0	13.0	13.0
	$a = 5\%$	21.0	20.8	20.2	20.2	21.0	17.0	18.0	17.0
	$a = 1\%$	26.5	28.0	25.8	24.0	24.0	24.0	24.3	24.0
$\kappa = 10$	$a = 10\%$	17.0	15.5	16.0	15.5	17.5	12.5	13.8	12.5
	$a = 5\%$	20.0	21.0	20.7	20.0	21.0	15.8	17.0	15.8
	$a = 1\%$	30.0	26.0	27.0	26.0	25.5	21.5	23.0	21.5

The ARLs range from 14.4 to 15.5, 19.5 to 20.2 and 24.0 to 26.0 for $\alpha = 10\%$, 5% and 1% respectively when the multivariate approach is applied. Keeping in mind that the true outbreak has actually occurred at $t = 170$ and that the monitoring phase is the period $t = 150, \ldots, 200$, we conclude that if a false alarm is triggered, this is expected to happen around a week earlier than the true outbreak when $\alpha = 10\%$, around the time of the true outbreak when $\alpha = 5\%$ and a bit later than the time of the true outbreak when $\alpha = 1\%$. Fitting three independent INAR(1) models to the data results in generally lower ARLs that range between 12.5 and 13.0 when $\alpha = 10\%$, between 15.8 and 17.5 when $\alpha = 5\%$

and between 21.5 and 24.5 when $\alpha = 1\%$. The consistently higher ARLs obtained by the multivariate approach indicate its superiority over the univariate modelling approach in terms of this specific evaluation measure. The estimated detection rates and false alarm rates are summarized in Table 4.4 below.

Table 4.4: Detection rates (DR) and false alarm rates (FAR) for different outbreak sizes κ and different significance levels α. The reported numbers have been multiplied by 100.

		trivariate INAR(1)		independent INAR(1)	
Outbreak size	Sign. level	DR	FAR	DR	FAR
$\kappa = 5$	$\alpha = 10\%$	87.5	1.20	87.0	2.80
	$\alpha = 5\%$	78.6	0.30	76.9	0.90
	$\alpha = 1\%$	53.8	0.01	48.2	0.09
$\kappa = 8$	$\alpha = 10\%$	98.9	1.25	98.8	3.70
	$\alpha = 5\%$	98.2	0.30	97.5	1.45
	$\alpha = 1\%$	92.8	0.01	90.8	0.22
$\kappa = 10$	$\alpha = 10\%$	99.5	1.35	99.6	4.10
	$\alpha = 5\%$	99.6	0.38	99.5	1.85
	$\alpha = 1\%$	98.0	0.02	97.6	0.35

As expected, the larger the size of the outbreak is, the higher the achieved detection rate. This conclusion holds for both the trivariate and the indepedent INAR(1) modelling approaches that are equivalently effective in terms of the estimated detection rates. However, the multivariate approach has an obvious superiority in terms of the false alarm rates that are consistently lower than the corresponding false alarms rates achieved for all κ 's and α 's under the univariate approach. The outperformance of the multivariate approach in terms of false alarm rates is not surprising since the independent INAR(1) models ignore the cross-correlation between the series resulting in narrower prediction intervals and thus increasing the number of false alarms.

Focusing on the multivariate approach, it is evident that the false alarm rates are generally low without any particular pattern with regard to the outbreak size. Regarding the role of the significance level α, we observe that decreasing α results in lowering both the detection rates and false alarm rates. The degree of reduction depends however on the true oubreak size. In particular, the conservative $\alpha = 1\%$ proves to be too strict for $\kappa = 5$ as it achieves a detection rate of around 53.8% contrary to $\alpha = 5\%$ or 10% that achieve detection rates of 78.6% and 87.5% respectively. However, the detection rates achieved at different significance levels improve considerably for larger outbreak sizes even reaching 99.6% for $\kappa = 10$ and $\alpha = 5\%$. For $\alpha = 1\%$ the corresponding detection rate is equal to 98.0% but with a false alarm rate of 0.02% that is much smaller than those corresponding to $\alpha = 5\%$ or 10%(0.38% and 1.35% respectively). In the following section, the choice of the significance level to be used for outbreak detection purposes is the conservative $\alpha = 1\%$.

4.2 Application to COVID-19 Syndromic Surveillance

Among various aspects of health surveillance, syndromic surveillance is considered as an important tool since it is based on symptoms rather than diagnosis and hence it can create alerts faster. For example, syndromic surveillance systems for detection of biologic terrorism after the terrorist attack of September 11, 2001 have been launched in New York city (Das et al. [6]). In addition, during large athletic events such surveillance systems can be useful to quickly detect threats for the public health, and they have been used in winter Olympic Games in Salt Lake City 2002 (Gesteland et al. [9], Mundorff et al. [22]) and Athens 2004 Olympic Games (Dafni et al. [5]). Syndromic surveillance data are by nature low count data, especially if they refer to incidences of diseases and symptoms that are not so common. In such cases, the usual normal approximation is not appropriate and the data should rather be treated as discrete-valued time series. Moreover, when the collected data involve several related variables, this brings forward the need to consider multivariate surveillance techniques.

Unlike traditional surveillance systems that generally rely on voluntary reports from providers, syndromic surveillance systems continuously acquire data through protocols or automated routines, making them valuable for early detection, monitoring, and investigation of outbreaks, including bioterrorism-related events. The data used in this study is the syndromic surveillance data of the most commonly experienced clinical symptoms among hospitalized COVID-19 patients provided by the Brazilian Ministry of Health available for download at OpenDataSUS SRAG. The dataset has been organized on a daily basis. In this study, we considered three distinct symptoms of COVID-19 that are significantly correlated to each other (cross-correlations ranging from 0.20 to 0.45). In particular, we shall consider fever, cough, and dyspnea(a sensation of difficulty breathing or shortness of breath).

Our Set up phase was between March 11, 2020 and August 7, 2020 while our monitoring period started on August 8, 2020 and ended on September 26, 2020. During both periods symptoms were recorded on a daily basis so that the historical and surveillance data consist of $t_0 = 150$ and $t_1 = 50$ observations respectively.

The time series plots of the three series during the set-up and monitoring phases are included in Figure 4.2. Table 4.5 summarizes basic descriptive statistics. The plots of autocorrelations and partial autocorrelations of the three series during the set-up phase are shown in Figure 4.3 and Figure 4.4. The exponentially decaying autocorrelation functions indicate the appropriateness of an AR-type modelling approach whilst the partial autocorrelation functions suggest an order of dependence around one or two.

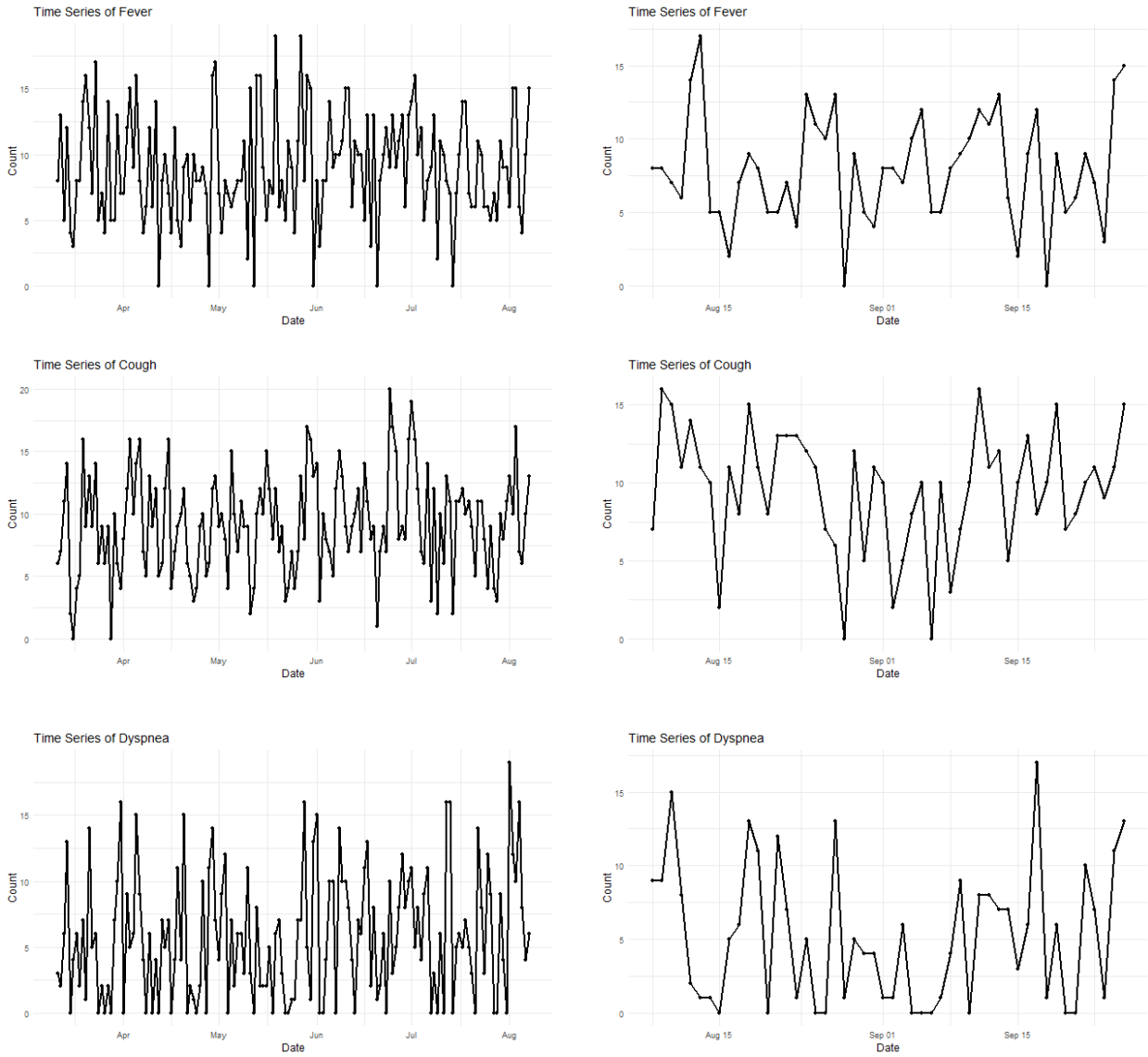

Figure 4.2: Time series plot for distinct symptoms for covid-19, fever, cough and dyspnea at Set-up phase(left panel) and Operational phase(Right panel)

Table 4.5: Mean, variance and coefficient of variation (CV) for the three syndromes during the set-up and monitoring phases.

	Set-up phase			Monitoring phase		
	Mean	Variance	CV	Mean	Variance	CV
Fever	8.96	17.80	47%	7.94	14.42	48%
Cough	9.19	16.65	44%	9.56	15.52	41%
Dyspnea	5.76	22.75	83%	5.18	22.15	91%

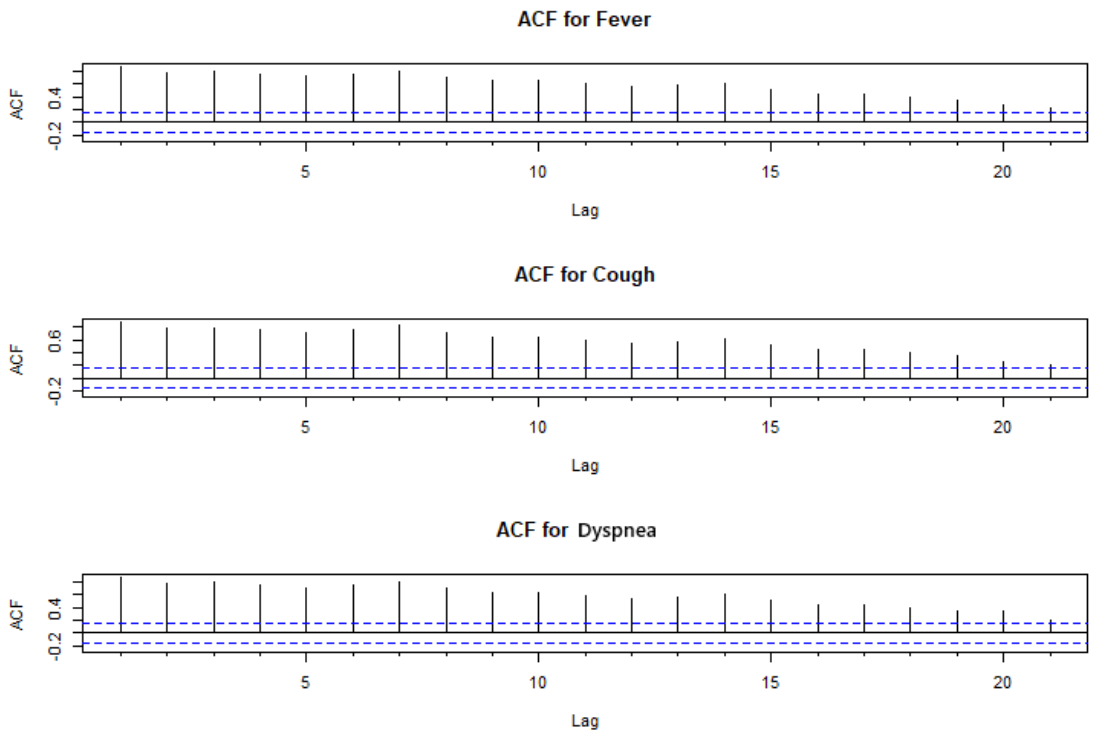

Figure 4.3: Plots of the autocorrelations of the historical data (set-up phase) for an INAR(1) time series model shows an exponential decay validating the suitability of the model to the given data showing exponential decay indicate the appropriateness of an AR-type modelling approach.

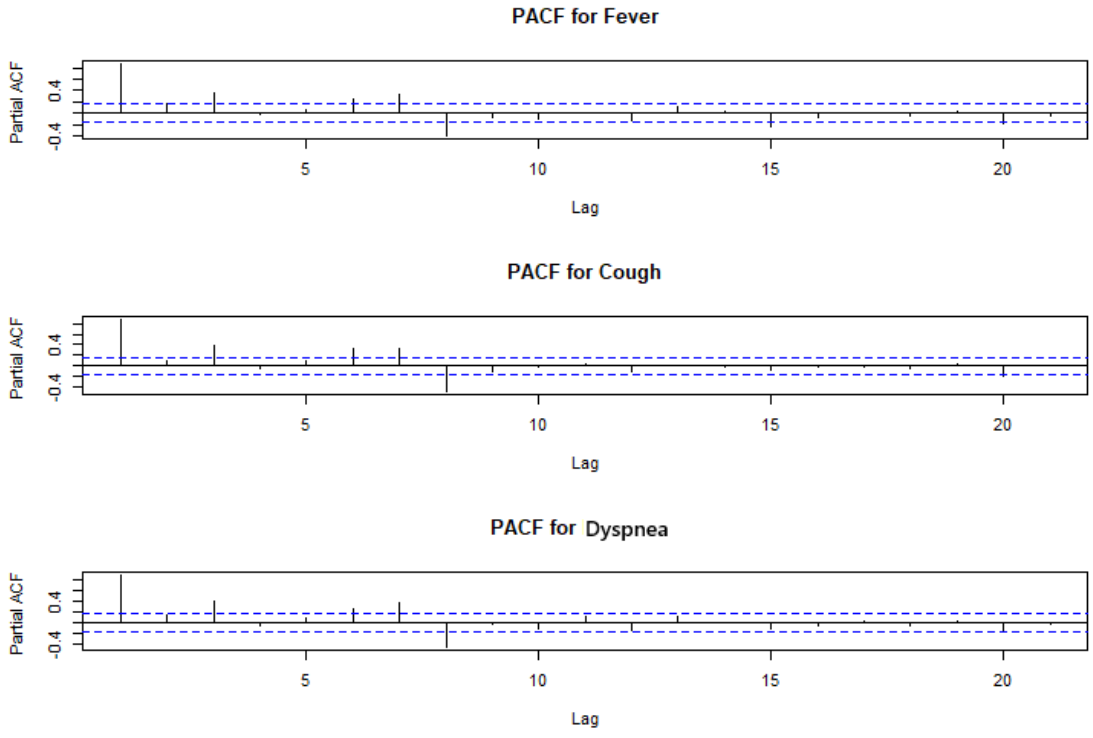

Figure 4.4: Plots of the patial autocorrelations of the historical data (set-up phase)

In the following we apply the approach of Section 3.4, i.e. we fit a trivariate INAR(1) model with indepedent Poisson innovations for modelling and prediction using the historical syndromic surveillance data. To account for regressors usually related to infectious disease data we express the expectation of the innovation series as function of the available covariate information, i.e. $E(\varepsilon_{it}) = \exp\left(\boldsymbol{z}_t^T\boldsymbol{\beta}\right), i = 1,2,3$, where $\boldsymbol{z}_t$ as a vector of covariates with associated regression parameters $\boldsymbol{\beta}$ (see Pedeli and Karlis [27]). As candidate covariates we consider terms for seasonality and a binary indicator for the day of the week on which the recording of syndromes was implemented (weekdays vs. weekends). We don't consider time trends since Figure 4.2 does not suggest the presence of any trend in our data. Therefore, each marginal series is modeled as $X_{it} = \sum_{j=1}^{3} \alpha_{ij} \circ X_{j,t-1} + \varepsilon_{it}, i = 1,2,3$, where ε_{it} are independent Poisson random variables with mean

$$E(\varepsilon_{it}) = \exp\left\{\beta_{i0} + \beta_{i1}\ \text{Weekday}\ + \beta_{i2}\cos\left(\frac{2\pi t}{365}\right) + \beta_{i3}\sin\left(\frac{2\pi t}{365}\right)\right\} \tag{4.2}$$

for $t = 1,\ldots,t_0$. Note that in the trigonometric terms that have been employed to capture seasonal patterns, we consider a seasonal period equal to 365. For comparison purposes we also employ a univariate surveillance approach based on fitting three indepedent INAR(1) regression models with Poisson innovations. Covariate information is incorporated in the univariate models in the same way, i.e. through equation 4.2. With both approaches, the marginal one-step-ahead predictive distributions are used for the construction of $(1-\alpha)\%$ prediction intervals, the upper bounds of which serve as thresholds for outbreak detection. We assume a component-wise type I error rate of $\alpha = 0.01$ and for the overall alarm we set a rule of 2/3 that is an alarm is triggered if at least two out of the three series flag an alarm at the same point in time. The parameter estimates and corresponding standard errors obtained with the two modelling approaches are summarized in Table 4.6 and 4.7.

Correlation Parameters	Trivariate INAR(1)	Independent INAR(1)
$\hat{\alpha}_{11}$	0.32(0.044)	0.41(0.039)
$\hat{\alpha}_{12}$	0.28(0.043)	-
$\hat{\alpha}_{13}$	0.35(0.054)	-
$\hat{\alpha}_{21}$	0.42(0.040)	-
$\hat{\alpha}_{22}$	0.45(0.045)	0.39(0.041)
$\hat{\alpha}_{23}$	0.30(0.048)	-
$\hat{\alpha}_{31}$	0.41(0.039)	-
$\hat{\alpha}_{32}$	0.22(0.039)	-
$\hat{\alpha}_{33}$	0.34(0.047)	0.43(0.045)

Table 4.6: Maximum likelihood estimates (standard errors) of the correlation parameters obtained from fitting three independent Poisson INAR(1) or a trivariate INAR(1) regression model with independent Poisson innovations to the historical data.

Regression Parameters	Trivariate INAR(1)	Independent INAR(1)
$\tilde{\beta}_{10}$	0.890(0.153)	0.806(0.099)
$\hat{\beta}_{11}$	−0.355(0.145)	−0.378(0.110)
$\hat{\beta}_{12}$	−0.359(0.118)	−0.322(0.078)
$\hat{\beta}_{13}$	−0.318(0.098)	−0.340(0.073)
$\hat{\beta}_{20}$	0.797(0.135)	0.896(0.096)
$\hat{\beta}_{21}$	−0.367(0.133)	−0.318(0.102)
$\hat{\beta}_{22}$	0.311(0.121)	0.356(0.070)
$\hat{\beta}_{23}$	0.348(0.110)	0.396(0.068)
$\hat{\beta}_{30}$	0.690(0.155)	0.746(0.109)
$\hat{\beta}_{31}$	0.347(0.142)	0.346(0.113)
$\hat{\beta}_{32}$	−0.374(0.099)	−0.312(0.072)
$\hat{\beta}_{33}$	−0.398(0.090)	−0.346(0.071)

Table 4.7: Maximum likelihood estimates (standard errors) of the regression parameters obtained from fitting three independent Poisson INAR(1) or a trivariate INAR(1) regression model with independent Poisson innovations to the historical data.

Results in Table 4.6 and 4.7 indicate significant first-order autocorrelations under both fittings. The cross-correlation parameters estimated by the trivariate INAR(1) model are also significant indicating the appropriateness of the multivariate approach. Figure 4.5 below shows the correlograms of the residuals obtained by the two modelling approaches.

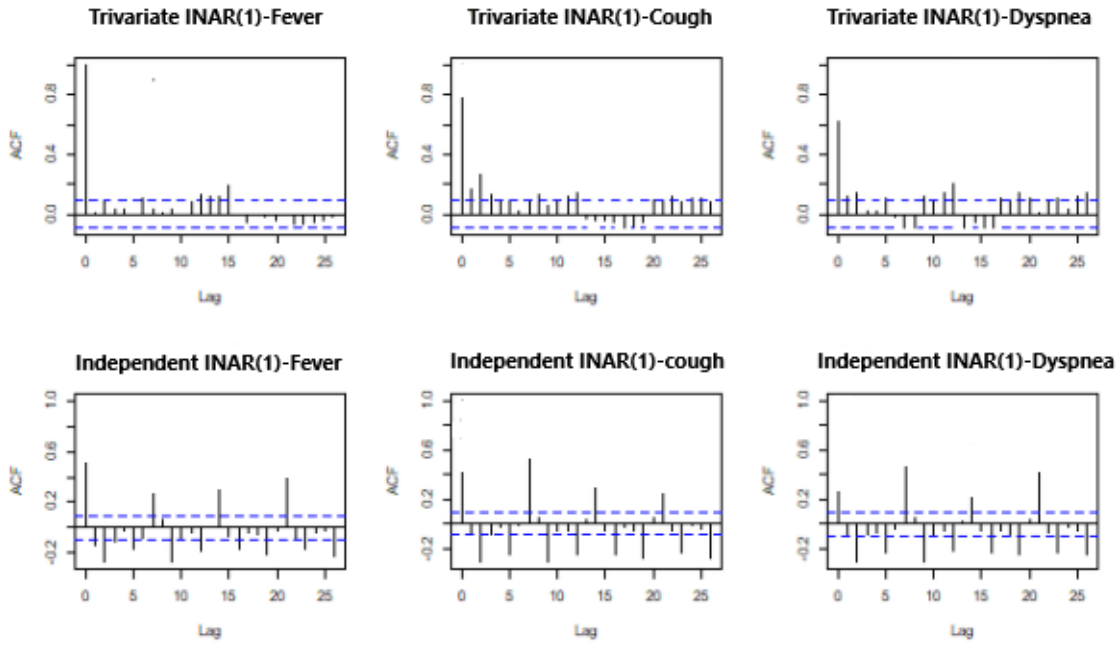

Figure 4.5: Plots of the autocorrelations of the residuals obtained by the trivariate INAR(1) (first row) and the independent INAR(1) regression models (second row).

The trivariate INAR(1) model shows fewer significant autocorrelations in the residuals. This implies that the residuals can be considered approximately white noise, indicating that the model fit is adequate. In contrast, the independent models show more significant autocorrelations, indicating a less precise fit compared to the trivariate approach.

However, Figure 4.6 below reveals some remaining cross-correlations with the two modelling approaches equally capturing cross-correlations effectively.

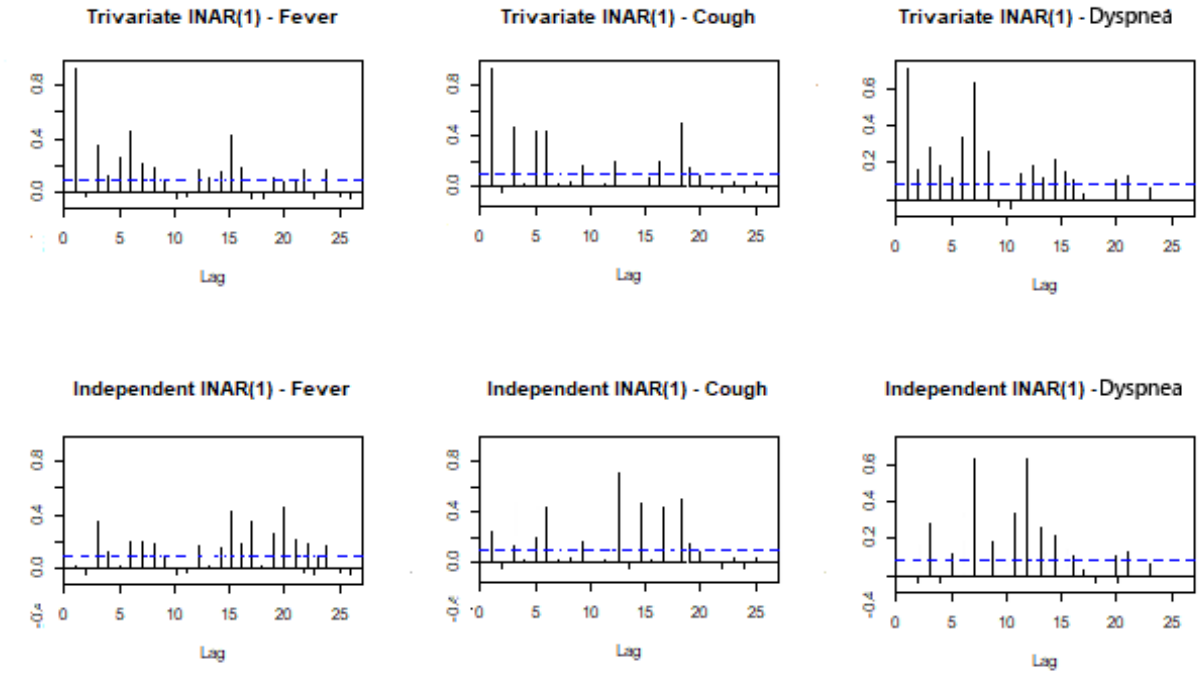

Figure 4.6: Plots of the cross-correlations of the residuals obtained by the trivariate INAR(1) (first row) and the independent INAR(1) (second row) regression models.

The surveillance plots for the two modelling approaches are illustrated in Figure 4.7. In these plots, red lines represent the upper bounds of the 99% prediction intervals, which define the thresholds for detecting alarms. Blue markers indicate the specific time points at which an overall alarm is raised.

The analysis shows that the four alarms detected using the trivariate INAR(1) model are also identified when applying the independent INAR(1) models to the historical data. This indicates that the alarms flagged by the trivariate model are consistent with those detected by the independent models. However, trivariate INAR(1) model raises an additional alarm that is not detected by the independent INAR(1) model. This suggests that while both models identify some common alarms, the trivariate INAR(1) model is more sensitive, revealing an extra alarm that may not be captured by the independent INAR(1) model.

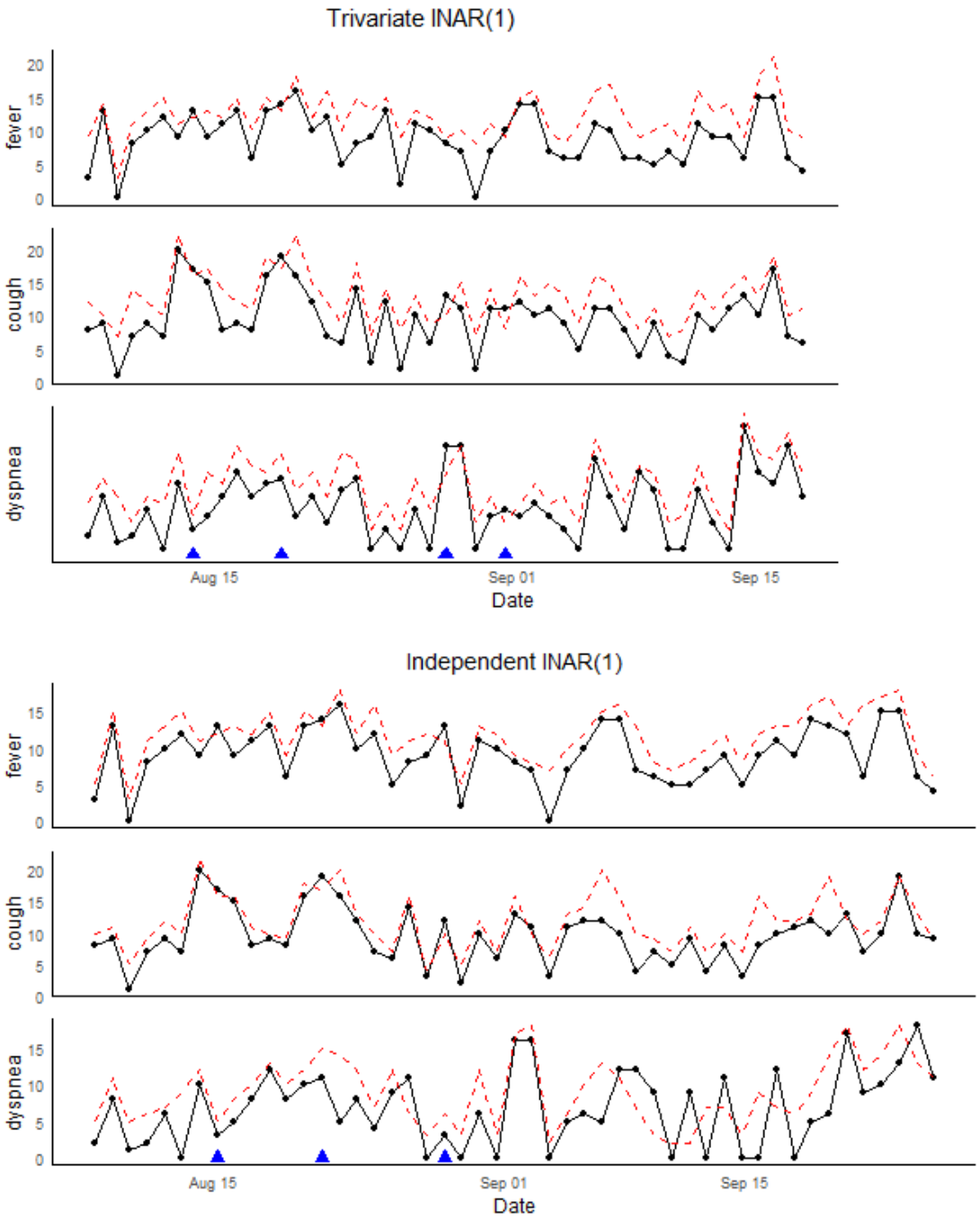

Figure 4.7: Surveillance plots as obtained after fitting a trivariate INAR(1) regression model with independent Poisson innovations (first three rows) or three independent Poisson INAR(1) regression models (last three rows) to the historical data. Statistical alarms (blue markers) are raised when at least two series exceed the upper bounds of the corresponding 99% prediction intervals (red lines).

Chapter 5

Discussion

In this chapter, we discuss the results and implications of the first-order multivariate Poisson integer-valued autoregressive (MPINAR(1)) model as a practical alternative to the more general multivariate integer-valued autoregressive (MINAR(1)) process for modelling multivariate count time series data in public health surveillance. Section 5.1 covers the formulation and key statistical properties of the MPINAR(1) model. Section 5.2 discusses a simulation study conducted to assess the model's performance under various outbreak scenarios, while Section 5.3 discusses the application of the model to real syndromic surveillance data.

5.1 On the Structure of the MPINAR(1) Model

This study has proposed the MPINAR(1) model as a practical alternative to the more general MINAR(1) process for modelling multivariate count time series data in public health surveillance. While the MINAR(1) model offers a flexible framework for capturing interdependencies among multiple count series, its reliance on correlated innovations introduces significant computational challenges during parameter estimation. These challenges become more pronounced as dimensionality increases and distributional assumptions about the correlation structure of the innovations become more difficult to justify in real-world surveillance settings. To address this, the MPINAR(1) model simplifies the innovation structure by assuming that the innovation terms follow independent Poisson distributions. This assumption reduces the complexity of the likelihood function, making conditional maximum likelihood estimation more tractable. Importantly, the assumption of independence does not compromise the model's ability to accommodate overdispersion, a common feature in public health count data, since dispersion parameters can be estimated through residual-based methods.

Moreover, the MPINAR(1) model preserves the key structural and probabilistic properties of the MINAR(1) model. These include the binomial thinning-based autoregressive structure and the stationarity conditions based on the spectral radius of the thinning matrix. By retaining these features, the MPINAR(1) model maintains theoretical robustness while enhancing practical usability. In the context of public health surveillance, this simplified framework facilitates more efficient implementation of monitoring and outbreak detection systems. The two-step methodology outlined in Section 3.4 leverages the fitted model from clean historical data to generate one-step-ahead forecasts and define predictive thresholds. This setup allows for timely detection of abnormal deviations in observed data, enabling early warning signals for potential outbreaks.

While the independence assumption simplifies inference, it may limit the model's ability to capture cross-correlations in innovations that may be present in practice. Nevertheless, this trade-off is justified in applications where computational efficiency, interpretability, and timely detection are prioritised. Future extensions of this work could explore flexible dependence structures, such as copula-based innovations or hierarchical models, that preserve tractability while allowing for limited correlation among innovation terms.

5.2 On the Simulation Study

To evaluate the practical effectiveness of the MPINAR(1) model in outbreak detection, a simulation study was conducted using trivariate count time series with controlled outbreak events. The results confirmed the model's ability to detect outbreaks reliably, especially for larger deviations from the in-control state. For small outbreaks ($\kappa = 5$), the model exhibited weaker detection performance due to the challenge of differentiating small anomalies from routine variability. However, detection performance improved significantly with increased outbreak sizes ($\kappa = 8$ and 10), as reflected in higher detection rates and lower empirical probabilities of missed detection.

The MPINAR(1) model consistently outperformed independent univariate INAR(1) models across key performance metrics, including average run length (ARL) and false alarm rate (FAR). The multivariate structure leveraged cross-series information to generate more stable predictions and appropriately wide prediction intervals, resulting in fewer false alarms. For example, at $\kappa = 8$ and $\alpha = 5\%$, the MPINAR(1) model achieved a detection rate of 98.2% and a FAR of only 0.30%, compared to a higher FAR of 1.45% under the univariate approach. These findings underscore the importance of accounting for multivariate dependencies in surveillance data.

The simulation study also examined the role of the significance level α in balancing detection sensitivity and specificity. A conservative threshold ($\alpha = 1\%$) reduced false alarms but also lowered the detection rate for smaller outbreaks. Nonetheless, for larger outbreaks, $\alpha = 1\%$ still provided high detection performance, suggesting it may be an appropriate choice when minimising false positives is a priority.

Several limitations of the simulation design should be noted. The outbreak was assumed to occur at a fixed point in time within a fixed-length series, which may not fully reflect real-world conditions. Furthermore, the number of simulation replicates (1000) may limit the precision of ARL estimates. The chosen majority rule (2/3) for overall alarm generation is one of several possible strategies and could be refined or replaced in future work. Additional investigations could consider a broader range of parameter values, outbreak timings, and alternative decision rules.

5.3 On the Real Data Application

The application of the MPINAR(1) model to syndromic surveillance data from hospitalized COVID-19 patients provides valuable insights into its practical utility for real-time outbreak monitoring. The dataset, consisting of daily counts of three key symptoms fever, cough, and dyspnea exhibited moderate cross-correlations and clear signs of temporal dependence, justifying the use of a multivariate integer-valued autoregressive framework.

By fitting a trivariate INAR(1) model with independent Poisson innovations, we were able to incorporate relevant covariate information, including day-of-week effects and seasonal components, to account for systematic variability in the innovation process. The parameter estimates indicated significant first-order autocorrelations and cross-correlations, reinforcing the appropriateness of the multivariate structure. In contrast, univariate INAR(1) regression models failed to capture the cross-series dependence, resulting in narrower prediction intervals and a higher potential for false alarms.

The residual analysis supported the improved fit of the multivariate approach. The correlograms of residuals from the trivariate model showed fewer significant autocorrelations, suggesting a better approximation to white noise. Meanwhile, residuals from the independent models exhibited more pronounced autocorrelation, indicating that these models did not fully capture the structure in the data. Interestingly, the cross-correlation plots of the residuals showed that both approaches performed similarly in capturing cross-series dynamics, though the trivariate model offered slight improvements.

Surveillance performance was evaluated using one-step-ahead predictive intervals with a 99% confidence level. The trivariate MPINAR(1) model triggered five alarms, compared to four from the univariate approach. Notably, all alarms raised by the univariate model were also detected by the trivariate model, but the latter identified an additional potential outbreak that was not captured by the independent models. This added sensitivity illustrates a key advantage of the multivariate approach: its ability to detect subtle, correlated deviations that might be missed when monitoring each series separately.

The use of a 2/3 majority rule for issuing an overall alarm proved to be an effective compromise between sensitivity and specificity. It ensured that isolated spikes in individual series did not lead to unwarranted alerts while still allowing for timely detection when multiple symptoms simultaneously showed abnormal behaviour.

Overall, the application highlights the strengths of the MPINAR(1) model in real-world settings. Its flexibility in incorporating covariate effects, ability to capture interdependencies, and suitability for low-count data make it a robust tool for modern syndromic surveillance. While the independent univariate models offer simplicity, they may overlook meaningful multivariate patterns that are crucial for early outbreak detection.

Future work could explore the inclusion of more granular covariates, alternative alarm rules,

or extensions to account for potential overdispersion or missing data. Nonetheless, the current findings strongly support the use of MPINAR(1) models in public health monitoring systems.

Chapter 6

Conclusion and Recommendations

In this final chapter, we summarise the main findings of the dissertation and highlight their implications for both theoretical and practical applications. Based on these insights, we offer recommendations to guide future research and potential improvements. Section 6.1 presents the summary, and Section 6.2 outlines the recommendations.

6.1 Conclusion

This dissertation has developed and evaluated the MPINAR(1) model as a practical and theoretically robust framework for modelling multivariate count time series data in public health surveillance. By simplifying the more general MINAR(1) process through the assumption of independent Poisson-distributed innovations, the model achieves a significant reduction in computational complexity while preserving critical structural properties, such as the binomial thinning mechanism and conditions for stationarity. This innovation enhances computational efficiency while maintaining the methodological rigour required for accurate and reliable analysis.

The comprehensive simulation study in Chapter 4 underscores the model's effectiveness in outbreak detection, demonstrating its superior performance over univariate INAR(1) models across key metrics, including average run length, false alarm rate, and detection rate. Furthermore, the study reveals how the choice of significance level influences the trade-off between sensitivity and specificity, providing valuable insights for optimizing outbreak detection systems. These findings highlight the model's potential to improve the timeliness and accuracy of public health responses.

The real-data application to syndromic surveillance during the COVID-19 pandemic further solidifies the model's practical utility. The MPINAR(1) model delivers improved model fit, superior residual diagnostics, and enhanced detection of multivariate anomalies. Its flexibility in incorporating covariates and adaptability to real-world surveillance conditions make it an indispensable tool for modern public health monitoring. This application not only validates the model's theoretical foundations but also demonstrates its readiness for deployment in real-time surveillance systems.

In conclusion, the MPINAR(1) model represents a significant advancement in the field of multivariate count time series analysis, bridging the gap between methodological rigour and computational efficiency. Its design enables real-time implementation, making it a powerful tool for early warning systems and informed decision-making in public health surveillance.

By addressing critical challenges in outbreak detection and monitoring, this work contributes to the ongoing effort to safeguard public health in an increasingly complex data-driven world.

6.2 Recommendations

The MPINAR(1) model has demonstrated its effectiveness in modelling multivariate count time series data for public health surveillance, as evidenced by the simulation study, which highlighted its superior outbreak detection performance, and the real-data application, which validated its practical utility in COVID-19 syndromic surveillance. However, challenges such as detecting small outbreaks and handling missing data, as well as opportunities to extend the model's methodological framework, have been identified during this research. To address these challenges and leverage these opportunities, the following recommendations are proposed to extend the methodological framework, enhance outbreak detection performance, and broaden the model's applicability in real-world settings.

1. **Enhancing the Innovation Structure**
   While the assumption of independent Poisson innovations simplifies computation, it may limit the ability of the model to capture cross-correlations and overdispersion in real-world data. Future work could explore alternative innovation structures, such as copula-based models or hierarchical frameworks, to allow for limited dependence among innovation terms while maintaining computational tractability. Moreover, the use of alternative distributions, such as negative binomial or zero-inflated Poisson, could be investigated to better handle overdispersed or zero-inflated data. These improvements would further enhance the model's flexibility and applicability.

2. **Extending the Methodological Framework**
   The current MPINAR(1) framework focuses on first-order autoregressive processes, but future work could extend the framework to higher-order models, such as MPINAR(p) for $p > 1$, to capture more complex temporal dependencies. Additionally, a Bayesian implementation of the MPINAR(1) model could provide a natural framework for incorporating prior knowledge, handling uncertainty, and enabling real-time updating of parameter estimates as new data become available. Furthermore, comparative studies should be conducted to evaluate the MPINAR(1) model against other multivariate count time series models, such as Bayesian networks, machine learning approaches, or state-space models, to identify the specific contexts where each method excels.

3. **Exploring Alternative Estimation Techniques**
   While conditional maximum likelihood estimation was effective for parameter estimation in the MPINAR(1) model, future work could explore alternative approaches, such as Bayesian inference or Expectation-Maximization (EM) algorithms. These methods may offer improved robustness, particularly in scenarios with limited data or complex dependence structures. Additionally, a comparative study of estimation techniques could provide insights into their relative strengths and limitations.

4. **Improving Outbreak Detection Performance**
   The simulation study revealed that the MPINAR(1) model struggles to detect small outbreaks ($\kappa = 5$) due to the challenge of differentiating subtle anomalies from routine variability. Future research could explore adaptive thresholding techniques or machine learning-based anomaly detection methods to enhance sensitivity for small outbreaks. Moreover, the simulation study assumed fixed outbreak timings, which may not reflect real-world conditions. Future work should investigate the model's performance under varying outbreak timings and durations to better assess its robustness in dynamic settings. Lastly, the 2/3 majority rule used in this study is one of many possible decision rules. Future research could evaluate alternative rules, such as weighted voting schemes or Bayesian decision frameworks, to optimize the balance between sensitivity and specificity.

5. **Expanding the Real-Data Application**
   The real-data application demonstrated the model's ability to incorporate covariates like day-of-week effects and seasonal components. Future work could explore the inclusion of more granular covariates, such as weather data, vaccination rates, or mobility patterns, to further improve predictive accuracy. Additionally, while the model was validated using COVID-19 syndromic surveillance data, its application to other infectious diseases (e.g., influenza, dengue, or tuberculosis) should be investigated to assess its generalizability and versatility. Furthermore, real-world surveillance data often contain missing values, and future research could extend the MPINAR(1) model to incorporate imputation methods or robust estimation techniques for handling incomplete data.

Appendix A

Data set for the Application

In Section 4.2, the MPINAR(1) model is applied to a trivariate syndromic surveillance dataset related to COVID-19. The dataset comprises daily counts of three key symptoms: fever, cough, and dyspnea, recorded in Brazil from March 11, 2020, to September 26, 2020. The data, sourced from the Brazilian Ministry of Health and available on OpenDataSUS SRAG, is divided into two phases: the setup phase (March 11, 2020, to August 7, 2020) with 150 observations, and the monitoring phase (August 8, 2020, to September 26, 2020) with 50 observations. Observations were recorded daily, and cross-correlations among symptoms ranged from 0.20 to 0.45. This appendix provides the dataset used in the analysis.

Test date	**Fever**	**Cough**	**Dyspnea**	**Test date**	**Fever**	**Cough**	**Dyspnea**
11/03/2020	8	6	3	29/04/2020	16	12	14
12/03/2020	13	7	2	30/04/2020	17	13	7
13/03/2020	5	11	6	01/05/2020	7	9	4
14/03/2020	12	14	13	02/05/2020	4	10	9
15/03/2020	4	2	0	03/05/2020	8	8	12
16/03/2020	3	0	4	04/05/2020	7	4	0
17/03/2020	8	4	6	05/05/2020	6	15	7
18/03/2020	8	5	2	06/05/2020	7	10	2
19/03/2020	14	16	7	07/05/2020	8	7	6
20/03/2020	16	9	1	08/05/2020	8	11	6
21/03/2020	12	13	14	09/05/2020	11	9	3
22/03/2020	7	9	5	10/05/2020	2	9	11
23/03/2020	17	14	6	11/05/2020	15	2	3
24/03/2020	5	6	0	12/05/2020	0	4	0
25/03/2020	7	9	2	13/05/2020	16	10	8
26/03/2020	4	6	0	14/05/2020	16	12	2
27/03/2020	14	9	2	15/05/2020	9	10	2
28/03/2020	5	0	0	16/05/2020	5	15	2
29/03/2020	5	10	7	17/05/2020	8	12	5
30/03/2020	13	6	10	18/05/2020	7	8	0
31/03/2020	7	4	16	19/05/2020	19	12	6
01/04/2020	7	8	0	20/05/2020	6	7	7
02/04/2020	12	12	9	21/05/2020	8	9	3

Test date	Fever	Cough	Dyspnea	Test date	Fever	Cough	Dyspnea
03/04/2020	15	16	5	22/05/2020	5	3	0
04/04/2020	9	10	6	23/05/2020	11	4	0
05/04/2020	16	14	15	24/05/2020	9	7	1
06/04/2020	8	16	9	25/05/2020	4	4	1
07/04/2020	4	7	4	26/05/2020	11	7	7
08/04/2020	6	5	0	27/05/2020	19	13	7
09/04/2020	12	13	6	28/05/2020	8	8	16
10/04/2020	6	9	0	29/05/2020	16	17	5
11/04/2020	14	12	4	30/05/2020	15	16	1
12/04/2020	0	5	0	31/05/2020	0	13	13
13/04/2020	8	6	7	01/06/2020	8	14	15
14/04/2020	10	12	5	02/06/2020	3	3	0
15/04/2020	7	16	7	03/06/2020	8	10	0
16/04/2020	4	4	0	04/06/2020	8	8	4
17/04/2020	12	7	4	05/06/2020	14	7	10
18/04/2020	5	9	11	06/06/2020	9	5	10
19/04/2020	3	10	4	07/06/2020	10	12	0
20/04/2020	9	12	15	08/06/2020	10	15	14
21/04/2020	10	6	0	09/06/2020	11	13	10
22/04/2020	5	5	2	10/06/2020	15	9	10
23/04/2020	10	3	1	11/06/2020	15	7	8
24/04/2020	8	4	0	12/06/2020	6	9	4
25/04/2020	8	9	2	13/06/2020	11	10	0
26/04/2020	9	10	10	14/06/2020	10	12	7
27/04/2020	7	5	0	15/06/2020	10	7	6
28/04/2020	0	6	11	16/06/2020	5	14	11
17/06/2020	13	11	13	07/08/2020	15	13	6
18/06/2020	3	8	2	08/08/2020	8	7	9
19/06/2020	13	9	8	09/08/2020	8	16	9
20/06/2020	0	1	1	10/08/2020	7	15	15
21/06/2020	8	7	2	11/08/2020	6	11	8
22/06/2020	10	9	6	12/08/2020	14	14	2
23/06/2020	12	7	0	13/08/2020	17	11	1
24/06/2020	9	20	10	14/08/2020	5	10	1
25/06/2020	13	17	3	15/08/2020	5	2	0
26/06/2020	9	15	5	16/08/2020	2	11	5
27/06/2020	11	8	8	17/08/2020	7	8	6
28/06/2020	13	9	12	18/08/2020	9	15	13

Test date	Fever	Cough	Dyspnea	Test date	Fever	Cough	Dyspnea
29/06/2020	6	8	8	19/08/2020	8	11	11
30/06/2020	13	16	10	20/08/2020	5	8	0
01/07/2020	14	19	11	21/08/2020	5	13	12
02/07/2020	16	16	5	22/08/2020	7	13	7
03/07/2020	10	12	8	23/08/2020	4	13	1
04/07/2020	12	7	4	24/08/2020	13	12	5
05/07/2020	5	6	9	25/08/2020	11	11	0
06/07/2020	8	14	11	26/08/2020	10	7	0
07/07/2020	9	3	0	27/08/2020	13	6	13
08/07/2020	13	12	3	28/08/2020	0	0	1
09/07/2020	2	2	0	29/08/2020	9	12	5
10/07/2020	11	10	6	30/08/2020	5	5	4
11/07/2020	10	6	0	31/08/2020	4	11	4
12/07/2020	8	13	16	01/09/2020	8	10	1
13/07/2020	7	11	16	02/09/2020	8	2	1
14/07/2020	0	2	0	03/09/2020	7	5	6
15/07/2020	7	11	5	04/09/2020	10	8	0
16/07/2020	10	11	6	05/09/2020	12	10	0
17/07/2020	14	12	5	06/09/2020	5	0	0
18/07/2020	14	10	7	07/09/2020	5	10	1
19/07/2020	7	11	5	08/09/2020	8	3	4
20/07/2020	6	9	3	09/09/2020	9	7	9
21/07/2020	6	5	0	10/09/2020	10	10	0
22/07/2020	11	11	14	11/09/2020	12	16	8
23/07/2020	10	11	8	12/09/2020	11	11	8
24/07/2020	6	8	3	13/09/2020	13	12	7
25/07/2020	6	4	12	14/09/2020	6	5	7
26/07/2020	5	9	9	15/09/2020	2	10	3
27/07/2020	7	4	0	16/09/2020	9	13	6
28/07/2020	5	3	0	17/09/2020	12	8	17
29/07/2020	11	10	9	18/09/2020	0	10	1
30/07/2020	9	8	4	19/09/2020	9	15	6
31/07/2020	9	11	0	20/09/2020	5	7	0
01/08/2020	6	13	19	21/09/2020	6	8	0
02/08/2020	15	10	12	22/09/2020	9	10	10
03/08/2020	15	17	10	23/09/2020	7	11	7
04/08/2020	6	7	16	24/09/2020	3	9	1
05/08/2020	4	6	8	25/09/2020	14	11	11
06/08/2020	10	10	4	26/09/2020	15	15	13

Appendix B

R code for the Simulation Study

In Section 4.1, a simulation study is conducted to evaluate the performance of the proposed MINAR(1) model for detecting disease outbreaks in multivariate health surveillance data. The simulation involves generating synthetic trivariate time series data with independent Poisson innovations, assessing the model's ability to capture serial and cross-correlations while addressing overdispersion. The study includes 200 observations, with the first 150 forming the setup phase (without outbreaks) and the last 50 constituting the monitoring phase (with simulated outbreaks).

This appendix contains the complete R code used to implement the simulation study. It includes scripts for generating synthetic data, estimating model parameters using conditional maximum likelihood, and evaluating performance metrics such as detection rates, false alarm rates, and average run lengths. The code is structured into sections corresponding to the simulation tasks, enabling reproducibility and further exploration of the methods.

B.1 Section 4.1: Generating Simulation Data

on simulation study to come up with data $n = 200$

```
set.seed(123) # For reproducibility

# Parameters
n <- 200
alpha <- matrix(c(0.4, 0.1, 0.3,
0.3, 0.4, 0.2,
0.3, 0.2, 0.1),
nrow = 3, byrow = TRUE)
lambda <- c(1, 1, 1)
kappa <- 5  # Example with kappa = 5

# Initialize the series
X <- matrix(0, nrow = 3, ncol = n)
```

```
# Generate the series
for (t in 2:n) {
if (t == 170) {
epsilon <- rpois(3, lambda + kappa)
} else {
epsilon <- rpois(3, lambda)
}
X[, t] <- rowSums(alpha * X[, t-1]) + epsilon
}

# Convert to data frame for easier manipulation and visualization
X_df <- data.frame(t(X))
colnames(X_df) <- c("X1", "X2", "X3")

# Display the entire dataset
print(X_df)
```

B.2 Table 4.1: Calculating Expected Values

```
# Define the parameters
alpha <- matrix(c(0.4, 0.1, 0.3,
0.3, 0.4, 0.2,
0.3, 0.2, 0.1),
nrow = 3, byrow = TRUE)
lambda <- c(1, 1, 1)
kappa_values <- c(5, 8, 10)  # Example outbreak sizes

# Identity matrix
I <- diag(3)

# Function to compute the expected values at the time of the outbreak
expected_values <- function(kappa, alpha, lambda) {
term1 <- solve(I - alpha) %*% lambda
term2 <- lambda + kappa
result <- alpha %*% term1 + term2
return(result)
}
```

```
# Compute expected values for each outbreak size
results <- sapply(kappa_values, function(kappa) {
expected_values(kappa, alpha, lambda)
})

# Convert results to a data frame
results_df <- data.frame(
kappa = kappa_values,
E_X1t = results[1, ],
E_X2t = results[2, ],
E_X3t = results[3, ]
)

# Print the results
print(results_df)
```

B.3 Table 4.2: Calculating Empirical Probabilities

```
# Define the parameters
alpha <- matrix(c(0.4, 0.1, 0.3,
0.3, 0.4, 0.2,
0.3, 0.2, 0.1),
nrow = 3, byrow = TRUE)
lambda <- c(1, 1, 1)
kappa_values <- c(5, 8, 10)
n <- 200
iterations <- 10000  # Number of iterations for simulation

# Function to compute the expected values at the time of the outbreak
expected_values <- function(kappa, alpha, lambda) {
I <- diag(nrow(alpha))
term1 <- solve(I - alpha) %*% lambda
term2 <- lambda * kappa
result <- term1 + term2
return(result)
}

# Function to simulate the trivariate INAR(1) series
simulate_inar <- function
(n, alpha, lambda) {
series <- matrix(0, n, ncol(alpha))
```

```
for (t in 2:n) {
series[t, ] <- rpois(ncol(alpha), lambda) + t(apply(alpha, 1,
function(a) rbinom(ncol(alpha), series[t - 1, ], a)))
}
return(series)
}

# Calculate empirical probabilities
calculate_empirical_prob <- function(kappa_values, alpha, lambda, n,
iterations) {
p <- ncol(alpha)
empirical_probs <- matrix(0, length(kappa_values), p)

for (k in 1:length(kappa_values)) {
kappa <- kappa_values[k]
expectations <- expected_values(kappa,
alpha, lambda)

max_values <- matrix(0, iterations, p)
for (i in 1:iterations) {
series <- simulate_inar(n, alpha, lambda)
max_values[i, ] <- apply(series[-170, ], 2, max)
}

for (j in 1:p) {
empirical_probs[k, j] <- mean(max_values[, j] > expectations[j])
}
}

return(empirical_probs)
}

# Calculate empirical probabilities for the given kappa values
empirical_probs <- calculate_empirical_prob(kappa_values,
alpha, lambda, n, iterations)

# Convert the results to a data frame for better presentation
empirical_probs_df <- data.frame(
kappa = kappa_values,
‘i=1‘ = empirical_probs[, 1],
```

```
`i=2` = empirical_probs[, 2],
`i=3` = empirical_probs[, 3]
)

# Print the results
print(empirical_probs_df)
```

B.4 Figure 4.1: Plotting Cumulative Sum

```
# Load required libraries
library(ggplot2)

# Parameters
n <- 200
alpha <- matrix(c(0.4, 0.1, 0.3,
0.3, 0.4, 0.2,
0.3, 0.2, 0.1),
nrow = 3, byrow = TRUE)
lambda <- c(1, 1, 1)

# Function to generate a single INAR(1) series
generate_inar1_series <- function(n, alpha, lambda) {
X <- matrix(0, nrow = 3, ncol = n)
for (t in 2:n) {
epsilon <- rpois(3, lambda)
X[, t] <- rowSums(alpha * X[, t-1]) + epsilon
}
return(X)
}
# Generate 10,000 series and find the maximum values
num_series <- 10000
max_values <- matrix(0, nrow = num_series, ncol = 3)
for (i in 1:num_series) {
X <- generate_inar1_series(n, alpha, lambda)
max_values[i, ] <- apply(X, 1, max)
}

# Convert to data frame for ggplot2
max_values_df <- data.frame(max_values)
colnames(max_values_df) <- c("X1", "X2", "X3")
```

```
# Plot cumulative distribution using step functions
plot_cumulative_step <- function(data, variable) {
ggplot(data, aes_string(x = variable)) +
stat_ecdf(geom = "step", color = "blue", size = 1) +
labs(title = paste("Cumulative Distribution of Maximum Values for", variable),
x = "Maximum Value",
y = "Cumulative Probability") +
theme_minimal()
}

# Plot for each series
plot1 <- plot_cumulative_step(max_values_df, "X1")
plot2 <- plot_cumulative_step(max_values_df, "X2")
plot3 <- plot_cumulative_step(max_values_df, "X3")

# Display plots
print(plot1)
print(plot2)
print(plot3)
```

Appendix C

R code for the Application

In Section 4.2, the proposed MINAR(1) model is applied to real-world syndromic surveillance data to assess its performance in modelling and forecasting occurrences of key COVID-19 symptoms, namely fever, cough, and dyspnea. The model's ability to capture serial and cross-correlations in the data while addressing overdispersion is demonstrated. Parameters were estimated using conditional maximum likelihood, and one-step-ahead forecasts were generated to evaluate the model's effectiveness in detecting potential outbreaks.

This appendix provides the complete R code used for the application of the MPINAR(1) model. The code includes scripts for data preprocessing, parameter estimation, plotting autocorrelation functions, generating forecasts, and visualizing results. Each section of the code corresponds to a specific step in the analysis, ensuring reproducibility and facilitating further exploration of the methodology.

C.1 Table 4.6: Parameter Estimation

```
library(tscount)

# ‘brazil_covid‘ is a data frame with columns fever, cough, and dyspnea
data <- brazil_covid

#Independent INAR(1) Models
# Fit independent INAR(1) models to each series
inar1_fever <- tsglm(data$fever, model = list(past_obs = 1), distr = "poisson")
inar1_cough <- tsglm(data$cough, model = list(past_obs = 1), distr = "poisson")
inar1_dyspnea <- tsglm(data$dyspnea, model = list(past_obs = 1), distr = "poisson")

# Extract coefficients and standard errors for independent INAR(1)
independent_coeffs <- c(coef(inar1_fever), coef(inar1_cough), coef(inar1_dyspnea))
independent_se <- c(sqrt(diag(vcov(inar1_fever))),
sqrt(diag(vcov(inar1_cough))), sqrt(diag(vcov(inar1_dyspnea))))
```

```
# Print results for independent INAR(1)
print("Independent INAR(1) Model:")
print(data.frame(Parameter = c("alpha11_fever", "alpha22_cough",
"alpha33_dyspnea"),
Estimate = independent_coeffs, StdError = independent_se))

#Trivariate INAR(1) Model
# Define a likelihood function for trivariate INAR(1) model
log_likelihood_trivar_inar <- function(params, data) {
  alpha11 <- params[1]
  alpha12 <- params[2]
  alpha13 <- params[3]
  alpha21 <- params[4]
  alpha22 <- params[5]
  alpha23 <- params[6]
  alpha31 <- params[7]
  alpha32 <- params[8]
  alpha33 <- params[9]

  # Initialize log-likelihood
  log_likelihood <- 0

  # Loop through time series data
  for (t in 2:nrow(data)) {
    # Model for fever, cough, and dyspnea (trivariate INAR(1))
    mean_fever <- alpha11 * data$fever[t-1] + alpha12 * data$cough[t-1] + alpha13 * da
    mean_cough <- alpha21 * data$fever[t-1] + alpha22 * data$cough[t-1] + alpha23 * da
    mean_dyspnea <- alpha31 * data$fever[t-1] + alpha32 * data$cough[t-1] + alpha33 *

    # Log likelihood for each series (assuming Poisson distribution for each)
    log_likelihood <- log_likelihood +
      dpois(data$fever[t], lambda = mean_fever, log = TRUE) +
      dpois(data$cough[t], lambda = mean_cough, log = TRUE) +
      dpois(data$dyspnea[t], lambda = mean_dyspnea, log = TRUE)
  }

  return(-log_likelihood)  # Negative log-likelihood for optimization
}
```

```
# Set initial parameter for optimization
initial_params <- rep(0.5, 9)  # Initial guesses for alpha parameters

# Perform optimization (maximum likelihood estimation)
optim_results <- optim(par = initial_params,
fn = log_likelihood_trivar_inar, data = data, method = "BFGS")

# Extract parameter estimates for the trivariate INAR(1) model
trivar_inar_coeffs <- optim_results$par

# Print results for trivariate INAR(1)
print("Trivariate INAR(1) Model:")
print(data.frame(Parameter = c("alpha11", "alpha12", "alpha13",
"alpha21", "alpha22", "alpha23", "alpha31", "alpha32", "alpha33"),
                 Estimate = trivar_inar_coeffs))
```

C.2 Figure 4.5: Plotting ACF for Residuals

```
library(tscount)
library(forecast)

#Inputing Data
data <- brazil_covid[, c("fever", "cough", "dyspnea")]

#data columns converted to numeric
data <- data.frame(lapply(data, as.numeric))

# Fit independent Poisson INAR(1) models to each series
inar1_fever <- tsglm(data$fever, model = list(past_obs = 1), distr = "poisson")
inar1_cough <- tsglm(data$cough, model = list(past_obs = 1), distr = "poisson")
inar1_dyspnea <- tsglm(data$dyspnea,
    model = list(past_obs = 1), distr = "poisson")

# Extract residuals for independent INAR(1) models
residuals_fever_ind <- residuals(inar1_fever)
residuals_cough_ind <- residuals(inar1_cough)
residuals_dyspnea_ind <- residuals(inar1_dyspnea)
```

```
# Define likelihood function for trivariate INAR(1) model
log_likelihood_trivariate_inar1 <- function(params, data) {
n <- nrow(data)
alpha <- matrix(params, nrow = 3, ncol = 3, byrow = TRUE)
lambda <- colMeans(data)

log_likelihood <- 0
for (t in 2:n) {
mu_t <- alpha %*% as.numeric(data[t-1, ])
log_likelihood <- log_likelihood + sum(dpois(as.numeric(data[t, ]),
            lambda = mu_t + lambda, log = TRUE))}

return(-log_likelihood)  # Negative log-likelihood for minimization}

# Initial parameter estimates for optimization
init_params <- rep(0.2, 9)  # Example: Initial values for a 3x3 matrix

# Fit trivariate INAR(1) model using optim
fit <- optim(par = init_params,
fn = log_likelihood_trivariate_inar1,
data = data,
hessian = TRUE,
method = "BFGS"
)

# Extract residuals for trivariate INAR(1) model
alpha_hat <- matrix(fit$par, nrow = 3, ncol = 3)
lambda <- colMeans(data)
residuals_trivar <- matrix(0, nrow = nrow(data), ncol = ncol(data))
for (t in 2:nrow(data)) {
mu_t <- alpha_hat %*% as.numeric(data[t-1, ])
residuals_trivar[t, ] <- as.numeric(data[t, ]) - (mu_t + lambda)
}
residuals_fever_trivar <- residuals_trivar[, 1]
residuals_cough_trivar <- residuals_trivar[, 2]
residuals_dyspnea_trivar <- residuals_trivar[, 3]

# Plot autocorrelations of residuals
par(mfrow = c(2, 3))
```

```
# Trivariate INAR(1) model residuals
acf(residuals_fever_trivar, main = "Trivariate INAR(1) - Fever")
acf(residuals_cough_trivar, main = "Trivariate INAR(1) - Cough")
acf(residuals_dyspnea_trivar, main = "Trivariate INAR(1) - dyspnea")

# Independent INAR(1) model residuals
acf(residuals_fever_ind, main = "Independent INAR(1) - Fever")
acf(residuals_cough_ind, main = "Independent INAR(1) - Cough")
acf(residuals_dyspnea_ind, main = "Independent INAR(1) - dyspnea")

par(mfrow = c(1, 1))  # Reset to default
```

C.3 Figure 4.6:Plotting Cross-Correlation correlogram

```
library(tscount)
library(forecast)

# ‘brazil_covid‘ is a data frame with columns fever, cough, and dyspnea
data <- brazil_covid[, c("fever", "cough", "dyspnea")]

# Ensure the data columns are numeric
data <- data.frame(lapply(data, as.numeric))

# Fit independent Poisson INAR(1) models to each series
inar1_fever <- tsglm(data$fever, model = list(past_obs = 1), distr = "poisson")
inar1_cough <- tsglm(data$cough, model = list(past_obs = 1), distr = "poisson")
inar1_dyspnea <- tsglm(data$dyspnea, model = list(past_obs = 1),
distr = "poisson")

# Extract residuals for independent INAR(1) models
residuals_fever_ind <- residuals(inar1_fever)
residuals_cough_ind <- residuals(inar1_cough)
residuals_dyspnea_ind <- residuals(inar1_dyspnea)

# Define likelihood function for trivariate INAR(1) model
log_likelihood_trivariate_inar1 <- function(params, data) {
  n <- nrow(data)
  alpha <- matrix(params, nrow = 3, ncol = 3, byrow = TRUE)
  lambda <- colMeans(data)
  log_likelihood <- 0
  for (t in 2:n) {
    mu_t <- alpha %*% as.numeric(data[t-1, ])
    log_likelihood <- log_likelihood + sum(dpois(as.numeric(data[t, ]),
    lambda = mu_t + lambda, log = TRUE))
  }
  return(-log_likelihood) # Negative log-likelihood for minimization
}

# Initial parameter estimates for optimization
init_params <- rep(0.2, 9) # Example: Initial values for a 3x3 matrix

# Fit trivariate INAR(1) model using optim
fit <- optim(
```

```
  par = init_params,
  fn = log_likelihood_trivariate_inar1,
  data = data,
  hessian = TRUE,
  method = "BFGS"
)
# Extract residuals for trivariate INAR(1) model
alpha_hat <- matrix(fit$par, nrow = 3, ncol = 3)
lambda <- colMeans(data)
residuals_trivar <- matrix(0, nrow = nrow(data), ncol = ncol(data))
for (t in 2:nrow(data)) {
  mu_t <- alpha_hat %*% as.numeric(data[t-1, ])
  residuals_trivar[t, ] <- as.numeric(data[t, ]) - (mu_t + lambda)
}
residuals_fever_trivar <- residuals_trivar[, 1]
residuals_cough_trivar <- residuals_trivar[, 2]
residuals_dyspnea_trivar <- residuals_trivar[, 3]

# Plot partial autocorrelations (PACF) and cross-correlograms (CCF)
par(mfrow = c(3, 2)) # 3 rows, 2 columns of plots

# Cross-correlograms for Trivariate INAR(1) Model Residuals
ccf(residuals_fever_trivar, residuals_cough_trivar,
main = "Trivariate INAR(1) - Fever vs Cough")
ccf(residuals_fever_trivar,
residuals_dyspnea_trivar, main = "Trivariate INAR(1) - Fever vs dyspnea")
ccf(residuals_cough_trivar, residuals_dyspnea_trivar,

main ="Trivariate INAR(1) - Cough vs dyspnea")

# Cross-correlograms for Independent INAR(1) Model Residuals
ccf(residuals_fever_ind, residuals_cough_ind,
main = "Independent INAR(1) - Fever vs Cough")
ccf(residuals_fever_ind, residuals_dyspnea_ind,
main = "Independent INAR(1) - Fever vs dyspnea")
ccf(residuals_cough_ind, residuals_dyspnea_ind,
main = "Independent INAR(1) - Cough vs dyspnea")

par(mfrow = c(1, 1)) # Reset to default plot layout
```

References

[1] Al-Osh, M., Alzaid, A. (1987), *First-Order Integer-Valued Autoregressive Process*, Journal of Time Series Analysis **8** 261–275.

[2] Cardinal, M., Roy, R., Lambert, J. (1999), *On the Application of Integer-Valued Time Series Models for the Analysis of Disease Incidence*, Statistics in Medicine **18** 2025–2039.

[3] Choi, K. (1999), *An Evaluation of Influenza Mortality Surveillance*, Time series, 1962–1979.

[4] Czado, C., Gneiting, T., Held, L. (2009), *Model Assessment for Count Data*, Biometrics **65** 1254–1261.

[5] Dafni, U., Tsiodras S., D. Panagiotakos, K. Gkolfinopoulou, G. Kouvatseas, Z. Tsourti, and G. Saroglou (2004). *Algorithm for Statistical Detection of Peaks - Syndromic Surveillance System for the Athens 2004 Olympic Games*. MMWR. Morbidity and mortality weekly report **53** (Suppl), 86–94.

[6] Das, D., D. Weiss, F. Mostashari, T. Treadwell, J. McQuiston, L. Hutwagner, A. Karpati, K. Bornschlegel, M. Seeman, R. Turcios, P. Terebuh, R. Curtis, R. Heffernan, and S. Balter (2003), Enhanced Drop-in Syndromic Surveillance in New York City Following September 11, 2001. Journal of Urban Health: Bulletin of the New York Academy of Medicine **80**(Suppl1), i76–i88.

[7] Dunsmuir, W. T. M., & Scott, D. J. (1979). *The Poisson Autoregressive Model: Some Properties and Applications*. Biometrika, **66**(2), 191–201.

[8] Franke, J., Rao T. S. (1993), *Multivariate First-Order Integer-Valued Autoregressions.*Technical Report, Forschung Universitat Kaiserslautern.

[9] Gesteland, P. H., R. M. Gardner, F.-C. Tsui, J. U. Espino, R. T. Rolfs, B. C. James, W. W. Chapman, A. W. Moore, and M. M. Wagner (2003). A*utomated Syndromic Surveillance for the 2002 Winter Olympics*. Journal of the American Medical Informatics Association **10**, 547–554.

[10] Gurland, J. (1957), *Some Interrelations Among Compound and Generalized Distributions*, Biometrika **44** 265–268.

[11] Heilbron, D. (1994), *Zero-Altered and Other Regression Models for Count Data with Added Zeros*, Biometrical Journal **36**(5) 531–547.

[12] Held, L., Höhle, M., Hofmann, M. (2005), *A Statistical Framework for the Analysis of Multivariate Infectious Disease Surveillance Counts*, Statistical Modelling **5** 187–199.

[13] Heinen, A. and Rengifo, E. (2007), *Multivariate autoregressive modeling of time series count data using copulas*. Journal of Empirical Finance **14**, 564—583 83.

[14] Paul, M., Held, L. and Toschke, A. (2008), *Multivariate Modelling of Infectious Disease Surveillance Data*, Statistics in Medicine **27** 6250–6267.

[15] Paul, M., Held, L. (2011), *Predictive Assessment of a Non-Linear Random Effects Model for Multivariate Time Series of Infectious Disease Counts*, Statistics in Medicine **30** 1118–1136.

[16] Helfenstein, U. (1986), *Box-Jenkins's Modelling of Some Viral Infectious Diseases. Statistics in Medicine* **5** 37–47.

[17] Kemp, C., Kemp, A. (1965), *Some Properties of the Hermite Distribution*, Biometrika **52** 381–394.

[18] Latour, P. (1997), *The Multivariate GINAR(p) Process. Advances in Applied Probability* **29**(1) 228–248.

[19] Marshall A.W., Olkin, I. (1990), *Multivarirate Distribution Generated from Mixtures of Convolution and Product Families*, Topics in Statistical Dependence, Block, Sampson y Sanits (Eds), Institute of Mathematical Statistics 371–393.

[20] Meyer, S., Held, L., Höhle, M. (2017), *Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package Surveillance*, Journal of Statistical Software **77**(11) 1–55.

[21] McKenzie, E. (1988), *Some ARMA Models for Dependent Sequences of Poisson Counts*, Adv. Appl. Prob. **20** 822–835.

[22] Mundorff, M., P. Gesteland, M. Haddad, and R. Rolfs (2004). *Syndromic Surveillance using Chief Complaints from Urgent-Care Facilities During the Salt Lake 2002 Olympic Winter Games*. MMWR. Morbidity and mortality weekly report **53** (Suppl), 254.

[23] Noufaily, A., D. G. Enki, P. Farrington, P. Garthwaite, N. Andrews, and A. Charlett (2013). *An improved algorithm for outbreak detection in multiple surveillance systems*. Statistics in Medicine **32**(7), 1206–1222.

[24] Okhrin, Schmid W. (2007), *Surveillance of Univariate and Multivariate Nonlinear Time Series*, Financial Surveillance, M. Frisén ed., Wiley, Chichester, 2007, 153–177.

[25] Pedeli, X., Karlis, D. (2011), *A Bivariate INAR(1) Process with Application*, Statistical Modelling **11** 325–349.

[26] Pedeli, X., Karlis, D. (2013a), *On Composite Likelihood Estimation of a Multivariate INAR(1) Model*. Journal of Time Series Analysis **34** 206–220.

[27] Pedeli, X., Karlis, D. (2013), *Some Properties of Multivariate INAR(1) Processes*, Computational Statistics and Data Analysis, Elsevier **67**(8) 213–225.

[28] Quoreshi, A. M. M. S. (2006), *Bivariate time series modeling of financial count data*. Communications in Statistics – Theory and Methods **35**, 1343-–58.

[29] Sonesson, C., Bock, D. (2003), *A Review and Discussion of Prospective Statistical Surveillance in Public Health*, Journal of the Royal Statistical Society: Series A **166** 5–21.

[30] Shmueli, G., Burkom, H. (2010), *Statistical Challenges Facing Early Outbreak Detection in Biosurveillance*, Technometrics **52** 39–51.

[31] Unkel, S., Farrington, C. and Garthwaite, P. (2012), *Statistical Methods for the Prospective Detection of Infectious Disease Outbreaks. Journal of the Royal Statistical Society* **175** 49–82.

[32] Vial, F., Wei, C.H., Held, L. (2016), *Methodological Challenges to Multivariate Syndromic Surveillance: A Case Study Using Swiss Animal Health Data*, BMC Veterinary Research **12** 288.

[33] Watier, L., Richardson, S., Hubert, B. (1991), *A Time Series Construction of an Alert Threshold: Application to S.bovismorbificans in France*, Statistics in Medicine **10** 1493–1509.

[34] Weiß, C. H. (2008), *Thinning operations for modelling Time Series of Counts* , Advances in Statistical Analysis **92** 319–341.

[35] Weiß, C. (2009), *EWMA Monitoring of Correlated Process of Poisson counts*. Quality Technology & Qualitative Management **6** 137–153.

[36] Weiß, C., Testik, C. (2009), *CUSUM Monitoring of First-Order Integer-Valued Autoregressive Processes of Poisson Counts*, Journal of Quality Technology **41** 389–400.

[37] Weiß, C. H. (2009), *Modelling Time Series of Counts with Overdispersion*, Statistical Methods and Applications **18** 507–519.

[38] Weiß, C. (2011), *Detecting Mean Increases in Poisson INAR(1) Process with EWMA Control Charts*. Journal of Applied Statistics **38** 383–398.

[39] Weiß, C. (2015), *SPC Methods for Time-Dependent Processes of Count—A Literature Review*. Cogent Mathematics **2** 1116.

[40] Weiß, C. H. (2018), *An Introduction to Discrete-Valued Time Series*, John Wiley & Sons.