

Machine learning for nowcasting and data correction: applications in epidemiological surveillance and credit risk

João Flávio Andrade Silva

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em
Estatística (PIPGEs)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

João Flávio Andrade Silva

Machine learning for nowcasting and data correction: applications in epidemiological surveillance and credit risk

Thesis submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Rafael Izbicki

USP – São Carlos
June 2026

Folha de Aprovação

Defesa de Tese de Doutorado do candidato João Flávio Andrade Silva, realizada em 09/06/2026.

Comissão Julgadora:

Prof. Dr. Rafael Izbicki (UFSCar)

Prof. Dr. Helton Graziadei de Carvalho (UFSCar)

Prof. Dr. Marcelo Ferreira da Costa Gomes (FIOCRUZ)

Prof. Dr. Leonardo Soares Bastos (FIOCRUZ)

Prof. Dr. Jochen Einbeck (DU)

João Flávio Andrade Silva

Aprendizado de máquina para nowcasting e correção de dados: aplicações em vigilância epidemiológica e risco de crédito

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Rafael Izbicki

USP – São Carlos
Junho de 2026

*To my parents, Risomar Andrade Silva and João Nogueira da Silva, my foundation.
And in loving memory of Carvão, my beloved dog, whose companionship remained with me
throughout much of this journey.*

ACKNOWLEDGEMENTS

To my supervisor, Rafael Izbicki, for his trust, the intellectual freedom granted, and the fundamental support provided at every stage of this research. His guidance was essential in the maturation of the ideas presented here.

To professors and collaborators Leonardo Bastos, Luis Salasar, Marcelo Gomes, Jochen Einbeck, John Gosling, and Helton Carvalho for the valuable discussions and contributions that significantly enriched the chapters of this work.

To my colleagues, in particular Guilherme Soares, for their companionship, the constant exchange of ideas, and the excellent spirit of cooperation.

To my colleagues at Banco do Brasil—especially Lilian Uezu, Fioravante Mietto, and Flávio Gonçalves—and to the institution itself, for the support provided through its internal research program, which allowed for the successful integration of professional practice and academic rigor.

To my family, my friends, and my girlfriend, for their unconditional support, patience, and understanding throughout these years of intense dedication to this doctoral degree.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

ABSTRACT

SILVA, J. F. A. **Machine learning for nowcasting and data correction: applications in epidemiological surveillance and credit risk.** 2026. 143 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

In complex operational systems, the inherent latency between the occurrence of an event and its official registration creates severe structural vulnerabilities. This thesis explores the shared probabilistic foundation between epidemiological reporting delays and financial verification latency, recognizing both as problems of aggregating heterogeneous conditional probabilities under temporal censoring and distributional shifts. To overcome the computational bottlenecks of conventional parametric frameworks and the point-prediction limitations of standard machine learning, this research introduces a unified, scalable conformal nowcasting methodology. By synthesizing the non-linear pattern recognition of Gradient Boosting with Conformal Prediction, the methodology generates robust prediction intervals. While it leverages specific probabilistic structures to model heteroscedastic variance, the final predictive bounds maintain mathematically valid marginal coverage independent of these underlying distributional assumptions.

The framework's efficacy is demonstrated across three core contributions. First, we develop the ViDAS and SViDAS models to estimate the proportional landscape of respiratory viruses, overcoming the systematic absence of negative test records in public health databases. Second, we introduce a bottom-up conformal architecture to nowcast absolute Severe Acute Respiratory Infection (SARI) counts. By modeling the variance through Negative Binomial convolutions, the framework achieves highly calibrated heteroscedastic bounds while reducing execution time from hours to seconds compared to traditional spatio-temporal models. Finally, we adapt this architecture to the credit risk domain, modeling portfolio defaults via the Poisson-Binomial distribution. During the 2021 macroeconomic shock, the proposed framework identified structural *Prior Shift* at an early $h = 4$ months horizon, successfully correcting the projected peak default rate (from 12.0% to 15.0%) exactly 8 months before traditional 12-month models. By applying a marginal log-odds translation, the methodology allows institutions to dynamically adjust Expected Credit Loss (ECL) provisions in an anticipatory manner, preserving the model's rank-ordering without triggering rigid retraining validation cycles. Ultimately, this thesis provides scalable methodological frameworks—spanning from open-source epidemiological pipelines to advanced Model Risk Management (MRM) applications—that substantially mitigate the operational vulnerabilities caused by information latency.

Keywords: Nowcasting, Conformal Prediction, Machine Learning, Epidemiological Surveillance, Credit Risk.

RESUMO

SILVA, J. F. A. **Aprendizado de máquina para nowcasting e correção de dados: aplicações em vigilância epidemiológica e risco de crédito**. 2026. 143 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Em sistemas operacionais complexos, a latência inerente entre a ocorrência de um evento e o seu registro oficial cria vulnerabilidades estruturais severas. Esta tese explora a fundação probabilística compartilhada entre atrasos de notificação epidemiológica e a latência de verificação financeira, reconhecendo ambos como problemas de agregação de probabilidades condicionais heterogêneas sob censura temporal e mudanças de distribuição. Para superar os gargalos computacionais de abordagens paramétricas convencionais e as limitações de predição pontual do aprendizado de máquina padrão, esta pesquisa introduz uma metodologia unificada e escalável de *nowcasting* conformal. Ao sintetizar o reconhecimento de padrões não-lineares do *Gradient Boosting* com *Conformal Prediction*, a metodologia gera intervalos de predição robustos. Embora o *framework* utilize estruturas probabilísticas específicas para modelar a variância heteroscedástica, os limites preditivos finais mantêm uma cobertura marginal matematicamente válida, independente dessas premissas distribucionais subjacentes.

A eficácia do *framework* é demonstrada por meio de três contribuições centrais. Primeiro, desenvolvemos os modelos ViDAS e SViDAS para estimar o cenário proporcional de vírus respiratórios, contornando a ausência sistemática de registros de testes negativos em bases de dados de saúde pública. Segundo, introduzimos uma arquitetura conformal *bottom-up* para o *nowcasting* de contagens absolutas de Síndrome Respiratória Aguda Grave (SRAG). Ao modelar a variância através de convoluções Binomiais Negativas, o *framework* atinge limites heteroscedásticos altamente calibrados, reduzindo o tempo de execução de horas para segundos em comparação com modelos espaço-temporais tradicionais. Finalmente, adaptamos essa arquitetura para o domínio de risco de crédito, modelando a inadimplência de portfólios via distribuição Poisson-Binomial. Durante o choque macroeconômico de 2021, o *framework* proposto identificou um *Prior Shift* estrutural no horizonte precoce de $h = 4$ meses, corrigindo com sucesso a taxa de pico de inadimplência projetada (de 12,0% para 15,0%) com exatos 8 meses de antecedência em relação aos modelos tradicionais de 12 meses. Ao aplicar uma translação marginal de *log-odds*, a metodologia permite que instituições ajustem as provisões de Perda Esperada de Crédito (ECL) de maneira antecipada, preservando a ordenação de risco (*rank-ordering*) do modelo sem acionar ciclos rígidos de validação de retreinamento. Em suma, esta tese fornece *frameworks* metodológicos escaláveis — desde *pipelines* epidemiológicos em código aberto até aplicações avançadas de Gestão de Risco de Modelo (MRM) — que mitigam substancialmente os pontos cegos operacionais causados pela latência de informação.

Palavras-chave: Nowcasting, Conformal Prediction, Machine Learning, Vigilância Epidemiológica, Risco de Crédito.

LIST OF FIGURES

Figure 1 – Illustrative Example: Estimated Viral Distribution by Municipality and Age (2024)	27
Figure 2 – Estimated Weekly Proportion of SARS-CoV-2 by Age Group (São Paulo, 2022)	40
Figure 3 – Estimated Weekly Proportion of RSV by Age Group (São Paulo, 2022) . . .	41
Figure 4 – Conceptual Diagram of the Delay Analysis Methodology	45
Figure 5 – Impact of Reporting Lag on Viral Estimation Performance	46
Figure 6 – Performance of NViDAS vs. Naive Model under Data Scarcity	47
Figure 7 – Predictor Variable Importance based on Total Gain	48
Figure 8 – Learned Effects of Age and Epidemiological Week on Viral Probabilities . .	49
Figure 9 – Spatial ICE Plot for Influenza A (30-Year-Old Patient)	51
Figure 10 – Spatial ICE Plot for RSV (Neonate Patient)	52
Figure 11 – Practical Objective of the Nowcasting Framework	59
Figure 12 – Empirical Distribution of Prediction Errors Across Reporting Delays	78
Figure 13 – Empirical Marginal Coverage of Conformal Methods Across Reporting Delays	80
Figure 14 – Ablation Study of Prediction Interval Efficiency Across Reporting Delays .	82
Figure 15 – Out-of-Time Evaluation of Empirical Marginal Coverage	87
Figure 16 – Out-of-Time Evaluation of Prediction Interval Efficiency	88
Figure 17 – Real-Time Nowcasting Simulation for Major Health Regions	91
Figure 18 – Global Feature Importance for the Reporting Delay Classifier	92
Figure 19 – Partial Dependence and ICE Plots for Patient Age and Population Size . . .	94
Figure 20 – Spatial Partial Dependence Plot of Reporting Delays	96
Figure 21 – Conformal Drift Detection: Baseline Maturation vs. Early Framework . . .	111
Figure 22 – Nowcasting Accuracy: Comparative Evaluation of Adapted Predictions vs. Baseline	113
Figure 23 – Recursive Partitioning of the Brazilian Geographic Space	132

LIST OF TABLES

Table 1	– Example illustrating the structure of dataset $\mathbb{D}$.	32
Table 2	– Observed virus proportion for the period from 2022 to 2024	38
Table 3	– Mean Brier Score (and standard deviation) for each virus	43
Table 4	– Mean Brier Score (and standard deviation) for the four most incident viruses, stratified by age range	43
Table 5	– Minimum Stabilization Lag for Main Viruses by Region	45
Table 6	– Diagram representing right-censored surveillance data (the run-off triangle).	66
Table 7	– Point Estimate Accuracy (MAE and RMSE) Across Reporting Delays	83
Table 8	– Out-of-Time Point Estimate Accuracy (MAE and RMSE)	89

CONTENTS

1	INTRODUCTION	21
1.1	Context and Motivation: The Paradox of Real-Time Information	21
1.2	The Methodological Gap	22
1.3	The Central Thesis and Shared Probabilistic Foundation	23
1.4	Research Objectives and Contributions	23
1.5	Thesis Structure	24
2	GRANULAR ESTIMATION OF VIRAL PROPORTIONS	25
2.1	Context	25
2.2	Related Work	28
2.3	Methodology	29
2.3.1	<i>Surveillance Challenges</i>	30
2.3.2	<i>Data</i>	31
2.3.3	<i>Statistical Framework</i>	33
2.3.3.1	<i>Virus distribution estimator</i>	33
2.3.3.2	<i>Estimating virus distribution in age intervals</i>	34
2.3.3.3	<i>Explaining Individual Predictions via Embedding Space Analysis</i>	36
2.3.4	<i>Summary of Proposed Estimators: The ViDAS Framework</i>	36
2.4	Results	37
2.4.1	<i>Performance Evaluation of the Proposed Methodology</i>	38
2.4.2	<i>Analysis of the Effect of Delays on Estimates</i>	44
2.4.3	<i>Performance in Data-Scarcity Scenarios</i>	46
2.4.4	<i>Interpretation and Explanation of the NViDAS Model</i>	47
2.5	Discussion	53
3	NOWCASTING EPIDEMIOLOGICAL COUNTS	57
3.1	Context	57
3.2	Related work	61
3.3	Methodology	64
3.3.1	<i>Dataset and Feature Engineering</i>	65
3.3.2	<i>Data Representation: The Run-off Triangle</i>	66
3.3.3	<i>Statistical Modeling: Conditional Estimates and Individual Aggregation</i>	67
3.3.4	<i>Uncertainty Quantification and Conformal Predictors</i>	70

3.4	Results	75
3.4.1	<i>Analysis of Residuals and Non-Conformity Scores</i>	76
3.4.2	<i>Internal Evaluation of Conformal Predictors</i>	79
3.4.3	<i>Comparison with the State-of-the-Art (Bayesian Benchmark)</i>	85
3.4.4	<i>Interpretability of the Delay Classifier</i>	92
3.5	Discussion	95
4	DOMAIN ADAPTATION: CORRECTING CREDIT DEFAULT RISK	101
4.1	Context	101
4.2	Related Work	102
4.3	Methodology	104
4.3.1	<i>Problem Formulation and Temporal Recalibration</i>	104
4.3.2	<i>Statistical Formalization: The Poisson-Binomial Estimator</i>	104
4.3.3	<i>Early Drift Detection via Conformal Prediction</i>	105
4.3.4	<i>Predictive Adaptation and Vintage Nowcasting</i>	106
4.3.5	<i>Parameter Estimation and Data Splitting Strategy</i>	108
4.4	Experimental Evaluation	108
4.4.1	<i>Dataset Description and Experimental Setup</i>	109
4.4.2	<i>Comparative Analysis: Baseline Maturation vs. Early Drift Detection</i>	110
4.4.3	<i>Nowcasting Accuracy: Correcting the Structural Bias</i>	112
4.5	Discussion	113
5	CONCLUSION	117
5.1	Synthesis of the Central Thesis	117
5.2	The Research Journey: Summary of Critical Findings	117
5.3	Practical and Operational Impact	118
5.4	Limitations and Future Directions	119
	BIBLIOGRAPHY	121
APPENDIX A	MODEL CONFIGURATION AND TRAINING STRATEGY	129
APPENDIX B	GEOGRAPHIC COORDINATES AND TREE-BASED MODELS	131
APPENDIX C	DERIVATION OF THE VIRUS DISTRIBUTION CONDITIONED ON AGE INTERVALS	133
APPENDIX D	ESTIMATION OF CONDITIONAL DENSITY PARAMETERS VIA GBDT	135

APPENDIX E	NUMERICAL INTEGRATION USING THE PROBABILITY-WEIGHTED MIDPOINT RULE	137
APPENDIX F	DEFINITION OF THE BRIER SCORE	139
APPENDIX G	MOMENT-MATCHING EQUATIONS FOR HEURIS-TIC QUANTILES	141
G.1	Negative Binomial Moment Matching	141
G.2	Gamma Moment Matching	142

INTRODUCTION

1.1 Context and Motivation: The Paradox of Real-Time Information

We live in an era defined by the digitalization of complex systems, where public health agencies and financial institutions alike rely on massive streams of data to guide critical operations. The modern expectation is that strategic decisions—whether allocating intensive care unit (ICU) beds during a viral outbreak or adjusting credit limits for millions of consumers—can be executed in real-time, driven by live dashboards and high-frequency indicators.

However, this expectation is fundamentally challenged by the friction of the physical world. Information is rarely instantaneous; it is inherently bound by operational, administrative, and biological latency. In public health surveillance, the true epidemiological event of interest begins with the onset of a patient’s symptoms. Yet, the time required to seek medical care, and process administrative records ensures that a Severe Acute Respiratory Infection (SARI) case initiating today will only be fully consolidated in official databases weeks later ([BASTOS *et al.*, 2019](#); [SILVA *et al.*, 2022](#)). Similarly, in the financial domain, a credit card issued under current macroeconomic conditions requires a full maturation window—typically 12 months—before the institution can definitively verify whether the operation resulted in a default ([BELLINI, 2019](#)).

This inherent lag between the occurrence of an event and its official confirmation creates severe structural vulnerabilities, even though manifesting differently across domains. In epidemiology, the ongoing accumulation of delayed reports generates a right-censored data structure ([LAWLESS, 1994](#); [HÖHLE; HEIDEN, 2014](#)); when decision-makers observe the most recent weeks on a raw surveillance dashboard, the incidence curve systematically exhibits an artificial downward trend. Conversely, in the financial sector, the mandatory maturation window imposes a chronological blind spot. Because definitive performance metrics are only available for credit cohorts originated a year in the past, risk managers are effectively evaluating a lagged reality

that may no longer reflect the current economic environment.

Naively trusting these uncorrected or delayed observations induces a dangerous false sense of security. Public health officials might falsely conclude that an epidemic wave has already passed due to the artificial drop in recent counts, while financial institutions might continue to grant billions in credit under the assumption that current risk levels accurately mirror the success of outdated, fully matured vintages. Consequently, making operational decisions without mathematically accounting for this verification latency leads to substantial misallocations of resources and delayed preventive actions.

1.2 The Methodological Gap

To correct these data truncation issues and estimate the true current state of a system—a process known as *nowcasting*—the scientific literature has traditionally relied on parametric statistical frameworks. In epidemiology, spatio-temporal Bayesian hierarchical models (often inferred via Integrated Nested Laplace Approximations, or INLA) represent the state-of-the-art (BASTOS *et al.*, 2019; SALMON *et al.*, 2015). While statistically robust, these models suffer from the curse of dimensionality. As spatial granularity increases to hundreds of health regions, the rigid requirements for predefined spatial adjacency matrices and the heavy computational burden render these models computationally expensive for timely, high-resolution surveillance.

In response to these computational bottlenecks, the literature has increasingly turned to Machine Learning (ML) algorithms, such as Gradient Boosting Machines (GBM) (FRIEDMAN, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). These highly flexible algorithms excel at extracting non-linear patterns from complex, high-dimensional data. However, classical ML approaches introduce a critical flaw for high-stakes domains: they predominantly suffer from point-prediction limitations. Without rigorous Uncertainty Quantification (UQ) (IZBICKI, 2025), they fail to provide the statistical safety margins (prediction intervals) required to make safe decisions.

Furthermore, adjusting to rapid distribution shifts—such as a sudden viral mutation or a sharp macroeconomic interest rate hike—typically requires fully retraining the ML model (QUIÑONERO-CANDELA *et al.*, 2008; POLO *et al.*, 2023). In highly regulated environments like the financial industry, retraining an algorithm triggers prolonged Model Risk Management (MRM) validation cycles, substantially hindering the agility required to survive an economic shock.

1.3 The Central Thesis and Shared Probabilistic Foundation

This thesis posits that the delay in epidemiological reporting and the verification latency in credit risk are not merely analogous challenges, but are structurally bound by a **shared probabilistic foundation**. Both phenomena are fundamentally problems of aggregating heterogeneous conditional probabilities under temporal right-censoring and distributional shifts (GRZENDA; GOMES; BIFET, 2020; GOMES *et al.*, 2022).

By recognizing this shared structural foundation, this research introduces a unified, scalable analytical framework capable of overcoming the limitations of both Bayesian inference and traditional machine learning. The core proposition is the synthesis of **Gradient Boosting**—to flexibly learn highly granular probability surfaces from individual-level data—with **Conformal Prediction (CP)**—to rigorously calibrate the residual uncertainty.

Unlike heuristic error margins or strict parametric bounds, Conformal Prediction provides distribution-free, finite-sample coverage guarantees (VOVK; GAMMERMAN; SHAFER, 2022; ANGELOPOULOS; BARBER; BATES, 2024). By employing locally adaptive non-conformity scores (ROMANO; PATTERSON; CANDÈS, 2019; IZBICKI; SHIMIZU; STERN, 2020) and structurally modeling the underlying variance through convolutions of independent and heterogeneous random variables (Negative Binomials for health counts, and Poisson-Binomials for credit defaults (WANG, 1993)), the proposed framework demonstrates *model-based robustness*. This allows the methodology to natively absorb macroscopic distribution shifts, dynamically translating the risk surface based solely on short-term empirical evidence without the need to retrain the underlying non-linear models.

1.4 Research Objectives and Contributions

The primary objective of this thesis is to develop, validate, and generalize a conformal machine learning framework for real-time nowcasting and early drift detection under severe reporting latency. To achieve this, the research is structured around three main contributions, progressively advancing from conditional probability estimation to macroscopic uncertainty quantification, and finally, to cross-domain generalization.

1. **Estimating Proportional Landscapes under Incomplete Data:** The first contribution addresses the fundamental epidemiological question of *which* pathogens are circulating. By modeling viral etiologies as conditional probabilities, we propose a methodology to overcome the pervasive absence of negative test records in public health databases, enabling the granular tracking of respiratory viruses (Chapter 2).

2. **Conformal Nowcasting of Absolute Counts:** The second contribution solves the operational question of *how many* severe cases are occurring. We introduce a bottom-up framework that aggregates individual-level reporting delay probabilities into regional incidence curves. By applying Split Conformal Prediction over Negative Binomial convolutions, the framework achieves highly calibrated, heteroscedastic prediction intervals, surpassing the computational efficiency and calibration of traditional Bayesian spatial models (Chapter 3).
3. **Predictive Adaptation and Early Drift Detection in Finance:** The final contribution proves the broad generalizability of the framework by translating it to the credit risk domain. We formalize the financial verification latency through the Poisson-Binomial distribution and demonstrate that the conformal framework can successfully anticipate structural macroeconomic shocks (*Prior Shift*) exactly 8 months before traditional 12-month models. Furthermore, we introduce a log-odds adaptation mechanism that corrects long-term risk projections in an anticipatory manner, substantially mitigating the operational frictions of model retraining (Chapter 4).

1.5 Thesis Structure

The remainder of this thesis is organized as follows. Chapter 2 presents the foundational machine learning framework used to model the proportional distribution of respiratory viruses. Chapter 3 extends this logic to absolute count data, detailing the mathematical derivation of the aggregated delay probabilities and the empirical validation of the Conformal Prediction strategies against state-of-the-art Bayesian benchmarks. Chapter 4 generalizes the framework to the financial sector, demonstrating its efficacy in early drift detection and the nowcasting of credit portfolio defaults during macroeconomic crises. Finally, Chapter 5 provides a general conclusion, synthesizing the theoretical and operational impacts of this research and outlining promising frontiers for future work.

GRANULAR ESTIMATION OF VIRAL PROPORTIONS

Effective epidemiological surveillance requires not only tracking the total volume of severe infections but also understanding their underlying etiological composition. Before addressing the operational challenge of nowcasting absolute case counts—the focus of the subsequent chapter—it is fundamental to determine *which* respiratory pathogens are actively circulating across specific demographic and spatial contexts. This chapter introduces a machine learning framework designed to estimate the proportional distribution of respiratory viruses among hospitalized patients. By modeling these viral proportions as conditional probabilities, we present a robust methodology to overcome a pervasive limitation in public health databases: the systematic absence of negative test records.

2.1 Context

Severe Acute Respiratory Infection (SARI) is a medical condition characterized by an acute respiratory infection that lead to severe symptoms, including respiratory difficulty and low blood oxygen levels. SARI can be caused by a variety of infectious agents, including viruses (such as SARS-CoV-2, influenza, and respiratory syncytial virus (RSV)), bacteria, and other microorganisms. The condition gained particular relevance during the COVID-19 pandemic, as many infected patients exhibited SARI symptoms([World Health Organization, 2022](#)).

The World Health Organization (WHO) defines SARI for surveillance purposes as an acute respiratory infection in patients with a history of fever (or measured fever above 38°C), cough within the last 10 days, and requiring hospitalization. SARI can progress to severe respiratory failure, often requiring intensive medical interventions like mechanical ventilation in Intensive Care Units (ICU)([World Health Organization, 2022](#)).

In Brazil, universal surveillance of SARI was implemented in 2009, following the WHO's declaration of a Public Health Emergency of International Concern due to the influenza A(H1N1) pandemic. Since 2019, SARI cases are reported in the Influenza Epidemiological Surveillance Information System (SIVEP-Gripe), the official system used to monitor the clinical-epidemiological profile of hospitalized cases and deaths, as well as to track circulating viruses (Ministério da Saúde do Brasil, Secretaria de Vigilância em Saúde, 2022; Ministério da Saúde do Brasil, 2025; SILVA *et al.*, 2022).

Epidemiological data is essential for public health analyses to understand viral distribution and infectious disease spread. The Brazilian Ministry of Health makes these detailed SARI records publicly available — including dates, locations, demographic characteristics, and identification of the viral agents — updated weekly through an open data portal (SILVA *et al.*, 2022; Ministério da Saúde do Brasil, 2025).

Current surveillance systems, such as the InfoGripe panel (Fiocruz, FGV e Ministério da Saúde, 2023), are useful for monitoring which viruses are most incident across regions. However, they rely on aggregated empirical estimates of viral proportions and counts, pooling data over relatively long time windows (e.g., the previous four weeks), broad geographical areas, and predefined age groups. This aggregation leads to a loss of granular information, hindering the ability to provide specific, local estimates.

Our objective is to efficiently monitor the predominant viruses causing SARI across all Brazilian municipalities and age groups, with a particular focus on producing reliable estimates for the current epidemiological context. We aim to complement existing surveillance systems by providing local and timely estimates at a fine granularity (municipality and patient's age), enabling continuous monitoring and supporting healthcare professionals in detecting changes in viral circulation that are relevant for immediate prevention and control decisions.

To ensure that this continuous monitoring is readily accessible, we also provide a dedicated web application (<https://joaoflavio.shinyapps.io/Virus_Monitor>) that implements the proposed methodology (SILVA; AL., 2024b). As an illustrative example, Figure 1 presents the viral distribution estimated by our method across the 52 epidemiological weeks of 2024 for the municipalities of São José dos Campos (state of São Paulo) and Uberlândia (state of Minas Gerais), comparing patients aged 1 and 65 years. The figure highlights pronounced heterogeneity across both municipalities and age groups, revealing distinct local viral profiles that would be masked under aggregation at broader geographic levels. In support of this framework, we developed the Python package PYSRAG (SILVA; AL., 2024a), which provides a comprehensive set of tools for processing and analyzing SARI data in Brazil, including data preprocessing, feature engineering, and model training, thereby facilitating the application and reproducibility of the proposed method.

This work addresses several challenges arising from both data characteristics and statistical limitations. These include substantial delays between symptom onset and registration in

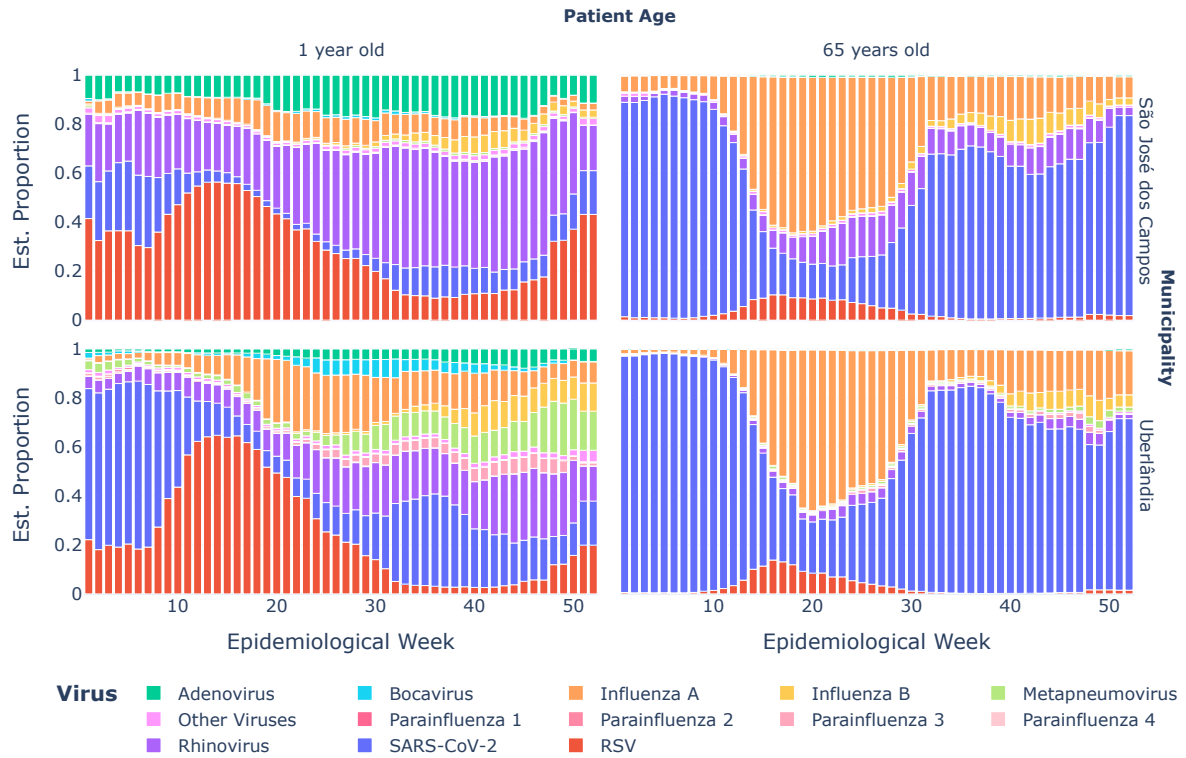


Figure 1 – Estimated distribution of respiratory viruses over the 52 epidemiological weeks of 2024 as inferred by our method, stratified by municipality (São José dos Campos and Uberlândia) and patient age (1 and 65 years). The panels reveal distinct, municipality-specific viral profiles across age groups: RSV and Rhinovirus predominate in infants (1 year), whereas SARS-CoV-2 and Influenza A are more prevalent among the elderly (65 years). This fine-scale heterogeneity would be obscured if data were aggregated across cities or locations, demonstrating the added value of our method in preserving spatial and demographic granularity.

SIVEP, the lack of information on negative test results for most viruses—which prevents direct estimation of true incidence—and the previously noted difficulty of producing local estimates without excessive aggregation or loss of information. In addition, computational efficiency is a key constraint, as the proposed system must support weekly updates and deliver results through a web application operating under limited resources (approximately 1 GB of RAM).

Our methodology quantifies the proportion of viral distribution by estimating the likelihood of detecting a particular virus, conditioned on specific covariates (location, population size, patient age, and time). To address the data challenges, we employ a probabilistic multi-class classification algorithm to estimate viral distribution among those who tested positive. This approach allows us to circumvent the absence of negative data and minimize the impact of information delay. By utilizing a properly selected probabilistic multi-class classification algorithm and effective variable treatment, we eliminate the necessity of defining prior groups, enabling us to obtain specific, contextualized, and generalizable estimates even when observations for specific configurations are absent.

2.2 Related Work

The surveillance of SARI in Brazil relies heavily on the SIVEP-Gripe database, a centralized system pivotal since the 2009 H1N1 pandemic (SILVA *et al.*, 2022). Current monitoring initiatives focus primarily on tracking the volume of cases; however, effective surveillance requires not only monitoring the volume of cases but also accurately attributing them to specific etiological agents in real-time. This section outlines how our work advances the state-of-the-art by addressing the limitations of current aggregation methods, traditional spatial modeling, and nowcasting techniques.

Surveillance Systems

Current monitoring initiatives differ significantly in scope and methodology. The InfoGripe system (Fiocruz, FGV e Ministério da Saúde, 2023; BASTOS *et al.*, 2019) employs robust Bayesian frameworks to estimate SARI incidence counts, correcting for reporting delays. However, its estimation of viral distribution — the proportion of cases attributed to specific pathogens — often relies on empirical aggregation over large geographical areas and extended time windows.

In contrast, the Instituto Todos pela Saúde (ITpS) (ITpS - Instituto Todos pela Saúde, 2022) leverages data from a network of private laboratories to monitor pathogen circulation. Recently, ITpS has begun explicitly juxtaposing these private positivity rates against SARI (SIVEP-Gripe) curves. However, this approach presents two significant limitations for the specific goal of SARI surveillance. First, similar to the InfoGripe panel, ITpS estimates are typically aggregated over large geographical areas (national, regional, or state levels), which obscures the heterogeneity of viral circulation at the municipality level. Second, the populations differ: SARI surveillance targets severe cases requiring hospitalization, whereas private laboratory data are based on individuals tested in partner laboratories, covering a broader clinical spectrum from mild to severe respiratory infections. Consequently, the viral distribution observed in private exams may not accurately mirror the etiological profile of severe infections, as pathogens associated with mild symptoms may appear with different frequencies than those causing severe respiratory failure.

Our work bridges this gap by providing a granular, patient-level estimator that relies exclusively on SARI notification cases. Unlike empirical aggregation, which obscures local heterogeneity (ROCHA *et al.*, 2021), our approach allows for specific viral attribution even in municipalities with sparse data.

Spatiotemporal Modeling

To estimate disease risk in areas with sparse data, the literature has traditionally leveraged spatial and spatiotemporal dependence via model-based geostatistics (DIGGLE; GIORGI, 2019;

LIMA *et al.*, 2021). Bayesian hierarchical models and Generalized Linear Mixed Models (GLMMs) are frequently employed to smooth incidence rates, explicitly incorporating spatial dependence through neighborhood adjacency matrices or distance-based covariance functions (CHAMMARTIN *et al.*, 2014; COURA-VITAL *et al.*, 2021).

However, these methods often require a prior specification of spatial structures and linear interaction terms. While efficient approximations like INLA exist, they can be sensitive to modeling assumptions (WONG *et al.*, 2023) and computationally demanding for high-dimensional, real-time surveillance systems.

Consequently, recent literature has explored Machine Learning (ML) alternatives (ALBRECHT *et al.*, 2024; SILVA *et al.*, 2024). Our work adopts this paradigm by utilizing Gradient Boosting Decision Trees (GBDT). Our approach treats geolocation (latitude and longitude) as continuous covariates, allowing the estimator to adaptively partition the spatial domain—effectively learning data-driven neighborhoods—and to capture complex, non-linear interactions between location, time, and patient age without requiring explicit interaction terms.

Complementing Nowcasting Efforts

Finally, this work complements the extensive “nowcasting” literature (MCGOUGH *et al.*, 2020; BASTOS *et al.*, 2019) by shifting the target from incidence counts to viral proportions. This shift allows for the characterization of the etiological landscape even when total case counts are still being consolidated.

Beyond estimation, we address the “black box” criticism often leveled at ML models in public health (VILLANUEVA-MIRANDA; XIAO; XIE, 2025). We introduce a model-specific interpretability strategy based on the estimator’s internal embedding space, identifying influential training instances to explain predictions. Furthermore, we extend the standard point-estimation framework to propose SViDAS, a semi-parametric estimator that integrates viral probabilities over continuous age distributions, enabling precise estimation for arbitrary age intervals (e.g., 0-2 years) crucial for pediatric surveillance.

2.3 Methodology

This work proposes a statistical framework designed to provide granular and timely estimates of respiratory viral distribution. We model viral proportions not as raw aggregates but as conditional probabilities estimated via a machine learning approach. This section outlines the inherent challenges of surveillance data, the dataset composition, and the specific statistical procedures employed to generate robust estimates.

2.3.1 Surveillance Challenges

To achieve a fast and granular estimator of viral proportions, we need to address many challenges:

1. **Information Delays:** Delays in data availability arise from the time required for it to pass through several stages, from an individual's first symptom perception to system reporting with virus identification. This includes seeking medical care, undergoing testing, obtaining results, and entering the results into the system (BASTOS *et al.*, 2019).
2. **Incomplete Data on Negative Tests:** The database records a virus only if the test result is positive, except for influenza. This leads to a lack of data on negative test results: the data does not distinguish individuals with negative results from those not tested for a given virus.
3. **Data Sparsity in Some Regions:** Information collection and availability vary significantly across different areas. For example, in São José dos Campos, as shown in Figure 1, gathering enough data on 1-year-old patients is highly unlikely. This data sparsity can result from factors such as lower population density, limited healthcare resources, or infrequent reporting in these specific demographics.
4. **Computational Efficiency:** Due to the requirement for weekly updates of the models and the final deployment to a web application with limited memory resources (e.g., 1 GB RAM), efficiency is a critical constraint.

Our methodology quantifies the viral distribution by estimating the likelihood of detecting a particular virus, conditioned on specific covariates (location, population size, patient age, and time).

To circumvent issues related to reporting delays (1) and missing negative data (2), we estimate the proportion of each viral agent among SARI patients with at least one positive viral test. This estimand should therefore be interpreted as the viral composition among detected positive cases, rather than as the absolute incidence or positivity rate of each virus in the underlying SARI population.

Regarding reporting delays, this proportion is expected to be less sensitive than absolute case counts when delays affect viral agents approximately uniformly, so that the relative composition of positive cases is not systematically distorted. In the results section, we empirically assess this condition by comparing estimates obtained under different reporting lags with the consolidated data. This analysis identifies the reporting lags after which the estimated viral distribution becomes sufficiently stable for surveillance purposes.

The issue of data sparsity (3) is generally solved by aggregating data across various regions and age groups, which unfortunately leads to a loss of information. To overcome this,

our methodology models the proportion through the probability of observing a specific virus, conditioned on specific covariates. We use a probabilistic multi-class classification algorithm to obtain these estimates, which leverages the patterns identified across all available data to generate specific, contextualized, and generalizable estimates even in regions with sparse data.

We addressed the issue of computational efficiency (4) by selecting and setting up an algorithm properly, ensuring rapid performance during both training and application (see [Appendix A](#) for detailed model configuration and training strategy).

2.3.2 Data

Since its implementation in Brazil in 2009, SARI universal surveillance mandates the collection and reporting of all hospitalized cases and/or deaths due to respiratory viruses ([Ministério da Saúde do Brasil, Secretaria de Vigilância em Saúde, 2022](#); [SILVA *et al.*, 2022](#)). The current clinical criteria established by the Brazilian Ministry of Health for characterizing SARI define it as an individual presenting with an influenza-like illness (ILI), which must include at least two of the following signs and symptoms: fever, chills, sore throat, headache, cough, runny nose, olfactory or gustatory disturbances. In addition to these symptoms, the individual must also exhibit one or more of the following: shortness of breath/respiratory discomfort, persistent chest pressure or pain, oxygen saturation lower than 95% in ambient air, bluish coloration (cyanosis) of the lips or face ([Ministério da Saúde do Brasil, Secretaria de Vigilância em Saúde, 2022](#)).

While the SARI criterion established by the Brazilian Ministry of Health is fundamental for national surveillance, its formulation was historically directed at monitoring Influenza and was more recently adapted for the COVID-19 pandemic. However, the growing relevance of other respiratory pathogens, such as Respiratory Syncytial Virus (RSV), has led to discussions in global health organizations about the need to broaden case definitions to improve the detection of these agents ([World Health Organization, 2019](#); [HIRVE *et al.*, 2020](#)).

Studies demonstrate that fever, a central component of the traditional criterion, is an inconsistent clinical sign in RSV infections, often being absent or blunted in vulnerable populations like the elderly ([URCHUEGUÍA-FORNES *et al.*, 2025](#); [NORMAN, 2000](#)), or serving as a non-specific symptom common in young children for various other causes ([KLINK *et al.*, 2020](#)). To enhance surveillance sensitivity for RSV and other respiratory viruses and to maintain data comparability across different seasonal and epidemiological periods, this study adopts a modified criterion that excludes the mandatory presence of fever. This approach aligns with the methodology employed by the InfoGripe surveillance system (Fiocruz), in accordance with updated Ministry of Health guidelines ([Fundação Oswaldo Cruz \(Fiocruz\), 2025](#)). Accordingly, the datasets were filtered using the following logical rule: (cough OR sore throat) AND (dyspnea OR oxygen saturation < 95% OR difficulty breathing) AND (hospitalized OR deceased). This strategy prioritizes respiratory and severity signs that exhibit greater consistency across different etiological agents.

Data were retrieved from the Brazilian Ministry of Health’s Open Data Portal ([Ministério da Saúde do Brasil, 2024](#)). To incorporate spatial and demographic dimensions, we enriched the records with geographical coordinates from the IBGE Selected Localities Register ([Instituto Brasileiro de Geografia e Estatística \(IBGE\), 2010](#)) and municipality population sizes based on the 2022 Demographic Census ([Instituto Brasileiro de Geografia e Estatística \(IBGE\), 2022](#)).

Formally, our dataset is denoted by $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^n$, where x_i consists of covariate vectors containing information about the i -th individual. This information includes the epidemiological week of the first symptoms, the patient’s age in years (computed from the reported age value and unit), geographical coordinates (latitude and longitude of the city, state, and region), and the population size of the notifying municipality. The selection of these covariates reflects specific epidemiological dynamics. Patient age is included because the prevalence and impact of different respiratory viruses vary distinctively across age groups. Furthermore, incorporating geolocation and temporal data is essential given Brazil’s continental proportions, where viral distribution can differ significantly between regions at any given time. These choices align with literature identifying age as a critical risk factor and highlighting the role of regional disparities in SARI incidence ([ROCHA *et al.*, 2021](#); [ARAUJO *et al.*, 2020](#)).

The response variable, y_i , indicates which virus the patient’s test result was positive for: SARS-CoV-2, Influenza A/B Virus, Respiratory Syncytial Virus, Human Parainfluenza Virus 1/2/3/4, Human Adenovirus, Human Metapneumovirus, Human Bocavirus, Human Rhinovirus, and/or Others (other viruses or unspecified). The dataset only contains patients who tested positive for at least one virus. In cases of viral co-detection, the patient record was expanded into multiple observations, one for each detected virus. Refer to [Table 1](#) for a sample representation of the dataset $\mathbb{D}$.

Table 1 – Example illustrating the structure of dataset $\mathbb{D}$.

Virus	Week	Patient’s Age	City		State		Region		Population Size
	1st Sympt.		Lat.	Long.	Lat.	Long.	Lat.	Long.	
FLUA	1	68	-23.87	-47.99	-23.54	-46.63	-21.64	-43.53	32,062
SARS2	3	56	-23.54	-46.63	-23.54	-46.63	-21.64	-43.53	12,200,180
SARS2	4	72	-10.91	-37.07	-10.91	-37.07	-7.31	-37.99	605,309
RSV	5	0.17	-16.82	-49.24	-16.67	-49.25	-17.12	-51.98	500,760
METAP	12	2	-23.54	-46.63	-23.54	-46.63	-21.64	-43.53	12,200,180
SARS2	26	46	-23.59	-48.05	-23.54	-46.63	-21.64	-43.53	166,959
FLUA	42	8	-25.42	-49.27	-25.42	-49.27	-27.68	-49.68	1,871,789
Others	46	68	-14.67	-40.47	-12.97	-38.51	-7.31	-37.99	23,290
SARS2	47	43	-30.03	-51.22	-30.03	-51.22	-27.68	-49.68	1,404,269
RSV	48	0.83	-23.54	-46.63	-23.54	-46.63	-21.64	-43.53	12,200,180

2.3.3 Statistical Framework

This section details the statistical methodology employed to estimate the conditional distribution of viral infections. We first define the multiclass classification framework used to estimate the class-conditional probability distribution over all viruses given a patient's covariate profile. Subsequently, we extend this framework to estimate the marginal virus distribution over specific patient age intervals, employing a semiparametric integration strategy. Finally, we describe the embedding-based approach used to interpret individual estimator predictions by identifying influential training instances.

2.3.3.1 Virus distribution estimator

Our goal is to estimate the proportion of patients who test positive for a specific virus among positive patients, conditioned on x . We approach this task as a multiclass classification problem, where the variable Y can take values y representing different viruses being considered. The proportion is modeled by the conditional probability, denoted as $\mathbb{P}_Y(y|x)$, which indicates the likelihood of a patient testing positive for a specific virus given the value of x . To estimate this probability, we employ a probabilistic classifier, denoted as $\hat{g}(\cdot)$, obtained based on the data in $\mathbb{D}$.

For the function $\hat{g}(\cdot)$, we chose the Gradient Boosting Machine (GBM) estimator based on decision trees, widely known as Gradient Boosting Decision Trees (GBDT; see [Appendix A](#) for detailed model configuration and training strategy). In a multi-class problem like predicting viruses, the GBDT constructs a separate additive model for each possible virus class k . Each of these models produces a raw score (or logit), which represents the evidence for that specific class. The estimated score for a given virus k , which we denote as $\hat{h}_k(x)$, is calculated as the sum of the predictions from all the trees built for that class:

$$\hat{h}_k(x) = \sum_{m=1}^M \alpha_m \hat{b}_{mk}(x),$$

where M is the total number of trees in the ensemble for each class, $\hat{b}_{mk}(\cdot)$ is the m -th decision tree assigned to the k -th virus class, and α_m is the learning rate, adjusting each tree's contribution ([HASTIE; TIBSHIRANI; FRIEDMAN, 2009](#)).

These raw scores, $\hat{h}_k(x)$, are not probabilities; they are unbounded real numbers. To convert the vector of scores $[\hat{h}_1(x), \hat{h}_2(x), \dots, \hat{h}_K(x)]$ into a valid probability distribution, we use the *softmax function*. The final estimated probability for a specific virus k , denoted $\hat{g}_k(x)$, is given by substituting the additive estimator directly into the softmax equation:

$$\hat{g}_k(x) = \frac{\exp(\hat{h}_k(x))}{\sum_{j=1}^K \exp(\hat{h}_j(x))}. \quad (2.1)$$

Here, the numerator is the exponentiated score for the virus class k , and the denominator is the sum of exponentiated scores across all possible virus classes $j \in \{1, \dots, K\}$. This

transformation ensures that each probability $\hat{g}_k(x)$ is between 0 and 1 and that their sum across all classes equals 1, forming a proper probability distribution. The complete estimator, $\hat{g}(x)$, is the function that outputs this full vector of probabilities. This estimator was trained using the multi-class logarithmic loss (multi-logloss), which is suitable for multiclass classification as it works in conjunction with the softmax function to ensure the predicted values from $\hat{g}(x)$ approximate these valid class-conditional probabilities (BISHOP, 2006).

The choice of a tree-based estimator is due to its effective partitioning of the covariate space. It allows detailed analysis across geographical and temporal dimensions and interaction with other relevant variables, eliminating the need for predefined groups and offering a more flexible and precise data analysis approach. For a theoretical justification on the use of geographic coordinates over indicator variables for spatial partitioning in tree-based models, refer to [Appendix B](#).

This method was also chosen for its practical advantages: it allows for fast deployment due to low computational cost at prediction time and has low memory requirements for storing the estimator. Furthermore, GBDT performs implicit variable selection during the boosting process. The estimator's structure also offers high interpretability; its tree-based nature supports global variable importance analysis, while its ensemble nature allows for the specific local prediction explanations (based on an internal embedding space) demonstrated in the [section 2.4](#).

2.3.3.2 Estimating virus distribution in age intervals

The estimator $\hat{g}(x)$ provides estimates for the distribution of viruses conditioned on the vector x , which in this context means estimating the virus distribution for a patient with SARI given a specific patient age, specific location, and specific epidemiological week of the first symptoms. However, it is often of interest to estimate the virus distribution not for a particular patient age, but rather for a patient age range. This extension is motivated by scenarios in which the focus is not on a single patient profile, but rather on understanding the expected virus distribution for groups defined by age intervals — for instance, infants under 1 year of age or adults between 60 and 70 years old.

In this scenario, our goal is to estimate the conditional probability:

$$\mathbb{P}_Y(y|X_j \in [a, b], x_{-j}) = \frac{\int_a^b \mathbb{P}_Y(y|x, x_{-j}) f_{X_j}(x|x_{-j}) dx}{F_{X_j}(b|x_{-j}) - F_{X_j}(a|x_{-j})}, \quad (2.2)$$

representing the conditional probability of virus y , given that patient age (X_j) belongs to the interval $[a, b]$, holding constant the other predictor variables (x_{-j}). Here, $f_{X_j}(\cdot|x_{-j})$ denotes the conditional density function of X_j given the remaining covariates x_{-j} , and $F_{X_j}(\cdot|x_{-j})$ is the corresponding cumulative distribution function (CDF). The derivation of the expression (2.2) can be found in [Appendix C](#).

To estimate this probability efficiently, we proceed in two steps: first, we estimate the conditional density of the patient's age, and second, we integrate this density with the virus estimator (2.1) using a numerical approximation.

Step 1: Estimation of the conditional density of age

The true conditional density $f_{X_j}(x_j|x_{-j})$ is unknown and must be estimated. Alternative approaches for conditional density estimation include nonparametric methods such as FlexCode or quantile regression-based techniques (IZBICKI; LEE, 2017; DALMASSO *et al.*, 2020; IZBICKI, 2025). While flexible, these methods typically require training a large ensemble of estimators—often necessitating dozens of distinct models to determine basis coefficients or specific quantile levels. Given our application's requirement for rapid response times and computational efficiency, we opted for a semiparametric approach. We simplify the estimation process by assuming a parametric distribution, which reduces the computational burden to estimating only the parameters of the distribution.

Since the patient's age is a strictly positive variable, we assume it follows a lognormal distribution conditioned on location variables and the epidemiological week of the first symptoms (DEGROOT; SCHERVISH, 2012; FORBES *et al.*, 2010). That is,

$$X_j|x_{-j} \sim LN(\mu(x_{-j}), \sigma^2(x_{-j})).$$

Under this assumption, the conditional density of patient age can be estimated by determining only the conditional mean and variance parameters (IZBICKI, 2025). We adopted an approach based on (WASSERMAN, 2006) to estimate each of these parameters using a GBDT estimator (see Appendix D for details). Therefore, the estimated conditional density is given by

$$\hat{f}_{X_j}(x_j|x_{-j}) = \frac{1}{x_j \sqrt{2\pi \hat{\sigma}^2(x_{-j})}} \exp\left(-\frac{(\log(x_j) - \hat{\mu}(x_{-j}))^2}{2\hat{\sigma}^2(x_{-j})}\right),$$

and the estimated cumulative distribution function is given by

$$\hat{F}_{X_j}(x_j|x_{-j}) = \Phi\left(\frac{\log(x_j) - \hat{\mu}(x_{-j})}{\hat{\sigma}(x_{-j})}\right) \quad (2.3)$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution $N(0, 1)$.

Step 2: Numerical Approximation of the Estimator

Having estimated the conditional distribution of viruses using the probabilistic estimator $\hat{g}(x)$ and the conditional density of patient age through a log-normal model, we now describe how to integrate both estimators to obtain the virus distribution across a range of patient ages. To address this, we construct an estimator for the conditional probability in (2.2) by integrating the pointwise estimator $\hat{g}(x)$ over the estimated conditional density of age, $\hat{f}_{X_j}(x_j|x_{-j})$, as follows:

$$\hat{\mathbb{P}}_Y(y|X_j \in [a, b], x_{-j}) = \frac{\int_a^b \hat{g}(x, x_{-j}) \hat{f}_{X_j}(x|x_{-j}) dx}{\hat{F}_{X_j}(b|x_{-j}) - \hat{F}_{X_j}(a|x_{-j})}. \quad (2.4)$$

This estimator reflects the average predicted virus distribution over the age interval $[a, b]$, weighted by the likelihood of observing each age within that interval, conditional on the remaining covariates. The denominator term, $\hat{F}_{X_j}(b|x_{-j}) - \hat{F}_{X_j}(a|x_{-j})$, in (2.4), can be easily computed using the cumulative distribution function of the standard normal distribution, as shown in (2.3). However, to estimate the numerator, a numerical approximation is required to evaluate the integral (see [Appendix E](#) for details). Accordingly, the conditional probability in (2.4) can be approximated as:

$$\hat{\mathbb{P}}_Y(y|X_j \in [a, b], x_{-j}) \approx \frac{\sum_{i=1}^{s-1} \hat{g}\left(\frac{r_i + r_{i+1}}{2}, x_{-j}\right) [\hat{F}_{X_j}(r_{i+1}|x_{-j}) - \hat{F}_{X_j}(r_i|x_{-j})]}{\hat{F}_{X_j}(b|x_{-j}) - \hat{F}_{X_j}(a|x_{-j})}$$

where the interval $[a, b]$ is partitioned into $s - 1$ subintervals using a sequence of ordered points $a = r_1 < r_2 < \dots < r_s = b$.

2.3.3.3 Explaining Individual Predictions via Embedding Space Analysis

To provide a rationale for individual estimator predictions, particularly for hypothetical cases in data-scarce contexts, we employ a nearest-neighbor analysis within the estimator's internal embedding space. This approach identifies which training instances the estimator considers most similar to a given case, thus revealing the basis for its prediction.

The similarity between any two observations is defined by their representation within the Gradient Boosting Decision Tree (2.1). For a given observation x , its estimated embedding vector $\hat{v}_x$ is constructed from the weighted outputs of each tree in the ensemble:

$$\hat{v}_x = [\alpha_1 \hat{b}_1(x), \alpha_2 \hat{b}_2(x), \dots, \alpha_M \hat{b}_M(x)]$$

where $\hat{b}_m(x)$ is the prediction from the m -th tree and α_m is its corresponding weight (learning rate). The similarity between a new case and instances in the training data is then quantified by calculating the Euclidean distance between their respective embedding vectors. This technique allows us to identify the training observations that are likely most influential in the estimator's decision-making process for a specific prediction, thereby providing a clear, data-driven explanation for its output.

2.3.4 Summary of Proposed Estimators: The ViDAS Framework

Based on the statistical framework developed in the previous subsections, we formally define our proposed methodology, named **ViDAS (Viral Distribution estimation in Acute respiratory infection Surveillance)**. It is implemented through three distinct variants:

- **NViDAS (Nonparametric ViDAS):** As detailed in [subsubsection 2.3.3.1](#), this approach estimates the conditional probability of a patient testing positive for a specific virus, given the patient's individual information.

- **ANViDAS (Aggregated Nonparametric ViDAS):** To obtain an estimate of the virus proportion for a specific municipality in each epidemiological week, the NViDAS predictive model is applied to each observed case in the database. Then, the average of the estimated probabilities is calculated for each week. This procedure results in a weekly estimate of the virus proportion that considers the empirical age distribution observed in the data for that week and municipality.
- **SViDAS (Semiparametric ViDAS):** As described in [subsubsection 2.3.3.2](#), this approach is used to estimate the virus proportion for a specific age range in a given location and epidemiological week. It employs the estimator combination methodology, which estimates the average virus probability for the age range of interest by weighting the predictions of the virus probability estimator by the estimated age distribution (log-normal) for the context. The result is a weekly estimate that reflects the average proportion for an age group, based on an estimated age distribution.

In summary, both ANViDAS and SViDAS are extensions of the base NViDAS model. While NViDAS operates at the individual level, estimating the probability for a patient with a specific profile (age, location, and week), ANViDAS and SViDAS aggregate these predictions to provide proportion estimates for populations defined by a municipality, an age range, and an epidemiological week.

2.4 Results

This section presents the results of the evaluation of the proposed methodology for monitoring viral infections in patients with Severe Acute Respiratory Infection in Brazil. Four main objectives will be addressed:

1. To evaluate the performance of the methodology in estimating the distribution of respiratory viruses, comparing it with a standard approach and complete data (Consolidated Data).
2. To analyze the impact of data availability delays on the estimates.
3. To investigate the behavior of the methodology in scarce data scenarios.
4. To interpret the model, highlighting the importance of covariates and analyzing the influence of key predictor variables, particularly age, on the predicted outcome.

The analyses were conducted using data released weekly between January 2022 and May 2025, referring to the years 2022, 2023, and 2024.

[Table 2](#) presents the observed virus distribution for the period from 2022 to 2024, based on the data released on May 5, 2025. A predominance of SARS-CoV-2 is observed, accounting for

approximately 44.7% of the total positive cases. Following this are Respiratory Syncytial Virus (RSV), Rhinovirus, and Influenza A, with 22.1%, 13.9%, and 8.7% of detections, respectively. Together, these four viral agents account for almost 90% of the occurrences in the analyzed period.

Table 2 – Observed virus proportion for the period from 2022 to 2024, based on SIVEP-Gripe data released by the Brazilian Ministry of Health on May 5, 2025.

Virus	Proportion
SARS-CoV-2	0.4469
RSV	0.2212
Rhinovirus	0.1388
Influenza A	0.0868
Adenovirus	0.0276
Metapneumovirus	0.0237
Other Viruses	0.0160
Influenza B	0.0158
Parainfluenza 3	0.0109
Bocavirus	0.0083
Parainfluenza 1	0.0018
Parainfluenza 4	0.0013
Parainfluenza 2	0.0010

For the comparative analysis, we evaluate the proposed **ViDAS framework** (NViDAS, ANViDAS, and SViDAS, as defined in [subsection 2.3.4](#)) against a standard model, referred to here as the **Naive Model**. This standard model is based on the empirical proportion of each virus by epidemiological week and State. Using only these variables, its main limitation is the inability to incorporate the granularity of age or municipality, producing a single estimate for each state per week.

The following subsections detail the analyses and discussions for each of the proposed objectives.

2.4.1 Performance Evaluation of the Proposed Methodology

This section aims to evaluate the performance of the proposed models in estimating age-stratified viral proportions. Given this emphasis, the evaluation focuses on the ANViDAS and SViDAS models, which are capable of generating estimates for age intervals, a functionality not supported by the NViDAS model. The performance is investigated through two complementary approaches: a graphical analysis for a case study and a systematic quantitative evaluation of estimation errors.

The evaluation was designed to compare the performance of the proposed models (ANViDAS and SViDAS) and a Naive Model, using the Consolidated Data as a reference.

The **Consolidated Data**, representing the most accurate available epidemiological information, was constructed using the most complete and recent database (released on May 5, 2025). For the graphical analysis, 95% confidence intervals for the weekly proportions were calculated using the Clopper-Pearson method, an exact binomial approach for proportions ([CLOPPER; PEARSON, 1934](#)).

The graphical analysis focuses on the municipality of São Paulo during the first 26 epidemiological weeks of 2022. The choice of a metropolis like São Paulo is strategic, as the volume of data allows for the construction of a robust Consolidated Data for each analyzed stratum. The estimates from the ANViDAS, SViDAS, and Naive models were generated using data available on July 4, 2022 (epidemiological week 27), simulating a nowcasting scenario. The analysis was segmented by age ranges (overall, < 2, 2-5, 5-15, 15-50, 50-65, and 65+ years) for the SARS-CoV-2 and RSV viruses, selected due to their high incidence.

The quantitative evaluation complements the graphical analysis by simulating the operational use of the models. For this, the models were retrained with each new data release between January 1, 2022, and May 5, 2025, and their estimates were evaluated for the epidemiological weeks within the 2022-2024 period. For each data release, a new set of estimators was generated and used to predict the 14-week window preceding its creation date; this window focuses on where data is most incomplete and subject to updates. Performance was measured using the Brier Score ([BRIER, 1950](#)). The Brier score is a proper scoring rule that measures the quality of probabilistic predictions by calculating the mean squared difference between the predicted probabilities and the actual outcomes (1 for positive, 0 for negative) ([IZBICKI, 2025](#)); see [Appendix F](#) for details. This metric has significance to assessing model calibration, with scores ranging from 0 to 1, where 0 indicates a perfect score and 1 represents the most unfavorable score. The estimates were compared against the individual outcomes from the Consolidated Data. To ensure comparability, imputation strategies were adopted for cases of missing data: for ANViDAS, the midpoint of the age range was used to generate the estimate; for the Naive Model, in the rare event of a total lack of data for a State and week, a discrete uniform distribution was assumed for the virus distribution.

Graphical Analysis: Viral Dynamics and Model Performance

[Figure 2](#) and [Figure 3](#), which present the viral proportions for SARS-CoV-2 and RSV, reveal markedly distinct temporal and demographic dynamics. A clear negative correlation is observed between the proportions of the two viruses, where the increase of one coincides with the decrease of the other. This dynamic is a consequence of the proportional data structure and the predominant incidence of these two agents during the period, reflecting the circulation of the viruses in different subpopulations and time frames, rather than direct viral competition.

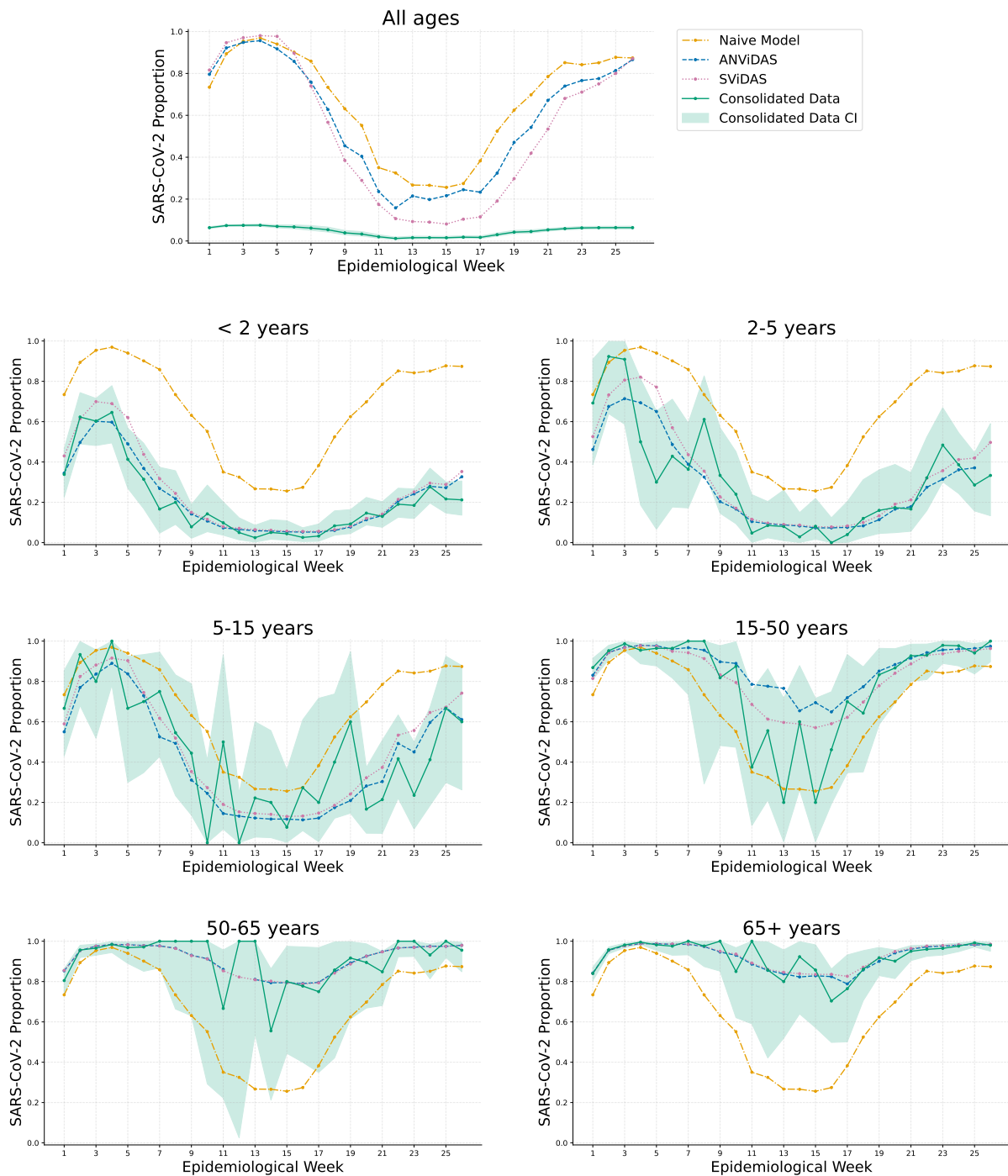


Figure 2 – Estimated weekly proportion of SARS-CoV-2 among virus-positive SARI cases in São Paulo for the first 26 epidemiological weeks of 2022, stratified by age group. The figure compares the estimates from the proposed models (**ANViDAS** and **SViDAS**) and the Naive Model with the observed proportions from the Consolidated Data. Shaded areas represent the 95% confidence interval for the Consolidated Data. Both **ANViDAS** and **SViDAS** estimate the Consolidated Data more accurately than the Naive Model.

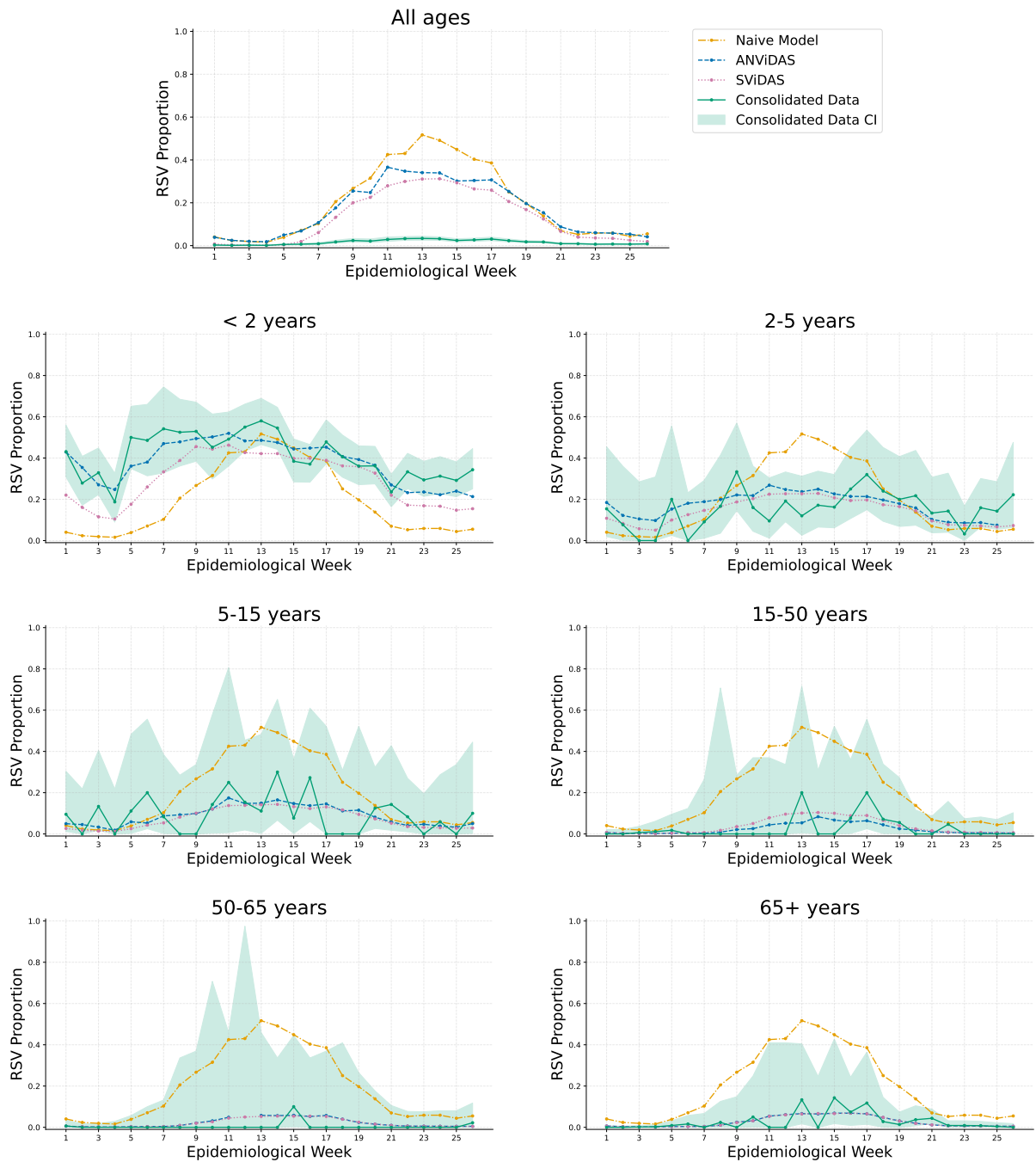


Figure 3 – Estimated weekly proportion of RSV among virus-positive SARI cases in São Paulo for the first 26 epidemiological weeks of 2022, stratified by age group. The figure compares the estimates from the proposed models (**ANViDAS** and **SViDAS**) and the Naive Model with the observed proportions from the Consolidated Data. Shaded areas represent the 95% confidence interval for the Consolidated Data. Both **ANViDAS** and **SViDAS** estimate the Consolidated Data more accurately than the Naive Model.

The proposed models, ANViDAS and SViDAS, demonstrate consistently superior performance compared to the Naive Model, with estimates that closely track the Consolidated Data curve, often falling within its confidence interval. SViDAS, by modeling age continuously, generates notably smooth curves, proving robust in scenarios of data scarcity where the Consolidated Data exhibits high volatility and wide confidence intervals due to a low number of notifications. ANViDAS, based on the empirical distribution, also produces regular estimates, although it more directly reflects weekly fluctuations and cannot generate predictions in the absence of observations—a limitation evidenced in the graphs for week 12 in the 50-65 age range (Figure 2 and Figure 3).

As expected, the Naive Model showed systematic and age-dependent errors. The visual analysis confirms that the model overestimates the proportion of RSV in adults and the elderly and underestimates it in the pediatric population, while the opposite occurs for SARS-CoV-2. Age stratification also reveals critical insights, such as the predominance of RSV in the pediatric population and the persistence of SARS-CoV-2 as the agent with the highest proportion among the elderly, even during periods of low overall circulation.

Quantitative Analysis: Validation of Estimation Errors

The quantitative evaluation, summarized in Table 3 and Table 4, corroborates and deepens the findings of the graphical analysis. Table 3 confirms the consistent superiority of the proposed models for all 13 viruses, with significant error reductions for the most incident agents. For SARS-CoV-2, for example, the average Brier Score drops by more than 40% compared to the Naive Model. Additionally, the ANViDAS and SViDAS models show very similar performance, with ANViDAS having a marginal advantage in most cases, suggesting that both approaches to handling age distribution (empirical vs. modeled) are effective. The reduction in the standard deviation (SD) for the most prevalent viruses also indicates that the predictions of the proposed models are not only more accurate but also more reliable over time.

Table 4, in turn, details the source of this improvement by stratifying the error by age range. The analysis of RSV best illustrates the inadequacy of the Naive Model, which shows a very high error (0.2549) in the 0-2 age range, whereas the proposed models drastically reduce it (to ≈ 0.18). Even for adults and the elderly, where RSV is rare, the proposed models reduce the Naive Model's error from ≈ 0.08 to ≈ 0.03 . For SARS-CoV-2, the table quantifies the Naive Model's systematic error in the elderly population (65+ years), where it reaches its highest value (0.1897), an error that the proposed models correct, reducing it to approximately 0.102. For Rhinovirus and Influenza A, the proposed models also outperform the Naive Model across all age ranges, with the largest error for Rhinovirus occurring between 2 and 15 years, a pattern that ANViDAS and SViDAS capture more accurately. These results quantify the systematic errors and validate the importance of age stratification.

Table 3 – Mean Brier Score (and standard deviation) for each virus. The lowest error for each virus is highlighted in bold.

Virus	Naive Model	ANViDAS	SViDAS
	Mean (SD)	Mean (SD)	Mean (SD)
SARS-CoV-2	0.1703 (0.0840)	0.0997 (0.0489)	0.1011 (0.0499)
RSV	0.1299 (0.0879)	0.0825 (0.0691)	0.0840 (0.0708)
Rhinovirus	0.1252 (0.0837)	0.1025 (0.0657)	0.1032 (0.0661)
Influenza A	0.1021 (0.0876)	0.0811 (0.0608)	0.0817 (0.0610)
Adenovirus	0.0308 (0.0337)	0.0287 (0.0314)	0.0290 (0.0317)
Metapneumovirus	0.0238 (0.0209)	0.0226 (0.0200)	0.0227 (0.0201)
Influenza B	0.0211 (0.0357)	0.0187 (0.0297)	0.0189 (0.0302)
Other viruses	0.0211 (0.0142)	0.0200 (0.0136)	0.0201 (0.0136)
Parainfluenza 3	0.0102 (0.0100)	0.0096 (0.0097)	0.0097 (0.0098)
Bocavirus	0.0091 (0.0103)	0.0085 (0.0099)	0.0086 (0.0100)
Parainfluenza 1	0.0023 (0.0022)	0.0021 (0.0021)	0.0022 (0.0021)
Parainfluenza 4	0.0018 (0.0024)	0.0016 (0.0023)	0.0017 (0.0023)
Parainfluenza 2	0.0016 (0.0018)	0.0015 (0.0016)	0.0015 (0.0017)

Table 4 – Mean Brier Score (and standard deviation) for the four most incident viruses, stratified by age range. The lowest error for each scenario is highlighted in bold.

Age Range	Virus	Naive Model	ANViDAS	SViDAS
		Mean (SD)	Mean (SD)	Mean (SD)
< 2 years	Influenza A	0.0412 (0.0251)	0.0296 (0.0149)	0.0301 (0.0152)
	Rhinovirus	0.1231 (0.0463)	0.1164 (0.0434)	0.1174 (0.0437)
	SARS-CoV-2	0.1722 (0.1072)	0.0802 (0.0380)	0.0814 (0.0389)
	RSV	0.2549 (0.0561)	0.1821 (0.0256)	0.1873 (0.0264)
2-5 years	Influenza A	0.0690 (0.0400)	0.0625 (0.0359)	0.0626 (0.0358)
	Rhinovirus	0.1985 (0.0388)	0.1685 (0.0299)	0.1694 (0.0299)
	SARS-CoV-2	0.1627 (0.0982)	0.0782 (0.0411)	0.0789 (0.0419)
	RSV	0.1642 (0.0579)	0.1454 (0.0505)	0.1475 (0.0510)
5-15 years	Influenza A	0.1199 (0.0613)	0.1032 (0.0523)	0.1040 (0.0522)
	Rhinovirus	0.2397 (0.0399)	0.1747 (0.0255)	0.1759 (0.0252)
	SARS-CoV-2	0.1579 (0.0797)	0.0958 (0.0445)	0.0980 (0.0466)
	RSV	0.1102 (0.0612)	0.0759 (0.0372)	0.0769 (0.0376)
15-50 years	Influenza A	0.1371 (0.0992)	0.1059 (0.0629)	0.1067 (0.0632)
	Rhinovirus	0.0812 (0.0513)	0.0720 (0.0451)	0.0728 (0.0458)
	SARS-CoV-2	0.1666 (0.0508)	0.1269 (0.0405)	0.1290 (0.0415)
	RSV	0.0846 (0.0631)	0.0295 (0.0235)	0.0296 (0.0237)
50-65 years	Influenza A	0.1354 (0.1111)	0.1009 (0.0699)	0.1012 (0.0702)
	Rhinovirus	0.0586 (0.0442)	0.0473 (0.0360)	0.0475 (0.0362)
	SARS-CoV-2	0.1725 (0.0665)	0.1160 (0.0514)	0.1163 (0.0517)
	RSV	0.0837 (0.0655)	0.0322 (0.0263)	0.0323 (0.0264)
65+ years	Influenza A	0.1102 (0.1032)	0.0848 (0.0684)	0.0857 (0.0691)
	Rhinovirus	0.0501 (0.0418)	0.0358 (0.0318)	0.0363 (0.0321)
	SARS-CoV-2	0.1897 (0.0871)	0.1015 (0.0562)	0.1029 (0.0574)
	RSV	0.0819 (0.0659)	0.0301 (0.0288)	0.0303 (0.0290)

Conclusion of the Analysis

In summary, the combined evaluation validates the ANViDAS and SViDAS methodologies, demonstrating their ability to reveal and quantify complex viral dynamics that are invisible to non-stratified approaches. The precision in identifying specific demographic patterns, such as the predominance of RSV in children and the persistence of SARS-CoV-2 in the elderly, are examples of critical insights for epidemiological surveillance that can only be realistically provided by models that incorporate age stratification.

2.4.2 Analysis of the Effect of Delays on Estimates

In this subsection, we evaluate the impact of inherent data notification delays on the performance of the estimates generated by the proposed methodology. To this end, we conducted an experiment that simulates the operational use of the NViDAS model over time, mimicking the challenge of nowcasting, in which predictions are made with partial information that is later consolidated. The analysis was based on weekly data releases from January 2022 to May 2025. For each release, a new NViDAS estimator was trained, reflecting the knowledge available at that specific point in time. The performance of the estimates for each epidemiological week was evaluated as a function of the “lag” defined as the difference in weeks between the data release date and the epidemiological week for which the estimate was generated.

To ensure a robust analysis of the most critical period, where data is incomplete and subject to updates, the analysis focused on a retrospective window of up to 14 weeks of lag, and only epidemiological weeks with estimates for at least seven (including the first) of the possible 14 lags were included. Performance was measured using the Brier Score, comparing the predicted probabilities against the individual outcomes from the consolidated data of May 5, 2025. To facilitate comparison across different epidemiological weeks, the Brier Scores for each virus were standardized (Min-Max scaled) for each week. The conceptual design of this evaluation is illustrated in [Figure 4](#).

The results, presented in [Figure 5](#), show the average behavior of the Standardized Brier Scores for each lag for the four most incident viruses: SARS-CoV-2, RSV, Rhinovirus, and Influenza A. A consistent pattern is observed across Brazil and its five macro-regions: a sharp reduction in estimation error occurs within the first few weeks, followed by a stabilization. This rapid improvement in performance may not only reflect the increase in data volume but also a potential hierarchy in laboratory testing processes, where cases with higher clinical probability are confirmed first, and results from less probable cases are subsequently incorporated into the surveillance system. As more data is consolidated, this initial bias is reduced, and the estimates become more precise.

The line graphs in [Figure 5](#) indicate that, for the main viruses, the estimation error tends to stabilize after a lag of approximately 3 to 5 weeks. [Table 5](#) quantifies this stabilization point,

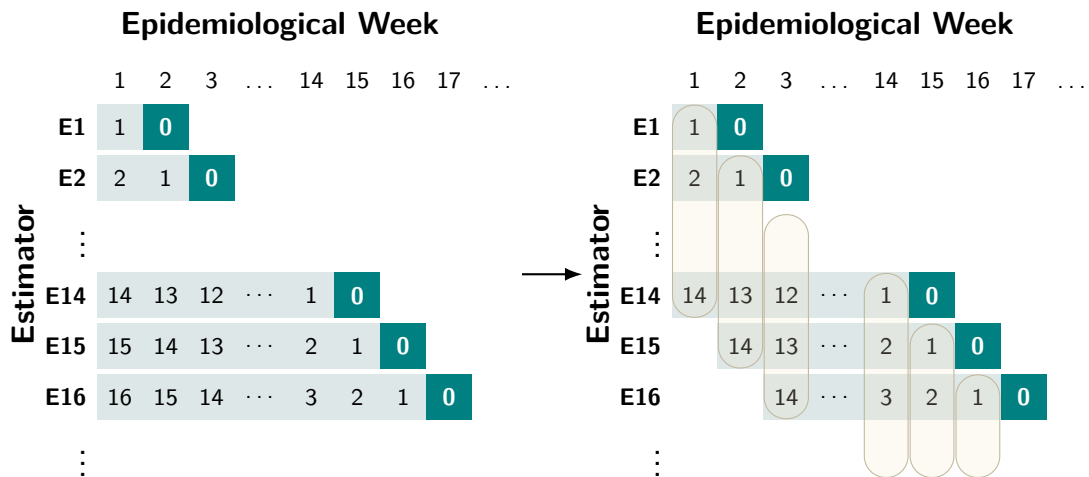


Figure 4 – Conceptual diagram of the delay analysis methodology. The diagram on the left illustrates the estimation process, where each row (E1, E2,...) represents an estimator trained on a new weekly data release. The columns are epidemiological weeks, and the numbers within the cells indicate the lag (in weeks) of each estimate, with lag 0 highlighted as the week of the data release. The diagram on the right represents the final dataset used for the analysis, which is obtained after applying the following steps: (1) estimates with a lag greater than 14 weeks are filtered out; (2) only epidemiological weeks with at least seven estimates, including the one at lag 1, are retained; and (3) the resulting Brier scores are standardized for each epidemiological week.

highlighting for each virus and region the smallest lag at which the bootstrapped 95% confidence interval (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) of the error estimate overlaps with that of the 14-week lag, suggesting that the error has reached a stable level of reliability. For Brazil as a whole, for example, estimates for SARS-CoV-2 and Rhinovirus stabilize at a lag of 5 weeks, while Influenza A stabilizes earlier, at 4 weeks.

Table 5 – Minimum Stabilization Lag for Main Viruses by Region. The value represents the minimum lag at which the estimate's performance is considered stable. Stability is defined as the first point where the 95% confidence interval of the standardized Brier score for a given lag overlaps with the 95% CI of the 14-week lag.

Region	Virus			
	SARS-CoV-2	RSV	Rhinovirus	Influenza A
Brazil	5	8	5	4
North	3	4	4	4
Northeast	5	6	5	3
Central-West	3	3	3	2
Southeast	5	6	4	3
South	3	3	4	3

Furthermore, it is generally observed that the mean standardized score reaches a level between 0.2 and 0.4, on a scale of 0 to 1, from lag 3 in nearly all cases. The minimum value for this stabilized mean error is typically observed in the 0.1 to 0.3 range. This demonstrates that despite the challenges of data delay, the methodology provides robust and reliable estimates for

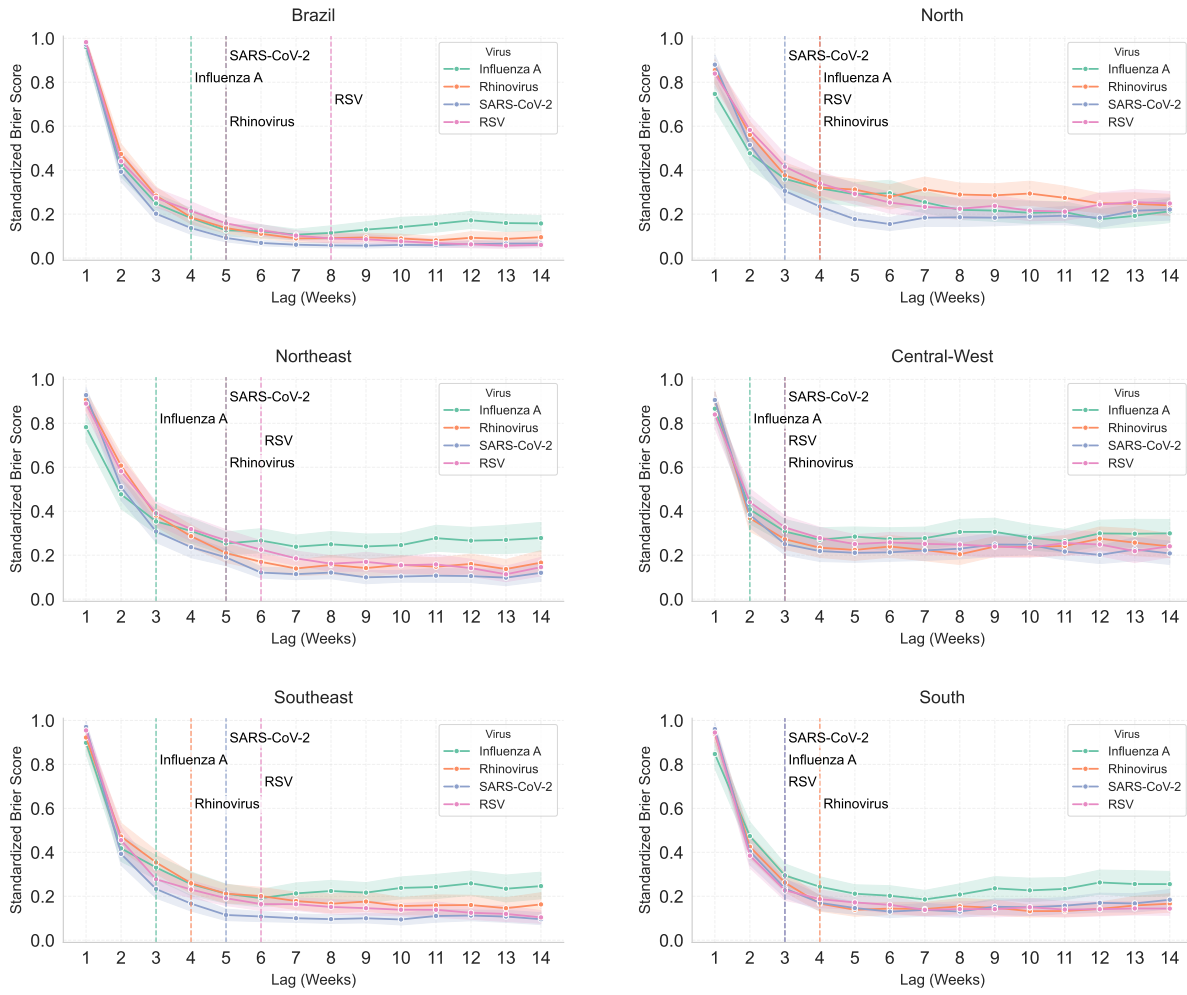


Figure 5 – Impact of data reporting lag on the performance of virus estimates, measured by the Standardized Brier Score. Each panel shows the mean score (solid line) and its 95% confidence interval (shaded area) as a function of the lag in weeks for the four most incident viruses. The analysis is presented for Brazil as a whole and for its five macro-regions. The vertical dashed lines mark the stabilization point for each virus, corresponding to the values in Table 5. A consistent pattern is observed across all regions, with a sharp improvement in performance (decrease in error) within the first 3-5 weeks, after which the estimates reach a stable level of reliability.

recent epidemiological weeks, a critical feature for effective and timely public health surveillance.

2.4.3 Performance in Data-Scarcity Scenarios

This subsection evaluates the performance of the proposed methodology in a data-scarcity scenario, a common challenge in epidemiological surveillance. To simulate this condition, we conducted an analysis using all weekly datasets from 2022 to 2024, released between January 2022 and May 2025. For each dataset released, the models were trained on a small random subsample of 10% of the data and evaluated on the remaining 90%. The performance of our NViDAS methodology was compared against the Naive Model, with the Brier Score used as the evaluation metric to assess the precision of the probabilistic estimates.

The results of this analysis are summarized in Figure 6. The findings indicate that the NViDAS model consistently outperforms the Naive Model, particularly for the most incident viruses. As shown by the lower Brier scores, the advantage is most pronounced for SARS-CoV-2 and RSV, where NViDAS provides more precise estimates and exhibits less variance than the Naive Model. Conversely, for viruses with very low incidence, such as the Parainfluenza group (PARA1-4) and Bocavirus (BOCA), the performance difference between the two models is less evident, as both achieve very low error scores.

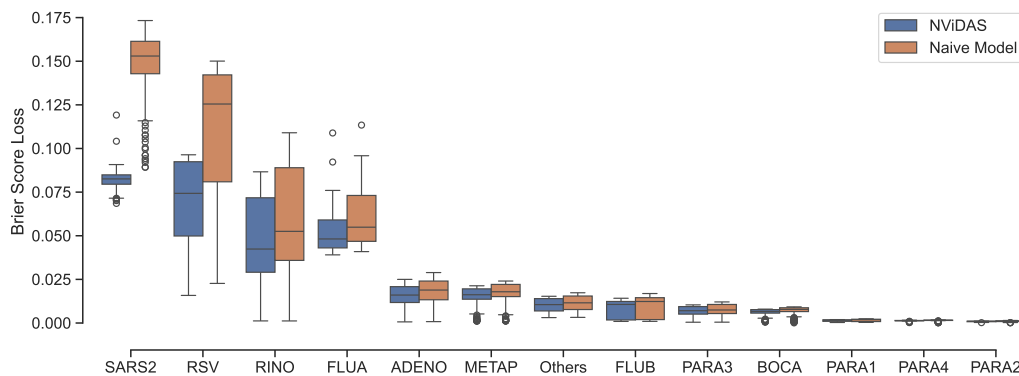


Figure 6 – Performance comparison between the **NViDAS** and the Naive Model in a data-scarcity scenario. The models were trained on 10% of the available data and evaluated on the remaining 90%. The boxplots show the distribution of Brier Score Loss for each virus, where lower scores indicate better performance.

In conclusion, this analysis demonstrates the robustness of the proposed methodology. By leveraging relationships among covariates, NViDAS can generate more reliable estimates than simpler models even when trained on limited data, making it a valuable tool for surveillance in data-scarcity settings.

2.4.4 Interpretation and Explanation of the NViDAS Model

This subsection aims to interpret the NViDAS model, highlighting the importance of predictor variables and analyzing the influence of key features, particularly age and epidemiologic week, on the predicted outcome. The analysis was conducted using the most recent estimator, trained on the dataset released in May 2025.

Analysis of Predictor Variable Importance

First, we evaluated the global importance of each predictor variable by quantifying its total gain, which represents the overall improvement in prediction accuracy contributed by that feature across all the branches where it is used in a tree-based model such as GBDT (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). The analysis reveals a clear hierarchy among the variables, as shown in Figure 7. Patient's Age emerges as the most influential predictor by a substantial margin, followed by the Epidemiological Week of the first symptoms. Geographical variables exhibit

a tiered importance corresponding to their granularity, with city-level coordinates being more influential than state-level, which in turn are more important than region-level coordinates. This pattern indicates that the model effectively captures localized viral transmission dynamics. The Population Size of the notifying municipality also ranks as a significant predictor, underscoring the role of urban density in disease transmission.

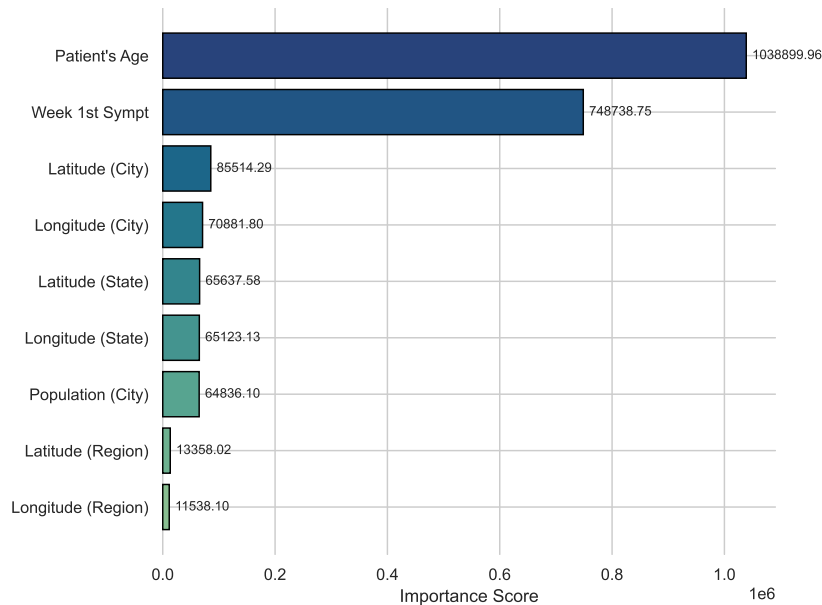


Figure 7 – Predictor Variable Importance based on Total Gain. The chart illustrates the relative importance of each feature, with Patient's Age and Epidemiological Week emerging as the most influential predictors.

Visualizing Predictor Effects with PDP and ICE Plots

To visualize how the most influential variables affect the model's outputs, we generated Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) plots (FRIEDMAN, 2001; GOLDSTEIN *et al.*, 2015), using data not included in the model training. The PDP illustrates the average effect of a feature on the predicted outcome, computed using a random sample of 1000 observations. To reveal the heterogeneity of this effect, ICE plots show the prediction for individual instances; for visual clarity, these plots were generated for a subsample of 100 observations.

Figure 8 presents these visualizations for the four main viruses, revealing distinct and clinically recognized epidemiological profiles.

- **SARS-CoV-2:** The model learned a strong positive correlation between age and the probability of SARS-CoV-2. The predicted likelihood is low for infants and increases steadily with age, reaching its peak in the elderly population (65+ years). This age-dependent risk profile is further corroborated by the analysis of temporal trends. During periods of high viral circulation, the epidemiological week plot consistently shows that the

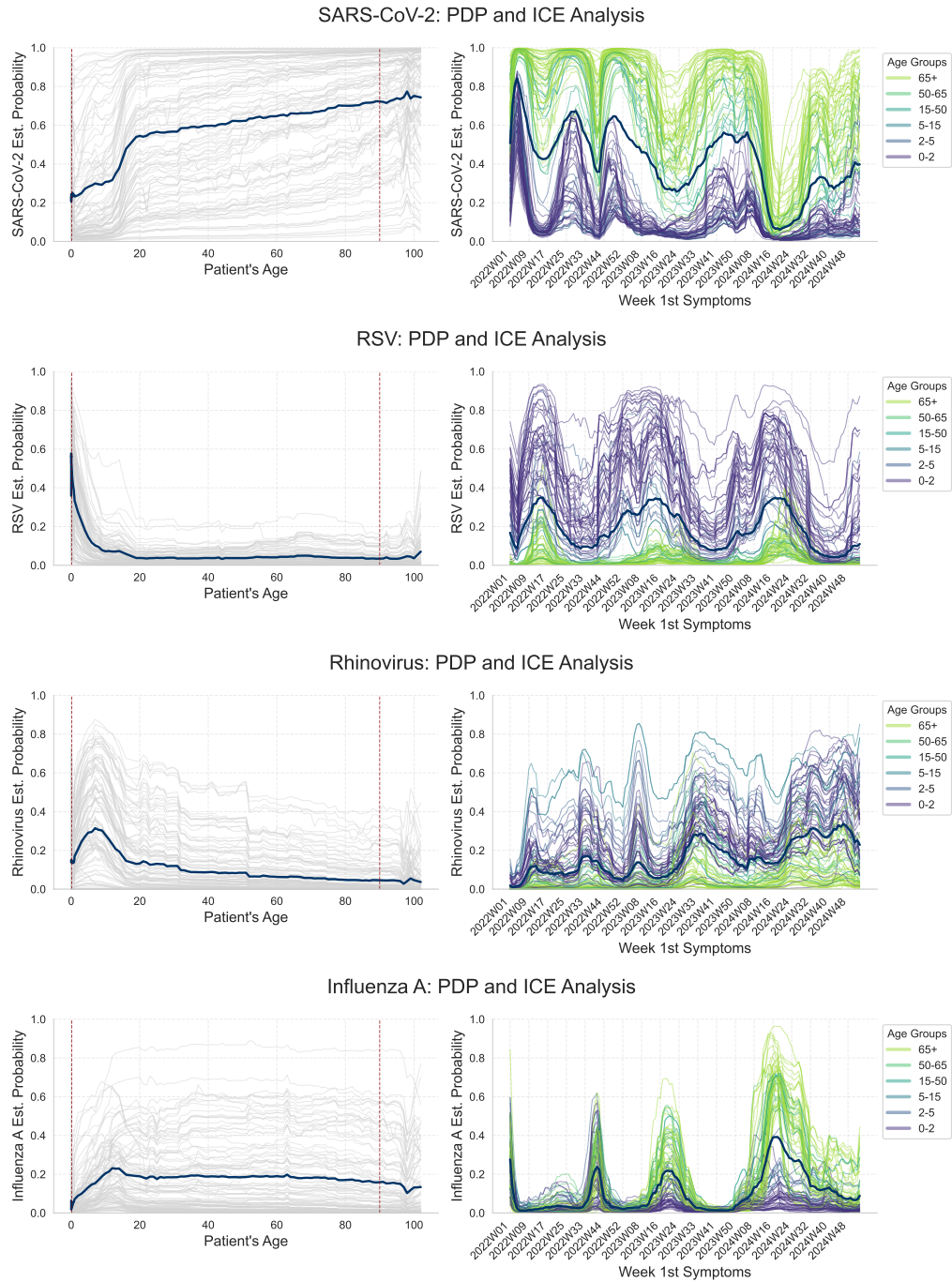


Figure 8 – Learned effects of patient age and epidemiological week on the estimated probabilities for the four main viruses. For each virus, the left panel displays the effect of age, with the bold line representing the Partial Dependence Plot (PDP), faint lines showing Individual Conditional Expectation (ICE) plots, and red dashed vertical lines indicating the 5th and 95th percentiles of observed patient age, which delimit the central range of the empirical age distribution and highlight age regions with fewer observations. The right panel displays the effect of the epidemiological week, where the bold line is the PDP, and the colored ICE plots illustrate the interaction with different age groups as detailed in the legend. The plots reveal that the model has successfully learned distinct and clinically recognized epidemiological profiles for each virus.

probability curves for older age groups (light green) are positioned at the top, confirming their heightened susceptibility.

- **Respiratory Syncytial Virus (RSV):** In stark contrast to SARS-CoV-2, the model identified a powerful negative correlation between age and RSV probability. The likelihood of infection is exceptionally high for neonates and infants—a vulnerability that drops sharply within the first few years of life before stabilizing at a low baseline for adolescents and adults. This finding aligns with established RSV epidemiology, where primary infections are most severe in the immunologically naive pediatric population. The epidemiological week plot reinforces this age stratification, showing that pediatric age groups (dark purple and blue) consistently dominate the seasonal peaks of RSV activity.
- **Rhinovirus:** The model captured a more complex, non-linear relationship for Rhinovirus, with the probability peaking in early-to-mid childhood (approximately 5-15 years) before gradually declining through adulthood. This pattern likely reflects the combined effects of intense social mixing in school environments and the gradual acquisition of immunity to numerous Rhinovirus serotypes over a lifetime. The corresponding epidemiological week plot supports this observation, showing a clear stratification where the probability curves for pediatric and adolescent groups remain generally higher than those for adults across different seasons.
- **Influenza A:** While the predicted probability of Influenza A shows a slight increase during childhood before leveling off, the time-series plot reveals an extremely strong seasonal signal, with sharp, distinct peaks across all age groups. This indicates that for Influenza A, the epidemiological week is a more dominant predictor than the patient's age.

These visualizations confirm that the model has not simply identified age and time as important predictors, but has learned how their interaction shapes the probability of each virus, capturing distinct epidemiological signatures. In practice, these patterns corroborate the empirical observation that age-stratified weekly SARI trends can provide clues about the predominant virus: increases concentrated among infants are more compatible with RSV, increases among children and adolescents with Rhinovirus, increases concentrated among the elderly with SARS-CoV-2, and increases observed across most age groups with Influenza A.

Local Interpretation in a Zero-Data Scenario

To demonstrate the model's decision-making process in a data-scarcity scenario, we analyzed its predictions for hypothetical cases in Araporã, a municipality with a small population (approx. 8,500) located in the Triângulo Mineiro region of Minas Gerais, for which no SARI cases were recorded in 2024 after applying the study's filters. We investigated two distinct profiles for epidemiological week 22 of 2024.

Case 1: Influenza A in a 30-Year-Old Adult. For a hypothetical 30-year-old patient, the model assigned a high probability of 78.0% to Influenza A. To explain this, we applied the nearest-neighbor analysis in the model’s embedding space, as described in Section 2.2.4. The 10 most similar training instances were from four nearby municipalities within the same state (Minas Gerais). Within this group, the observed proportion of Influenza A was 70.0%. A rule-based filter, defined by the minimum and maximum observed values for the neighbors’ covariates (age: 15-83 years, population: 21k-102k, epidemiological weeks: 2024W22-2024W30, and geographic location in the Triângulo Mineiro region; see [Figure 9](#)), identified a slightly larger group of 13 cases, where the observed proportion of Influenza A was 76.9%, closely matching the model’s prediction.

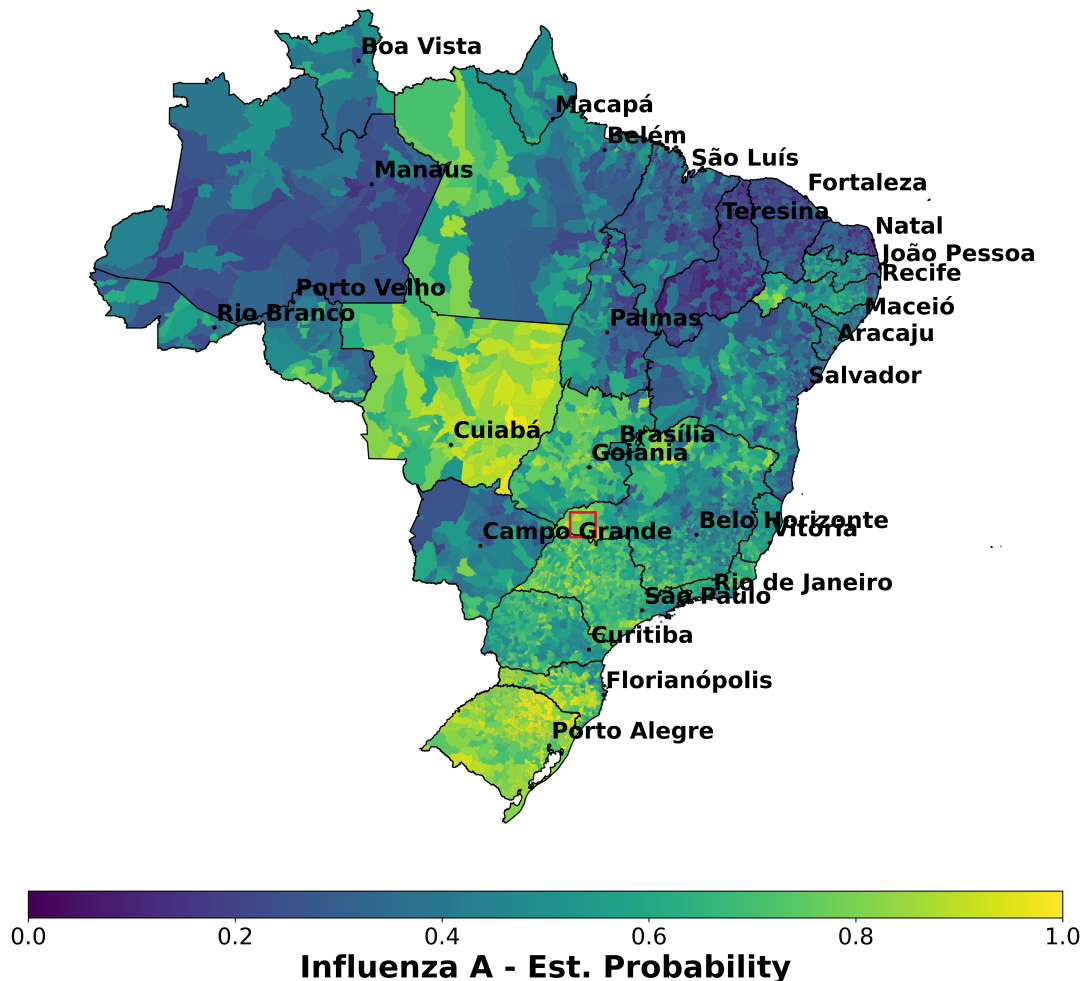


Figure 9 – Spatial Individual Conditional Expectation (ICE) plot showing the predicted probability of Influenza A across geographic coordinates for a hypothetical 30-year-old patient. The overlaid rectangle delineates the spatial boundaries derived from the min-max rule applied to the nearest neighbors in the model’s embedding space. The model anchors its high-probability prediction for this zero-data municipality by identifying a highly localized spatial cluster of similar adult cases within the Triângulo Mineiro region.

Case 2: Respiratory Syncytial Virus (RSV) in a Neonate. For a hypothetical neonate

(age 0.01 years), the model predicted a 54.1% probability of RSV. The nearest-neighbor analysis identified a different set of influential cases: 10 infants from six other cities in Minas Gerais, where the observed RSV proportion was 60.0%. The corresponding rule-based filter, derived from the neighbors' characteristics (age: approx. 1 month, population: 81k-713k, epidemiological weeks: 2024W22-2024W23, and a geographic location in a wider region containing the Triângulo Mineiro and other areas in central and southern Minas Gerais, see [Figure 10](#)), identified a larger set of 23 infants, in which the observed RSV proportion was 52.2%, again aligning closely with the model's prediction.

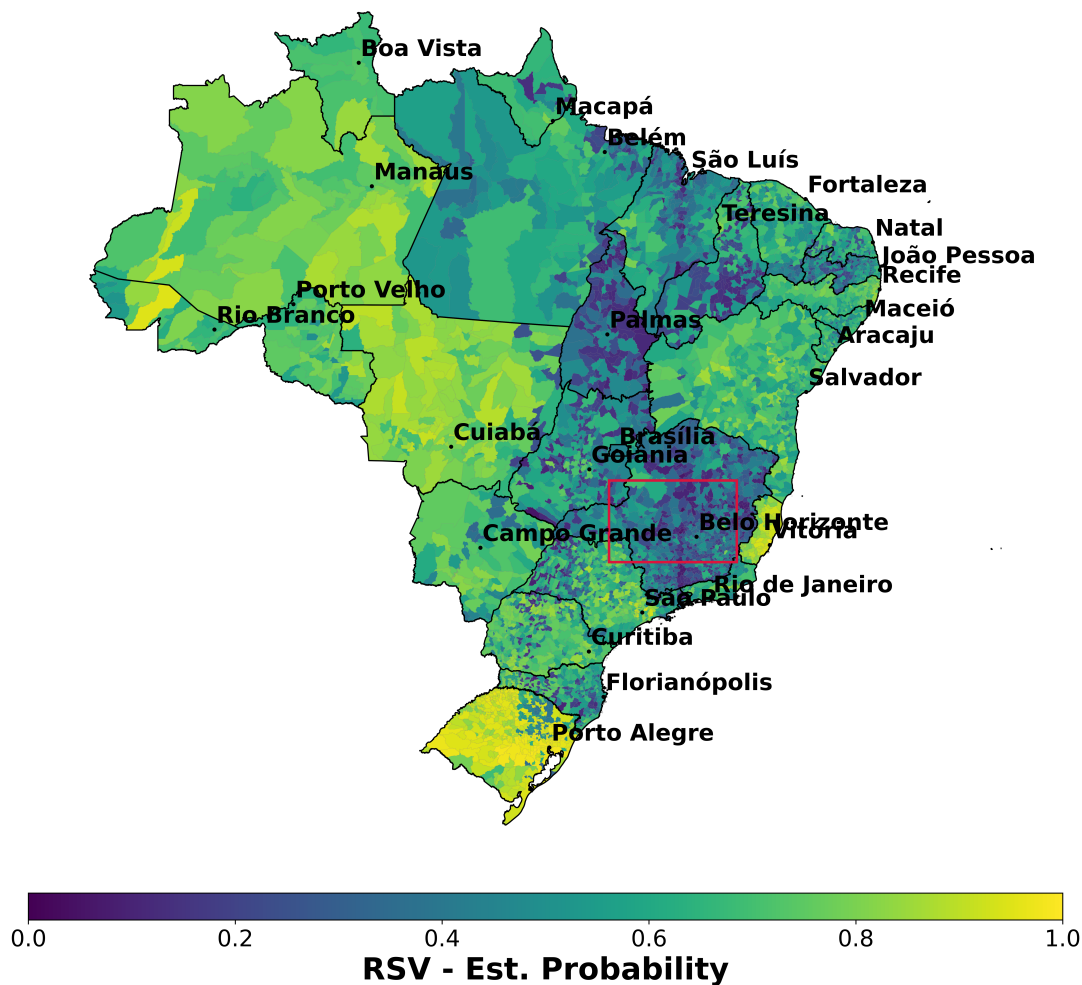


Figure 10 – Spatial Individual Conditional Expectation (ICE) plot illustrating the predicted probability of Respiratory Syncytial Virus (RSV) across geographic coordinates for a hypothetical neonate (0.01 years old). The bounding rectangle represents the spatial filter defined by the most influential nearest neighbors. In contrast to the highly localized adult profile in Case 1, the model draws upon a broader regional area across central and southern Minas Gerais. Notably, rather than showing a dense concentration of high-probability hotspots within this bounding box, the plot reveals a more diffuse and generalized distribution of risk.

This case study reveals that in the absence of direct evidence, the model bases its predictions on data from demographically, temporally, and geographically similar profiles. It correctly inferred that the regional trend for an adult pointed to Influenza A, while for a

neonate, the dominant risk was RSV, showcasing its ability to produce nuanced, age-appropriate predictions even in a zero-data context.

2.5 Discussion

Interpretation of Main Findings

This study presented a computational framework for estimating the granular distribution of respiratory viruses among SARI patients in Brazil. Our results indicate that modeling viral proportions as conditional probabilities, using GBDT, significantly outperforms traditional empirical aggregation methods. Specifically, the proposed models (ANViDAS and SViDAS) achieved a reduction in estimation error (Brier Score) of approximately 40% for SARS-CoV-2 and 36% for RSV compared to the benchmark.

These findings directly address the research objective of efficiently monitoring predominant viruses across municipalities and age groups. The analysis of feature importance revealed that patient age is the most critical predictor of viral etiology, followed by the epidemiological week. This confirms established biological understandings — such as the high susceptibility of infants to RSV and the elderly to SARS-CoV-2 — but extends current knowledge by quantifying these risks continuously across the lifespan. The Partial Dependence Plots demonstrated that the model successfully learned these non-linear biological relationships without explicit programming, distinguishing the sharp decline of RSV risk in early childhood from the monotonic increase of SARS-CoV-2 risk with age.

Comparison with Related Work and State-of-the-Art

Our approach represents a methodological shift from aggregate-level surveillance to patient-level estimation. While systems like InfoGripe rely on Bayesian frameworks to correct incidence counts and aggregate viral proportions over large geographical areas, our method preserves local heterogeneity. By treating geolocation as continuous covariates (latitude and longitude), the model effectively learns data-driven neighborhoods. This overcomes the limitations of traditional spatiotemporal models (e.g., GLMMs or INLA), which often require pre-specified adjacency matrices and can be computationally prohibitive for high-dimensional real-time systems.

Furthermore, contrasting with the ITpS approach which utilizes private laboratory data, our methodology focuses exclusively on hospitalized SARI cases. This is crucial because pathogen distribution in severe cases often differs from that in spontaneous testing demands. The ability of SViDAS to integrate viral probabilities over continuous log-normal age distributions constitutes a novel semi-parametric contribution. This allows for precise estimation in arbitrary age intervals (e.g., neonates vs. toddlers), offering a granularity that fixed-bin aggregation cannot

match. Separately, to address the “black box” limitations often leveled at machine learning models, we introduced an interpretability strategy based on the estimator’s internal embedding space. By identifying the specific training instances that influence a prediction, this approach provides data-driven explanations for the model’s outputs, even in data-scarce contexts.

Methodological Considerations and Robustness

The choice of a GBDT architecture was justified by its ability to handle missing data and capture complex interactions between space, time, and demographics without the need for extensive feature engineering or predefined grouping. Furthermore, GBDT is recognized for its computational efficiency in both training and inference when properly configured. In this study, using a standard notebook (Windows 11 Pro, AMD Ryzen 7 5700U, 16GB RAM), the models for the final epidemiological weeks of 2024—which comprised the largest dataset volume (2022–2024)—required less than two minutes to train. The decision to model the proportion of positive cases rather than total incidence was a strategic response to the absence of negative test records for non-influenza viruses in the SIVEP-Gripe database. This formulation renders the metric robust to variations in testing volume, provided that the reporting lag is uniform across viral agents.

The robustness of this approach is evidenced by the “zero-data” experiments. The case study of Araporã demonstrated that even in the total absence of local observations, the model could leverage the embedding space to identify similar training instances—geographically nearby and demographically comparable—to generate clinically coherent predictions. Additionally, the stability analysis revealed that the model achieves a substantial reduction in estimation error as early as the second week of lag, with the Standardized Brier Score fully stabilizing within 3 to 5 weeks. This suggests the model is a viable tool for “nowcasting” viral proportions, filling the information gap during the period when diagnostic results are still being consolidated.

Age-Density Specification in SViDAS

The SViDAS estimator uses a lognormal working model for the conditional distribution of patient age, chosen to enable computationally efficient estimation of viral proportions over arbitrary age intervals. In the empirical evaluation, SViDAS achieved performance very close to that of ANViDAS, which aggregates predictions from the same virus-probability model using the empirical age distribution. This close agreement suggests that, for the settings and age intervals evaluated in this study, the lognormal specification provides an adequate practical approximation for the final estimates.

At the same time, the age distribution of SARI cases, including its conditional distribution in specific locations and epidemiological periods, may display multimodal patterns. Exploring flexible conditional density estimators or mixture-based models would therefore be a valuable extension, particularly for settings with more complex age structures.

Scope and Interpretation of the Estimates

To ensure the accurate application of this framework, it is essential to consider the context in which these viral proportions are estimated. First, the understanding of viral distribution requires caution due to its sensitivity to testing strategies and resource availability. Comparing distributions across different localities presupposes a uniformity in testing policies, assuming that individuals with similar clinical conditions are tested equivalently in all compared areas. Similarly, contrasting proportions over time within the same locality assumes that testing protocols have remained stable. This implies a stability in the testing ratio for different viruses during the analyzed period, without shifts in the prioritization or exclusion of specific viral targets. Consequently, differential testing protocols may bias the observed viral detections, particularly when specific viral targets are prioritized or omitted (e.g., if hospitals test for SARS-CoV-2 but not RSV). Therefore, the estimates should be interpreted as the relative distribution among detected positive cases under the prevailing testing practices.

Second, the proportion of different viruses among SARI cases should be interpreted as the etiological profile of severe respiratory infections, rather than as a direct measure of viral circulation in the general population. This distinction arises because respiratory viruses differ in their transmission dynamics and in their probability of progressing to severe disease. A less virulent pathogen may be more prevalent in the general population than among hospitalized SARI cases, whereas a pathogen associated with more severe outcomes may be overrepresented among SARI notifications. This severe-disease profile is precisely the intended target for hospital surveillance and resource planning, but it should not be interpreted as a direct proxy for general community circulation.

Computational Considerations

From an operational perspective, while the computational efficiency was optimized for deployment with limited resources, the retraining strategy requires access to the full historical dataset. As the SIVEP-Gripe database grows, the retraining time may eventually necessitate incremental learning strategies or more aggressive data subsampling.

Practical Implications and Future Directions

The practical utility of this framework is demonstrated by its implementation in the PySRAG package and a dedicated web application. For public health authorities, the ability to query viral proportions for specific age (e.g., ‘1 year’) allows for targeted interventions, such as the timing of RSV prophylaxis or pediatric ICU allocation, which are often obscured in aggregate “all-ages” reports.

Future work should focus on integrating this viral proportion estimator with total volume nowcasting models. By combining the estimated proportion of a virus (from ViDAS) with the estimated count of SARI cases (from systems like InfoGripe), it would be possible to estimate

the absolute incidence of specific viruses in real-time. Additionally, incorporating meteorological data could further refine the spatial-temporal capture of seasonal respiratory viruses. Finally, extending the interpretability analysis using the embedding space could provide automated textual explanations for surveillance alerts, increasing trust and adoption among non-technical health officials.

NOWCASTING EPIDEMIOLOGICAL COUNTS

This chapter builds directly upon the epidemiological context established in [Chapter 2](#). While the previous chapter focused on estimating the proportional distribution of respiratory viruses to determine which pathogens are circulating, this chapter addresses the critical operational question of how many severe cases are occurring. To achieve this, we shift our methodological focus from quantifying the relative viral landscape to solving the problem of reporting delays in real-time surveillance. We introduce a novel machine learning framework combined with Conformal Prediction to nowcast SARI hospitalizations, providing absolute case counts with locally adaptive, finite-sample uncertainty bounds designed to be robust against the non-stationary dynamics of epidemic reporting.

3.1 Context

The previous chapter addressed the critical challenge of identifying the etiological landscape of Severe Acute Respiratory Infections (SARI). By developing a machine learning framework to model viral proportions as conditional probabilities, we effectively overcame the limitations imposed by the absence of negative test records in surveillance systems. This approach provided a granular understanding of *which* respiratory viruses—such as SARS-CoV-2, RSV, or Influenza—are predominantly circulating across specific demographic profiles and geographic locations.

However, while estimating the relative distribution of viral agents is essential for guiding targeted epidemiological interventions, it is fundamentally insufficient for the broader operational management of healthcare systems. For effective public health resource planning—such as the timely allocation of intensive care unit (ICU) beds, mechanical ventilators, and medical personnel—decision-makers require absolute metrics. Furthermore, obtaining a realistic absolute count is crucial for accurately defining the current epidemiological stage, enabling health author-

ities to formally identify an outbreak or declare an epidemic wave. Knowing that RSV accounts for the majority of pediatric infections is only actionable if we also know the total volume of admissions. Therefore, it is imperative to estimate precisely *how many* severe cases are occurring in real time.

However, monitoring SARI incidence in real-time is inherently challenged by the phenomenon of reporting delays. In universal surveillance systems such as the Brazilian SIVEP-Gripe, a SARI case is defined syndromically through clinical symptoms, meaning an initial notification does not strictly depend on a laboratory test. Nevertheless, identifying the specific etiological agent still requires laboratory confirmation. Furthermore, Brazilian health regulations, such as Portaria GM/MS No. 10.175 ([Ministério da Saúde do Brasil, 2026a](#); [Ministério da Saúde do Brasil, 2026b](#)), mandate the immediate notification (within 24 hours) of SARI cases associated with Coronaviruses, explicitly encompassing highly pathogenic strains such as SARS-CoV, MERS-CoV, and SARS-CoV-2.

Despite these rigorous regulatory frameworks, the registration of a case is rarely instantaneous in practice. The operational lag between the onset of a patient's symptoms and the administrative digitization of the case into the official SIVEP-Gripe database is driven by a complex combination of logistical bottlenecks, infrastructure limitations, variable times in seeking medical care, and laboratory turnaround times. Consequently, cases occurring in the present may take weeks to be fully consolidated in public records.

From a statistical and operational perspective, this reporting latency introduces severe right-censoring into the most recent epidemiological weeks of the time series. Because a SARI case can be reported several days or even weeks late—especially in regions with limited infrastructure or overburdened healthcare systems—the actual number of recent cases is systematically delayed. Consequently, when public health officials examine the raw surveillance dashboard at any current moment, the incidence curve will consistently appear to exhibit an abrupt downward trend. This artificial drop is a dangerous artifact of the data collection process rather than a true epidemiological decline. Relying on these uncorrected counts can induce a false sense of security, directly impacting the early detection of outbreaks, accurate trend monitoring, critical resource allocation, and the implementation of timely interventions required to contain the disease's spread.

To visually illustrate this operational challenge and the practical objective of this chapter, [Figure 11](#) presents a real-time surveillance scenario. The raw, currently reported cases (dotted red line) exhibit a severe artificial plummet in the most recent epidemiological weeks due to the aforementioned reporting delays. The primary goal of our proposed framework is to accurately bridge the gap between these incomplete, right-censored notifications and the true incidence curve (solid blue line). By generating delay-adjusted model predictions (dotted green line) accompanied by rigorous 95% conformal prediction intervals (shaded green area), the methodology equips decision-makers with a reliable, mathematically bounded expectation of

the actual case volume. This effectively neutralizes the right-censoring artifact before it can misguide public health resource allocation.

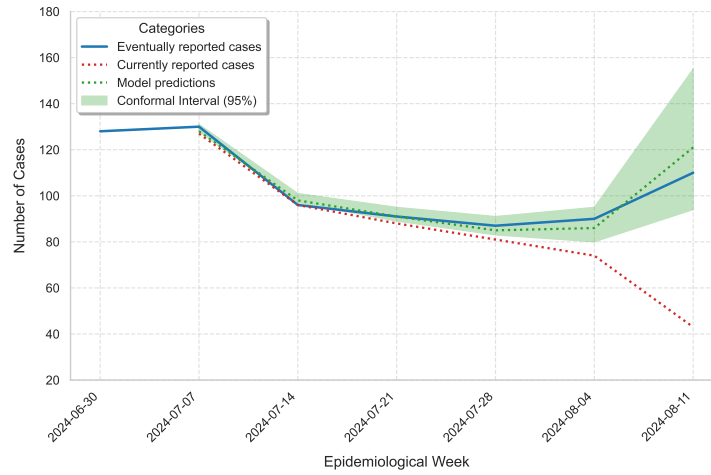


Figure 11 – Practical illustration of the nowcasting objective. The graph highlights the severe right-censoring phenomenon, where currently reported cases artificially drop in recent weeks (dotted red line). The proposed framework corrects this latency by generating nowcast predictions (dotted green line) for the eventually reported cases, which are approximated by the consolidated data released on May 5, 2025 (blue line), and providing a calibrated 95% conformal interval (shaded green area) to support safe epidemiological decision-making.

To address the challenge of real-time SARI nowcasting, the traditional epidemiological literature employs a variety of statistical methods. While simpler models are computationally efficient, they often yield poor predictive performance when dealing with complex reporting dynamics. Consequently, the state-of-the-art heavily relies on spatio-temporal hierarchical Bayesian models. These approaches generally produce highly accurate estimators by rigorously accounting for spatial correlations and temporal trends. However, they are highly susceptible to the curse of dimensionality. As the number of covariates and spatial locations increases—such as modeling hundreds of distinct health regions—these Bayesian frameworks suffer from prohibitive computational costs and convergence issues, limiting their operational viability for real-time, high-granularity surveillance.

To overcome these dimensional limitations, this chapter proposes a novel, highly scalable framework based on Machine Learning (ML). Rather than relying on rigid prior distributions and predefined spatial adjacency matrices, we adopt Gradient Boosting algorithms to estimate reporting delay probabilities at the individual level. Specifically, the highly flexible ML model estimates the probability that a given case is reported within a specific delay, conditioned on individual and municipal features such as the patient’s age, municipal population size, and exact continuous spatial coordinates (latitude and longitude). Given these estimated probabilities, the total number of consolidated cases for each partially observed instance can be modeled under a Negative Binomial distribution, yielding individual-level estimates for expectation and variance. Finally, to achieve the required macro-level insights, these individual estimates are aggregated at the health region level. Utilizing the linearity of expectation, the regional expected count is

simply the sum of the individual expectations. Furthermore, under the assumption that individual reporting events are conditionally uncorrelated given the estimated reporting probabilities, the variance of the regional sum straightforwardly reduces to the sum of the individual variances.

However, in epidemiological surveillance, providing isolated point estimates is insufficient. A point estimate represents only a single expected value—or the most likely outcome—but inherently carries a prediction error that remains unknown if not explicitly modeled. In the context of public health, where administrative decisions directly impact human lives, relying solely on an expectation without understanding its potential deviation is a high-risk strategy. For instance, allocating hospital resources based on a point estimate of 50 required intensive care beds offers no safety margin if the true demand, due to natural epidemiological variance, could realistically surge to 80. Therefore, quantifying the exact magnitude of this uncertainty—knowing the reliable upper and lower bounds of the prediction error—is what provides the necessary safety margin that the healthcare domain demands for effective capacity planning.

To provide this rigorous uncertainty quantification, the core methodological innovation of this chapter is the integration of Conformal Predictors. Using the aggregated expectation and variance at the health region level, we apply conformal inference to generate rigorously calibrated prediction intervals. While classical conformal guarantees rely on strict data exchangeability—an assumption inherently violated by the non-stationary, temporal dynamics of an epidemic—our framework is specifically designed to be robust against these macroscopic distribution shifts. By employing locally adaptive non-conformity scores and utilizing structural model-based aggregation to absorb the macroscopic conditional drift, the conformal procedure delivers finite-sample, distribution-free uncertainty bands. This effectively bridges the gap between the flexible, high-dimensional pattern recognition of Gradient Boosting and the rigorous statistical reliability required in public health decision-making.

The primary contributions of this chapter are demonstrated through the following key empirical findings:

- **Computational Scalability and Speed:** The proposed ML framework drastically reduces the computational burden traditionally associated with high-dimensional spatio-temporal nowcasting. It enables extremely fast model fitting and real-time generation of estimates across hundreds of health regions, scaling more efficiently than even fast approximate Bayesian inference methods like Integrated Nested Laplace Approximations (INLA).
- **Temporal Robustness:** The conformal nowcasting model demonstrated exceptional temporal generalization. By structurally adapting to covariate shifts within the incoming data stream, it maintained valid empirical coverage and stable predictive performance over a continuous six-month prospective test period, entirely without the need for intermediate retraining.

- **Resilience to Severe Censoring:** The framework effectively corrected incidence curves even under extreme reporting lag conditions. Notably, it successfully recovered SARI counts in scenarios where delays systematically exceeded the standard six-week threshold, a critical phenomenon empirically observed in major demographic hubs like São Paulo.
- **Calibrated Uncertainty and Efficiency:** When benchmarked against the state-of-the-art INLA-based spatial model, the results revealed a clear methodological trade-off. While the Bayesian baseline yielded slightly more accurate point estimates, the conformal predictors demonstrated superior interval efficiency and calibration. Specifically, the conformal approach generated tighter prediction intervals and achieved empirical coverage tightly aligned with the nominal 95% target, successfully avoiding the conservative over-coverage observed in the Bayesian model.

To facilitate operational deployment, the conformal nowcasting framework introduced herein has been implemented within the PySRAG Python library (SILVA; AL., 2024a), a package maintained by our research team. Additionally, these delay-adjustment algorithms are currently being embedded into the Virus Monitor platform (SILVA; AL., 2024b). Together, these software tools establish an end-to-end surveillance pipeline, merging rigorous statistical count corrections with interactive visualizations to support real-time public health monitoring.

The remainder of this chapter is organized as follows. The [section 3.2](#) reviews the traditional nowcasting literature and the emerging intersection of machine learning with conformal inference. The [section 3.3](#) details the proposed methodology, formalizing the individual-level probability model and the conformal predictors. The [section 3.4](#) presents the experimental setup, the internal ablation study, and the benchmark against the Bayesian baseline. Finally, [section 3.5](#) provides a concluding discussion on the practical implications for public health management.

3.2 Related work

The correction of reporting delays in epidemiological surveillance systems requires the integration of statistical modeling, real-time nowcasting, and uncertainty quantification. This section reviews the evolution of these methodologies, highlighting the transition from traditional statistical models to modern machine learning approaches, and identifies the current gap regarding rigorous uncertainty quantification under non-stationary, right-censored count data.

The Traditional Paradigm: Statistical and Bayesian Nowcasting

The correction of reporting delays is fundamentally a right-censoring problem. The foundational literature often approaches this challenge using the run-off triangle data structure, a concept originally developed in actuarial sciences for claims reserving via the chain-ladder

method (MACK, 1993; RENSHAW; VERRALL, 1998; ENGLAND; VERRALL, 2002). In epidemiology, this structure has been successfully adapted using generalized linear models (GLMs) and conditional hierarchical models to estimate occurred-but-not-yet-reported cases (HÖHLE; HEIDEN, 2014; SALMON *et al.*, 2015).

Currently, the state-of-the-art in epidemiological nowcasting heavily relies on spatio-temporal Bayesian hierarchical models. These frameworks, which power national monitoring systems such as the Brazilian InfoGripe, excel at capturing complex temporal trends and spatial correlations (BASTOS *et al.*, 2019; MCGOUGH *et al.*, 2020; SCHNEBLE *et al.*, 2020; HALLIDAY *et al.*, 2026). However, they exhibit limitations in the context of high-resolution, real-time surveillance. First, because they model aggregated counts, they inherently restrict the inclusion of granular, individual-level patient features (BASTOS *et al.*, 2019). Furthermore, these Bayesian models require rigid parametric assumptions (e.g., assuming Poisson or Negative Binomial data-generating processes) and depend on predefined spatial adjacency matrices. As the spatial granularity increases to hundreds of health regions, these frameworks severely suffer from the curse of dimensionality. Consequently, even when employing deterministic approximations such as INLA to bypass traditional MCMC convergence issues, the computational burden becomes a critical operational bottleneck for high-resolution, rapid-response surveillance.

Machine Learning in Epidemiological Surveillance

To overcome the rigidity of traditional statistical models, recent literature has increasingly adopted Machine Learning (ML) techniques—such as Gradient Boosting Machines (GBM), Random Forests, and Neural Networks—for epidemiological forecasting and real-time surveillance (SILVA *et al.*, 2024; ALBRECHT *et al.*, 2024; KOEMEN *et al.*, 2025). A primary advantage of these algorithms is their native ability to handle high-dimensional covariate spaces and capture complex, non-linear interactions without strict *a priori* specifications. Crucially, in spatial epidemiology, ML models can learn spatial dynamics directly from continuous geographical coordinates (latitude and longitude), entirely bypassing the need for rigid, predefined neighborhood matrices (WONG *et al.*, 2023; UTAZI *et al.*, 2025; MATAZI; GOGNET; KAKAI, 2023).

Despite these predictive advantages, ML approaches face a major critique in public health applications: they predominantly generate deterministic point estimates. When uncertainty quantification is attempted, it frequently relies on heuristic methods. Common heuristic approaches include uncalibrated quantile regression lacking finite-sample guarantees, variance estimation across bootstrapped ensembles (such as Bagging or Random Forests), which primarily captures epistemic uncertainty, and global homoscedastic Gaussian approximations (IZBICKI, 2025; ANGELOPOULOS; BARBER; BATES, 2024). Because these heuristics lack rigorous mathematical guarantees of coverage, they fail to provide the reliable safety margins required for critical health resource allocation.

Conformal Inference and Uncertainty Quantification

In the broader field of statistical learning, Conformal Prediction (CP) has emerged as an elegant and robust framework for uncertainty quantification. Unlike traditional probabilistic models or heuristic ML bounds, conformal inference is strictly model-agnostic and distribution-free, classically providing mathematically proven coverage guarantees in finite samples under the foundational assumption of data exchangeability (VOVK; GAMMERMAN; SHAFER, 2022; ANGELOPOULOS; BARBER; BATES, 2024; IZBICKI, 2025).

However, in time-series forecasting and epidemiology, the non-stationary nature of disease transmission and evolving demographic profiles inherently violate strict exchangeability. Consequently, recent literature has pivoted toward robust conformal adaptations to handle distribution drift. Fundamental theoretical advances include weighted conformal prediction to rigorously correct for covariate shift (TIBSHIRANI *et al.*, 2020), as well as broader non-exchangeable frameworks that bound the coverage gap under general temporal drift by assigning decaying weights to historical data (BARBER *et al.*, 2023). Most importantly for operational, large-scale forecasting, theoretical guarantees have been extended to standard *Split Conformal Prediction* under suitable forms of temporal dependence and non-exchangeability (OLIVEIRA *et al.*, 2024). By analyzing the split-conformal architecture—the highly scalable framework employed in our methodology—Oliveira *et al.* (2024) derive approximate coverage bounds under concentration and decoupling conditions, with stationary β -mixing processes providing an important sufficient setting.

Complementing these algorithmic advances, modern approaches such as Conformalized Quantile Regression (CQR) (ROMANO; PATTERSON; CANDÈS, 2019), as well as Dist-split and CD-split (IZBICKI; SHIMIZU; STERN, 2020), allow for the construction of asymmetric prediction intervals that dynamically adjust to the heteroscedasticity of the data. Furthermore, recent foundational literature advocates for a *model-based perspective* on conformal inference (ANGELOPOULOS; BARBER; BATES, 2024). Rather than relying solely on post-hoc distribution-free corrections, this paradigm structurally embeds domain knowledge into the baseline estimator to proactively absorb primary distribution shifts—achieving what is termed model-based robustness to violations of exchangeability. Together, these structural and algorithmic adaptations maintain valid empirical coverage and sharp intervals even when strict exchangeability is compromised. Yet, while robust conformal inference is rapidly gaining traction in general time-series forecasting, its specific application to the epidemiological nowcasting problem—specifically addressing non-stationary, right-censored count data organized within delay triangles—remains critically underexplored.

The Research Gap and Chapter Contributions

This chapter addresses a significant gap in the literature by synthesizing the computational scalability of machine learning with the mathematical rigor of conformal inference for real-time

epidemiological nowcasting. Specifically, the proposed framework contributes to the state-of-the-art by combining:

1. The flexibility of Gradient Boosting (FRIEDMAN, 2001) to model complex, individual-level reporting delay surfaces based on granular patient features and continuous spatial coordinates.
2. A novel aggregation mechanism that scales individual-level predictions to regional expectation and variance estimators for the total number of eventually reported cases. Notably, for the expectation, this approach conceptually aligns with the classical Inverse Probability Weighting (IPW) estimator (HORVITZ; THOMPSON, 1952), effectively addressing the right-censoring inherent in reporting delays while structurally absorbing macroscopic demographic shifts.
3. The rigorous application of Conformal Prediction—systematically comparing residual-based and quantile-regression strategies—to provide distribution-agnostic prediction intervals. By utilizing locally adaptive scores to handle non-stationarity, the framework delivers robust, finite-sample uncertainty bands that maintain valid empirical coverage, effectively overcoming the parametric constraints and computational bottlenecks of traditional spatio-temporal Bayesian models.

3.3 Methodology

This section details the proposed analytical framework for correcting reporting delays in real-time epidemiological surveillance. To overcome the limitations of traditional statistical models, the proposed methodology synthesizes the probabilistic modeling of reporting delays with modern computational techniques and robust inference. By leveraging Machine Learning to estimate individual-level reporting probabilities and Conformal Prediction to calibrate the residual uncertainty, the framework is specifically designed to structurally absorb macroscopic drift and deliver robust empirical uncertainty bounds.

To provide a clear conceptual roadmap, the methodology unfolds through four interconnected steps, designed to bridge the gap between individual patient data and macro-level public health decisions:

Step 1: Dataset and Feature Definition. We begin by establishing the granular foundation, defining the individual-level features (e.g., demographics and continuous spatial coordinates) required to capture the heterogeneities in reporting behavior.

Step 2: Formalization of the Run-off Triangle. Next, we structure the right-censored data into a classical temporal triangle. This allows us to train a flexible Machine Learning model capable of estimating the conditional probability of delay for each specific case.

Step 3: Regional Aggregation of Moments. Because public health planning requires absolute volume metrics rather than individual probabilities, the third step bridges the micro-macro gap. We aggregate the individual predictions to derive regional expectation and variance estimators for the total number of eventually reported cases, a process that structurally absorbs the epidemic’s macroscopic drift.

Step 4: Conformal Uncertainty Quantification. Finally, acknowledging that even the most robust point estimates carry inherent prediction errors, we utilize these aggregated moments to construct locally adaptive non-conformity scores. This last step leverages Conformal Prediction to calibrate the residual uncertainty, transforming the baseline estimates into robust, distribution-agnostic prediction intervals vital for safe decision-making.

3.3.1 Dataset and Feature Engineering

The empirical analysis is conducted using data extracted from the Influenza Epidemiological Surveillance Information System (SIVEP-Gripe). Consistent with the data curation process established in [Chapter 2](#), the raw dataset was filtered using a specific logical rule to standardize case inclusion: (cough OR sore throat) AND (dyspnea OR oxygen saturation < 95% OR difficulty breathing) AND (hospitalized OR deceased). As previously discussed, this strategy prioritizes respiratory and severity signs that exhibit greater consistency across different etiological agents, rather than relying strictly on the varying administrative definitions of SARI. However, unlike the analysis in the previous chapter—which focused exclusively on the proportional distribution of specific viruses and thus restricted the sample to positive laboratory results—this chapter addresses the absolute volume of hospital demand. Consequently, the dataset utilized here encompasses all records that satisfy the aforementioned clinical rule, regardless of their virological test status.

To capture the spatial dynamics and administrative reality of the disease, each patient is mapped to their corresponding *Regional de Saúde* (Health Region)—official groupings of municipalities used for health resource allocation in Brazil. These regions serve as the essential geographic boundaries for the macro-level aggregation performed in the later stages of our framework. Consistent with [Chapter 2](#), the individual-level probability estimates for the proposed machine learning framework are conditioned on a granular feature space ($\mathbf{X}$). This space encompasses patient-level demographics, specifically age in years (IDADE_ANO), and municipal-level characteristics. The latter includes the continuous spatial coordinates of the municipality’s centroid (LATITUDE and LONGITUDE), retrieved from the IBGE Selected Localities Register (2010), and the municipality population size (POPULACAO), based on the 2022 Demographic Census. By utilizing these precise point coordinates, the Gradient Boosting algorithm can organically learn complex spatial patterns and distance-based correlations without relying on rigid boundary definitions.

To facilitate the mapping to the Health Region level and to structure the spatial components of the benchmark model, we utilized the *geobr* R/Python package, maintained by the Institute for Applied Economic Research (IPEA). This package was exclusively used to map each municipality to its respective Health Region and to extract the exact spatial polygons of these regions. In contrast to our proposed machine learning approach, the hierarchical Bayesian model—employed as the benchmark in [section 3.4](#)—fundamentally relies on these explicitly defined spatial polygons to construct the predefined spatial adjacency matrices (neighborhood graphs) strictly required for its spatial random effects.

3.3.2 Data Representation: The Run-off Triangle

The correction of reporting delays is inherently a problem of right-truncated or right-censored data, a methodological challenge well-documented in real-time epidemiological surveillance ([BASTOS *et al.*, 2019](#); [HÖHLE; HEIDEN, 2014](#)). Let t denote the time of the epidemiological event (e.g., the epidemiological week of symptom onset), where $t = 1, 2, \dots, T$, and let d denote the reporting delay in weeks, where $d = 0, 1, \dots, D$.

The count of cases that occurred at time t and were reported with a delay d is denoted as $n_{t,d}$. At any current observation time T , the surveillance system only records cases for which the sum of the occurrence time and the reporting delay does not exceed the present moment, meaning $t + d \leq T$ ([BASTOS *et al.*, 2019](#)). This observational constraint creates a two-way incomplete data structure known as a *run-off triangle*, illustrated in [Table 6](#). Originally developed in actuarial sciences to predict outstanding insurance claims ([MACK, 1993](#); [ENGLAND; VERRALL, 2002](#)), this representation is now the standard framework for formulating the occurred-but-not-yet-reported problem in infectious disease nowcasting ([HÖHLE; HEIDEN, 2014](#); [BASTOS *et al.*, 2019](#)).

Table 6 – Diagram representing right-censored surveillance data (the run-off triangle).

Time (t)	Delay (d)						
	0	1	2	$\dots$	$D-2$	$D-1$	D
1	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$	$\dots$	$n_{1,D-2}$	$n_{1,D-1}$	$n_{1,D}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$T-D-1$	$n_{T-D-1,0}$	$n_{T-D-1,1}$	$n_{T-D-1,2}$	$\dots$	$n_{T-D-1,D-2}$	$n_{T-D-1,D-1}$	$n_{T-D-1,D}$
$T-D$	$n_{T-D,0}$	$n_{T-D,1}$	$n_{T-D,2}$	$\dots$	$n_{T-D,D-2}$	$n_{T-D,D-1}$	$n_{T-D,D}$
$T-D+1$	$n_{T-D+1,0}$	$n_{T-D+1,1}$	$n_{T-D+1,2}$	$\dots$	$n_{T-D+1,D-2}$	$n_{T-D+1,D-1}$	$n_{T-D+1,D}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$T-1$	$n_{T-1,0}$	$n_{T-1,1}$	$n_{T-1,2}$	$\dots$	$n_{T-1,D-2}$	$n_{T-1,D-1}$	$n_{T-1,D}$
T	$n_{T,0}$	$n_{T,1}$	$n_{T,2}$	$\dots$	$n_{T,D-2}$	$n_{T,D-1}$	$n_{T,D}$

The shaded cells in Table 6 indicate the observed SARI counts, while the unshaded cells represent the unobserved future reports. The primary statistical goal of the nowcasting framework is to estimate these missing values in the lower-right segment of the triangle up to a predefined maximum delay D . By predicting these unobserved components, we can reliably compute the total consolidated cases for any recent epidemiological week, $N_t = \sum_{d=0}^D n_{t,d}$, thereby recovering the true incidence curve.

3.3.3 Statistical Modeling: Conditional Estimates and Individual Aggregation

The proposed framework translates the aggregated run-off triangle into a supervised machine learning problem, operating through three distinct stages to derive the regional expectations and variances of the total eventual case counts from individual-level probabilities.

Step 1: Defining the Delay Threshold and Absorbing State. First, a maximum reporting delay D is established (e.g., $D = 6$ weeks). While this threshold is empirically chosen to encompass the vast majority of historical reporting lags, epidemiological data frequently exhibit extreme, long-tail delays. To ensure the model captures these late reports without unnecessarily expanding the dimensionality of the classification problem, the final delay class D is defined as an absorbing upper bound. Consequently, the framework models the final category as encompassing all delays greater than or equal to the threshold ($\geq D$), a robust truncation strategy adopted in the nowcasting literature (HÖHLE; HEIDEN, 2014).

Step 2: Probabilistic Delay Estimation. To learn the complex reporting dynamics, we employ a Gradient Boosting Machine (GBM) (FRIEDMAN, 2001) probabilistic classifier. The theoretical foundation and the hyperparameter configuration of the algorithm are analogous to those utilized in Chapter 2 (see subsection 2.3.3.1 and Appendix A), with the primary distinction being the prediction target: instead of classifying viral etiologies, the model is trained to predict the reporting delay. The algorithm minimizes a multi-class logarithmic loss over a historical sample of fully observed data (the upper segment of the triangle) to estimate the probability mass function across all possible delay classes conditioned on its individual feature vector $\mathbf{x}_i$ (comprising patient age, municipal latitude, longitude, and population size), $\mathbb{P}(Y = j \mid \mathbf{X} = \mathbf{x}_i)$ for $j \in \{0, 1, \dots, \geq D\}$. From these raw multi-class estimates, we compute the cumulative probability p_i that a specific case is reported within a given delay d by aggregating the probabilities up to d :

$$p_i = \mathbb{P}(Y \leq d \mid \mathbf{X} = \mathbf{x}_i) = \sum_{j=0}^d \mathbb{P}(Y = j \mid \mathbf{X} = \mathbf{x}_i).$$

To implement our model-based robustness strategy, explicitly temporal variables (such as the epidemiological week) are deliberately excluded from the feature space $\mathbf{x}_i$. This enforces a stationary conditional reporting rule, structurally preventing the model from over-fitting to transient temporal dynamics. By relying solely on these individual and municipal features,

this flexible tree-based model naturally captures heterogeneities in the delay surface across different age groups and locations without requiring predefined spatial matrices. Furthermore, the proposed probabilistic classifier organically accommodates the absorbing state $\geq D$ without introducing the statistical artifacts common in traditional models that enforce smooth temporal transitions for the delay effects.

Step 3: Aggregated Expectation, Variance, and Estimators. Once the cumulative individual probabilities are estimated, the problem shifts to predicting the unobserved cases for an arbitrary epidemiological week t evaluated at a current delay d . For notational simplicity throughout this step, we omit the temporal and delay subscripts, focusing on the spatial aggregation for a given region r (i.e., a specific Health Region).

The tree-based ensemble architecture of the GBM inherently partitions the covariate space into a highly granular set of disjoint subsets. Specifically, individuals who are routed to the exact same sequence of terminal nodes across all individual trees in the boosting ensemble fall into the same subset. Let $\mathcal{B}_r$ represent the set of these unique, ensemble-defined subgroups that contain observations belonging to region r . For any specific subgroup $b \in \mathcal{B}_r$, all individuals within this partition share the exact same estimated reporting probability, denoted as p_b .

Let $n_b \geq 1$ be the number of partially observed cases falling into subgroup b . The theoretical total eventual number of cases N_b (observed plus unobserved) originating from this specific subgroup can be rigorously modeled by a Negative Binomial distribution representing the total number of trials required to achieve n_b successes:

$$N_b \sim \text{NegBin}(n_b, p_b), \quad \text{for } N_b \in \{n_b, n_b + 1, \dots\} \text{ and } p_b \in (0, 1].$$

The total consolidated case count for the entire region r is the sum of the cases across all its internal subgroups:

$$N_r = \sum_{b \in \mathcal{B}_r} N_b.$$

By leveraging the linearity of expectation, the theoretical aggregate expected case count $\mathbb{E}[N_r]$ is the sum of the expected values of each Negative Binomial subgroup:

$$\mathbb{E}[N_r] = \sum_{b \in \mathcal{B}_r} \mathbb{E}[N_b] = \sum_{b \in \mathcal{B}_r} \frac{n_b}{p_b}.$$

In practice, the true probabilities p_b are unknown. By substituting them with the algorithmic estimates $\hat{p}_b$, we define the point estimator for the regional total as:

$$\hat{N}_r = \sum_{b \in \mathcal{B}_r} \frac{n_b}{\hat{p}_b}.$$

Conceptually, this method treats each partially observed subgroup as a representative sample of a larger unobserved population. If a subgroup b has an estimated observation probability of $\hat{p}_b$, its n_b reported cases conceptually represent $n_b / \hat{p}_b$ total individuals sharing that

specific covariate profile. Summing these fractional contributions yields the point estimate $\hat{N}_r$. This aggregation is formally equivalent to the classical Inverse Probability Weighting (IPW) or Horvitz-Thompson estimator for population totals (HORVITZ; THOMPSON, 1952; COCHRAN, 1977). A fundamental property of the Horvitz-Thompson estimator is that it is strictly unbiased for the population total if the true observation probabilities p_b are known. While the use of estimated probabilities $\hat{p}_b$ introduces a potential for finite-sample bias, the estimator is constructed to be asymptotically unbiased. In our framework, this property is achieved by ensuring that the probability estimator converges to the true conditional probability. We accomplish this by training the machine learning classifier using the multi-class logarithmic loss (log-loss) function, which is mathematically equivalent to maximizing the log-likelihood of the underlying multinomial distribution for the observed delays.

Within the specific literature of reporting delays, this approach generalizes the foundational delay-adjusted event count estimator proposed by Lawless (1994). While Lawless applied a single macroscopic probability weight to an entire time cohort under strict temporal stationarity assumptions, our tree-based framework computes the expected total by dynamically weighting the observed counts in each highly granular subgroup by their specific, covariate-conditioned inverse probability of observation.

This bottom-up aggregation grants the framework an organic temporal adaptability. The predictive model assumes the absence of *conditional shift*, meaning the learned rules mapping the covariate space to the delay probabilities remain invariant over time ($P(Y | \mathbf{X})$ is constant). However, the macroscopic delay distribution for any given region is dynamically marginalized over the changing profile of the infected population. By naturally accommodating *covariate shift* (or compositional shift) within the incoming data stream, the aggregated delay surface inherently exhibits non-stationary behavior. As the spatial or demographic characteristics of the outbreak evolve week-by-week, this covariate shift induces a natural *prior shift* in the marginal delay distribution ($P(Y)$), capturing macroscopic temporal variations without the need for explicit autoregressive time-series components.

Furthermore, under the assumption that the reporting events across different subgroups are conditionally uncorrelated given their respective probabilities, the theoretical covariance matrix is diagonal. Thus, the variance of the aggregated total straightforwardly reduces to the sum of the subgroup variances:

$$\sigma_r^2 = \text{Var}[N_r] = \sum_{b \in \mathcal{B}_r} \text{Var}[N_b] = \sum_{b \in \mathcal{B}_r} \frac{n_b(1 - p_b)}{p_b^2}.$$

Correspondingly, the variance estimator utilized to evaluate the intrinsic aleatoric dispersion is given by:

$$\hat{\sigma}_r^2 = \widehat{\text{Var}}[N_r] = \sum_{b \in \mathcal{B}_r} \frac{n_b(1 - \hat{p}_b)}{\hat{p}_b^2}.$$

3.3.4 Uncertainty Quantification and Conformal Predictors

Aleatoric and Epistemic Uncertainty in Count Aggregation

While aggregated point estimates, such as $\mathbb{E}[N_r]$, provide operational guidance, real-time epidemiological decision-making requires a rigorous quantification of the uncertainty surrounding these predictions. In predictive modeling, the total predictive uncertainty associated with an estimate is fundamentally decomposed into two distinct components: aleatoric and epistemic uncertainty (IZBICKI, 2025).

In the proposed framework, these components map directly to the macroscopic aggregation of the reporting process. The *aleatoric uncertainty* captures the inherent, irreducible randomness of the events. At the regional level, it is formally represented by the theoretical variance of the sum of the Negative Binomial subgroups, σ_r^2 . Even if the true conditional reporting probabilities p_b were known with absolute certainty, the final aggregated count N_r would still fluctuate around its expectation due to the stochastic nature of human reporting behavior.

Conversely, *epistemic uncertainty* arises from a lack of knowledge about the true data-generating process (IZBICKI, 2025). In this context, it represents the variability and potential bias introduced because the true probabilities p_b are unknown and must be replaced by predictions $\hat{p}_b$ derived from the Gradient Boosting ensemble. This uncertainty is driven by the finite size of the training sample, the tree-based approximation of the delay surface, and structural assumptions, such as the conditional uncorrelation of reporting events across subgroups. While aleatoric uncertainty represents a hard limit on predictability, epistemic uncertainty is reducible and decreases as more surveillance data becomes available (IZBICKI, 2025).

Classical statistical approaches attempt to construct prediction intervals by analytically combining these two sources of variance (e.g., via the Law of Total Variance). However, deriving the exact probability mass function for a sum of independent Negative Binomial random variables with heterogeneous, machine-learned parameters $\hat{p}_b$ is analytically intractable. Because the convolution of Negative Binomial distributions with distinct probabilities lacks a tractable closed-form expression, obtaining the exact aggregated distribution requires computationally prohibitive numerical methods. Conversely, relying on simple asymptotic approximations (e.g., Gaussian limits) often inadequately captures the heavy-tailed dispersion typical of epidemiological count data. Furthermore, deriving exact analytical bounds for the epistemic variance of a highly dimensional, non-parametric ensemble like GBM is notoriously difficult, often leading to heuristic approximations that fail to guarantee coverage in finite samples.

The Split Conformal Prediction Framework

To circumvent these mathematical limitations and rigorously capture the total predictive uncertainty, we employ Split Conformal Prediction (CP) (ANGELOPOULOS; BARBER; BATES, 2024; IZBICKI, 2025). Conformal inference is a robust, distribution-free framework

that leverages an independent calibration set to construct prediction intervals with guaranteed marginal coverage. Essentially, rather than applying a naive, globally symmetric error margin that assumes homoscedasticity, our framework utilizes the theoretically derived regional variance (σ_r^2) to construct *locally adaptive* conformal intervals. By evaluating empirical, variance-scaled non-conformity scores on unseen data, this adaptive procedure naturally encapsulates both the aleatoric dispersion of the Negative Binomial convolutions and the ensemble’s epistemic errors into valid, heteroscedastic prediction intervals.

To formally construct these intervals, it is necessary to distinguish between the data domains used throughout the modeling pipeline. The underlying Gradient Boosting classifier is trained exclusively on the raw, individual-level dataset, consisting of patient covariates $\mathbf{x}_i$ and their respective observed reporting delays y_i . However, the conformal prediction procedure operates strictly at the macroscopic level.

Rather than evaluating a single region in isolation, the framework partitions the entire surveillance space into a set of mutually exclusive (disjoint) Health Regions. For a specific evaluation delay d , we construct a comprehensive aggregated dataset spanning a sequence of historical epidemiological weeks $t \in \mathcal{T}$ and across all disjoint regions. For each specific region-week combination (r, t) , we extract the currently observed partial count $n_{r,t}$ and aggregate the individual probabilities to compute the expected regional count $\hat{N}_{r,t}$ and its theoretical variance $\hat{\sigma}_{r,t}^2$, matching them alongside the true, fully observed final case count $N_{r,t}$.

This two-stage architecture yields a macroscopic dataset $\mathcal{D}_{agg} = \{(N_{r,t}, n_{r,t}, \hat{N}_{r,t}, \hat{\sigma}_{r,t}^2)\}$. To apply the Split Conformal procedure while actively mitigating the temporal autocorrelation inherent in epidemiological time series, we randomly partition the sequence of historical epidemiological weeks $\mathcal{T}$ into two disjoint sets: training weeks and calibration weeks ($\mathcal{T}_{cal}$). Rather than utilizing a strict chronological cutoff—which could inadvertently restrict the calibration set to a narrow, unrepresentative phase of the epidemiological curve (e.g., exclusively the waning tail of an outbreak)—we randomly sample the historical weeks. This random partitioning ensures that the calibration set $\mathcal{D}_{cal}$ captures a highly representative mixture of the underlying transmission dynamics, incorporating both high-variance peaks and low-variance baselines. This structural diversity is essential for the locally adaptive scores to properly calibrate the heteroscedasticity of the delay surface. The macroscopic calibration set is then strictly composed of all region-week instances that belong to the selected calibration weeks. To proceed with the conformal calculations without notational clutter, let k index a unique, flattened region-week instance within this specific calibration subset, forming $\mathcal{D}_{cal} = \{(N_k, n_k, \hat{N}_k, \hat{\sigma}_k^2)\}_{k=1}^{m_{cal}}$, where m_{cal} represents the total number of region-week instances reserved for calibration.

Next, we define a generic non-conformity score function, denoted as $s(\hat{N}_k, \hat{\sigma}_k^2, N_k) \in \mathbb{R}$. This function evaluates the discrepancy between the true aggregated count N_k and the algorithmic expected moments. By design, higher values of s indicate a worse fit, meaning the macroscopic observation is less conforming to the model’s expected behavior.

For each instance in the independent calibration set, we compute its corresponding empirical non-conformity score, resulting in a sequence of scores $S = \{s_1, s_2, \dots, s_{m_{cal}}\}$. Given a user-specified miscoverage rate $\alpha \in (0, 1)$ (e.g., $\alpha = 0.05$ for a target coverage of 95%), we calculate the critical threshold $\hat{q}$ as the empirical quantile of these calibration scores. Specifically, the conformal threshold is precisely defined as the $\lceil (m_{cal} + 1)(1 - \alpha) \rceil$ -th smallest value in the sorted set S (VOVK; GAMMERMAN; SHAFER, 2022; ANGELOPOULOS; BARBER; BATES, 2024; IZBICKI, 2025).

Finally, for a new, ongoing epidemiological week and region instance k_{new} currently under surveillance, the conformal prediction set $C(k_{new})$ is constructed by including all possible candidate values N whose hypothetical non-conformity score is less than or equal to the calibrated threshold:

$$C(k_{new}) = \{N \in \mathbb{N} : s(\hat{N}_{k_{new}}, \hat{\sigma}_{k_{new}}^2, N) \leq \hat{q}\}.$$

Under ideal conditions of perfect exchangeability, this procedure rigorously guarantees marginal coverage in finite samples, such that $\mathbb{P}(N_{k_{new}} \in C(k_{new})) \geq 1 - \alpha$. In the context of our framework, this guarantee is structurally reinforced by what recent literature terms *model-based robustness* (ANGELOPOULOS; BARBER; BATES, 2024). Because we actively model the heteroscedasticity and macroscopic drift using Gradient Boosting and adaptive scores, the framework exhibits a “double-robustness” property. If our baseline estimator successfully approximates the true conditional distribution—effectively extracting the temporal signal—the resulting non-conformity scores become inherently exchangeable, securing the exact $1 - \alpha$ coverage without relying on the innate stationarity of the raw epidemiological data.

However, recognizing that unobserved epidemiological shocks or administrative delays may still introduce mild residual non-stationarity at the prediction point, our reliance on the Split Conformal architecture provides an additional robustness mechanism. As shown by Oliveira et al. (2024) (OLIVEIRA *et al.*, 2024), the standard, unmodified Split Conformal Prediction algorithm can retain approximate coverage when applied to non-exchangeable processes under suitable temporal dependence conditions. In this setting, unmodeled temporal dependence translates into a mathematically bounded penalty on the theoretical coverage limit (ϵ). Therefore, even if the model-based assumptions partially fail and strict exchangeability is compromised, the framework can retain coverage of at least $1 - \alpha - \epsilon$, where the penalty ϵ depends on the temporal mixing structure.

Locally Adaptive Non-Conformity Scores

Building upon this foundational conformal framework, the critical design choice lies in the definition of the non-conformity score s . While the straightforward approach of using the absolute residual score ($s_{res} = |N_k - \hat{N}_k|$) is common in standard regression, it yields a single, globally constant margin of error. In epidemiological surveillance, this homoscedastic assumption is fundamentally flawed, as predictive uncertainty naturally scales with the magnitude of the

epidemiological curve and the specific delay profile of the affected population. To address this inherent heteroscedasticity and the right-skewness of the count data, we evaluate and contrast four distinct *locally adaptive* conformal strategies. Throughout these strategies, the expected total $\hat{N}_k$ and the theoretical variance $\hat{\sigma}_k^2$ derived in Step 3 serve as the foundational inputs.

1. Residual-Based Conformal Predictors

In this family of approaches, the expectation itself ($\hat{N}_k$) serves as the point estimate and anchor, while non-conformity scores are constructed using locally adaptive (standardized) residuals to account for variance inflation. We evaluate two variants:

- *Locally Adaptive Symmetric Score*: The non-conformity score is defined as the absolute standardized error,

$$s_{sym} = \frac{|N_k - \hat{N}_k|}{\hat{\sigma}_k}.$$

By extracting the $(1 - \alpha)$ empirical quantile $\hat{q}_{sym}$ from the calibration set, the resulting interval is

$$C(k_{new}) = [\hat{N}_k - \hat{q}_{sym}\hat{\sigma}_k, \hat{N}_k + \hat{q}_{sym}\hat{\sigma}_k].$$

While computationally efficient, forcing symmetry around the mean risks under-covering the heavy right tails characteristic of outbreak dynamics.

- *Locally Adaptive Asymmetric Score*: To naturally accommodate the right-skewness of SARI counts, this variant separates the calibration of the lower and upper bounds. We define two distinct directional scores:

$$s_{low} = \frac{\hat{N}_k - N_k}{\hat{\sigma}_k} \quad \text{and} \quad s_{high} = \frac{N_k - \hat{N}_k}{\hat{\sigma}_k}.$$

By extracting the independent empirical quantiles $\hat{q}_{low}$ and $\hat{q}_{high}$ at the $(1 - \alpha/2)$ level, the framework yields a flexible, asymmetric interval:

$$C(k_{new}) = [\hat{N}_k - \hat{q}_{low}\hat{\sigma}_k, \hat{N}_k + \hat{q}_{high}\hat{\sigma}_k].$$

2. Distributional Conformalized Quantile Regression (CQR)

Rather than defining margins around the expected mean, Conformalized Quantile Regression (CQR) (ROMANO; PATTERSON; CANDÈS, 2019) constructs intervals by adjusting pre-estimated heuristic bounds and naturally adopts the conditional median as its point estimate. Instead of training a separate quantile-based machine learning model, we strategically derive these initial bounds ($Q_{\alpha^*/2}$ and $Q_{1-\alpha^*/2}$) by mapping our derived expected moments ($\hat{N}_k$ and $\hat{\sigma}_k^2$) to known parametric probability distributions via the method of moments.

Here, α^* denotes the heuristic error level used strictly for generating the initial parametric quantiles. In this framework, we set this initial level equal to the target conformal error rate ($\alpha^* = \alpha$) to structurally align the dispersion of the base distribution with the desired final coverage. We test two distributional assumptions for generating these base quantiles:

- *CQR - Negative Binomial*: The expected mean and variance are mapped to the parameters of a Negative Binomial distribution, strictly respecting the discrete nature of the count data.
- *CQR - Gamma*: As a continuous alternative for heavily overdispersed counts, the moments are matched to the shape and scale parameters of a Gamma distribution.

The complete derivation of the moment-matching equations for both distributions is detailed in [Appendix G](#). Once the heuristic quantiles $Q_{\alpha^*/2}$ and $Q_{1-\alpha^*/2}$ are parametrically extracted, the CQR non-conformity score evaluates how far the true calibration counts fall outside these proposed bounds:

$$s_{cqr} = \max\{Q_{\alpha^*/2} - N_k, N_k - Q_{1-\alpha^*/2}\}.$$

By calculating the $(1 - \alpha)$ empirical quantile $\hat{q}_{cqr}$ of these scores, the initial parametric bounds are rigorously conformalized (expanded or shrunk) to guarantee exact marginal coverage:

$$C(k_{new}) = [Q_{\alpha^*/2} - \hat{q}_{cqr}, Q_{1-\alpha^*/2} + \hat{q}_{cqr}].$$

As previously discussed, the true aggregated response is an analytically intractable convolution of heterogeneous Negative Binomials, meaning any single parametric distribution mapped via expected moments is inherently misspecified. However, because the conformal framework is strictly distribution-free, the mathematical validity of the final interval does not depend on the accuracy of these initial heuristic quantiles. The conformal procedure acts as an empirical correction mechanism, dynamically shifting the misspecified bounds to secure valid marginal coverage regardless of the underlying convolution. This dual-stage strategy inherits the structural shape (right-skewness and heavy tails) of the parametric approximations while relying entirely on conformal inference to neutralize any epistemic misspecification introduced in the moment-matching step.

Boundary Conditions and Practical Interpretation

Handling the Zero-Observation Edge Case

A critical operational scenario arises when a Health Region records zero partial observations in a given week ($n_k = 0$). In standard survey sampling literature, the foundational estimator for an empty sample natively evaluates to zero. Consequently, our expected point estimate and the conformal lower bound are logically bounded to zero ($\hat{N}_k = 0$). However, because $\hat{N}_k = 0$ yields a theoretical standard deviation of zero ($\hat{\sigma}_k = 0$), calculating the non-conformity scores natively fails for all variants: applying the standardized residuals (s_{sym} or s_{asym}) results in an undefined division by zero, while mapping these null moments to parametric distributions for CQR yields degenerate point masses, causing parameterization singularities. To ensure the framework remains mathematically robust for small, sparsely populated municipalities without observed cases, the conformal algorithm dynamically defaults to an unscaled, homoscedastic

upper bound for these specific regions, reducing the non-conformity score to the raw count itself ($s = N_k$). This safeguards the coverage guarantee while circumventing the mathematical singularities.

Domain Enforcement and Discretization

Ultimately, epidemiological decision-making requires discrete and logically coherent case counts. Therefore, we map the continuous outputs of the conformal framework back into the operational domain of natural numbers ($\mathbb{N}$). Let L and U be the continuous lower and upper bounds derived from any of the conformal strategies. Because the true aggregated response N_k is strictly an integer, any continuous interval $[L, U]$ contains the exact same set of valid integer predictions as its inner discretized bounds. Thus, to translate the continuous bounds into a discrete interval without altering the exact set of valid integers—and without inducing unnecessary overcoverage—we apply the ceiling function to the lower bound and the floor function to the upper bound.

Furthermore, we enforce a strict operational constraint: the final predictive bounds and point estimates can never be smaller than the number of cases already observed and reported to the surveillance system (n_k). Consequently, the final operational interval is explicitly defined as $[\max(n_k, \lceil L \rceil), \max(n_k, \lfloor U \rfloor)]$. Similarly, the final point estimate is discretized to the nearest integer, strictly bounded by the observed counts, $\max(n_k, \text{round}(\hat{N}_k))$.

This locally adaptive formulation is highly advantageous for public health decision-making. During periods of low viral transmission or in regions with fast and reliable reporting systems (where $\hat{\sigma}_k^2$ is inherently small), the prediction intervals naturally narrow, providing precise estimates. Conversely, during major outbreaks or when the disease spreads to regions with historically delayed reporting architectures (where $\hat{\sigma}_k^2$ is large), the intervals automatically expand. This dynamic widening safely encapsulates the increased aleatoric variance and the ensemble’s epistemic uncertainties, securing valid marginal coverage without compromising the structural heavy tails of the count data.

3.4 Results

The spatial domain of this study encompasses the entirety of the Brazilian territory, which is administratively subdivided into 438 health regions. These regions serve as the fundamental spatial units for epidemiological planning and resource allocation by the Ministry of Health. Consequently, all SARI notifications extracted from the SINAN database were spatially aggregated at this resolution.

To evaluate the framework under realistic surveillance conditions, the SARI notification dataset (2022–2024) was partitioned into historical and future evaluation periods. The second semester of 2024 (2S2024) was isolated as a strict out-of-time *Test* set. The remaining historical data, spanning from the first epidemiological week of 2022 to the end of the first semester of 2024 (1S2024), was reserved for model training and calibration.

To ensure the calibration process captured a representative mixture of the underlying transmission dynamics—incorporating both high-variance outbreak peaks and low-variance endemic baselines—the historical weeks were randomly shuffled and partitioned into an 80% *Training* set and a 20% *Calibration* set, rather than utilizing a chronological split.

Empirical analysis of the Training set revealed that over 98% of all notifications are formally registered within six weeks of symptom onset. Consequently, we defined the maximum evaluation delay as $D = 6$ weeks. To retain the information from extreme outliers without discarding data, any reporting delays exceeding this threshold in the training set were capped at 6 weeks. The target variable Y for the Gradient Boosting classifier was thus categorized into seven mutually exclusive classes: $y \in \{0, 1, 2, 3, 4, 5, \geq 6\}$.

The underlying probability classifier was trained using a strictly spatial and demographic feature space ($\mathbf{x}_i$), consisting of the patient’s age group, the population size of the reporting municipality, and its geographic coordinates (latitude and longitude). Once the reporting surface was learned and calibrated on the historical data, the complete nowcasting pipeline was deployed and evaluated week-by-week on the out-of-sample 2S2024 dataset.

Following this experimental design, the results are presented in three consecutive stages. First, we examine the empirical distribution of the predictive residuals and the dynamics of the theoretical non-conformity scores. Second, we present an internal 5-fold cross-validation evaluation on the historical data to assess the coverage validity, interval efficiency, and point estimate accuracy of the proposed conformal configurations, ultimately selecting the optimal framework. Finally, we conduct an out-of-time evaluation on the unseen 2S2024 dataset, benchmarking the selected machine learning pipeline against an established Bayesian hierarchical spatial model.

3.4.1 Analysis of Residuals and Non-Conformity Scores

To understand how the predictive uncertainty evolves across different reporting delays, we evaluated the empirical distribution of the standardized prediction errors derived from the Calibration dataset, alongside the empirical critical cutoffs of the residual-based scores. For the core residual analysis, we explicitly isolated regions with at least one observed partial notification ($n_k > 0$). The specific zero-observation edge case ($n_k = 0$) fundamentally lacks a variance scale ($\hat{\sigma}_k^2 = 0$) and was therefore analyzed separately.

Standardized Residuals and Score Dynamics

While locally adaptive conformal scores are structurally designed to address the inherent heteroscedasticity of epidemiological count data, analyzing the standardized prediction errors $\left(\frac{N_k - \hat{N}_k}{\hat{\sigma}_k}\right)$ also reveals a severe distribution asymmetry (Figure 12). At the initial evaluation point ($d = 0$), the distribution exhibits moderate right-skewness (Skewness = 1.66). However, as the delay increases, the skewness intensifies dramatically, peaking at 10.15 for $d = 3$ (Figure 12).

This long right tail indicates the presence of extreme instances where the true final cases substantially exceed the algorithmic expectation.

The impact of this heavy-tailed distribution is evident when compared to standard parametric expectations. If one were to rely solely on an uncorrected Normal approximation, a 95% prediction interval would typically span approximately ± 1.96 standard deviations. In contrast, the empirically calibrated Symmetric bounds (Sym-CP) required consistently wider margins to guarantee the same nominal coverage, expanding from ± 2.28 at $d = 0$ to ± 2.48 at $d = 2$ (Figure 12).

This pronounced asymmetry also highlights the practical differences between the Symmetric (Sym-CP) and Asymmetric (Asym-CP) conformal strategies. Because Sym-CP utilizes the absolute standardized error ($s_{sym} = |N_k - \hat{N}_k| / \hat{\sigma}_k$), it extracts a single critical threshold ($\hat{q}_{sym}$) dictated entirely by the extreme right tail. Consequently, it applies this large margin equally to the left, often pushing the lower bound down to the deterministic floor of already observed cases (n_k), which can create an uninformative lower interval.

Conversely, Asym-CP calculates independent thresholds for the lower and upper tails. For instance, at $d = 0$, the Asym-CP bounds natively adapt to the skewness, resulting in a tighter left margin and an extended right margin $(-1.74, 2.90)$, compared to the forced ± 2.28 of the Sym-CP. However, this flexibility introduces a statistical trade-off regarding interval width and efficiency. While both methods yield comparable total interval widths at early delays ($d \in \{0, 1, 2\}$), Asym-CP produces noticeably wider intervals at higher delays. For $d = 5$, the average width of Asym-CP was 3.98 standard deviations, compared to 3.01 for Sym-CP. This efficiency loss occurs because Asym-CP independently calibrates the $\alpha/2$ quantiles on both tails. In heavily bounded distributions, this independent calibration tends to be statistically conservative (similar to a Bonferroni correction), occasionally yielding an empirical coverage strictly greater than the nominal $1 - \alpha$ target, at the cost of wider margins.

CQR Calibration and Distributional Fit

While residual-based scores manipulate standard deviations, the Conformalized Quantile Regression (CQR) strategies rely on heuristic parametric distributions. The empirical scores extracted during calibration reveal the efficacy of the initial moment-matching step.

The conformal threshold ($\hat{q}_{cqr}$) dictates how much the heuristic bounds must be shifted to achieve the exact $1 - \alpha$ coverage. For the CQR-Negative Binomial strategy, the calibration process yielded $\hat{q}_{cqr} = 0.0$ for all delays $d \geq 1$. This signifies that the theoretical quantiles derived solely from our estimated macroscopic moments ($\hat{N}_k$ and $\hat{\sigma}_k^2$) naturally achieved at least 95% marginal coverage without requiring any empirical conformal expansion. For the highly uncertain delay $d = 0$, the conformal procedure applied a minimal, discrete correction of precisely 1.0 case.

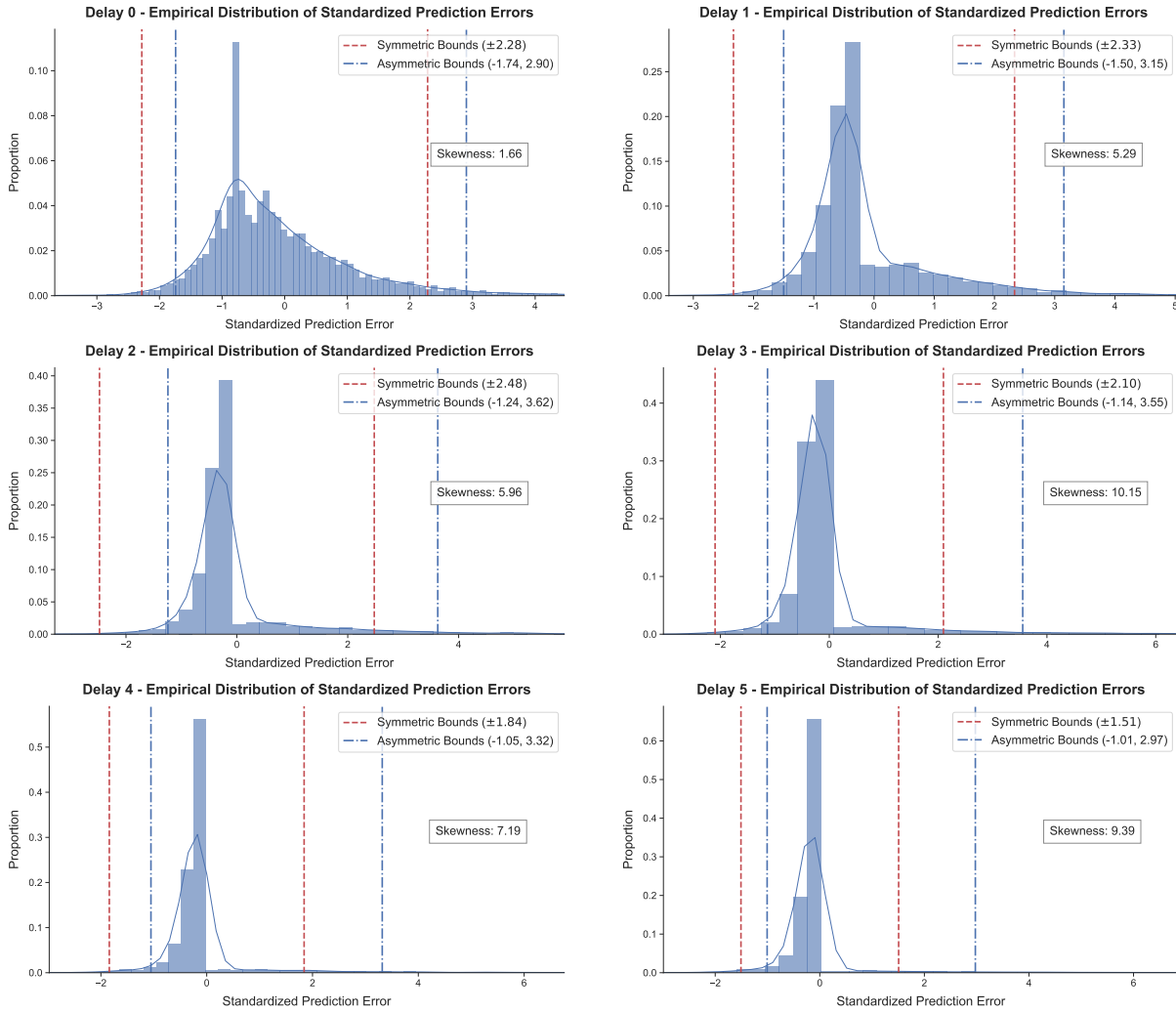


Figure 12 – Empirical distribution of standardized prediction errors across different reporting delays ($d = 0$ to $d = 5$). The dashed red lines indicate the Symmetric conformal bounds, while the dot-dashed green lines represent the independent Asymmetric bounds. Notably, as the reporting delay increases, the distribution exhibits severe right-skewness, demonstrating how asymmetric strategies dynamically adapt to heavy-tailed epidemiological counts, whereas symmetric bounds are forced to become unnecessarily conservative on the lower tail.

Similar behavior was observed for the CQR-Gamma strategy. Because the Gamma distribution operates over continuous positive reals, the conformal corrections were highly granular but equally negligible (e.g., $+1.29$ for $d = 0$, and $+0.04$ for $d = 4$). These near-zero scores suggest that the dual-stage methodology is highly robust: the parametric moment-matching adequately captures the structural heavy tails of the convolution, while the conformal framework acts as an efficient mathematical safety net, triggering corrections when the expected dispersion slightly underestimates the true epidemiological variance.

The Zero-Observation Edge Case

Finally, we analyzed the fallback behavior of the framework for unpopulated or entirely silent municipalities during a given epidemiological week ($n_k = 0$). In these scenarios, the

non-conformity score collapses simply to the raw number of final cases ($s = N_k$).

The empirical thresholds extracted for these empty regions provide a highly actionable operational rule for public health officials. At $d = 0$, the conformal cutoff was calibrated at 3 cases. This implies that if a region reports zero cases in the current week, its true final tally will rarely exceed 3 cases with 95% marginal confidence. At $d = 1$, this uncertainty threshold drops to 1 case. Most notably, for any delay $d \geq 2$, the calibrated threshold reaches 0.

Epidemiologically, this suggests that the reporting system’s latency is generally resolved within the first two weeks for initially "silent" regions. If a Health Region has registered zero partial notifications by the second week of evaluation, the conformal framework bounds the final count to safely remain at zero.

3.4.2 Internal Evaluation of Conformal Predictors

Before deploying the framework to the out-of-time evaluation set (2S2024), we conducted an internal assessment to determine which of the four proposed conformal configurations—residual-based (Sym-CP and Asym-CP) and quantile-based (CQR-NegBin and CQR-Gamma)—provided the optimal balance for the nowcasting task. The strategies were evaluated across four main axes: marginal coverage validity, interval efficiency (width), and point estimate accuracy. To rigorously assess the stability of these methods, the internal evaluation was conducted through a 5-fold cross-validation framework over the combined historical data. Consistent with the initial partitioning methodology, these folds were constructed by randomly shuffling the temporal index (epidemiological weeks) to ensure a representative mixture of transmission dynamics within each split.

Coverage Validation

A fundamental requirement of any Conformal Prediction framework is to guarantee that the empirical marginal coverage reaches the nominal $1 - \alpha$ level, which is set to 95%.

Figure 13 illustrates the empirical marginal coverage across different reporting delays. The error bars denote the 95% confidence intervals for the mean coverage across the 5 folds, derived using the Student’s t-distribution to account for the cross-validation scheme.

The results highlight significant behavioral differences among the evaluated non-conformity scores, particularly regarding over-coverage. The parametric moment-matched Negative Binomial approach (CQR-NegBin) exhibits systematic over-coverage that progressively exacerbates as the reporting delay increases, reaching approximately 97.7% for delays of 5 weeks. In the context of conformal inference, this pronounced over-coverage indicates that the discrete nature of the Negative Binomial distribution forces the algorithm to generate prediction intervals that are unnecessarily conservative.

Conversely, the remaining three methods (Sym-CP, Asym-CP, and CQR-Gamma) suc-

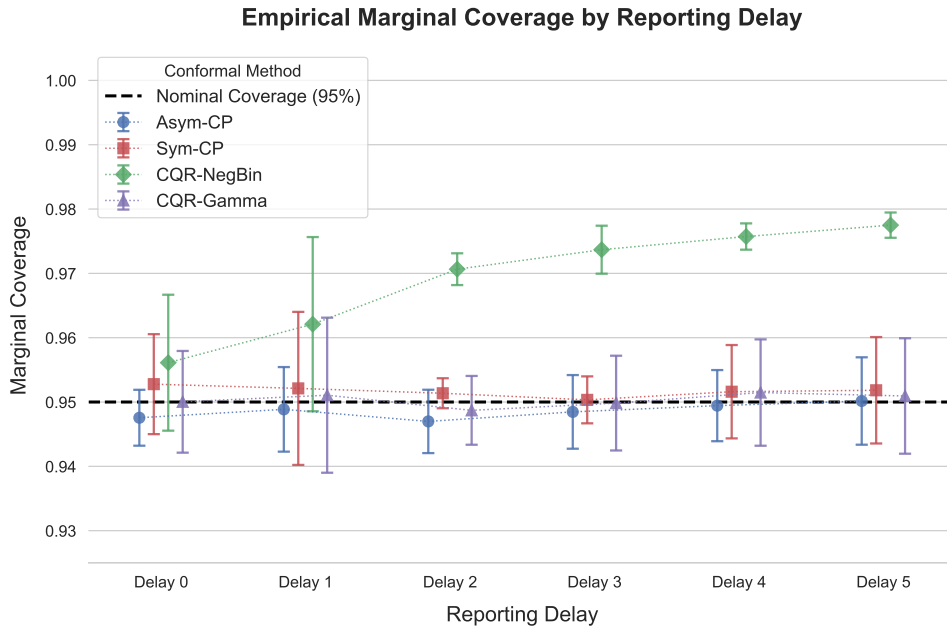


Figure 13 – Empirical marginal coverage of the 95% prediction intervals across reporting delays, evaluated via 5-fold cross-validation. Error bars represent the 95% confidence intervals for the mean coverage. While the continuous (CQR-Gamma) and residual-based methods (Sym-CP, Asym-CP) successfully center around the 95% nominal target, the discrete moment-matched strategy (CQR-NegBin) exhibits systematic and progressive over-coverage. Furthermore, despite its independent tail calibration, the Asym-CP method maintains valid coverage without becoming overly conservative, demonstrating consistent stability across different calibration folds.

cessfully centered their average empirical coverage near the 95% nominal target across the delay strata. However, their stability across the cross-validation folds varied. While Sym-CP occasionally achieved the tightest confidence interval at specific delays (e.g., $d = 2$), Asym-CP demonstrated the most consistent variance of coverage overall.

Notably, despite the theoretical risk of over-coverage associated with Asym-CP—arising from the independent calibration of the lower and upper tails, as discussed in the residual analysis—this empirical evaluation confirms that the method maintains the 95% target without becoming overly conservative. This suggests that accommodating the inherent right-skewness of the count distributions via independent tail calibration allows the asymmetric strategy to achieve valid marginal coverage while preserving cross-fold stability.

Interval Efficiency and Ablation of Physical Constraints

While conformal calibration guarantees valid marginal coverage, the practical utility of a prediction interval in epidemiological surveillance heavily depends on its efficiency, defined by its width. A valid interval is only actionable if it is sufficiently narrow to inform decision-making. To rigorously evaluate this, we analyzed the mean interval width across the 5-fold cross-validation scheme. Because the absolute widths at later delays ($d \geq 3$) become highly compressed, we evaluated the relative performance of the methods. Specifically, we computed the

width difference for each health region and epidemiological week relative to the CQR-NegBin strategy, and subsequently averaged these individual differences for each delay.

Furthermore, to isolate how domain-specific restrictions affect the final efficiency of each non-conformity score, we conducted an ablation study tracking the intervals through three states: Raw (unconstrained), Constrained (physically bounded by observed cases), and Rounded (discrete integer mapping) (Figure 14).

In the raw, unconstrained continuous space, the moment-matched methods (CQR-NB and CQR-Gamma) appeared structurally more efficient. Conversely, the Sym-CP method exhibited the widest intervals at early delays, averaging roughly 2.2 cases wider than the baseline at $d = 0$ (Figure 14b). This initial inefficiency of Sym-CP is a mathematical consequence of its absolute residual construction. By accumulating uncertainty into a single absolute margin ($|N_k - \hat{N}_k|$) dictated by the extreme right tail of the outbreak, Sym-CP applies an equal margin to the left tail. Consequently, it routinely generates lower bounds that extend below the already reported cases (n_k).

Because the true total count (N_k) cannot be smaller than the partial cumulative cases already reported, we enforced a physical lower-bound restriction. The raw intervals $[L, U]$ were truncated using $\max(L, n_k)$ and $\max(U, n_k)$.

Applying this operational constraint altered the efficiency landscape (Figure 14c and Figure 14d). The Sym-CP method eliminated its invalid predictive region on the left tail, undergoing a notable reduction in interval width. Its relative difference to the baseline decreased from +2.2 to approximately +0.4 cases at $d = 0$. This highlights a theoretical feature of the symmetric absolute residual: while it is initially wider in the unbounded continuous space, once the lower bound is restricted by the observed cases, the accumulated residual on the right tail provides an efficient upper bound.

Finally, because epidemiological counts are strictly discrete ($N_k \in \mathbb{N}$), the physically bounded continuous limits were mapped to their nearest valid integers. To maintain the theoretical distribution-free coverage guarantee, the continuous lower bound was mapped to its ceiling, and the upper bound to its floor, yielding the fully operational interval: $[\max(\lceil L \rceil, n_k), \max(\lfloor U \rfloor, n_k)]$.

Following these adjustments, the continuous methods shed their remaining fractional excess (Figure 14f). The Asym-CP method remains the widest approach across early delays (approximately 1.2 cases wider than the baseline at $d = 0$). This increased width reflects its sensitivity to the data's asymmetry: by independently calibrating the upper and lower bounds to achieve 97.5% nominal coverage in each direction, the method shifts the upper bound higher to accommodate the long right tail, resulting in a wider interval to achieve the overall 95% marginal coverage.

Conversely, CQR-Gamma and Sym-CP stabilize as the most efficient methods, producing intervals narrower than the baseline (roughly -0.4 and -0.2 , respectively, at $d = 0$). While

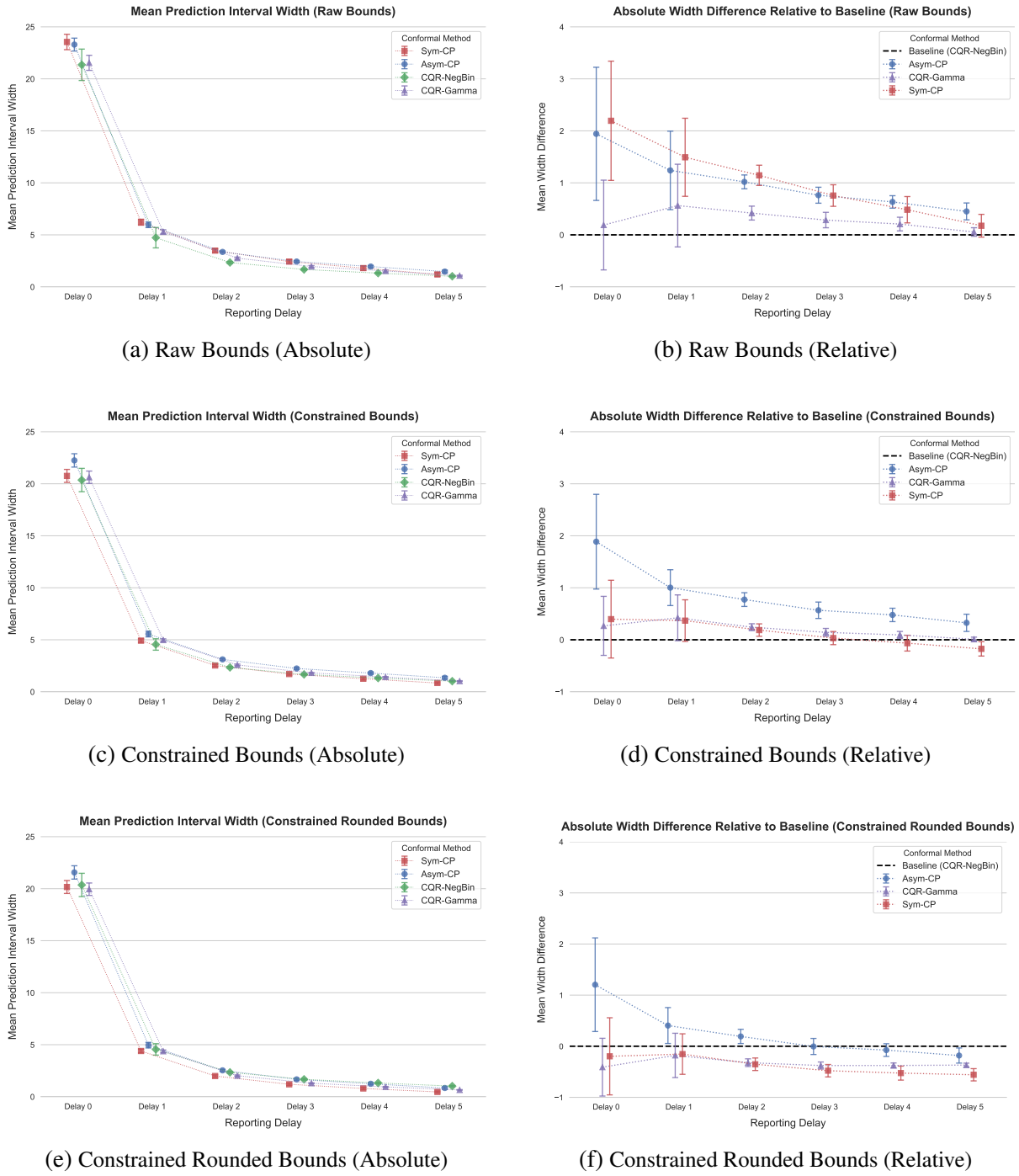


Figure 14 – Ablation study of prediction interval efficiency across reporting delays. The left column displays the mean absolute interval width, while the right column presents the mean width difference relative to the CQR-NB baseline. Rows depict the progressive application of domain constraints: unconstrained continuous bounds (a, b), physically constrained bounds ensuring values are $\geq n_k$ (c, d), and fully constrained rounded bounds mapped to discrete integers (e, f). While Sym-CP initially appears highly inefficient in the unconstrained space (b), applying physical lower bounds drastically amputates its invalid probability mass, neutralizing the gap (d). Under fully operational conditions (f), Sym-CP and CQR-Gamma emerge as the most efficient strategies, dropping below the discrete baseline.

a mean reduction of a fraction of a case is operationally small in high-volume regions, it demonstrates that when physical constraints and discrete domain mappings are applied, the residual-based and continuous conformal frameworks deliver valid $1 - \alpha$ marginal coverage with improved efficiency compared to the strictly discrete CQR-NB baseline.

Point Estimate Accuracy

While the primary objective of the conformal framework is to quantify uncertainty through valid prediction bounds, the accuracy of the underlying central estimate remains essential for epidemiological baseline tracking. To evaluate this, we analyzed the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) across the 5-fold cross-validation scheme (Table 7).

Table 7 – Point Estimate Accuracy (MAE and RMSE) Across Reporting Delays. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) evaluated via 5-fold cross-validation, stratified by reporting delay. The 95% confidence intervals are provided in parentheses. The lowest error for each metric within a delay stratum is highlighted in bold. Note that Sym-CP and Asym-CP share the exact same point estimates, as both are anchored on the conditional mean of the underlying regression model.

Delay	Conformal Method	MAE	RMSE
		Mean (95% CI)	Mean (95% CI)
Delay 0	Sym/Asym-CP	4.8956 (4.6693, 5.1219)	10.0735 (9.2063, 10.9407)
	CQR-NB	4.7596 (4.5103, 5.0088)	10.0507 (9.1408, 10.9605)
	CQR-Gamma	4.7446 (4.4917, 4.9975)	10.0497 (9.1301, 10.9693)
Delay 1	Sym/Asym-CP	1.2233 (1.0917, 1.3549)	3.3885 (2.7895, 3.9875)
	CQR-NB	1.1830 (1.0441, 1.3219)	3.3935 (2.7827, 4.0043)
	CQR-Gamma	1.1729 (1.0330, 1.3128)	3.3968 (2.7847, 4.0089)
Delay 2	Sym/Asym-CP	0.6153 (0.5666, 0.6640)	2.1878 (2.0017, 2.3739)
	CQR-NB	0.5924 (0.5458, 0.6391)	2.1887 (2.0021, 2.3753)
	CQR-Gamma	0.5870 (0.5410, 0.6330)	2.1896 (2.0018, 2.3773)
Delay 3	Sym/Asym-CP	0.4229 (0.3877, 0.4580)	1.7902 (1.5790, 2.0015)
	CQR-NB	0.4074 (0.3723, 0.4424)	1.7892 (1.5746, 2.0039)
	CQR-Gamma	0.4044 (0.3696, 0.4392)	1.7930 (1.5808, 2.0053)
Delay 4	Sym/Asym-CP	0.3304 (0.2996, 0.3612)	1.6049 (1.3904, 1.8194)
	CQR-NB	0.3211 (0.2912, 0.3511)	1.6078 (1.3902, 1.8254)
	CQR-Gamma	0.3176 (0.2893, 0.3458)	1.6094 (1.3918, 1.8269)
Delay 5	Sym/Asym-CP	0.2633 (0.2393, 0.2873)	1.4818 (1.2763, 1.6873)
	CQR-NB	0.2549 (0.2313, 0.2786)	1.4826 (1.2740, 1.6911)
	CQR-Gamma	0.2526 (0.2287, 0.2764)	1.4878 (1.2811, 1.6944)

It is important to note that the conformal calibration procedure strictly manipulates the predictive intervals. Consequently, the symmetric and asymmetric residual methods (Sym/Asym-CP) share the exact same point estimates, as both are anchored on the conditional expectation (the mean) generated by the underlying Gradient Boosting model. In contrast, the parametric

methods (CQR-NB and CQR-Gamma) derive their point estimates from the conditional median of their respective probability density functions, which are moment-matched to the algorithmic output.

Across all methodologies, the point estimate errors decrease rapidly as reporting data accumulates. At the initial evaluation point ($d = 0$), the MAE ranges from 4.74 to 4.89 cases. By the second week ($d = 1$), the absolute error drops to approximately 1.2 cases, eventually converging to roughly 0.25 cases at $d = 5$. This trajectory demonstrates that the core predictive model successfully learns the reporting delay surface, providing accurate central estimates prior to the application of the conformal bounds.

Furthermore, a comparative analysis of the error metrics reveals a statistical phenomenon governed by the central tendency measures utilized by each method. The continuous parametric approach (CQR-Gamma) consistently achieves the lowest MAE across all delay strata. This is mathematically consistent with its architecture, as the median of a distribution is the theoretical estimator that minimizes the absolute error (L_1 loss).

Conversely, the residual-based approaches (Sym/Asym-CP) generally yield the lowest RMSE, outperforming the parametric methods in delays 1, 2, 4, and 5. Because the RMSE heavily penalizes large estimation errors (via the L_2 loss), it is particularly sensitive to the right-skewness and extreme outbreak peaks characteristic of SARI notifications. The Sym/Asym-CP methods excel under this metric because they are anchored on the conditional mean, which is the optimal estimator for minimizing squared errors.

Ultimately, these results indicate that all configurations yield highly competitive point estimates. The selection between them relies on the specific priorities of the surveillance system: minimizing standard absolute deviations (favoring CQR-Gamma) versus minimizing the penalty for extreme underestimations during sudden outbreaks (favoring Sym/Asym-CP).

Model Selection

The internal cross-validation assessment demonstrated that while all proposed methods—with the exception of the strictly discrete CQR-NB—successfully achieved the 95% nominal coverage, they presented distinct operational trade-offs.

The CQR-NB configuration is excluded from the final selection primarily due to its systematic over-coverage, which fails to tightly calibrate to the strict $1 - \alpha$ marginal target. Conversely, the Asym-CP method successfully maintains exact nominal coverage but incurs the highest efficiency penalty. By independently calibrating its upper and lower bounds to accommodate the distribution's skewness, it consistently generates the widest intervals among the valid methods, which is less optimal for actionable epidemiological decision-making.

This leaves Sym-CP and CQR-Gamma as the most efficient contenders. While Sym-CP achieves highly competitive interval widths after operational constraints are applied, its raw

formulation relies on an artificial symmetric assumption, accumulating structurally impossible predictive mass on the negative tail before truncation. In contrast, CQR-Gamma operates natively over the positive real domain and inherently models the natural right-skewness of epidemiological counts. This inherent asymmetry provides a mathematically elegant and structurally coherent approximation of the underlying distribution before the final integer mapping.

Furthermore, the point estimate analysis revealed that CQR-Gamma consistently delivers the lowest Mean Absolute Error (MAE) across all delay strata. In the context of public health surveillance, minimizing the absolute deviation of the central estimate provides a highly stable and interpretable baseline for weekly epidemiological tracking.

Therefore, because it offers the optimal balance of exact marginal coverage, optimized structural efficiency, inherent distributional asymmetry, and superior point estimate accuracy, the **CQR-Gamma** strategy is selected as the definitive conformal configuration. This method will proceed to the final phase of this study: the strict out-of-time evaluation on the unseen 2S2024

3.4.3 Comparison with the State-of-the-Art (Bayesian Benchmark)

To validate the proposed framework under a real-world scenario, we conducted an out-of-time evaluation utilizing unseen data from the second semester of 2024 (2S2024). For this evaluation, all reporting instances—including initially silent municipalities where zero partial cases were observed ($n_k = 0$)—were included. Therefore, the proposed estimator evaluated in this section represents our complete conformal pipeline: the GBM-driven CQR-Gamma strategy combined with the deterministic Zero-Count Conformal Predictor for boundary edge cases.

To contextualize the performance of this machine learning-based approach, we compared it against a highly established methodology in epidemiological nowcasting: a Bayesian hierarchical spatial model.

The Bayesian Benchmark Model and Operational Differences

The chosen benchmark is the Bayesian framework formulated by [Bastos et al. \(2019\)](#), adapted for the spatial surveillance of Severe Acute Respiratory Infection (SARI) in Brazil. To maintain notational consistency with our proposed methodology, we denote the spatial index (health regions) as r . This generative model treats the delayed counts $n_{t,d,r}$ (representing the number of cases occurring at epidemiological week t , reported with delay d , in health region r) as realizations of a Negative Binomial distribution with mean $\lambda_{t,d,r}$ and overdispersion parameter ϕ . The logarithmic mean is structured through additive latent random effects, defined as:

$$n_{t,d,r} \sim \text{NegBin}(\lambda_{t,d,r}, \phi)$$

$$\log(\lambda_{t,d,r}) = \mu + \alpha_t + \beta_d + \gamma_{t,d} + \beta_{d,r} + \psi_r^{\text{IAR}} + \psi_r^{\text{ind}}$$

In this formulation, μ serves as the global intercept. The linear predictor incorporates

structural components to capture the reporting dynamics: a temporal evolution component (α_t) modeled as a random walk, a baseline delay structure (β_d), and a time-delay interaction ($\gamma_{t,d}$) to account for temporal shifts in the reporting mechanism. For spatial data resolved at the health region level, the model utilizes a Besag-York-Mollié (BYM) specification—combining an Intrinsic Autoregressive (IAR) process (ψ_r^{IAR}) to account for spatial correlation among neighboring administrative regions, and an unstructured spatial effect (ψ_r^{ind}) to capture region-specific heterogeneity. Additionally, it includes an unstructured region-delay interaction ($\beta_{d,r}$) to accommodate local administrative variations in reporting speeds. Inference is performed using the Integrated Nested Laplace Approximation (INLA) approach.

Before detailing the performance metrics, a fundamental operational distinction between the two methodologies must be highlighted. Our proposed framework relies on a historically static training and calibration set; once calibrated, it generates predictive bounds for any incoming data without requiring model updates. Conversely, owing to its structural design to infer missing counts conditionally up to the current moment, the Bayesian hierarchical model must be computationally refitted from scratch at each successive epidemiological week.

By comparing our gradient boosting-driven pipeline against this iteratively refitted Bayesian spatial model, we evaluate the trade-offs between predictive accuracy, uncertainty robustness, and computational efficiency during the out-of-sample temporal window.

Coverage Validation

The objective of this out-of-time evaluation is to assess whether the predictive frameworks sustain their validity when deployed sequentially on unseen future data. [Figure 15](#) illustrates the empirical marginal coverage of both models across the reporting delays. The 95% confidence intervals in this analysis are derived from the variance of the mean coverage across the 26 individual epidemiological weeks comprising the 2S2024 evaluation period.

The results indicate that both approaches provide valid prediction regions, as their empirical coverages remain above the 95% nominal target. Consequently, both methods exhibit varying degrees of over-coverage.

The Bayesian Benchmark displays a more pronounced over-coverage across all delay strata, consistently exceeding 98% and reaching approximately 98.5% for delays 1 through 4. This behavior reflects the accumulation of uncertainty from the multiple latent spatial and temporal components, combined with the discrete nature of the Negative Binomial marginal distribution.

The proposed CQR-Gamma method, while also conservative, maintains an empirical coverage closer to the nominal target. This difference in calibration is particularly evident at $d = 0$ —the most operationally critical and uncertain reporting delay—where the conformal coverage centers near 96.4%, compared to the benchmark’s $\approx 98\%$. At later delays, the CQR-

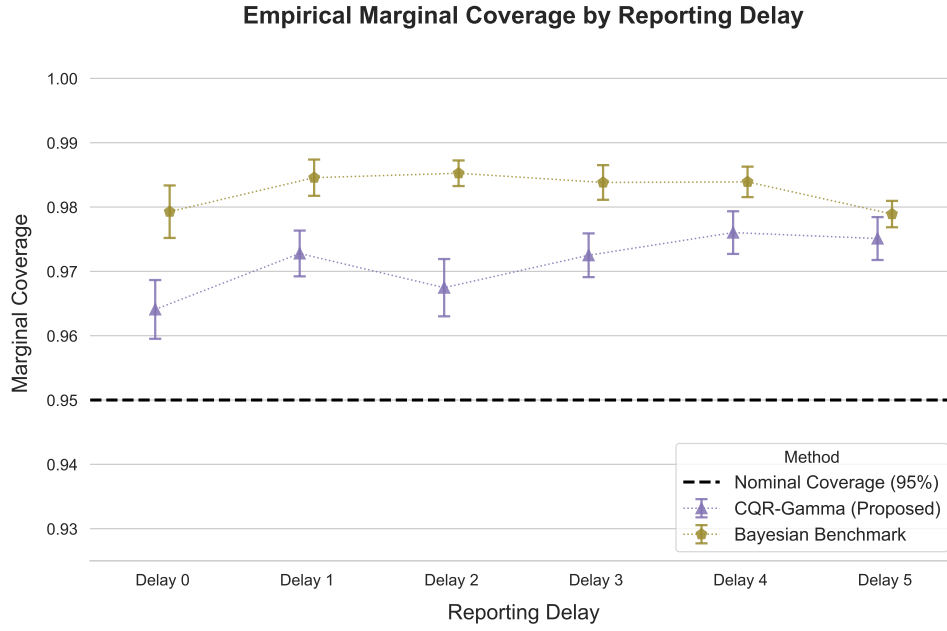


Figure 15 – Empirical marginal coverage of the 95% prediction intervals evaluated on the out-of-time test set (2S2024). Error bars represent the 95% confidence intervals derived from the variance across the 26 epidemiological weeks. Both methods achieve the $\geq 95\%$ validity requirement, thereby exhibiting over-coverage. However, the Bayesian Benchmark shows a more pronounced over-coverage ($\approx 98.5\%$), whereas the proposed CQR-Gamma framework provides improved calibration, particularly at the critical initial delay ($d = 0$).

Gamma coverage stabilizes near 97.5%.

It is important to note that the inclusion of initially silent municipalities (zero-count edge case, $n_k = 0$) naturally inflates the overall empirical coverage due to their deterministic boundary conditions. When isolating the evaluation strictly to active reporting regions (excluding $n_k = 0$), the calibration advantage of the proposed framework becomes even more pronounced. Specifically, at the critical initial delay ($d = 0$), the CQR-Gamma model achieves an empirical coverage of 95.04% ($\pm 1.07\%$), perfectly aligning with the nominal 95% target. In contrast, the Bayesian baseline maintains a conservative over-coverage of 96.78% ($\pm 1.54\%$) under the same conditions. This tight calibration of the proposed framework remains highly stable across subsequent delays (ranging from 95.5% to 96.3%), whereas the Bayesian benchmark systematically over-covers between 96.6% and 97.7%. This indicates that when outbreak dynamics are actively unfolding, the conformal procedure provides well-calibrated prediction intervals that closely track the nominal target, mitigating the tendency for over-coverage.

This performance in the out-of-time test set corroborates the theoretical property of *model-based robustness* (ANGELOPOULOS; BARBER; BATES, 2024) established in our framework. By actively modeling the macroscopic drift and heteroscedasticity through Gradient Boosting, the baseline estimator effectively extracts the temporal signal, rendering the resulting non-conformity scores approximately exchangeable.

Interval Efficiency Validation

While valid prediction regions are essential, the operational utility of a surveillance model also depends on its efficiency, defined by interval width. Wide intervals, though mathematically valid, may fail to provide actionable epidemiological thresholds. To evaluate efficiency, we analyzed the mean prediction interval width across reporting delays in the 2S2024 test set.

Figure 16 presents the absolute mean interval width for each method (left panel) and the mean paired width difference (Benchmark minus Proposed) at the health region level (right panel). Confidence intervals represent the temporal variability across the 26 evaluation weeks.

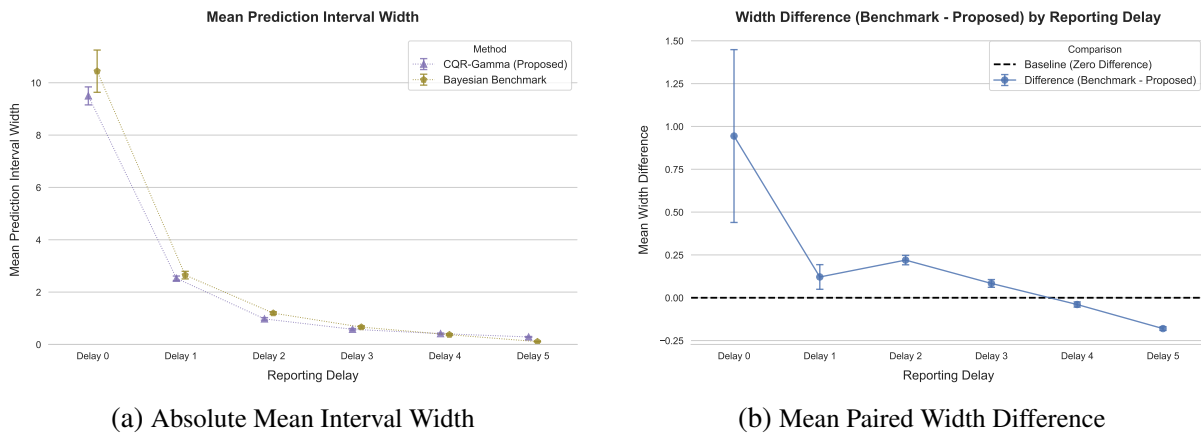


Figure 16 – Efficiency comparison of the 95% prediction intervals evaluated on the 2S2024 test set. The left panel (a) displays the absolute mean width decay across delays. The right panel (b) isolates the paired difference (Benchmark minus Proposed), where positive values indicate the proposed CQR-Gamma is narrower (more efficient). Error bars denote the 95% confidence intervals across the 26 test weeks. The CQR-Gamma framework is more efficient during the early delays ($d \leq 3$), reflecting its tighter calibration compared to the Bayesian baseline.

The absolute widths (Figure 16a) exhibit a nonlinear decay. At $d = 0$, where missing data is highest, the intervals are widest, compressing rapidly by $d = 1$ and approaching zero at later delays.

Analyzing the relative differences (Figure 16b), the CQR-Gamma framework is more efficient during the early delays ($d = 0$ to $d = 3$). At $d = 0$, the Bayesian Benchmark generates intervals that are, on average, 1.0 case wider than the proposed method. This difference persists through the first month. This behavior aligns with the coverage results: the Bayesian model's over-coverage corresponds to wider intervals, whereas CQR-Gamma achieves tighter calibration, reducing interval width while maintaining the 95% nominal target.

At delays $d = 4$ and $d = 5$, a minor inversion occurs, with the Benchmark becoming fractionally narrower (differences of roughly -0.05 and -0.20 cases, respectively). This reflects a methodological difference in handling long-tail reporting delays. The Bayesian baseline assumes all cases are fully observed within the maximum evaluation window. In contrast, our proposed methodology actively models cases reported at $d \geq 6$ as a distinct aggregated category

during training to account for extreme delays. Consequently, CQR-Gamma retains a slight residual uncertainty at later delays to cover these long-tail events. Operationally, this fractional width increase at $d = 5$ is a structural adjustment for late reports, while the efficiency gain at $d = 0$ provides sharper bounds for real-time detection.

Overall, the empirical results indicate that the CQR-Gamma pipeline is more efficient than the Bayesian benchmark during the early reporting window, providing tighter thresholds while satisfying the 95% marginal coverage guarantee.

Point Estimate Accuracy Validation

To complete the out-of-time evaluation, we assessed the accuracy of the central point estimates. Table 8 presents the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across the reporting delays for the 2S2024 test set.

Table 8 – Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) evaluated on the out-of-time test set (2S2024), stratified by reporting delay. The 95% confidence intervals, derived from the variance across the 26 epidemiological weeks, are provided in parentheses. The lowest error for each metric within a delay stratum is highlighted in bold.

Delay	Method	MAE	RMSE
		Mean (95% CI)	Mean (95% CI)
Delay 0	CQR-Gamma (Proposed)	2.0807 (1.9422, 2.2192)	5.4434 (4.9971, 5.8897)
	Bayesian Benchmark	1.6923 (1.5523, 1.8323)	4.2899 (3.8910, 4.6887)
Delay 1	CQR-Gamma (Proposed)	0.5083 (0.4764, 0.5402)	1.5065 (1.3718, 1.6412)
	Bayesian Benchmark	0.4940 (0.4612, 0.5269)	1.5948 (1.4245, 1.7651)
Delay 2	CQR-Gamma (Proposed)	0.2335 (0.2135, 0.2536)	0.8496 (0.7741, 0.9252)
	Bayesian Benchmark	0.2462 (0.2251, 0.2673)	1.1329 (0.9847, 1.2812)
Delay 3	CQR-Gamma (Proposed)	0.1419 (0.1308, 0.1531)	0.6382 (0.5762, 0.7003)
	Bayesian Benchmark	0.1697 (0.1533, 0.1860)	1.0688 (0.9467, 1.1910)
Delay 4	CQR-Gamma (Proposed)	0.1087 (0.0980, 0.1193)	0.6209 (0.5271, 0.7147)
	Bayesian Benchmark	0.1338 (0.1222, 0.1453)	1.1343 (1.0050, 1.2635)
Delay 5	CQR-Gamma (Proposed)	0.0839 (0.0737, 0.0941)	0.6229 (0.5027, 0.7431)
	Bayesian Benchmark	0.1082 (0.0984, 0.1181)	1.1611 (1.0078, 1.3144)

The results reveal a clear performance dynamic between the initial and subsequent delays. At $d = 0$, the Bayesian Benchmark achieves lower errors in both metrics. This initial advantage reflects an architectural difference between the models. The Bayesian framework is natively formulated and optimized at the aggregate health region level, directly fitting the regional counts. Conversely, the proposed GBM framework operates as a bottom-up estimator: it models delay probabilities at the individual patient level and aggregates these expectations to form the regional point estimate. While this individual-level approach captures fine-grained demographic and spatial signals, the aggregation is not directly optimized for regional-level loss minimization, resulting in a point estimate penalty at the period of maximum uncertainty ($d = 0$).

However, this aggregation effect does not compromise the final prediction regions. Because the conformal calibration (CQR-Gamma) is applied directly at the aggregate regional level, it corrects the aggregation bias. It translates the individual-level signals into the calibrated and efficient bounds demonstrated in the previous sections.

From $d = 1$ through $d = 5$, the proposed pipeline consistently outperforms the Bayesian baseline in both MAE and RMSE. As the initial reporting volume stabilizes, the individual-level features—such as patient age and exact spatial coordinates—provide a more precise estimation of the remaining delayed cases. Furthermore, by explicitly modeling the long-tail delays ($d \geq 6$) as an aggregated category during training, the machine learning estimator better captures the residual reporting dynamics.

In contrast, the parametric Bayesian model—which assumes all cases are bounded within the maximum delay window and lacks individual-level covariates—exhibits an error floor. Its RMSE stabilizes near 1.1, while the proposed method’s error continues to decay to approximately 0.6. Overall, the proposed framework leverages fine-grained individual data to deliver improved point estimate accuracy throughout the majority of the reporting cycle, while the conformal procedure ensures valid uncertainty quantification at the aggregate decision-making level.

Computational Scalability and Execution Time

Beyond predictive accuracy and interval efficiency, the operational deployment of a national-scale surveillance system is constrained by computational scalability. To quantify this aspect, we compared the execution times required to process the 2S2024 out-of-time evaluation for both methodologies.

The Bayesian hierarchical spatial model requires a complete structural refitting at each successive epidemiological week to incorporate new data and conditionally infer the missing counts. On average, estimating the posterior distributions for the 438 health regions via INLA took approximately 3 hours per weekly update. Consequently, processing the sequential nowcasting for the entire 26-week evaluation period demanded 78 hours (3.25 days) of computation.

In contrast, the proposed CQR-Gamma framework utilizes a static historical calibration approach, decoupling the learning phase from the prediction phase. The initial training of the Gradient Boosting probabilistic classifier on the historical set required 1 minute, followed by 10 seconds to calibrate the non-conformity scores via the Split Conformal procedure. Once calibrated, generating the point estimates and prediction bounds for the unseen data of all 26 future weeks required only 6 seconds of inference time. In total, the complete machine learning pipeline was executed in 1 minute and 16 seconds.

This contrast in execution time (approximately 1.25 minutes compared to 3.25 days) highlights a fundamental operational advantage. The proposed framework eliminates the computational bottleneck inherent to iterative spatial modeling, enabling near-instantaneous real-time

inference without sacrificing statistical validity.

Real-Time Nowcasting Simulation

To visually contextualize the operational implications of these performance metrics, we simulated a real-time surveillance scenario. Figure 17 illustrates the nowcasting performance for two major demographic hubs—the São Paulo and Rio de Janeiro Health Regions—across three distinct evaluation cutoffs (August 11, September 29, and November 17, 2024). These vertical cutoffs mimic the exact moment a public health official would consult the system.

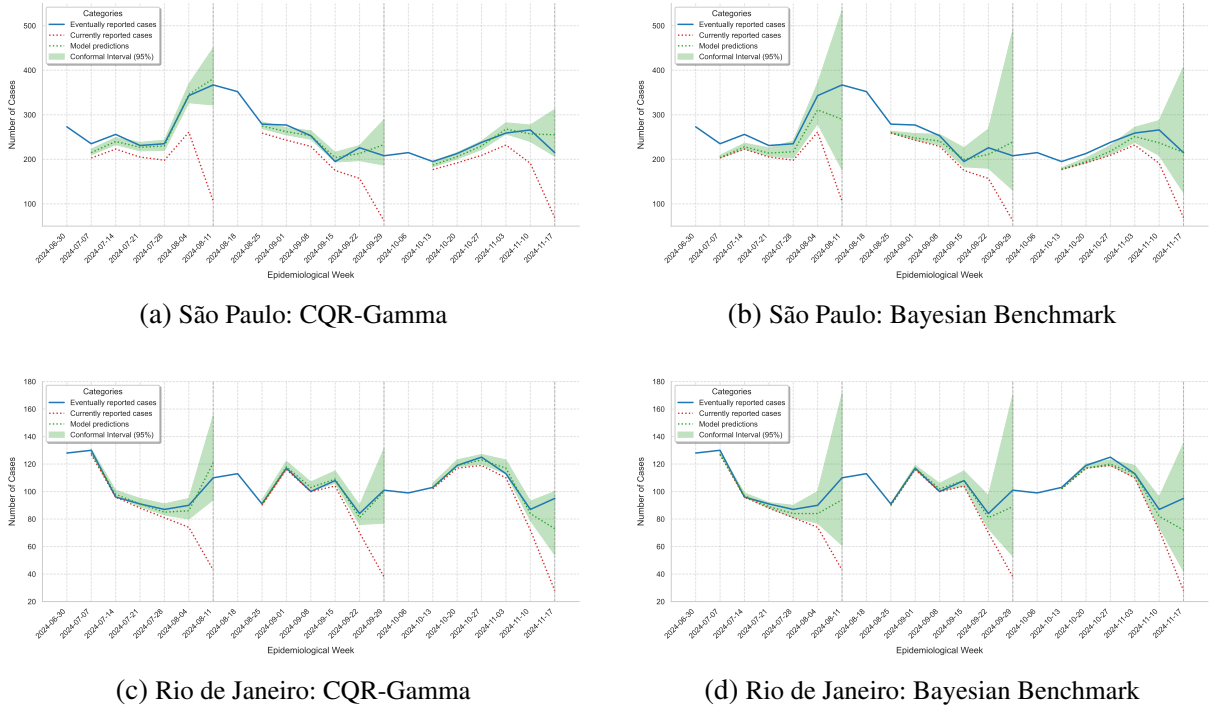


Figure 17 – Real-Time Nowcasting Simulation. The panels compare the predictive behavior of the proposed CQR-Gamma framework (left) and the Bayesian Benchmark (right) for the São Paulo and Rio de Janeiro Health Regions. Vertical dashed lines represent three distinct operational cutoffs (simulating the "present moment" on Aug 11, Sep 29, and Nov 17, 2024). Both methods effectively recover the artificially truncated incidence curves (dotted red line). However, the proposed conformal framework achieves this with significantly tighter and more efficient prediction intervals at the critical $d = 0$ prediction point, overcoming the wide, conservative uncertainty funnels characteristic of the Bayesian spatial model.

The raw delayed counts (dotted red lines) clearly demonstrate the severe right-censoring inherent in the surveillance system, exhibiting an artificial, steep drop just before each evaluation date. While both models successfully correct this incidence drop—their point estimates (dotted green lines) closely tracking the eventually reported cases (solid blue lines)—the distinct behaviors of their uncertainty bounds are evident. The Bayesian Benchmark systematically generates wide, funnel-like credible intervals at the critical $d = 0$ threshold. In contrast, the proposed CQR-Gamma framework leverages the locally adaptive non-conformity scores to deliver highly efficient, tightly calibrated conformal intervals. By avoiding unnecessary conservatism precisely

at the moment of prediction, the proposed framework provides sharper operational thresholds for immediate epidemiological response.

3.4.4 Interpretability of the Delay Classifier

While the macroscopic aggregation and conformal calibration provide the required statistical bounds for operational nowcasting, understanding the underlying drivers of reporting latency is essential for public health planning. To achieve this, we investigate the internal mechanics of the probabilistic Gradient Boosting classifier developed in Step 2 of the methodology, which models the cumulative delay probabilities at the individual patient level. By extracting feature importance and analyzing Partial Dependence, we can evaluate whether the algorithm has learned epidemiologically coherent rules regarding the Brazilian health system's reporting speed.

Global Feature Importance

We begin by evaluating the global contribution of the input variables to the classifier's predictive performance. Figure 18 presents the relative feature importance ranked by Total Gain, which measures the improvement in accuracy brought by a feature across all its splits in the tree ensemble.

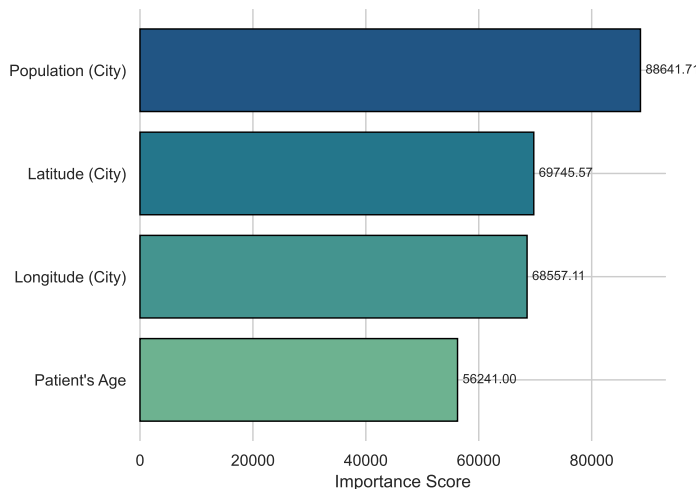


Figure 18 – Ranking of predictor variable significance according to total gain within the Gradient Boosting classifier. Geographic coordinates (Latitude and Longitude) and municipal population size strictly dominate the model's learning process, indicating that structural and regional logistics are the primary drivers of reporting delays, far outweighing individual demographic traits such as patient age.

The analysis reveals a strong hierarchy driven by logistical and structural factors rather than individual demographics. The continuous spatial coordinates (Latitude and Longitude) and the municipal population size emerge as the most impactful predictors. This distribution aligns closely with the operational reality of the Brazilian health system, where reporting delays are heavily influenced by geographic remoteness, regional infrastructure inequalities, and

the administrative capacity of different municipalities, whereas the patient's age plays only a marginal, fine-tuning role in the notification speed.

Local Interpretability: Demographics and Population Scale

To isolate and quantify how these specific variables influence the delay probabilities, we utilize Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) curves. [Figure 19](#) illustrates the marginal effects of patient age and municipal population size on the estimated cumulative probability of a case being reported within the critical early delays ($d \in \{0, 1, 2, 3\}$).

To properly interpret these curves, one must recall the aggregation mechanism defined in Step 3 of the methodology. Because the total unobserved count is estimated via Inverse Probability Weighting ($n_b/\hat{p}_b$), a lower cumulative probability implies a higher level of right-censoring, thus requiring a larger multiplier to correct the raw observations.

Focusing first on the critical initial evaluation point ($d = 0$), the baseline cumulative probability hovers around 0.4. Operationally, this means that, on average, the initial correction requires slightly more than doubling the observed values: for every 1 reported case, the framework estimates roughly 2.5 total cases ($1/0.4 = 2.5$) originating from that period.

Around this baseline, both variables exhibit specific non-linear behaviors. The partial dependence for patient age reveals a distinct, localized peak around 10 years of age (broadly between 5 and 15 years). This implies that SARI notifications for children in this bracket tend to be consolidated faster, thereby requiring the smallest average correction factor. Beyond this pediatric peak, the effect flattens significantly for the remainder of the lifespan, indicating that adult and elderly demographics do not introduce further structural variations in reporting speed. Conversely, municipal population size demonstrates a distinct, less smooth downward trend. For smaller municipalities, the estimated cumulative probability is generally higher, ranging from 0.40 to 0.45. However, as population size increases toward the largest observed municipalities (e.g., above 3 million inhabitants), the estimated cumulative probability declines and stabilizes at approximately 0.30. This pattern implies a significantly larger initial correction factor for these municipalities ($1/0.30 \approx 3.33$) compared with smaller cities.

As the temporal evaluation shifts to subsequent delays, the reporting information naturally consolidates, meaning the cumulative probability inherently approaches 1. This dynamic is rapidly evident by $d = 1$, where the baseline probability sharply increases to approximately 0.85 across the covariate space. At this stage, the average required correction drops drastically, adding only a marginal fractional weight to the observed counts. By $d = 2$ and $d = 3$, the curves flatten near the upper bound, indicating that the need for predictive adjustments is essentially neutralized as the raw surveillance system catches up to reality.

Spatial Partial Dependence: The Continuous Delay Surface

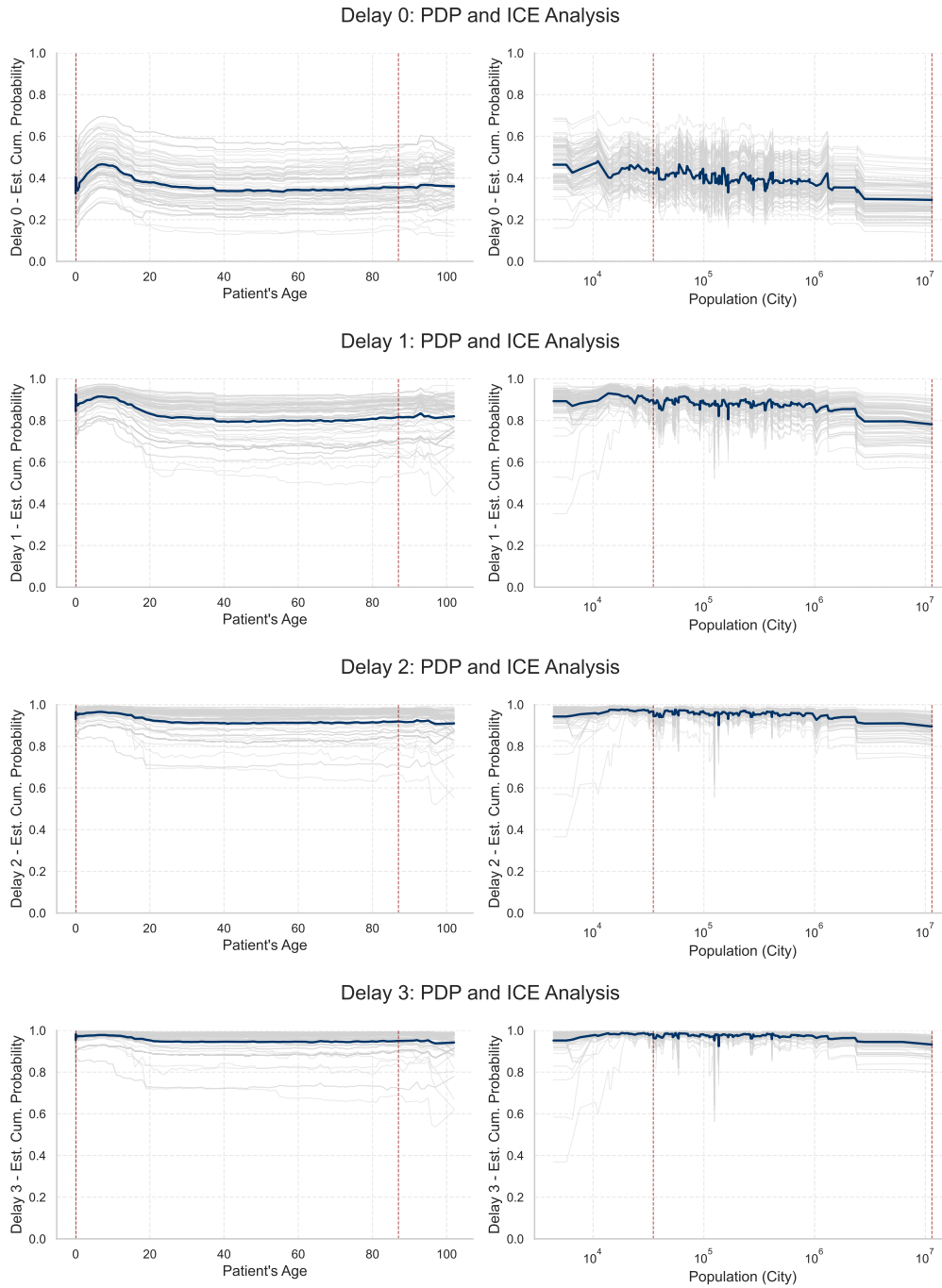


Figure 19 – Learned marginal effects of patient age and municipal population size on the estimated cumulative reporting probabilities for delays $d = 0$ to $d = 3$. For each delay, the left panel displays the effect of age (where the bold line represents the PDP, faint lines show ICE curves, and red dashed vertical lines indicate the 5th and 95th percentiles). The right panel displays the PDP for the municipal population size. The plots reveal nonlinear associations between the covariates and the estimated reporting probabilities. For both age and population size, the profiles vary more visibly at the earliest delays and become increasingly flat as the reporting information consolidates.

A primary methodological advantage of utilizing Gradient Boosting over traditional conditional autoregressive models is its ability to capture complex non-linear spatial patterns directly from continuous coordinates, entirely bypassing predefined regional boundaries. To visualize this learned spatial dynamic, we constructed a Spatial Partial Dependence Plot (Figure 20).

To construct this surface, rather than generating an arbitrary continuous grid, we utilized the exact latitude and longitude coordinates for all municipalities across the Brazilian territory. For each of these municipal geographic points, we computed the model’s average probability prediction across a representative random sample of 1,000 historical observations, effectively marginalizing the influence of all other variables (such as patient age and municipal population size).

The resulting spatial surfaces map the structural heterogeneities in Brazil’s surveillance network. With the color scales adjusted to isolate the core data distribution, the map at $d = 0$ reveals a stark logistical divide. The highest initial reporting probabilities (indicated by yellow and light green) are predominantly concentrated in the more developed South region and parts of the Midwest. In contrast, vast remote areas in the North (e.g., western Amazonas and Acre) and the deep interior of the Northeast exhibit the lowest probabilities (dark purple), reflecting severe geographic isolation and infrastructural delays. As the delay increases ($d \leq 1$), the overall probability mass shifts upward significantly across the territory, though some isolated remote pockets retain residual latency. By $d \leq 3$, the surface uniformly saturates near 1, demonstrating that the algorithm learned the extended consolidation timelines required for these structurally challenged regions.

3.5 Discussion

Timely epidemiological surveillance is fundamentally compromised by the inherent delays in the reporting of health events. In the context of Severe Acute Respiratory Infection (SARI) in Brazil, where spatial heterogeneity across 438 health regions severely complicates data consolidation, traditional parametric models often struggle to balance statistical rigor with operational agility. To address this, we proposed and evaluated a nowcasting framework combining Gradient Boosting with Split Conformal Prediction.

The empirical evaluations, culminating in an out-of-time assessment during the second semester of 2024, demonstrate that the proposed CQR-Gamma pipeline offers a robust alternative to established Bayesian hierarchical spatial models. The core achievement of this framework lies in its hybrid architecture. Instead of relying solely on exact parametric derivations for aggregated regional counts, our framework leverages machine learning to extract non-linear spatial and demographic signals at the individual patient level. While the subsequent regional aggregation incorporates baseline parametric approximations, the application of Split Conformal Prediction acts as a mathematical safeguard. It empirically corrects aggregation biases and unmodeled

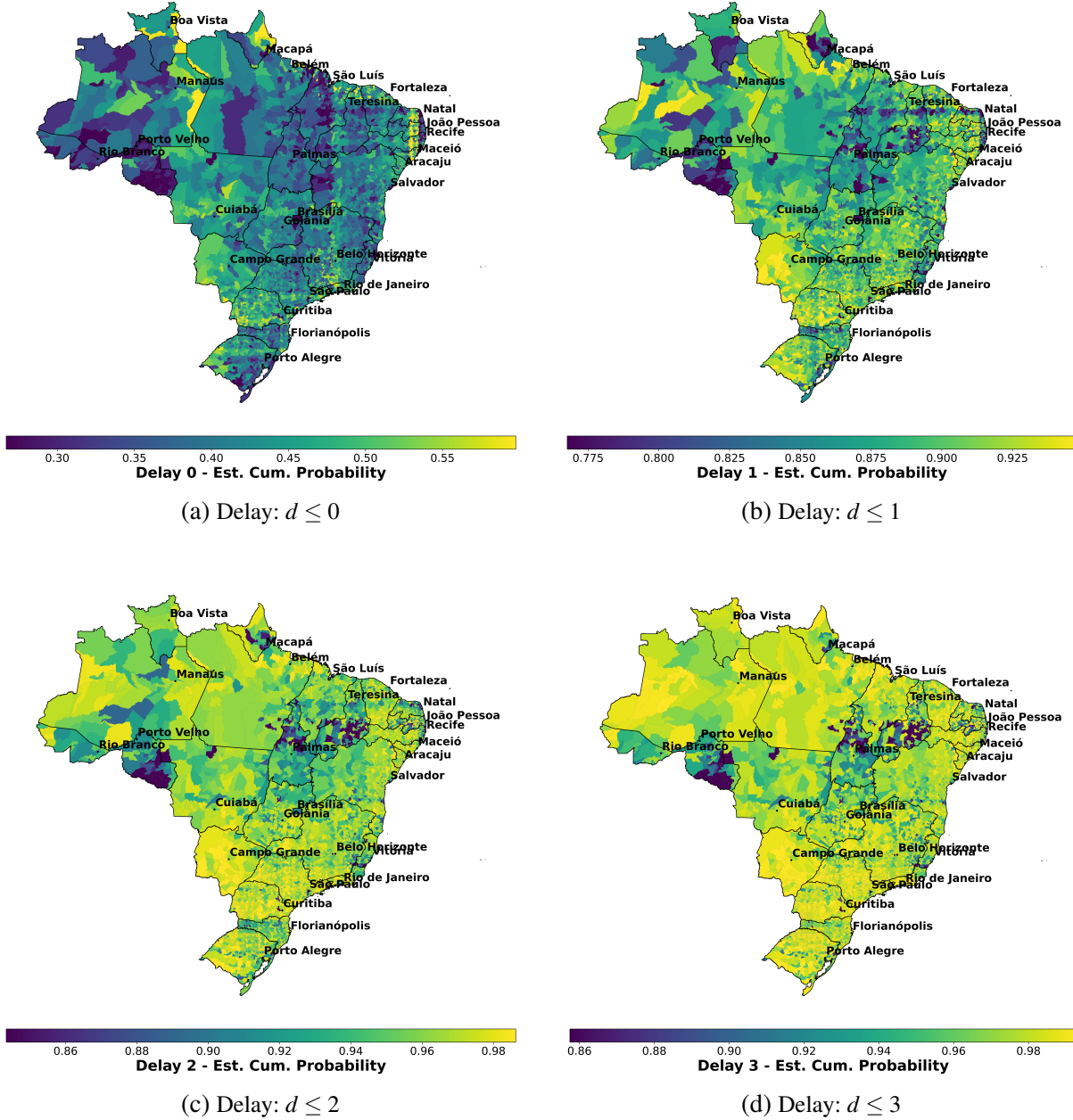


Figure 20 – Spatial Partial Dependence Plot mapping the isolated marginal effect of geography on cumulative reporting probabilities. By interpolating predictions over the exact municipal coordinates while marginalizing other covariates across a representative historical sample, the model reveals a stark logistical divide. The highest initial reporting probabilities ($d = 0$) are concentrated in the South and Midwest, while remote areas in the North and Northeast exhibit severe structural delays. Note that the color scales are dynamically adjusted to the bounds of each specific delay to highlight local gradients.

epistemic errors, ensuring valid coverage bounds adapted to the true data distribution.

The out-of-time evaluation over the second semester of 2024 provides empirical evidence that the proposed intervals remained well calibrated under the epidemiological and reporting conditions observed during that 26-week period. In particular, the coverage results indicate that the conformal correction was effective in absorbing residual uncertainty not captured by the baseline aggregation model. However, this empirical validation should be interpreted within the scope of that evaluation period. It does not directly quantify the temporal mixing penalty described by Oliveira *et al.* (2024), nor does it establish that the same level of calibration will necessarily hold during all epidemic regimes, especially under abrupt outbreak dynamics or major changes in reporting practices. Evaluating rolling out-of-time windows spanning multiple seasonal cycles, epidemic peaks, and administrative disruptions would provide a more comprehensive assessment of the stability of the conformal intervals under stronger temporal dependence.

Operationally, the framework demonstrates distinct efficiency during the critical early stages of the reporting cycle ($d \leq 3$). In real-time surveillance, excessively wide prediction intervals can mask emerging outbreaks and delay public health responses. By achieving tighter calibration precisely when uncertainty is highest, the conformal framework provides decision-makers with sharper, actionable thresholds without compromising coverage validity.

Furthermore, the operational viability of a nowcasting system depends critically on its computational scalability. The Bayesian spatial model, while accurate in its point estimates, requires computationally intensive refitting via INLA at each successive epidemiological week, averaging 3 hours per update. In contrast, the proposed pipeline relies on a historically static calibration set. Once the underlying probabilities and non-conformity scores are learned—a training process requiring approximately one minute—generating bounds for new incoming data involves only a fast forward pass and a simple quantile adjustment. This static-calibration property allowed the framework to process the entire 26-week evaluation with merely 6 seconds of inference time, rendering it highly suitable for national-scale automated surveillance dashboards.

Despite these substantial operational and computational advantages, the empirical results also highlight important methodological trade-offs inherent to the framework’s architecture. The point estimate analysis revealed that the Bayesian baseline achieved lower absolute and squared errors at the initial evaluation point ($d = 0$). This performance gap stems from the fundamental difference in the resolution of the estimators. The Bayesian hierarchical model is natively formulated at the aggregate level: its likelihood function is explicitly designed to minimize errors and fit the total regional counts, naturally optimizing the initial point estimate for that specific granularity. In contrast, the proposed Gradient Boosting pipeline employs a bottom-up approach, learning probabilities at the granular patient level before summing fractional expectations. While this individual-level modeling penalizes the macroscopic point estimate at the peak of missing data ($d = 0$), it becomes highly advantageous as partial reports accumulate. From $d = 1$ onward,

the inclusion of continuous spatial coordinates and specific demographic profiles allows the machine learning model to resolve the remaining reporting surface with progressively greater precision. Although the models perform competitively at $d = 1$, the predictive advantage of the bottom-up framework widens as the delay increases, ultimately establishing a clear superiority over the spatially parametric baseline in the subsequent weeks. This sustained accuracy is largely attributable to the absorbing state formulation ($D \geq 6$) employed during training. By aggregating all extreme delays into a single final category, the machine learning estimator explicitly captures the long-tail reporting dynamics, effectively preventing the chronic residual underestimation that typically afflicts models assuming a strict maximum delay window.

This bottom-up architecture also introduces specific theoretical limitations that must be acknowledged. To compute the theoretical regional variance prior to the conformal calibration, the framework assumes that reporting events are conditionally independent given their estimated probabilities. In reality, reporting delays may exhibit localized clustering—such as a specific hospital holding a batch of notifications—meaning the true covariance matrix is not strictly diagonal. Consequently, the baseline theoretical variance ($\hat{\sigma}_r^2$) may systematically underestimate the true aleatoric dispersion. However, the study demonstrates that the application of Split Conformal Prediction effectively neutralizes this limitation. Because the non-conformity scores are calibrated empirically on the macroscopic aggregate, the conformal threshold naturally inflates to absorb any unmodeled covariance, securing the target coverage regardless of the underlying independence violations.

Another critical consideration for practical deployment relates to the definition of the conformal targets. While this study evaluated standard two-sided prediction intervals, public health risk assessment is fundamentally asymmetric. Underestimating an outbreak (upper bound failure) carries severe consequences for capacity planning, whereas overestimating (lower bound failure) generally results only in conservative resource allocation. As demonstrated in the internal evaluation, while absolute symmetric scores are highly efficient, they risk miscalibrating the critical upper limit due to the severe right-skewness of SARI counts. Conversely, independent asymmetric tail calibration secures coverage but generates overly wide, inefficient intervals. Therefore, an operationally superior approach for outbreak detection might involve the use of one-sided conformal predictors. By dedicating the entire allowed error rate (α) to calibrating solely the upper bound, decision-makers could obtain a highly efficient and mathematically valid worst-case threshold tailored specifically for epidemiological alerts.

Looking forward, future research should explore methodologies to reconcile the granular flexibility of machine learning with the macro-level optimization of parametric models. One promising avenue is the development of custom loss functions for highly flexible algorithms—such as tree-based ensembles or neural networks—that can accommodate complex non-linear variable interactions while simultaneously penalizing individual probability errors and aggregate regional deviations during the training phase. Additionally, regarding the Confor-

malized Quantile Regression (CQR) strategies, further refinements could address the inherently homoscedastic nature of the conformal adjustment. By treating the heuristic error level used for the initial moment-matching—denoted as α^* , yielding base quantiles $Q_{\alpha^*/2}$ and $Q_{1-\alpha^*/2}$ —as a tunable hyperparameter, one could optimize the confidence level of the initial quantile estimator to natively approximate the target marginal coverage. By minimizing the required conformal score ($\hat{q}$), this approach ensures that the final predictive bounds rely more heavily on the locally adaptive heteroscedastic shape of the parametric approximation, and less on the uniform additive correction applied by the conformal procedure.

Finally, while the randomized calibration set successfully accommodated the macroscopic drift of the 2S2024 evaluation, deploying the model over multi-year horizons may benefit from adaptive conformal techniques. By assigning temporally decreasing weights to recent epidemiological weeks (BARBER *et al.*, 2023), the non-conformity threshold could react even faster to abrupt structural shocks in the surveillance system. Furthermore, this dynamic uncertainty adjustment should be coupled with active monitoring for conditional distribution shifts, triggering periodic retraining of the underlying probabilistic model to capture lasting changes in administrative reporting protocols or the emergence of novel pathogens.

In conclusion, this chapter demonstrates that synthesizing machine learning with conformal inference provides a highly scalable, robust, and mathematically valid solution to the epidemiological nowcasting problem. By decoupling granular probability estimation from macroscopic uncertainty quantification, the proposed framework equips public health authorities with tighter, real-time predictive bounds, facilitating faster and more reliable resource allocation across complex, continental-scale surveillance networks.

DOMAIN ADAPTATION: CORRECTING CREDIT DEFAULT RISK

Although the initial chapters focused on epidemiological surveillance, the mathematical framework developed for right-censored delays and uncertainty quantification is highly generalizable. This chapter extends the proposed methodology to the financial domain, specifically targeting credit risk assessment ([LANDO, 2004](#)). In this context, verification latency—the time required to confirm a credit default—creates an operational blind spot structurally similar to reporting delays in public health. By adapting our conformal inference and conditional probability techniques, we propose a robust methodology for the early detection of credit portfolio degradation.

4.1 Context

The fundamental premise connecting the previous chapters to the financial domain relies on a mathematical isomorphism: both epidemiological reporting delays and credit verification latency are, at their core, problems of aggregating heterogeneous probabilities under temporal censoring. Just as we must infer the true peak of an outbreak before all eventually reported cases are consolidated ([MCGOUGH *et al.*, 2020](#); [HÖHLE; HEIDEN, 2014](#)), financial institutions must infer the final failure rate of a credit portfolio long before the maturation window closes ([LANDO, 2004](#)).

In credit risk management, origination models (Application Scores) and behavioral models (Behavior Scores) are typically optimized to predict the Probability of Default (PD)—a statistical estimate of the likelihood that a borrower will fail to meet their contractual obligations over a specific timeframe ([LANDO, 2004](#)). Typically, this horizon is established at 12 months (PD_{12m}), where “default” is operationally defined by regulatory frameworks (such as the Basel Accords) or internal institutional policies ([LANDO, 2004](#); [BELLINI, 2019](#)). While a payment

delinquency exceeding a 90-day threshold (90 days past due, or DPD) is the most prevalent industry benchmark, this definition can vary depending on the asset class, local regulations, or specific contractual ‘unlikely to pay’ triggers. Operationally, the target variable is often defined as an indicator of whether a credit operation reaches a delinquency threshold, such as 90 days past due (DPD), observed within a fixed performance horizon of 12 months (PD_{12m}) (BELLINI, 2019). The PD is a foundational metric in modern banking, serving as a primary input to compute the Expected Credit Loss (ECL), alongside Exposure at Default (EAD) and Loss Given Default (LGD), typically expressed by the identity $ECL = PD \times LGD \times EAD$. This calculation is strictly required for establishing accounting provisions under standards such as IFRS 9 and CECL, determining regulatory capital requirements, and calibrating the institution’s overall risk appetite (LANDO, 2004; BELLINI, 2019). Consequently, definitively verifying the accuracy and calibration of a credit vintage requires a full year (12 months) of maturation.

This intrinsic latency creates a critical operational vulnerability (GRZENDA; GOMES; BIFET, 2020). If an abrupt macroeconomic shift—such as a sudden surge in inflation or unemployment—structurally elevates global credit risk, the financial institution will only confirm the model’s miscalibration months after the credit was granted, precluding timely preventive actions. Relying exclusively on the 12-month window to adjust risk appetite often results in delayed ECL provisioning and irrecoverable financial losses.

To address this, we propose an analytical framework for the early monitoring and predictive adaptation (nowcasting) of default rates. The approach is grounded in the temporal recalibration of the long-term model to short observation horizons (e.g., $h = 3$ months). By formalizing the statistical properties of the risk estimator through the lens of the Poisson-Binomial distribution (the convolution of heterogeneous binomials), we construct strict uncertainty bounds via Split Conformal Prediction (VOVK; GAMMERMAN; SHAFER, 2022; ANGELOPOULOS; BARBER; BATES, 2024). This enables not only the formulation of robust hypothesis tests for the early detection of *Prior Shift* (POLO *et al.*, 2023), but also the predictive adaptation of the final vintage projections well before the 12-month closure.

4.2 Related Work

The continuous monitoring of Machine Learning models in production environments is widely discussed in the literature under the paradigms of Concept Drift (LU *et al.*, 2019) and Dataset Shift (QUINONERO-CANDELA *et al.*, 2008). This continuous evaluation becomes critically complex when the operational environment imposes a structural verification latency, a challenge formally recognized in the machine learning literature as learning under delayed labels (GOMES *et al.*, 2022; PLASSE; ADAMS, 2016). In the credit industry, monitoring practices traditionally evaluate shifts in both the target response variable $P(Y)$ and the input covariates $P(X)$, utilizing metrics such as the Population Stability Index (PSI) as the pervasive standard for

both purposes (SIDDIQI, 2016). However, while $P(X)$ is immediately available, assessing the true outcome $P(Y)$ is severely hindered by the structural maturation latency.

While recent advancements in machine learning emphasize the importance of formally diagnosing distinct shift types via rigorous hypothesis testing (POLO *et al.*, 2023), the operational reality of credit risk remains constrained. Detecting alterations in $P(X)$ does not guarantee the detection of changes in the baseline risk rate $P(Y)$ —known as Prior Shift (or Response Shift)—and relying exclusively on the verification of $P(Y)$ strictly traps the institution in the latency blind spot.

Macroeconomic shocks frequently alter the marginal probability of default without altering the demographic distribution of credit applicants. Detecting Prior Shift strictly requires observing the target variable Y , as theoretical studies on delayed labeling have analytically demonstrated that uniquely identifying shifts in prior probabilities is impossible without access to the true outcomes (ZLIOBAITE, 2010). While advanced unsupervised drift detection frameworks attempt to circumvent label latency by monitoring structural changes exclusively within the unlabeled feature space (SETHI; KANTARDZIC, 2017), these techniques are fundamentally based on capturing changes in covariate distribution.

To mitigate the effects of verification latency, traditional methods often rely on monitoring X alone, apply instance reweighting techniques based on the density ratio between the new and historical covariate distributions, or resort to completely retraining the underlying model as new data arrives (QUIÑONERO-CANDELA *et al.*, 2008; ZLIOBAITE, 2010). Moreover, as comprehensively reviewed by Gomes *et al.* (2022), a prominent line of machine learning research explores methodologies for delayed partially labelled streams, leveraging techniques to dynamically update models. While theoretically sound, continuously retraining models or dynamically adjusting sample weights in highly regulated credit domains introduces severe practical frictions, fundamentally disrupting the baseline rank-ordering of clients consolidated by the primary long-term model.

This chapter advances the state-of-the-art by leveraging the partial empirical evidence already available—specifically, the actual default events observed well before the end of the full 12-month temporal horizon. Instead of relying on computationally heavy retraining pipelines, we promote a structural model adjustment that translates the risk surface by shifting exclusively the global intercept via a log-odds transformation (KING; ZENG, 2001). This non-destructive approach structurally adapts the predictions to the new macroeconomic reality without retraining, thereby preserving the complex non-linear covariate interactions and the original decision rule’s rank-ordering. This ensures predictive coherence while utilizing conformal inference (ANGELOPOULOS; BARBER; BATES, 2024) to shield the monitoring system against false positive alerts.

4.3 Methodology

4.3.1 Problem Formulation and Temporal Recalibration

To enable early monitoring without the necessity of training multiple independent models for each time horizon—which would fatally generate inconsistencies in the risk ordering of clients—we propose the marginal recalibration of the original probability.

Let $\mathcal{D}$ be a credit vintage—defined as a cohort of credit operations within a specific temporal window (e.g., a specific calendar month)—composed of n independent clients. We define the true default indicator variable for individual i over the full 12-month horizon as $Y_{i,12} \in \{0, 1\}$. The baseline risk model, already trained and deployed, provides the estimated long-term probability $\hat{p}_{i,12}$, which acts as an estimator for the true conditional probability $\mathbb{P}(Y_{i,12} = 1 \mid \mathbf{x}_i)$.

To conduct early monitoring, we define a partial observation horizon $h < 12$ (e.g., a maturation of $h = 3$ months), with its respective indicator variable $Y_{i,h} \in \{0, 1\}$. We assume that the probability of default within this reduced horizon is a monotonic function of the original long-term probability estimator. Preserving the discrimination capacity (rank-ordering) learned by the main model, we adjust only the probabilistic scale via a marginal logistic transformation:

$$\ln \left(\frac{\hat{p}_{i,h}}{1 - \hat{p}_{i,h}} \right) = \alpha_h + \ln \left(\frac{\hat{p}_{i,12}}{1 - \hat{p}_{i,12}} \right) \quad (4.1)$$

In this formulation, the horizon-specific intercept, α_h , is the sole free parameter. It is obtained empirically by optimizing the logarithmic loss function (*log-loss*) over mature historical vintages. This procedure translates the original probability surface to accurately reflect the expected reference rate at the h horizon, enabling early monitoring and reducing the operational blind spot inherent in the traditional 12-month maturation cycle.

4.3.2 Statistical Formalization: The Poisson-Binomial Estimator

To detect whether the current vintage under anticipated monitoring is experiencing drift, we must compare the empirically observed default rate at horizon h with the theoretical distribution expected by the recalibrated estimator.

Under the premise that there is no severe *Conditional Shift*—meaning the structural risk relationship $P(Y \mid X)$ learned by the model maintains its shape, undergoing only global variations in the baseline rate (*Prior Shift*)—we model the total number of defaults at horizon h as the sum of independent, non-identically distributed random variables. Since each client possesses a distinct probability of failure $\hat{p}_{i,h}$, we have $Y_{i,h} \sim \text{Bernoulli}(\hat{p}_{i,h})$.

The total count of events in the vintage, $N_h = \sum_{i=1}^n Y_{i,h}$, is characterized as the sum of Bernoulli variables with distinct success probabilities. The exact distribution of this convolution

of heterogeneous binomials is the **Poisson-Binomial** distribution (WANG, 1993). The aggregate default rate of the portfolio is defined by the proportion $R_h = N_h/n$.

By the linearity of expectation, the expected value of the default rate does not rely on homogeneity assumptions (ROSS, 1998):

$$\mathbb{E}[R_h] = \mathbb{E}\left[\frac{N_h}{n}\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_{i,h}] = \frac{1}{n} \sum_{i=1}^n \hat{p}_{i,h} = \bar{p}_h \quad (4.2)$$

Thus, the theoretical expected macroscopic rate is simply the arithmetic mean of the individual probabilities recalibrated for horizon h .

Assuming that default events are conditionally independent given the clients' feature vectors, the variance of the sum is the sum of the individual variances (ROSS, 1998):

$$\sigma_{R_h}^2 = \text{Var}\left(\frac{N_h}{n}\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_{i,h}) = \frac{1}{n^2} \sum_{i=1}^n \hat{p}_{i,h}(1 - \hat{p}_{i,h}) \quad (4.3)$$

It is instructive to contrast this formulation with the epidemiological framework developed in Chapter 3. In the context of SARI reporting delays, the true population of total eventual cases is unknown. As established previously, we modeled the total cases for any ensemble-defined subgroup (N_b) as the number of trials required to achieve a known number of successes (the partially observed cases, n_b), given an estimated reporting probability (p_b). This structural mechanism dictates that the regional aggregation naturally yields a sum of heterogeneous *Negative Binomial* distributions.

Here, however, the population size n is fixed, bounded, and known since the vintage's origination. Although the underlying phenomenon in both domains is driven by individual binary outcomes—a SARI case is either reported or not, just as a client either defaults or not—the probabilistic sampling is inverted. In epidemiological nowcasting, the number of successes is fixed while the total number of trials is a random variable. Conversely, in the credit domain, the total number of trials is fixed and the number of successes (defaults) is the random variable. Therefore, the aggregation strictly yields a sum of heterogeneous *Bernoullis* (the Poisson-Binomial distribution), rather than the Negative Binomial.

Explicitly defining this heterogeneous variance is the mathematical core of the methodology. The variance of a Poisson-Binomial distribution is strictly less than or equal to the variance of a simple Binomial distribution parameterized by the same mean $\bar{p}_h$ (WANG, 1993). Modeling the true heterogeneity of the portfolio (the natural dispersion of risk scores) reduces the aggregated theoretical uncertainty, providing tighter confidence bounds and increasing the statistical power to detect early pattern breaks.

4.3.3 Early Drift Detection via Conformal Prediction

At the temporal moment when the current vintage reaches exactly the maturation horizon h , we utilize the theoretical expectation ($\bar{p}_h$) and variance ($\sigma_{R_h}^2$) to construct rigorous conformal

prediction regions, mirroring the methodology developed in [Chapter 3](#).

The hypothesis test defines the null hypothesis (H_0) as the premise that the vintage under evaluation belongs to the same joint distribution as the historical calibration vintages. To evaluate this, we define a locally adaptive non-conformity score, $s(R_h^{obs}, \bar{p}_h, \sigma_{R_h}^2)$, analogous to the standardized residuals used in our epidemiological analysis:

$$s_h = \frac{|R_h^{obs} - \bar{p}_h|}{\sigma_{R_h}} \quad (4.4)$$

Let $\hat{q}$ be the $(1 - \alpha)$ empirical quantile of these non-conformity scores computed over a calibration set. If the empirically observed default rate R_h^{obs} yields a score that exceeds $\hat{q}$, it falls outside the conformal predictive interval, and we reject the null hypothesis.

As extensively discussed in [Chapter 3](#) regarding time-series data, financial observations inherently possess temporal correlation, implying that the strict exchangeability assumption required for standard Conformal Prediction is technically violated. However, our methodology addresses this through two complementary mechanisms. First, by actively standardizing the non-conformity scores using the theoretical Poisson-Binomial variance ($\sigma_{R_h}^2$), the framework exhibits *model-based robustness* ([ANGELOPOULOS; BARBER; BATES, 2024](#)), natively adjusting the conformal intervals to the specific heteroscedasticity of each portfolio. Second, regarding residual temporal dependence, we rely on the robustness of the Split Conformal architecture. As established in [Chapter 3](#), applying this procedure to non-exchangeable sequences still guarantees a valid empirical coverage bounded by $1 - \alpha - \varepsilon$, where ε is a penalty proportional to the temporal mixing dependence ([OLIVEIRA et al., 2024](#)). Crucially, in the context of early drift detection, this bounded coverage serves as a calibrated baseline: if a macroeconomic shock induces a severe *Prior Shift*, the affected vintage legitimately breaks this structural exchangeability bound, causing the empirical score to exceed $\hat{q}$ and triggering the operational alert.

The empirical connection between this short-term deviation and the long-term expectation is anchored in the structure of the cumulative default curve. A pure *Prior Shift* manifests as a systemic shock. Consequently, if the structural marginal probability rises for the 12-month horizon, this shock is proportionally reflected in the initial horizons. Rejecting H_0 at month h acts as a rigorous leading indicator that the accumulated trajectory has already decoupled from the original expectation.

4.3.4 Predictive Adaptation and Vintage Nowcasting

Once a significant *prior shift* is detected at an early horizon h (e.g., $h = 3$ months), relying on the original 12-month predictions becomes financially perilous. We must perform a nowcast of the final default rate (R_{12m}) by combining the partially observed data with a predictively adapted estimation for the remaining maturation period (the subsequent $12 - h$ months) to immediately adjust Expected Credit Loss (ECL) provisioning ([BELLINI, 2019](#)).

For clients who have already defaulted by month h ($Y_{i,h} = 1$), the outcome is deterministic ($Y_{i,12} = 1$). For the surviving population ($Y_{i,h} = 0$), we must estimate the conditional probability of default for the remaining months. To achieve this, we apply a structural log-odds correction factor (Δ_h) derived directly from the short-term anomaly.

The magnitude of the deviation at month h is calculated via the difference in logits:

$$\Delta_h = \ln \left(\frac{R_h^{obs} \cdot (1 - \bar{p}_h)}{(1 - R_h^{obs}) \cdot \bar{p}_h} \right) \quad (4.5)$$

This term represents the difference between the empirical log-odds of the observed defaults and the theoretical log-odds (KING; ZENG, 2001) of the expected defaults at the short-term horizon. Essentially, it quantifies the structural shock as a scalar effect, correcting the baseline risk trajectory based on the anomaly observed so far.

The adapted final default probability for a surviving client i , denoted as $\tilde{p}_{i,12}$, is projected by shifting the original 12-month estimator ($\hat{p}_{i,12}$) by the structural deviation Δ_h and a conditional survival intercept (α_{surv}):

$$\tilde{p}_{i,12} = \sigma \left(\alpha_{surv} + \Delta_h + \text{logit}(\hat{p}_{i,12}) \right) \quad (4.6)$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ denotes the standard logistic function.

Consequently, the adapted estimate for the final vintage default rate at the 12-month temporal horizon ($\hat{R}_{12}^{adapt}$) is formulated as:

$$\hat{R}_{12}^{adapt} = \frac{1}{n} \left(n \cdot R_h^{obs} + \sum_{i \in \text{survivors}} \tilde{p}_{i,12} \right) \quad (4.7)$$

Equipped with the updated projection $\hat{R}_{12}^{adapt}$, the baseline origination model can be immediately adjusted for new incoming vintages. By applying the same prior correction logic (KING; ZENG, 2001), the adapted individual probability offered by the credit engine is approximated as:

$$\hat{p}_{i,12}^{adapt} \approx \sigma \left(\ln \left(\frac{\hat{R}_{12}^{adapt} \cdot (1 - \bar{p}_{12})}{(1 - \hat{R}_{12}^{adapt}) \cdot \bar{p}_{12}} \right) + \text{logit}(\hat{p}_{i,12}) \right) \quad (4.8)$$

where $\bar{p}_{12}$ is the average rate originally expected for the vintage at 12 months.

The use of the approximation symbol ($\approx$) acknowledges that, due to the non-linearity of the logistic function and Jensen's inequality (CASELLA; BERGER, 2002), applying a constant scalar shift to individual logits does not mathematically guarantee that the adapted aggregate rate $\frac{1}{n} \sum_{i=1}^n \hat{p}_{i,12}^{adapt}$ will perfectly equal the updated projection $\hat{R}_{12}^{adapt}$. While exact algebraic equality could theoretically be forced using numerical root-finding algorithms, doing so is conceptually misaligned with the structural problem of verification latency (GOMES *et al.*, 2022). Because the true default target for the remaining months remains unobservable, $\hat{R}_{12}^{adapt}$ acts as a macroscopic stochastic estimator derived from partial evidence, not a deterministic ground truth. Consequently,

the closed-form log-odds shift $\hat{p}_{i,12}^{adapt}$ represents the feasible analytical approximation to project risk at the individual level, structurally adapting the predictions while respecting the inherent uncertainty of the unobserved temporal window.

4.3.5 Parameter Estimation and Data Splitting Strategy

To formalize the estimation of the parameters detailed above (α_h , α_{surv} , and the conformal quantile $\hat{q}$), we employ a robust data-splitting strategy similar to the one utilized in Chapter 3. Rather than utilizing a strict chronological cutoff—which could inadvertently restrict the calibration set to a narrow, unrepresentative phase of the economic cycle (e.g., exclusively the peak of a recession)—we treat the origination dates (vintages) as temporal indices and randomly partition the sequence of historical, fully matured vintages into two disjoint sets: training vintages and calibration vintages.

The Training Set is exclusively used to fit the probabilistic transformations: obtaining the base model estimator $\hat{p}_{i,12}$, calibrating the horizon-specific intercept α_h (via log-loss optimization), and estimating the survival intercept α_{surv} required for the $\tilde{p}_{i,12}$ projection. Once these parameters are fixed, the completely unseen Calibration Set is passed through the adapted models to generate the non-conformity scores s_h . The empirical quantile $\hat{q}$ is then extracted strictly from this calibration distribution, ensuring that the conformal prediction bounds remain valid while mitigating the risk of data leakage.

4.4 Experimental Evaluation

To transition the theoretical Poisson-Binomial conformal framework from its mathematical formulation into a practical risk management tool, this section presents a comprehensive empirical evaluation. The primary objective is to demonstrate the framework’s operational efficacy in anticipating and quantifying structural degradation within heterogeneous credit portfolios. Rather than relying on simulated environments, we assess the methodology using a large-scale, proprietary credit dataset. The experimental design deliberately focuses on an out-of-time testing period characterized by a severe macroeconomic shock, allowing us to stress-test the proposed statistical bounds under pronounced, real-world *Prior Shift* conditions.

The ensuing analysis is structured around two core operational objectives. First, after detailing the dataset and experimental setup, we systematically compare the early drift detection capabilities of our anticipatory $h = 4$ months conformal approach against the traditional 12-month baseline maturation, highlighting the substantial reduction in verification latency. Second, we evaluate the predictive accuracy of the proposed log-odds adaptation (Δ_4) in correcting structural bias. By doing so, we empirically validate the framework’s capacity to nowcast the final default rate months ahead of its consolidated realization, providing statistical evidence of its value for Expected Credit Loss (ECL) provisioning and the adjustment of credit origination policies.

4.4.1 Dataset Description and Experimental Setup

To empirically validate the proposed Poisson-Binomial conformal framework, we conducted an extensive evaluation using a proprietary, anonymized credit dataset provided by a financial institution. The dataset comprises $n = 17,727,080$ credit records, encompassing vintages originated between February 2017 and July 2023, structured as monthly reference snapshots. The target variable ($Y_{i,12}$) was defined under standard risk guidelines (BELLINI, 2019) as a record reaching 90 or more days past due (90+ DPD) within the first 12 months following its specific reference date.

The baseline risk model, responsible for estimating the original long-term probability of default ($\hat{p}_{i,12}$), was trained using a highly dimensional feature space of 214 standardized and anonymized covariates. These features encompass a wide array of risk indicators sourced from diverse pipelines, including demographic profiles, historical payment behavior, income metrics, and credit bureau inquiries. To process this massive volume of data efficiently, the base estimator utilizes a Gradient Boosting Decision Tree (GBDT) classification algorithm (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; FRIEDMAN, 2001). Consistent with the mathematical estimation process detailed in Appendix A, the algorithm’s hyperparameters were specifically adapted to handle the scale and ensure generalization: the maximum tree depth was constrained ($\text{max_depth} = 4$), the minimum number of observations per leaf and per bin was strictly enforced ($\text{min_data_in_leaf} = 5000, \text{min_data_in_bin} = 5000$), and the continuous feature discretization was optimized ($\text{max_bin} = 100$). Furthermore, during the internal training and validation phase of this base model, the longitudinal structure of the dataset necessitated a rigorous grouping strategy. To prevent data leakage—a scenario where multiple monthly snapshots of the same client could be inadvertently distributed across both training and validation folds (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)—the internal data split was strictly partitioned by unique client identifiers. This group-based splitting ensures that all historical records of a specific individual are completely isolated within a single fold, preserving the true out-of-sample integrity of the model’s performance metrics.

Following the parameter estimation strategy defined in Section 4.3, the temporal span was rigorously partitioned to simulate a real-world deployment and to evaluate the framework’s resilience to structural drift. The historical period, spanning from February 2017 to December 2020, was utilized for model training and conformal calibration. To ensure that the calibration bounds encapsulate a representative mixture of economic fluctuations rather than a narrow chronological artifact, these historical reference months were randomly shuffled and split equally: 50% of the months were assigned to the Training Set (to fit the GBDT estimator, the horizon intercept α_h , and the survival intercept α_{surv}), while the remaining 50% constituted the Calibration Set (to strictly extract the empirical conformal quantile $\hat{q}$).

Finally, the out-of-time Test Set spans from January 2021 to July 2023. This specific temporal window is of critical analytical importance, as it encompasses a severe macroeconomic

shock: the abrupt escalation of interest rates in Brazil during 2021, following the first year of the COVID-19 pandemic. This macroeconomic tightening induced a systemic degradation in credit quality across the financial sector, acting as a natural, highly pronounced *Prior Shift*. By evaluating the framework against this unseen data, we directly test its capacity to issue early conformal alerts and dynamically nowcast the deterioration of the 2021 vintages months before their 12-month maturation window concludes.

4.4.2 Comparative Analysis: Baseline Maturation vs. Early Drift Detection

To empirically demonstrate the operational value of the proposed framework, the primary analysis focuses on the out-of-time test set (vintages from January 2021 to July 2023). We systematically compare the drift detection capabilities of two distinct conformal monitoring strategies: the traditional long-term baseline approach (requiring a 12-month maturation window) and the proposed early detection framework (recalibrated for an $h = 4$ months horizon).

The traditional monitoring strategy relies on the original 12-month horizon estimator $\hat{p}_{i,12}$. Under this approach, the theoretical expected default rate is $\bar{p}_{12}$, and the conformal predictive bounds are constructed using the Poisson-Binomial variance $\sigma_{R_{12}}^2$ and its respective conformal quantile $\hat{q}_{12}$. The hypothesis test evaluates whether the true empirical default rate at maturation, R_{12}^{obs} , falls outside the predicted interval:

$$R_{12}^{obs} \notin [\bar{p}_{12} - \hat{q}_{12}\sigma_{R_{12}}, \bar{p}_{12} + \hat{q}_{12}\sigma_{R_{12}}] \quad (4.9)$$

While statistically rigorous, this approach suffers from severe verification latency. As illustrated in the top panel of [Figure 21](#), the macroeconomic shock of 2021 caused the true default rate R_{12}^{obs} to decisively breach the upper conformal bound across the entire year. The observed rate surged and peaked at approximately 15.0% for the late-2021 vintages, severely detaching from the expected baseline (which hovered near 12.0%) and its associated 95% conformal interval. However, because this system requires a full year of maturation, the operational alert for the January 2021 vintage, for example, is only triggered in January 2022 (upon the closure of its 12-month window). This represents a “post-mortem” analysis, confirming the miscalibration long after billions in credit limits have been granted under flawed risk assumptions.

In contrast, our proposed early monitoring framework evaluates the recalibrated default rate precisely at the $h = 4$ months horizon. The hypothesis test for early drift detection evaluates the short-term empirical rate, R_4^{obs} , against the recalibrated expectation $\bar{p}_4$ and the conformal bounds dictated by $\sigma_{R_4}^2$ and $\hat{q}_4$:

$$R_4^{obs} \notin [\bar{p}_4 - \hat{q}_4\sigma_{R_4}, \bar{p}_4 + \hat{q}_4\sigma_{R_4}] \quad (4.10)$$

The empirical superiority of this methodology is highly evident in the bottom panel of [Figure 21](#). For the exact same 2021 cohorts, the early monitoring framework detects a statisti-

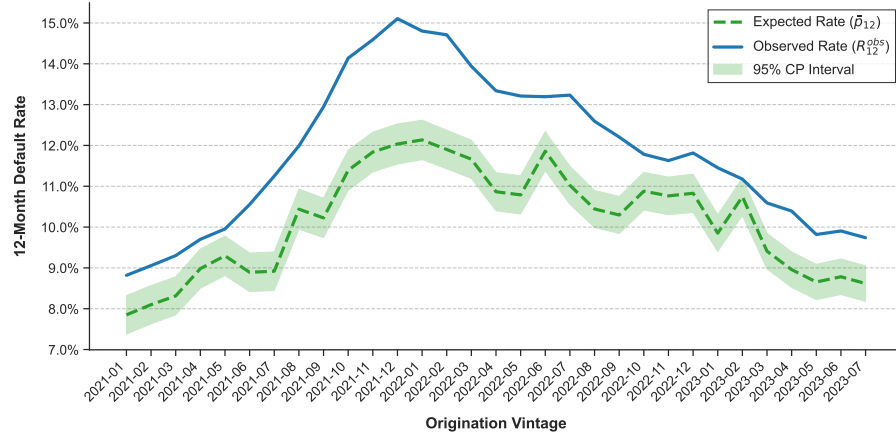
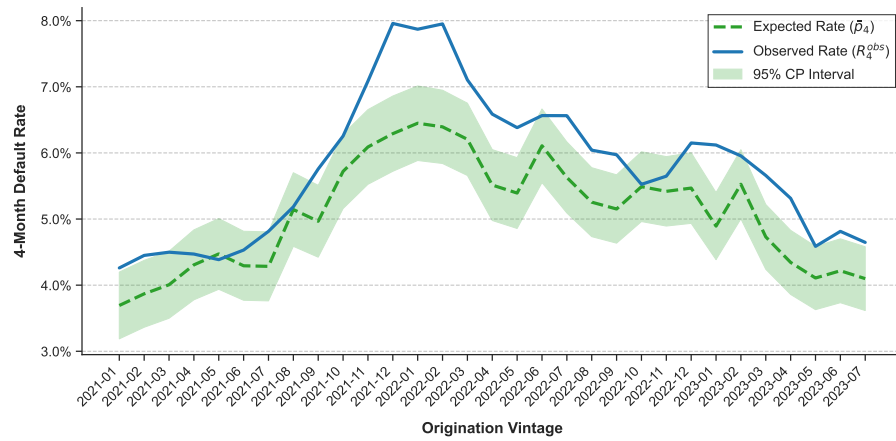
(a) Baseline Maturation Framework ($h = 12$ months)(b) Early Drift Detection Framework ($h = 4$ months)

Figure 21 – Conformal Drift Detection: Baseline Maturation vs. Early Framework. The charts illustrate the visual realization of the hypothesis test across the out-of-time test vintages (2021–2023). The dashed green line represents the expected default rate ($\bar{p}_h$), bounded by the 95% Poisson-Binomial conformal prediction interval (shaded region). The solid blue line represents the true empirical default rate (R_h^{obs}). During the 2021 macroeconomic shock, the empirical rate structurally breaches the upper conformal bound in both scenarios, rejecting the null hypothesis. While the traditional 12-month baseline (a) only flags this *Prior Shift* upon maturation (e.g., January 2022 for the 2021-01 cohort), the recalibrated early framework (b) triggers a statistically rigorous operational alert exactly 8 months earlier (in May 2021), neutralizing the verification latency.

cally significant *Prior Shift* organically and rapidly. The observed short-term rate R_4^{obs} sharply accelerates and crosses the upper conformal threshold, eventually peaking at approximately 8.0% (well above the expected baseline peak of $\approx 6.5\%$), providing strong statistical evidence that the vintage's risk profile has structurally degraded. Notably, for the February 2021 vintage, this statistical breach is clearly signaled as early as May 2021, the exact moment its 4-month maturation window closes.

By mapping the temporal evolution of the empirical rates against the conformal prediction bounds, the comparative analysis reveals that the 4-month conformal test successfully anticipates the long-term baseline failure by exactly 8 months. Crucially, because the interval is protected by the model-robust properties of Conformal Prediction (ANGELOPOULOS; BARBER; BATES, 2024), this early alert is not merely stochastic noise or an artifact of sample variance, but a mathematically rigorous signal of underlying distributional shift. This substantial reduction in verification latency equips risk managers with a critical leading indicator, allowing the institution to restrict credit policies or increase provisioning long before the peak of the macroeconomic crisis fully materializes in the 12-month balance sheets.

4.4.3 Nowcasting Accuracy: Correcting the Structural Bias

Detecting a systemic distributional shift early fulfills only the first objective of an effective monitoring system. Once the conformal hypothesis test rejects the stability of the current vintage at the $h = 4$ horizon, relying on the original 12-month predictions ($\bar{p}_{12}$) for portfolio management becomes statistically unreliable and financially hazardous. To mitigate impending losses, the framework must systematically project the final default rate by applying the structural log-odds correction (Δ_4) to the surviving population, yielding the adapted nowcast estimator $\hat{R}_{12}^{adapt}$.

To evaluate the precision and operational value of this predictive adaptation, Figure 22 plots the temporal evolution of three distinct metrics across the out-of-time test set: the original expected baseline ($\bar{p}_{12}$), the true empirical maturation rate (R_{12}^{obs}), and the proposed adapted nowcast ($\hat{R}_{12}^{adapt}$).

The empirical results clearly demonstrate the magnitude of the 2021 macroeconomic shock and the severe limitations of models relying on static calibration. The original baseline prediction (dashed gray line) fails to capture the sudden risk escalation, consistently underestimating the default rate throughout the crisis. For instance, while the unadapted baseline model projected a peak default rate of approximately 12.0% for the late-2021 vintages, the true empirical rate (solid blue line) soared to over 15.0%. In a large-scale credit portfolio, a systematic underestimation of 300 basis points in the Probability of Default translates into a massive, unaccounted deficit in Expected Credit Loss (ECL) provisioning. Furthermore, in the context of new credit originations, it implies that the institution is inadvertently absorbing higher-risk profiles, leading to incorrect interest rate pricing and adverse selection.

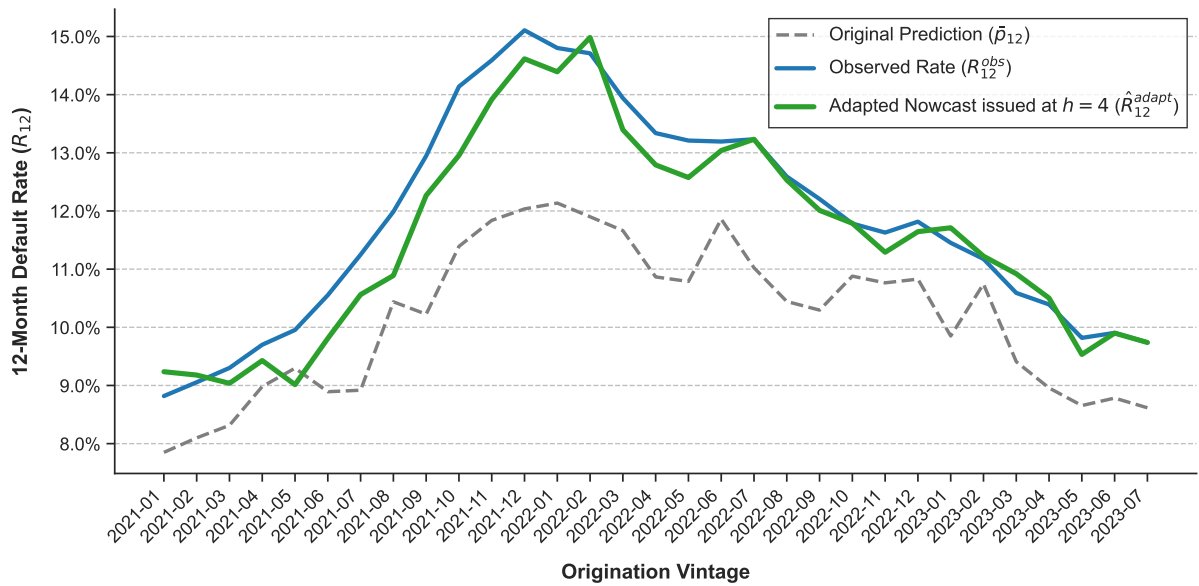


Figure 22 – Nowcasting Accuracy: Comparative Evaluation of Adapted Predictions vs. Baseline. The figure plots the temporal evolution of the final 12-month default rate (R_{12}) across the test vintages (2021–2023). The solid blue line represents the true empirical maturation rate (R_{12}^{obs}), and the dashed gray line depicts the original long-term predicted baseline ($\bar{p}_{12}$). The solid green line represents the predictively adapted nowcast issued early at the $h = 4$ months horizon ($\hat{R}_{12}^{adapt}$). Notably, during the steep macroeconomic deterioration starting in 2021, the original baseline substantially underestimates the risk. Conversely, by applying the proposed Δ_4 log-odds correction, the early adapted projection closely tracks the structural failure rate exactly 8 months ahead of its final realization, successfully reducing the verification latency.

Conversely, the adapted nowcast (solid green line) effectively corrects this structural bias. By capturing the Δ_4 deviation directly from the early 4-month conformal anomaly, the framework dynamically shifts the long-term risk surface. As clearly depicted in Figure 22, the $\hat{R}_{12}^{adapt}$ curve tightly shadows the true empirical rate, successfully anticipating the 15.0% peak behavior and correcting the trajectory exactly 8 months before the final data is consolidated.

This tight alignment confirms a fundamental theoretical premise of this chapter: a pure *Prior Shift* adaptation—adjusting only the global marginal intercept while preserving the complex non-linear covariate interactions (rank-ordering) learned by the original Gradient Boosting model—is highly effective for handling systemic macroeconomic shocks. By dynamically translating the risk surface based solely on short-term empirical evidence, the framework allows financial institutions to seamlessly correct portfolio projections and recalculate credit limits in an anticipatory manner, substantially mitigating the operational blind spot without the need for immediate, computationally expensive model retraining.

4.5 Discussion

The presented methodology successfully resolves the chronic problem of verification latency in credit risk modeling (PLASSE; ADAMS, 2016), allowing financial institutions to

anticipate vintage degradation in a fraction of the traditional time. By recalibrating the baseline risk estimator to an early $h = 4$ months observation horizon and modeling the probabilistic aggregation through the Poisson-Binomial distribution (WANG, 1993), the framework extracts analytical estimators for the expectation and variance of highly heterogeneous credit portfolios.

The empirical evaluation against the 2021 macroeconomic shock decisively demonstrated the operational magnitude of this approach. While the traditional 12-month baseline model remained systematically insensitive to the crisis—consistently underestimating the peak default rate by approximately 300 basis points (projecting 12.0% against a true rate of over 15.0%)—the proposed framework triggered statistically rigorous operational alerts exactly 8 months before the maturation window closed. The integration of the Poisson-Binomial formulation with Split Conformal Prediction established robust operational thresholds, effectively mitigating the emission of false positive alerts caused by random stochastic noise while swiftly detecting true *Prior Shift*.

Furthermore, the log-odds based mathematical architecture enabled a fluid predictive adaptation (nowcasting) of the ongoing vintages. By directly capturing the Δ_4 short-term anomaly and translating the long-term risk surface, the adapted projection ($\hat{R}_{12}^{adapt}$) seamlessly tracked the true 15.0% structural failure rate. This ensures that Expected Credit Loss (ECL) provisioning can be dynamically adjusted in an anticipatory manner. Beyond provisioning, this capability equips risk managers with the necessary lead time to correct interest rate pricing and restrict limits on new originations to prevent adverse selection, effectively addressing the operational blind spot.

For future work, there is a significant opportunity to empirically compare the efficacy of this Prior Shift adaptation (global intercept and log-odds adjustment) with more exhaustive Covariate Shift correction methodologies. Regarding *Conditional Shift*, while addressing it directly is operationally infeasible at early horizons—as it strictly requires the fully matured target variable (Y_{12}) to retrain the relationship $P(Y | X)$ —future research could explore advanced methodologies for learning under delayed labels (GOMES *et al.*, 2022; PLASSE; ADAMS, 2016). Adapting these streaming or semi-supervised architectures to approximate conditional shifts using only partial or delayed feedback could offer a promising, albeit highly complex, alternative to completely retraining the underlying predictive model. Although retraining a Machine Learning algorithm under Covariate or Conditional Shift is theoretically more flexible for capturing new non-linear interactions arising from economic shocks, it introduces severe practical frictions in the highly regulated financial market: instability in client rank-ordering, high computational costs, and prolonged Model Risk Management (MRM) governance validation cycles.

Additionally, future research should explore advanced conformal architectures to further optimize interval efficiency. While the implemented locally adaptive symmetric non-conformity procedure proved highly robust, testing alternative locally adaptive asymmetric quantile calibrations (ANGELOPOULOS; BARBER; BATES, 2024) could yield tighter predictive bounds

without compromising marginal coverage. Furthermore, deploying this framework over extended, multi-year horizons could significantly benefit from non-exchangeable conformal weighting strategies. By assigning temporally decaying weights to the calibration vintages—thereby granting higher statistical importance to the most recent macroeconomic data—the empirical conformal threshold could react even faster to abrupt structural shocks, dynamically adapting the institution’s safety margins to the prevailing economic cycle.

Another promising avenue for future research lies in expanding the temporal flexibility of the framework. While the current methodology demonstrates the efficacy of a fixed $h = 4$ months horizon, future studies could evaluate a multi-horizon approach (e.g., sequentially evaluating $h \in \{3, 4, 5, 6\}$) and formulate a mathematical optimization criterion to select the optimal evaluation point. Such a criterion would systematically balance the operational benefit of early detection (minimizing verification latency) against the statistical robustness of the prediction, given that the variance of the Poisson-Binomial aggregation inherently decreases as the vintage approaches its 12-month maturation. Furthermore, exploring alternative statistical mechanisms for estimating the *Prior Shift* adaptation represents a critical next step. Beyond the empirical log-loss optimization of the marginal intercept employed in this chapter, future work could investigate Bayesian updating schemes to continuously refine the prior probability.

In sharp contrast to computationally heavy retraining pipelines, the risk translation promoted by our methodology—recalibrating solely through marginal transformations—preserves the complex non-linear covariate interactions and the discriminating properties (rank-ordering) consolidated by the original Gradient Boosting estimator. This offers substantial practical advantages in terms of implementation simplicity within decision engines, adherence to strict regulatory compliance, and unmatched agility in adjusting credit origination policies during fast-paced economic crises.

CONCLUSION

5.1 Synthesis of the Central Thesis

This thesis addressed a fundamental vulnerability in the management of complex, high-stakes systems: the inherent latency of information. In domains ranging from public health surveillance to financial risk management, critical operational decisions are frequently compromised by right-censored data, where the true magnitude of an ongoing event is obscured by administrative, biological, or operational delays.

The central achievement of this research was the development of a unified, scalable analytical framework capable of overcoming these truncation issues without relying on rigid parametric assumptions for marginal coverage. By synthesizing the highly granular pattern-recognition capabilities of Machine Learning (Gradient Boosting) with the rigorous, distribution-free uncertainty calibration of Conformal Prediction, this thesis demonstrated that it is possible to free the nowcasting process from the computational bottlenecks and point-prediction limitations of traditional methodologies.

Crucially, this work validated a fundamental **shared probabilistic structure**: the underlying statistical challenge of temporal censoring and distributional shifts is structurally analogous across vastly different domains. We demonstrated that the same logical and mathematical architecture developed to estimate the true number of SARI cases—a metric vital for determining current epidemiological states and guiding ICU bed allocations—can be successfully adapted to the financial sector, effectively anticipating the structural degradation of a credit vintage.

5.2 The Research Journey: Summary of Critical Findings

The objectives of this thesis were achieved through a progressive methodological journey, documented across three core analytical chapters:

- **Chapter 2** tackled the challenge of identifying the etiological landscape of respiratory infections within highly incomplete databases characterized by the systematic absence of negative test records. By modeling viral etiologies as conditional probabilities, we developed the ViDAS framework. Furthermore, we introduced SViDAS, a novel semi-parametric extension that combines a point estimator with an age density estimator. This innovation allowed for personalized, continuous estimation of viral proportions across arbitrary age ranges, providing a level of demographic granularity unmatched by traditional aggregated panels.
- **Chapter 3** addressed the operational necessity of predicting absolute SARI case volumes. We proposed a bottom-up conformal nowcasting framework to correct right-censored epidemiological counts. When benchmarked against the state-of-the-art spatio-temporal Bayesian hierarchical model (INLA), the proposed machine learning architecture proved to be a highly competitive alternative. It offered exceptional computational performance—reducing execution time from hours to seconds—and delivered rigorously calibrated confidence intervals that natively adapted to the heteroscedasticity of the data, without requiring predefined spatial matrices.
- **Chapter 4** established the cross-domain generalizability of the methodology by translating the conformal framework to the credit risk industry. Focusing on early drift detection, the model successfully identified the severe 2021 macroeconomic interest rate shock. By evaluating the portfolio at an early $h = 4$ months horizon, the framework captured the structural *Prior Shift* and correctly projected the severe deterioration of the final default rate (from an expected 12.0% to an observed 15.0%). This provided operational alerts exactly 8 months in advance, substantially mitigating the chronic verification latency of traditional 12-month models.

5.3 Practical and Operational Impact

Beyond its theoretical and statistical contributions, this thesis was specifically designed to deliver tangible, operational tools for real-world decision-makers.

In the realm of public health, the methodologies developed in Chapters 2 and 3 have been operationalized through the creation of the PySRAG Python library and its integration into the interactive Virus Monitor dashboard. These open-source deployments provide health authorities and epidemiologists across Brazil with a ready-to-use, computationally efficient pipeline for the timely tracking of respiratory viruses, directly supporting localized public health interventions.

In the financial industry, the framework addresses a critical pain point in Model Risk Management (MRM). The methodology introduced in Chapter 4—utilizing a structural log-odds adaptation to correct baseline risk projections—allows banks to dynamically recalculate

Expected Credit Loss (ECL) provisions in an anticipatory manner. By applying a marginal translation based solely on early empirical anomalies, financial institutions can swiftly adapt to macroeconomic crises without losing the rank-ordering of their clients and without triggering the prolonged, rigid validation cycles required for full model retraining.

5.4 Limitations and Future Directions

While this research establishes a robust foundation for conformal nowcasting under non-stationary conditions, it also highlights mature frontiers for future investigation.

First, a natural methodological extension lies in the native integration of the absolute incidence counts (Chapter 3) with the conditional viral proportions (Chapter 2). Combining the estimated volume of total SARI cases with the specific probability distribution of circulating pathogens would enable the timely estimation of the absolute, pathogen-specific incidence (e.g., the volume of RSV hospitalizations among infants), providing a highly actionable metric for targeted epidemiological resource planning.

Second, regarding the Conformal Prediction architecture, future research should explore advanced non-exchangeable techniques for extended, multi-year time series. Implementing temporally decaying weights for the calibration data—granting higher statistical importance to the most recent epidemiological weeks or credit vintages—could allow the locally adaptive non-conformity scores to react even faster to abrupt structural shocks, further optimizing the efficiency of the predictive bounds.

Finally, while the global log-odds translation (*Prior Shift*) proved highly effective for correcting credit risk trajectories, a dynamic investigation of *Covariate Shift* and *Conditional Shift* represents a valuable future pursuit. By empirically comparing our marginal adjustment methodology against more exhaustive local correction algorithms, researchers could establish precise operational thresholds to determine when a simple prior adaptation is sufficient, and when the severity of an economic shock necessitates the computational friction of a complete model retraining.

BIBLIOGRAPHY

ALBRECHT, S.; BRODERICK, D.; DOST, K.; CHEUNG, I.; NGHIEM, N.; WU, M.; ZHU, J.; POONAWALA-LOHANI, N.; JAMISON, S.; RASANATHAN, D.; HUANG, S.; TRENHOLME, A.; STANLEY, A.; LAWRENCE, S.; MARSH, S.; CASTELINO, L.; PAYNTER, J.; TURNER, N.; MCINTYRE, P.; RIDDLE, P.; GRANT, C.; DOBBIE, G.; WICKER, J. S. Forecasting severe respiratory disease hospitalizations using machine learning algorithms. **BMC Medical Informatics and Decision Making**, v. 24, n. 1, Oct. 2024. ISSN 1472-6947. Citations on pages [29](#) and [62](#).

ANGELOPOULOS, A. N.; BARBER, R. F.; BATES, S. **Theoretical Foundations of Conformal Prediction**. [S.l.]: arXiv, 2024. Citations on pages [23](#), [62](#), [63](#), [70](#), [72](#), [87](#), [102](#), [103](#), [106](#), [112](#), and [114](#).

ARAÚJO, K. L. R. d.; AQUINO, É. C. d.; SILVA, L. L. S. d.; TERNES, Y. M. F. Fatores associados à síndrome respiratória aguda grave em uma região central do Brasil. **Ciência & Saúde Coletiva**, v. 25, p. 4121–4130, Oct. 2020. ISSN 1413-8123. Citation on page [32](#).

ATKINSON, K.; HAN, W. **Theoretical Numerical Analysis: A functional analysis framework**. 3. ed. New York, NY: Springer New York, 2009. (Texts in Applied Mathematics). ISSN 0939-2475. ISBN 9781441904584. Citation on page [137](#).

BARBER, R. F.; CANDÈS, E. J.; RAMDAS, A.; TIBSHIRANI, R. J. Conformal prediction beyond exchangeability. **Annals of Statistics**, v. 51, n. 2, p. 816–845, Apr. 2023. Citations on pages [63](#) and [99](#).

BASTOS, L. S.; ECONOMOU, T.; GOMES, M. F. C.; VILLELA, D. A. M.; COELHO, F. C.; CRUZ, O. G.; STONER, O.; BAILEY, T.; CODEÇO, C. T. A modelling approach for correcting reporting delays in disease surveillance data. **Statistics in Medicine**, v. 38, n. 22, p. 4363–4377, Jul. 2019. ISSN 1097-0258. Citations on pages [21](#), [22](#), [28](#), [29](#), [30](#), [62](#), [66](#), and [85](#).

BELLINI, T. **IFRS 9 and CECL Credit Risk Modelling and Validation: A Practical Guide with Examples Worked in R and SAS**. [S.l.]: Academic Press, 2019. ISBN 978-0-12-814940-9. Citations on pages [21](#), [101](#), [102](#), [106](#), and [109](#).

BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer, 2006. (Information Science and Statistics). ISBN 978-0387-31073-2. Citation on page [34](#).

BRIER, G. W. Verification of forecasts expressed in terms of probability. **Monthly Weather Review**, v. 78, n. 1, p. 1–3, Jan. 1950. ISSN 1520-0493. Citations on pages [39](#) and [139](#).

CASELLA, G.; BERGER, R. **Statistical Inference**. 2. ed. [S.l.]: Duxbury, 2002. ISBN 0-534-24312-6. Citation on page [107](#).

CHAMMARTIN, F.; HOUNGBEDJI, C. A.; HÜRLIMANN, E.; YAPI, R. B.; SILUÉ, K. D.; SORO, G.; KOUAMÉ, F. N.; N'GORAN, E. K.; UTZINGER, J.; RASO, G.; VOUNATSOU, P. Bayesian risk mapping and model-based estimation of *Schistosoma haematobium*–*Schistosoma*

mansoni co-distribution in Côte d'Ivoire. **PLoS Neglected Tropical Diseases**, v. 8, n. 12, p. e3407, Dec. 2014. ISSN 1935-2735. Citation on page 29.

CLOPPER, C. J.; PEARSON, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. **Biometrika**, v. 26, n. 4, p. 404–413, 1934. ISSN 0006-3444. Citation on page 39.

COCHRAN, W. G. **Sampling Techniques**. 3. ed. New York: John Wiley & Sons, 1977. (A Wiley publication in applied statistics). ISBN 047116240X. Citation on page 69.

COURA-VITAL, W.; CARDOSO, D. T.; KER, F. T. d. O.; MAGALHÃES, F. d. C.; BEZERRA, J. M. T.; VIEGAS, A. M.; MORAIS, M. H. F.; BASTOS, L. S.; REIS, I. A.; CARNEIRO, M.; BARBOSA, D. S. Spatiotemporal dynamics and risk estimates of COVID-19 epidemic in Minas Gerais state: analysis of an expanding process. **Revista do Instituto de Medicina Tropical de São Paulo**, v. 63, 2021. ISSN 1678-9946. Citation on page 29.

DALMASSO, N.; POSPISIL, T.; LEE, A. B.; IZBICKI, R.; FREEMAN, P. E.; MALZ, A. I. Conditional density estimation tools in Python and R with applications to photometric redshifts and likelihood-free cosmological inference. **Astronomy and Computing**, v. 30, p. 100362, Jan. 2020. ISSN 2213-1337. Citation on page 35.

DEGROOT, M. H.; SCHERVISH, M. J. **Probability and statistics**. 4. ed. Boston, Mass. [u.a.]: Addison-Wesley, 2012. ISBN 9780321500465. Citations on pages 35 and 135.

DIGGLE, P. J.; GIORGI, E. **Model-Based Geostatistics for Global Public Health: Methods and applications**. Milton: CRC Press LLC, 2019. (Chapman and Hall/CRC Interdisciplinary Statistics). ISBN 9781351743273. Citations on pages 28 and 29.

ENGLAND, P.; VERRALL, R. Stochastic claims reserving in general insurance. **British Actuarial Journal**, v. 8, n. 3, p. 443–518, Aug. 2002. ISSN 2044-0456. Citations on pages 62 and 66.

Fiocruz, FGV e Ministério da Saúde. **InfoGripe - Fiocruz**. 2023. Available: <<http://info.gripe.fiocruz.br/>>. Accessed: 2023-11-15. Citations on pages 26 and 28.

FORBES, C.; EVANS, M.; HASTINGS, N.; PEACOCK, B. **Statistical Distributions**. 4. ed. Hoboken, New Jersey: John Wiley & Sons, 2010. ISBN 9780470627242. Citation on page 35.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001. ISSN 00905364,21688966. Available: <<http://www.jstor.org/stable/2699986>>. Citations on pages 22, 48, 64, 67, and 109.

Fundação Oswaldo Cruz (Fiocruz). **Boletim InfoGripe: Monitoramento de casos de Síndrome Respiratória Aguda Grave (SRAG)**. 2025. Available: <https://agencia.fiocruz.br/sites/agencia.fiocruz.br/files/Resumo_InfoGripe_2025_46.pdf>. Citation on page 31.

GOLDSTEIN, A.; KAPELNER, A.; BLEICH, J.; PITKIN, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. **Journal of Computational and Graphical Statistics**, v. 24, n. 1, p. 44–65, Jan. 2015. ISSN 1537-2715. Citation on page 48.

GOMES, H. M.; GRZENDA, M.; MELLO, R.; READ, J.; NGUYEN, M. H. L.; BIFET, A. A survey on semi-supervised learning for delayed partially labelled data streams. **ACM Computing Surveys**, v. 55, n. 4, p. 1–42, Nov. 2022. ISSN 1557-7341. Citations on pages 23, 102, 103, 107, and 114.

GRZENDA, M.; GOMES, H. M.; BIFET, A. Delayed labelling evaluation for data streams. **Data Mining and Knowledge Discovery**, v. 34, n. 5, p. 1237–1266, Nov. 2020. ISSN 1573-756X. Citations on pages 23 and 102.

HALLIDAY, A.; STONER, O.; ECONOMOU, T.; BASTOS, L. S. Joint bayesian nowcasting of severe acute respiratory illness and COVID-19 positives in Brazil. **Statistics in Medicine**, v. 45, n. 8-9, p. e70529, 2026. Citation on page 62.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. 2. ed. New York: Springer New York, 2009. (Springer Series in Statistics). ISSN 2197-568X. ISBN 9780387848587. Citations on pages 22, 33, 45, 47, and 109.

HIRVE, S.; CRAWFORD, N.; PALEKAR, R.; ZHANG, W.; the WHO RSV surveillance Group. Clinical characteristics, predictors, and performance of case definition—interim results from the WHO global respiratory syncytial virus surveillance pilot. **Influenza and Other Respiratory Viruses**, v. 14, n. 6, p. 647–657, Oct. 2020. ISSN 1750-2659. Citation on page 31.

HÖHLE, M.; HEIDEN, M. an der. Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. **Biometrics**, v. 70, n. 4, p. 993–1002, Jun. 2014. ISSN 1541-0420. Citations on pages 21, 62, 66, 67, and 101.

HORVITZ, D. G.; THOMPSON, D. J. A generalization of sampling without replacement from a finite universe. **Journal of the American Statistical Association**, v. 47, n. 260, p. 663–685, Dec. 1952. ISSN 1537-274X. Citations on pages 64 and 69.

Instituto Brasileiro de Geografia e Estatística (IBGE). **Cadastro de Localidades Seleccionadas**. 2010. Available: <<https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/27385-localidades.html?edicao=27386>>. Accessed: 2024-11-08. Citation on page 32.

_____. **Censo Demográfico 2022**. 2022. Available: <<https://sidra.ibge.gov.br/tabela/4714>>. Accessed: 2024-11-08. Citation on page 32.

ITpS - Instituto Todos pela Saúde. **Monitoramento de Patógenos Respiratórios**. 2022. Available: <<https://www.itsp.org.br/pesquisa/monitoramento-de-patogenos-respiratorios>>. Accessed: 2023-11-15. Citation on page 28.

IZBICKI, R. **Machine Learning Beyond Point Predictions: Uncertainty Quantification**. 1. ed. [S.l.: s.n.], 2025. 260 p. ISBN 978-65-01-20272-3. Citations on pages 22, 35, 39, 62, 63, 70, 72, 135, and 139.

IZBICKI, R.; LEE, A. B. Converting high-dimensional regression to high-dimensional conditional density estimation. **Electronic Journal of Statistics**, v. 11, n. 2, p. 2800–2831, Jan. 2017. ISSN 1935-7524. Citation on page 35.

IZBICKI, R.; SHIMIZU, G.; STERN, R. Flexible distribution-free conditional predictive bands using density estimators. In: CHIAPPA, S.; CALANDRA, R. (Ed.). **Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics**.

PMLR, 2020. (Proceedings of Machine Learning Research, v. 108), p. 3068–3077. Available: <https://proceedings.mlr.press/v108/izbicki20a.html>. Citations on pages 23 and 63.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. LightGBM: a highly efficient gradient boosting decision tree: Annual conference on neural information processing systems. In: LUXBURG, U.; GUYON, I.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S. V. N.; GARNETT, R. (Ed.). **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Red Hook, NY: Curran Associates, Inc., 2017. (NIPS'17), p. 3149–3157. ISBN 9781510860964. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf. Citation on page 129.

KING, G.; ZENG, L. Logistic regression in rare events data. **Political Analysis**, v. 9, n. 2, p. 137–163, 2001. ISSN 1476-4989. Citations on pages 103 and 107.

KLINK, T.; RANKIN, D. A.; PIYA, B.; SPIEKER, A. J.; FAOURI, S.; SHEHABI, A.; WILLIAMS, J. V.; KHURI-BULOS, N.; HALASA, N. B. Evaluating the diagnostic accuracy of the WHO severe acute respiratory infection (SARI) criteria in Middle Eastern children under two years over three respiratory seasons. **PLoS ONE**, v. 15, n. 4, p. 1–15, Apr. 2020. ISSN 1932-6203. Citation on page 31.

KOEMEN, S.; FARIA, N. R.; BASTOS, L. S.; RATMANN, O.; AMARAL, A. V. R. Fast and trustworthy nowcasting of dengue fever: A case study using attention-based probabilistic neural networks in São Paulo, Brazil. **Epidemics**, v. 54, p. 100880, Dec. 2025. ISSN 1755-4365. Citation on page 62.

LANDO, D. **Credit Risk Modeling: Theory and Applications**. [S.l.]: Princeton University Press, 2004. ISBN 9780691089294. Citations on pages 101 and 102.

LAWLESS, J. Adjustments for reporting delays and the prediction of occurred but not reported events. **The Canadian Journal of Statistics**, v. 22, n. 1, p. 15–31, Mar. 1994. ISSN 03195724. Citations on pages 21 and 69.

LIMA, E. E. C. d.; GAYAWAN, E.; BAPTISTA, E. A.; QUEIROZ, B. L. Spatial pattern of COVID-19 deaths and infections in small areas of Brazil. **PLoS ONE**, v. 16, n. 2, p. e0246808, Feb. 2021. ISSN 1932-6203. Citations on pages 28 and 29.

LU, J.; LIU, A.; DONG, F.; GU, F.; GAMA, J.; ZHANG, G. Learning under concept drift: A review. **IEEE Transactions on Knowledge and Data Engineering**, Los Alamitos, CA, USA, v. 31, n. 12, p. 2346–2363, Dec. 2019. ISSN 1558-2191. Citation on page 102.

MACK, T. Distribution-free calculation of the standard error of chain ladder reserve estimates. **ASTIN Bulletin**, v. 23, p. 213–225, Nov. 1993. ISSN 0515-0361. Citations on pages 62 and 66.

MATAZI, A. K.; GOGNET, E. E.; KAKAI, R. G. Digital soil mapping: a predictive performance assessment of spatial linear regression, bayesian and ML-based models. **Modeling Earth Systems and Environment**, v. 10, n. 1, p. 595–618, Jun. 2023. ISSN 2363-6211. Citation on page 62.

MCGOUGH, S. F.; JOHANSSON, M. A.; LIPSITCH, M.; MENZIES, N. A. Nowcasting by bayesian smoothing: A flexible, generalizable model for real-time epidemic tracking. **PLoS Computational Biology**, v. 16, n. 4, p. e1007735, Apr. 2020. ISSN 1553-7358. Citations on pages 29, 62, and 101.

Microsoft Corporation. **LightGBM Documentation**. 2023. Available: <<https://lightgbm.readthedocs.io>>. Accessed: 2023-09-25. Citation on page 129.

Ministério da Saúde do Brasil. **Dados Abertos da Saúde**. 2024. Available: <<https://opendatasus.saude.gov.br/dataset/srag-2021-a-2024>>. Accessed: 2024-10-04. Citation on page 32.

_____. **Dados Abertos da Saúde**. 2025. Available: <<https://opendatasus.saude.gov.br/dataset/srag-2021-a-2025>>. Accessed: 2025-07-12. Citation on page 26.

_____. **Notificação Compulsória – Secretaria de Vigilância em Saúde e Ambiente**. 2026. Available: <<https://www.gov.br/saude/pt-br/composicao/svsa/notificacao-compulsoria>>. Accessed: 2026-03-18. Citation on page 58.

_____. **Portaria GM/MS Nº 10.175, de 23 de janeiro de 2026**. 2026. Diário Oficial da União. Available: <<https://www.in.gov.br/en/web/dou/-/portaria-gm/ms-n-10.175-de-23-de-janeiro-de-2026-683062951>>. Accessed: 2026-03-18. Citation on page 58.

Ministério da Saúde do Brasil, Secretaria de Vigilância em Saúde. **Guia de vigilância epidemiológica: emergência de saúde pública de importância nacional pela doença pelo coronavírus 2019 – COVID-19**. Brasília, Brasil: Ministério da Saúde, 2022. ISBN 978-65-5993-025-8. Available: <<https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/covid-19/publicacoes-tecnicas/guias-e-planos/guia-de-vigilancia-epidemiologica-covid-19>>. Accessed: 2024-09-12. Citations on pages 26 and 31.

NORMAN, D. C. Fever in the elderly. **Clinical Infectious Diseases**, v. 31, n. 1, p. 148–151, Jul. 2000. ISSN 1537-6591. Citation on page 31.

OLIVEIRA, R. I.; ORENSTEIN, P.; RAMOS, T.; ROMANO, J. V. Split conformal prediction and non-exchangeable data. **Journal of Machine Learning Research**, v. 25, n. 225, p. 1–38, 2024. Available: <<http://jmlr.org/papers/v25/23-1553.html>>. Citations on pages 63, 72, 97, and 106.

PLASSE, J.; ADAMS, N. **Handling Delayed Labels in Temporally Evolving Data Streams**: Ieee international conference on big data. [Piscataway, NJ]: IEEE, 2016. ISBN 9781467390057. Citations on pages 102, 113, and 114.

POLO, F. M.; IZBICKI, R.; LACERDA, E. G.; IBIETA-JIMENEZ, J. P.; VICENTE, R. A unified framework for dataset shift diagnostics. **Information Sciences**, v. 649, p. 119612, Nov. 2023. ISSN 0020-0255. Citations on pages 22, 102, and 103.

QUIÑONERO-CANDELA, J.; SUGIYAMA, M.; SCHWAIGHOFER, A.; LAWRENCE, N. D. (Ed.). **Dataset shift in machine learning**. Cambridge, Mass: The MIT Press, 2008. 229 p. (Neural information processing series). ISBN 978-0-262-17005-5. Citations on pages 22, 102, and 103.

RENSHAW, A.; VERRALL, R. A stochastic model underlying the chain-ladder technique. **British Actuarial Journal**, v. 4, n. 4, p. 903–923, Oct. 1998. ISSN 2044-0456. Citation on page 62.

ROCHA, R.; ATUN, R.; MASSUDA, A.; RACHE, B.; SPINOLA, P.; NUNES, L.; LAGO, M.; CASTRO, M. C. Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: a comprehensive analysis. **The Lancet Global Health**, v. 9, n. 6, p. e782–e792, Jun. 2021. ISSN 2214-109X. Citations on pages 28 and 32.

ROMANO, Y.; PATTERSON, E.; CANDÈS, E. J. Conformalized quantile regression. In: **Proceedings of the 33rd International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: [s.n.], 2019. Citations on pages 23, 63, and 73.

ROSS, S. M. **A first course in probability**. 5. ed. Upper Saddle River, NJ: Prentice Hall, 1998. ISBN 0137463146. Citation on page 105.

SALMON, M.; SCHUMACHER, D.; STARK, K.; HÖHLE, M. Bayesian outbreak detection in the presence of reporting delays. **Biometrical Journal**, v. 57, n. 6, p. 1051–1067, Aug. 2015. ISSN 1521-4036. Citations on pages 22 and 62.

SCHNEBLE, M.; NICOLA, G. D.; KAUERMANN, G.; BERGER, U. Nowcasting fatal COVID-19 infections on a regional level in Germany. **Biometrical Journal**, v. 63, n. 3, p. 471–489, Nov. 2020. ISSN 1521-4036. Citation on page 62.

SETHI, T. S.; KANTARDZIC, M. On the reliable detection of concept drift from streaming unlabeled data. **Expert Systems with Applications**, v. 82, p. 77–99, Oct. 2017. ISSN 0957-4174. Citation on page 103.

SIDDIQI, N. **Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards**. 2. ed. [S.l.]: Wiley, 2016. ISBN 9781119282396. Citation on page 103.

SILVA, A. D. d.; GOMES, M. F. d. C.; GREGIANINI, T. S.; MARTINS, L. G.; VEIGA, A. B. G. d. Machine learning in predicting severe acute respiratory infection outbreaks. **Cadernos de Saúde Pública**, v. 40, n. 1, 2024. ISSN 0102-311X. Citations on pages 29 and 62.

SILVA, A. D. d.; VEIGA, A. B. G. d.; CRUZ, O. G.; BASTOS, L. S.; GOMES, M. F. d. C. Severe acute respiratory infection surveillance in Brazil: the role of public, private and philanthropic healthcare units. **Health Policy and Planning**, v. 37, n. 9, p. 1075–1085, Jun. 2022. ISSN 1460-2237. Citations on pages 21, 26, 28, and 31.

SILVA, J. F. A.; AL. et. **Python package PySRAG**. 2024. Available: <<https://github.com/joao-1988/PySRAG>>. Accessed: 2026-02-12. Citations on pages 26, 61, and 130.

_____. **Virus Monitor: interactive monitoring of infectious diseases**. 2024. Available: <https://joaoflavio.shinyapps.io/Virus_Monitor>. Accessed: 2026-02-12. Citations on pages 26 and 61.

TIBSHIRANI, R. J.; BARBER, R. F.; CANDÈS, E. J.; RAMDAS, A. **Conformal Prediction Under Covariate Shift**. 2020. ArXiv preprint arXiv:1904.06019. Citation on page 63.

URCHUEGUÍA-FORNES, A.; MUÑOZ-QUILES, C.; MIRA-IGLESIAS, A.; LÓPEZ-LACORT, M.; MENGUAL-CHULIÁ, B.; LÓPEZ-LABRADOR, F. X.; DíEZ-DOMINGO, J.; ORRICO-SÁNCHEZ, A. Ten-year surveillance of respiratory syncytial virus hospitalizations in adults: Incidence rates and case definition implications. **The Journal of Infectious Diseases**, v. 231, n. 5, p. e830–e839, Feb. 2025. ISSN 1537-6613. Citation on page 31.

UTAZI, C. E.; YANKEY, O.; CHAUDHURI, S.; OLOWE, I. D.; DANOVARO-HOLLIDAY, M. C.; LAZAR, A. N.; TATEM, A. J. Geostatistical and machine learning approaches for high-resolution mapping of vaccination coverage. **Spatial and Spatio-temporal Epidemiology**, v. 54, p. 100744, Aug. 2025. ISSN 1877-5845. Citation on page 62.

VILLANUEVA-MIRANDA, I.; XIAO, G.; XIE, Y. Artificial intelligence in early warning systems for infectious disease surveillance: a systematic review. **Frontiers in Public Health**, v. 13, Jun. 2025. ISSN 2296-2565. Citation on page 29.

VOVK, V.; GAMMERMAN, A.; SHAFER, G. **Algorithmic Learning in a Random World**. 2. ed. Cham, Switzerland: Springer, 2022. 1–476 p. (Algorithmic Learning in a Random World, Second Edition). ISBN 9783031066498. Citations on pages 23, 63, 72, and 102.

WANG, Y. H. On the number of successes in independent trials. **Statistica Sinica**, v. 3, p. 295–312, 1993. Available: <<https://www.jstor.org/stable/24304959>>. Citations on pages 23, 105, and 114.

WASSERMAN, L. **All of Nonparametric Statistics**. 1. ed. New York, NY: Springer New York, 2006. (Springer Texts in Statistics Ser.). ISBN 9780387306230. Citations on pages 35 and 135.

WONG, S.; FLEGG, J. A.; GOLDING, N.; KANDANAARACHCHI, S. Comparison of new computational methods for spatial modelling of malaria. **Malaria Journal**, v. 22, n. 1, Nov. 2023. ISSN 1475-2875. Citations on pages 29 and 62.

World Health Organization. **WHO strategy for global respiratory syncytial virus surveillance project based on the influenza platform**. 2019. Available: <<https://cdn.who.int/media/docs/default-source/influenza/who-rsv-surveillance-strategy-phase-26mar2021.pdf>>. Accessed: 2025-10-14. Citation on page 31.

_____. **Clinical care for severe acute respiratory infection - Toolkit**. 2022. Available: <<https://www.who.int/publications/i/item/clinical-care-of-severe-acute-respiratory-infections-tool-kit>>. Accessed: 2024-09-28. Citation on page 25.

ZLIOBAITE, I. Change with delayed labeling: When is it detectable? In: FAN, W. (Ed.). **2010 IEEE International Conference on Data Mining Workshops**. Piscataway, NJ: IEEE Computer Society, 2010. p. 843–850. ISBN 9780769542577. ISSN 15504786. Citation on page 103.

MODEL CONFIGURATION AND TRAINING STRATEGY

To construct the estimator, we utilized the Gradient Boosting Decision Tree (GBDT) classification algorithm via the LightGBM API ([Microsoft Corporation, 2023](#)). This choice is based on the algorithm's ability to handle multidimensional data, model non-linearities, capture complex interactions among covariates, provide probabilistic classification, and deliver excellent computational performance ([KE *et al.*, 2017](#)).

The configuration of the LightGBM model involves defining several hyperparameters that are crucial for its performance. These settings are detailed below:

- `boosting_type="gbdt"`: Defines the type of boosting LightGBM uses. "gbdt" stands for "Gradient Boosted Decision Tree". This implies the algorithm is based on constructing decision trees to improve model performance. Trees are built in iterative stages, where each new tree is trained to correct errors made by previous ones.
- `objective="multiclass"`: Defines the model's objective function to be optimized during training. Here, "multiclass" indicates that LightGBM is configured to solve a multiclass classification problem, where data examples belong to more than two classes.
- `learning_rate=0.1`: Defines the algorithm's learning rate, i.e., how much the model learns from each added tree. In this case, the learning rate is set to 0.1.
- `n_estimators=5000`: Specifies the number of estimators (decision trees) to be created during training. The model is set to create up to 5000 trees.
- `max_depth=3`: Limits the maximum depth of decision trees. Setting "max_depth" to 3 prevents trees from growing beyond a depth of 3 levels, meaning each tree will have a maximum of 2^3 leaves.

- `min_data_in_leaf=30`: Defines the minimum number of data points (samples) required in a leaf (terminal node). A value of 30 means a leaf will only be created if it contains at least 30 samples.
- `feature_fraction=0.8`: Controls the fraction of features (covariates) selected randomly in each iteration (tree construction). With a value of 0.8, LightGBM will randomly select 80% of the features before training each tree.
- `min_data_in_bin=10`: Specifies the minimum number of data points in each bin used to construct the histogram during training. Histograms are used in gradient boosting histogram algorithms like LightGBM to accelerate training. A value of 10 means a bin will only be created if it contains at least 10 data points.

The choice of `boosting_type="gbdt"` allows the model to effectively partition the variable space, particularly regarding the segmentation of epidemiological weeks, geographic coordinates, municipal population size, and patient age. This results in the creation of multiple local models or groups, each focusing on specific areas of the variable space, enabling the algorithm to handle different behaviors and patterns in the data.

The selection of `learning_rate=0.1` was strategic to ensure the efficiency of our solution within the monitoring dashboard, as reducing it affects both training and inference performance. During training, a lower `learning_rate` means each additional tree contributes less to correcting errors from previous trees, resulting in slower training as more iterations are needed to reach the optimal solution. Conversely, during the application of the trained model, a lower `learning_rate` may require a larger number of trees to obtain precise predictions, potentially increasing prediction time.

To evaluate model performance and determine the optimal number of boosting iterations, we adopted a holdout validation strategy to ensure the computational efficiency required for weekly updates, initially splitting the data into training (80%) and validation (20%) sets. To prevent overfitting, we employed early stopping as a criterion to halt the boosting iterations. Additionally, we incorporated the `multi_logloss` (multi-class logarithmic loss) evaluation metric as part of the stopping criterion; training is interrupted when the metric stops improving and begins to increase on the validation set. Once the optimal number of iterations is identified through this validation process, the final model is retrained using this number of iterations and the full dataset (100% of the data) to maximize the information available for estimation.

It is worth noting that this classifier algorithm with these configurations is available in the Python package PySRAG ([SILVA; AL., 2024a](#)).

GEOGRAPHIC COORDINATES AND TREE-BASED MODELS

The decision to employ geographic coordinates, such as latitude and longitude, instead of indicator variables (one-hot encoding) to represent location, is grounded in considerations related to the application of tree-based estimators. This section outlines the justification underlying this approach and highlights the associated advantages.

The primary reason for utilizing geographic coordinates is the preservation of distance relationships. Employing coordinates is fundamental to maintaining the spatial relationship between points. The use of indicator variables would result in the loss of information regarding the spatial proximity between observations. Consequently, there would be no differentiation between nearby or distant municipalities, potentially leading to the inadequate grouping of municipalities without considering their geographic proximity.

It is important to emphasize that this section focuses exclusively on raw latitude and longitude variables to demonstrate that, even without complex transformations, tree-based models can autonomously generate geographically coherent partitions. For instance, consider c_{lat} as the number of partitions (splits) in the latitude coordinate and c_{lon} as the number of partitions in the longitude coordinate. The combination of these partitions results in a total of $c_{lat} + c_{lon} + 1$ distinct regions (groups). This approach creates coherent groups that respect the distance between municipalities in the geographic space, as demonstrated in Figure 23.

While Figure 23 illustrates the rectangular partitions characteristic of individual decision trees, the application of Gradient Boosting Decision Trees (GBDT) fundamentally refines this spatial representation. Although a single tree produces discrete “jumps” between regions, the boosting process—by iteratively adding trees that correct previous errors—smooths these boundaries. As the number of boosting iterations increases and the learning rate decreases, the initially jagged curves separating regions become increasingly fine-grained, allowing for a smoother and nuanced modeling of spatial risk that approximates a continuous surface, rather

(a) 2 groups, $c_{lat} = 1$ and $c_{lon} = 0$.(b) 4 groups, $c_{lat} = 2$ and $c_{lon} = 1$.(c) 6 groups, $c_{lat} = 3$ and $c_{lon} = 2$ (d) 8 groups, $c_{lat} = 4$ and $c_{lon} = 3$

Figure 23 – Recursive Partitioning of the Brazilian Geographic Space. Example of the partitioning of the Brazilian geographic space based on municipality latitude and longitude coordinates. We observe that this partitioning generates groups consistent with the distances between municipalities.

than relying on the abrupt changes typical of single trees.

Another reason for this choice is computational efficiency. Using indicator variables to represent every municipality (in the case of Brazil, 5,570 municipalities) would lead to the creation of a massive number of features, many of which would be sparse. This would significantly slow down the estimation process, depending on the algorithm used. Therefore, the use of geographic coordinates reduces the dimensionality of the data, making the process computationally more efficient.

In summary, the use of geographic coordinates and spatial partitioning offers an effective method for handling data scarcity in certain municipalities. Even when information is limited or absent for a specific municipality, it is possible to assign estimates based on information from neighboring municipalities within the same spatial partition.

DERIVATION OF THE VIRUS DISTRIBUTION CONDITIONED ON AGE INTERVALS

Our objective is to derive the expression for the conditional probability $\mathbb{P}_Y(y|X_j \in [a, b], X_{-j} = x_{-j})$, which represents the probability of observing virus y given that the predictor variable X_j (patient's age) falls within the interval $[a, b]$, holding constant the other covariates.

We define the notation where:

- $X = (X_1, X_2, \dots, X_p)$ is the vector of continuous predictor variables;
- Y is the categorical random variable representing the virus type;
- X_j is the specific variable of interest within the predictor set (patient's age);
- $X_{-j} = (X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p)$ represents the vector of all predictor variables excluding X_j ;
- $f_{X_j}(\cdot|x_{-j})$ denotes the conditional probability density function of X_j given the remaining covariates x_{-j} ;
- $F_{X_j}(\cdot|x_{-j})$ denotes the corresponding conditional cumulative distribution function.

The derivation proceeds as follows:

$$\begin{aligned}
 \mathbb{P}_Y(y|X_j \in [a, b], x_{-j}) &= \int_{-\infty}^{\infty} \mathbb{P}_Y(y|X_j = x, x_{-j}) f_{X_j}(x|X_j \in [a, b], x_{-j}) dx \\
 &= \int_{-\infty}^{\infty} \mathbb{P}_Y(y|X_j = x, x_{-j}) \frac{\mathbb{I}(x \in [a, b]) f_{X_j}(x|x_{-j})}{\mathbb{P}(X_j \in [a, b]|x_{-j})} dx \\
 &= \frac{\int_{-\infty}^{\infty} \mathbb{P}_Y(y|X_j = x, x_{-j}) \mathbb{I}(x \in [a, b]) f_{X_j}(x|x_{-j}) dx}{\mathbb{P}(X_j \in [a, b]|x_{-j})} \\
 &= \frac{\int_a^b \mathbb{P}_Y(y|X_j = x, x_{-j}) f_{X_j}(x|x_{-j}) dx}{F_{X_j}(b|x_{-j}) - F_{X_j}(a|x_{-j})}.
 \end{aligned}$$

ESTIMATION OF CONDITIONAL DENSITY PARAMETERS VIA GBDT

This appendix provides a detailed account of the statistical methodology employed for estimating the conditional density of patient age. As established in the main text, our approach is predicated on the assumption that patient age follows a lognormal distribution, a distribution fully characterized by its conditional mean and variance. This section elaborates on the two-stage regression procedure, utilizing Gradient Boosting Decision Trees (GBDT), to estimate these parameters. The process involves an initial estimator to determine the conditional mean, followed by a second estimator that regresses the logarithm of the squared residuals to determine the conditional variance.

Thus, if

$$X_j|x_{-j} \sim LN(\mu(x_{-j}), \sigma^2(x_{-j})),$$

then its logarithm, $W = \log(X_j)$, is normally distributed ([DEGROOT; SCHERVISH, 2012](#)):

$$W|x_{-j} \sim N(\mu(x_{-j}), \sigma^2(x_{-j})).$$

Under this hypothesis, we can estimate the conditional density of patient age by determining only the parameters of the conditional mean and variance ([IZBICKI, 2025](#)).

We will use an approach based on ([WASSERMAN, 2006](#)) to estimate the conditional mean and variance. Consider the model:

$$W_i = \mu(x_{-j,i}) + \sigma(x_{-j,i})\varepsilon_i,$$

where we define auxiliary variables:

$$Z_i = \log [(W_i - \mu(x_{-j,i}))^2], \quad \delta_i = \log(\varepsilon_i^2).$$

From these definitions, we have the relationship:

$$Z_i = \log(\sigma^2(x_{-j,i})) + \delta_i,$$

indicating that the estimate of $\log \sigma^2(x_{-j})$ can be obtained by regressing the squared residuals (on a logarithmic scale) on x_{-j} . Specifically, we proceed as follows:

1. Initially estimate the function $\mu(x_{-j})$, representing the conditional mean estimate, using a GBDT estimator with L_2 loss for the objective function, obtaining $\hat{\mu}(x_{-j})$.
2. Using this estimate, calculate the residuals and define:

$$Z_i = \log \left[(W_i - \hat{\mu}(x_{-j,i}))^2 \right].$$

3. Then regress the values Z_i on the explanatory variables $x_{-j,i}$, again using a GBDT estimator with L_2 loss for the objective function. This step will yield an estimated function $\hat{q}(x_{-j})$, corresponding to $\log \sigma^2(x_{-j})$. Finally, the estimate of the conditional variance is given by:

$$\hat{\sigma}^2(x_{-j}) = e^{\hat{q}(x_{-j})}.$$

Therefore, the estimated conditional density (normal) for the logarithm of the patient's age is given by

$$\hat{f}_{W_j}(w_j|x_{-j}) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2(x_{-j})}} \exp \left(-\frac{(w_j - \hat{\mu}(x_{-j}))^2}{2\hat{\sigma}^2(x_{-j})} \right).$$

Similarly, the estimated conditional density (lognormal) for the patient's age is given by

$$\hat{f}_{X_j}(x_j|x_{-j}) = \frac{1}{x_j \sqrt{2\pi\hat{\sigma}^2(x_{-j})}} \exp \left(-\frac{(\log(x_j) - \hat{\mu}(x_{-j}))^2}{2\hat{\sigma}^2(x_{-j})} \right),$$

and the estimated cumulative function distribution is given by

$$\hat{F}_{X_j}(x_j|x_{-j}) = \Phi \left(\frac{\log(x_j) - \hat{\mu}(x_{-j})}{\hat{\sigma}(x_{-j})} \right)$$

where $\Phi(\cdot)$ is the cumulative function distribution of the standard normal distribution, i.e. mean 0 and variance 1, $N(0, 1)$.

NUMERICAL INTEGRATION USING THE PROBABILITY-WEIGHTED MIDPOINT RULE

To estimate the numerator in (2.4), we must rely on a numerical approximation to evaluate the integral. One such approach consists of subdividing the interval $[a, b]$ into $s - 1$ subintervals using a sequence of ordered points $a = r_1 < r_2 < \dots < r_s = b$. Then, we write:

$$\int_a^b \hat{g}(x, x_{-j}) \hat{f}_{X_j}(x|x_{-j}) dx = \sum_{i=1}^{s-1} \int_{r_i}^{r_{i+1}} \hat{g}(x, x_{-j}) \hat{f}_{X_j}(x|x_{-j}) dx. \quad (\text{E.1})$$

Assuming that each subinterval has a sufficiently small length $h_i = r_{i+1} - r_i$, it is reasonable to approximate the function $\hat{g}(x, x_{-j})$ within the subinterval (r_i, r_{i+1}) by evaluating it at a point $x^* \in [r_i, r_{i+1}]$, such as the midpoint $x^* = (r_i + r_{i+1})/2$. Thus, we have:

$$\int_{r_i}^{r_{i+1}} \hat{g}(x, x_{-j}) \hat{f}_{X_j}(x|x_{-j}) dx \approx \hat{g}\left(\frac{r_i + r_{i+1}}{2}, x_{-j}\right) [\hat{F}_{X_j}(r_{i+1}|x_{-j}) - \hat{F}_{X_j}(r_i|x_{-j})].$$

Therefore, the integral in (E.1) can be approximated by:

$$\int_a^b \hat{g}(x, x_{-j}) \hat{f}_{X_j}(x|x_{-j}) dx \approx \sum_{i=1}^{s-1} \hat{g}\left(\frac{r_i + r_{i+1}}{2}, x_{-j}\right) [\hat{F}_{X_j}(r_{i+1}|x_{-j}) - \hat{F}_{X_j}(r_i|x_{-j})]. \quad (\text{E.2})$$

Consequently, the probability in (2.4) can be approximated as:

$$\begin{aligned} \frac{\int_a^b \hat{g}(x, x_{-j}) \hat{f}_{X_j}(x|x_{-j}) dx}{\hat{F}_{X_j}(b|x_{-j}) - \hat{F}_{X_j}(a|x_{-j})} &\approx \frac{\sum_{i=1}^{s-1} \hat{g}\left(\frac{r_i + r_{i+1}}{2}, x_{-j}\right) [\hat{F}_{X_j}(r_{i+1}|x_{-j}) - \hat{F}_{X_j}(r_i|x_{-j})]}{\hat{F}_{X_j}(b|x_{-j}) - \hat{F}_{X_j}(a|x_{-j})} \\ \hat{\mathbb{P}}_Y(y|X_j \in [a, b], x_{-j}) &\approx \frac{\sum_{i=1}^{s-1} \hat{g}\left(\frac{r_i + r_{i+1}}{2}, x_{-j}\right) [\hat{F}_{X_j}(r_{i+1}|x_{-j}) - \hat{F}_{X_j}(r_i|x_{-j})]}{\hat{F}_{X_j}(b|x_{-j}) - \hat{F}_{X_j}(a|x_{-j})}. \end{aligned}$$

This approximation is analogous to applying the *midpoint rule* (see (ATKINSON; HAN, 2009)) to a Riemann-Stieltjes integral. The key difference from the classical formulation arises here: rather than using subinterval widths $r_{i+1} - r_i$, we use the probability mass of the region $[r_i, r_{i+1}]$ under the estimated cumulative distribution function $\hat{F}_{X_j}(\cdot|x_{-j})$ as the integrator, and we evaluate $\hat{g}(\cdot, x_{-j})$ at the midpoint. In contrast, the classical midpoint rule over a Riemann integral would evaluate the product $\hat{g}(\cdot, x_{-j}) \hat{f}_{X_j}(\cdot|x_{-j})$ and multiply it by the subinterval width.

DEFINITION OF THE BRIER SCORE

The Brier score is a proper scoring rule used to assess the quality and calibration of probabilistic predictions (BRIER, 1950). For a multi-class classification problem, it is defined as the mean squared difference between the predicted probabilities and the actual outcomes. Given a sample of size n , the Brier score is calculated as follows (IZBICKI, 2025):

$$\text{Brier Score} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K (\hat{g}_k(x_i) - \mathbb{I}(y_i = k))^2$$

where

- n is the total number of observations.
- K is the total number of classes (viruses).
- x_i is the feature vector for observation i .
- y_i is the true class for observation i .
- $\hat{g}_k(x_i)$ is the predicted probability that observation i belongs to class k .
- $\mathbb{I}(y_i = k)$ is the indicator function, which is 1 if the true class of observation i is k , and 0 otherwise.

A lower Brier score indicates better estimator performance, with a score of 0 representing a perfect fit in terms of both prediction and calibration.

MOMENT-MATCHING EQUATIONS FOR HEURISTIC QUANTILES

In [subsection 3.3.4](#), the Conformalized Quantile Regression (CQR) strategy requires initial, heuristically proposed bounds ($Q_{\alpha^*/2}$ and $Q_{1-\alpha^*/2}$) before applying the empirical conformal correction. These base quantiles are derived by mapping the algorithmically expected moments to known parametric probability distributions via the method of moments.

To rigorously respect the boundary condition that the final estimated total cannot be smaller than the already observed partial count (n_k), we strategically apply the moment-matching procedure strictly to the *unobserved* cases. Let $X_k = N_k - n_k$ represent the theoretical unobserved count. The expectation of the unobserved count is defined as $\mu = \hat{N}_k - n_k$. Because the observed count n_k is a deterministic constant at the time of prediction, the variance of the unobserved cases is identical to the variance of the total aggregated cases, thus $\sigma^2 = \hat{\sigma}_k^2$.

G.1 Negative Binomial Moment Matching

Standard statistical software packages (such as `scipy.stats.nbinom` in Python or `stats::qnbinom` in R) parameterize the Negative Binomial distribution as the number of *failures* before achieving r successes with probability p . By modeling the unobserved cases directly under this formulation, $X_k \sim \text{NegBin}(r, p)$, the expectation and variance are mathematically defined as:

$$\mathbb{E}[X_k] = \frac{r(1-p)}{p} = \mu$$

$$\text{Var}[X_k] = \frac{r(1-p)}{p^2} = \sigma^2$$

To extract the specific parameters $\hat{r}$ and $\hat{p}$ implemented in our algorithm, we isolate the

components. By dividing the expectation by the variance, we obtain the success probability:

$$\frac{\mu}{\sigma^2} = \frac{\frac{r(1-p)}{p}}{\frac{r(1-p)}{p^2}} = p \implies \hat{p} = \frac{\mu}{\sigma^2}$$

To find the size parameter $\hat{r}$, we can substitute $\hat{p}$ back into the expectation equation:

$$\mu = \frac{r \left(1 - \frac{\mu}{\sigma^2}\right)}{\frac{\mu}{\sigma^2}} = \frac{r \left(\frac{\sigma^2 - \mu}{\sigma^2}\right)}{\frac{\mu}{\sigma^2}} = \frac{r(\sigma^2 - \mu)}{\mu}$$

Solving for r yields the precise formula utilized in the computational framework:

$$\hat{r} = \frac{\mu^2}{\sigma^2 - \mu}$$

Once $\hat{r}$ and $\hat{p}$ are parameterized, the algorithm extracts the heuristic base quantiles for the unobserved cases, denoted as $Q_{\alpha^*}^X$. Finally, to obtain the quantiles for the total eventual cases, the algorithm systematically shifts the distribution by adding the known deterministic constant of observed cases:

$$Q_{\alpha^*}^N = Q_{\alpha^*}^X(\text{size} = \hat{r}, \text{prob} = \hat{p}) + n_k$$

G.2 Gamma Moment Matching

As a continuous alternative that robustly handles heavy right-skewness and overdispersion, we evaluate mapping the macroscopic moments of the unobserved cases to a Gamma distribution. Let $X_k \sim \text{Gamma}(k, \theta)$, parameterized by a shape parameter $k > 0$ and a scale parameter $\theta > 0$. The expectation and variance are:

$$\mathbb{E}[X_k] = k\theta = \mu$$

$$\text{Var}[X_k] = k\theta^2 = \sigma^2$$

By solving this system for k and θ , we isolate the scale parameter by taking the ratio of the variance to the expectation:

$$\frac{\sigma^2}{\mu} = \frac{k\theta^2}{k\theta} = \theta \implies \hat{\theta} = \frac{\sigma^2}{\mu}$$

Substituting $\hat{\theta}$ back into the expectation equation yields the shape parameter (often denoted as a in computational libraries):

$$\hat{k} = \frac{\mu}{\hat{\theta}} = \frac{\mu}{\sigma^2/\mu} = \frac{\mu^2}{\sigma^2}$$

Following the same structural logic, the heuristic base quantiles $Q_{\alpha^*}^X$ are extracted using the continuous Gamma percent point function based on $\hat{k}$ and $\hat{\theta}$. The quantiles for the total cases are then computed by adding the observed partial count:

$$Q_{\alpha^*}^N = Q_{\alpha^*}^X(\text{shape} = \hat{k}, \text{scale} = \hat{\theta}) + n_k$$

Any fractional violations generated by the continuous nature of the Gamma distribution are gracefully discretized and overridden during the *Domain Enforcement* step of the overarching conformal procedure.

