Faculty of
Mathematics
and Informatics

VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

DATA SCIENCE STUDY PROGRAMME

Master's Thesis

Functional Data Analysis of COVID-19 Dynamics in Lithuania

Funkcinių duomenų analizė tiriant COVID-19 dinamiką Lietuvoje

Miglė Capė

Supervisor : Partn. Prof. Dr. Vaidotas Zemlys-Balevičius

Scientific advisor : Prof. Dr. Jurgita Markevičiūtė

Vilnius
2026

Acknowledgments

I would like to thank my thesis supervisor, Partn. Prof. Dr. Vaidotas Zemlys-Balevičius, for agreeing to supervise this work. I am especially grateful to Prof. Dr. Jurgita Markevičiūtė for her guidance, thorough review of the work, and valuable advice throughout the thesis preparation process. I also thank Vilnius University, in particular the Faculty of Mathematics and Informatics, for offering the Data Science study programme, which has provided me with a structured academic foundation and a broader conceptual understanding of the field in which I have been working for the past 10 years.

Abstract

This thesis applies Functional Data Analysis (FDA) to model and interpret the weekly spread of COVID-19 in Lithuanian municipalities from 2020 to 2022, treating all relevant processes as smooth functional trajectories. To better understand the influences behind COVID-19 dynamics over time, the analysis examines the roles of weather conditions (temperature and absolute humidity), vaccination rollout, and Lithuanian media sentiment. The methodological framework integrates Functional Principal Component Analysis (FPCA), Functional Canonical Correlation Analysis (FCCA), lagged FCCA, and function-on-function regression into a single workflow, representing a relatively rare application of a unified FDA approach to epidemic processes in Lithuania.

The results show that seasonal meteorological patterns were the dominant structural drivers of COVID-19 cases, that vaccination effects primarily reflected national synchronisation rather than municipality-level variation, and that media sentiment offered a modest but anticipatory signal of rising case numbers. These findings highlight how FDA can reveal coherent temporal relationships across epidemiological, environmental, and informational processes, and demonstrate its potential for improving epidemic monitoring in settings with rich time-series data.

Keywords: COVID-19, Media Sentiment, Functional Data Analysis, Functional Canonical Correlation Analysis, Function-on-Function Regression

Santrauka

Šiame darbe taikomi funkcinės duomenų analizės (FDA) metodai, siekiant modeliuoti ir interpretuoti savaitinį COVID-19 plitimą Lietuvos savivaldybėse 2020–2022 m., visus nagrinėjamus procesus traktuojant kaip glotnias funkcines trajektorijas. Norint geriau suprasti COVID-19 dinamiką laikui bėgant, tyrime analizuojamas oro sąlygų poveikis (temperatūros ir absoliučios drėgmės), vakcinacijos diegimo eiga bei Lietuvos žiniasklaidos nuotaikos. Metodologinis pagrindas apima funkcinę pagrindinių komponentų analizę (FPCA), funkcinę kanoninę koreliacijos analizę (FCCA), vėluojančią funkcinę kanoninę koreliacijos analizę (lagged FCCA) bei funkcinę regresiją funkcijai pagal funkciją (function-on-function regression), kurios kartu sudaro vientisą analizės sistemą. Toks integruotas FDA metodų taikymas epideminiams procesams Lietuvoje yra palyginti retas.

Gauti rezultatai rodo, kad sezoniniai meteorologiniai dėsningumai buvo pagrindiniai struktūriniai COVID-19 atvejų dinamikos veiksniai, vakcinacijos poveikis daugiausiai atspindėjo nacionalinio lygmens sinchronizaciją nei savivaldybių skirtumus, o žiniasklaidos nuotaikos suteikė silpną, tačiau ankstyvą signalą apie galimą atvejų augimą. Šie rezultatai parodo, kad FDA leidžia atskleisti nuoseklius laiko ryšius tarp epidemiologinių, aplinkos ir informacinių procesų bei pabrėžia šių metodų potencialą gerinant epidemijų stebėseną aplinkose, kuriose prieinami išsamūs laiko eilučių duomenys.

Reikšminiai žodžiai: COVID-19, žiniasklaidos nuotaikos, funkcinė duomenų analizė, funkcinė kanoninė koreliacijos analizė, funkcinė regresija funkcijai pagal funkciją

List of Figures

List of Tables

Contents

Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, was officially declared on 11 March 2020 [17]. Since then, more than 780 million cases have been reported globally [56], with Lithuania approaching 1.5 million confirmed infections. Understanding the drivers of COVID-19 transmission is critical for designing effective public health responses. Prior research has identified a wide range of environmental, demographic, and socio-behavioural factors that shape epidemic dynamics. Seasonal climate conditions, population structure, behavioural restrictions, and changes in mobility have all been linked to transmission patterns [18]. Government policies—including quarantine measures, mobility limitations, and vaccination campaigns—have substantially influenced the course of the pandemic [47]. However, these factors do not operate independently and rather interact through complex temporal pathways.

This thesis focuses on three different variables on COVID-19 dynamics in Lithuania: weather, vaccination rollout, and media sentiment. Like many countries, Lithuania experienced multiple epidemic waves and periods of strict societal restrictions through quarantine. Analysing the combined influence of meteorological conditions, vaccination trends, and public sentiment provides a more complete understanding of how these elements shaped the trajectory of the pandemic. The motivation arises from the need to explain why infection rates fluctuated over time and to what extent external environmental conditions (such as temperature and absolute humidity), biomedical interventions (vaccination), and societal factors (media coverage and sentiment) contributed to these fluctuations.

While many international studies have examined these drivers separately, relatively little attention has been given to an integrated, time-series analysis specifically for Lithuania—a country characterised by pronounced seasonal weather patterns, a highly coordinated national vaccination campaign, and an active media environment. Addressing this gap helps clarify Lithuania's pandemic dynamics and provides insights relevant to monitoring future outbreaks or emerging infectious diseases in comparable settings.

The aim of this thesis is to apply Functional Data Analysis (FDA) to model COVID-19 epidemiological trends in Lithuania and to investigate how weather variables, vaccination trajectories, and national media sentiment relate to the evolution of infection rates. More specifically, the thesis objectives are to:

1. model the COVID-19 infection rate as a smooth function of time using FDA;
2. examine how weekly temperature and absolute humidity, along with vaccination patterns, relate to infection dynamics using Functional Canonical Correlation Analysis (FCCA), including lagged correlations;
3. incorporate weekly national-level media sentiment as an additional functional predictor to explore its temporal relationship with epidemic behaviour.

FDA treats time-varying processes as continuous curves rather than discrete observations. This enables the analysis of how temporal fluctuations in one curve (for example, vaccination uptake) correspond to fluctuations in another (infection rate) over the same period. Representing all variables as functional objects provides a framework for studying shared temporal structure, alignment, and lead–lag relationships that are not visible in pointwise or discrete-time models.

The empirical analysis draws on a comprehensive dataset covering the 2020–2022 period. The dataset includes weekly COVID-19 case counts, daily meteorological observations (temperature and absolute humidity), vaccination data (first, second, booster, and total doses), and a national-level media sentiment index derived from major Lithuanian news portals. Using FDA techniques, smooth functional curves are constructed to represent the progression of each variable across time. The results of the study reflect how these processes jointly contributed to the large-scale temporal structure of the epidemic in Lithuania.

This thesis is structured as follows: the Literature Review summarises existing research on the roles of weather, vaccination, and media sentiment in shaping COVID-19 transmission, and introduces the theoretical foundations of FDA and its applications in epidemiology. The Data Preparation chapter describes data sources, preprocessing procedures, and construction of functional representations. The Functional Modelling chapter outlines the analytical framework, including FPCA, FCCA, lagged FCCA, and function-on-function regression. The Results chapter presents the empirical findings, and the Conclusions and Recommendations section summarises key insights and discusses their implications for epidemic monitoring and future research.

1 Literature Review

1.1 Meteorological Drivers of COVID-19 Transmission

A substantial amount of researchers have examined whether COVID-19 follows seasonal patterns similar to other respiratory viruses. Evidence shows that pathogens such as influenza spread more efficiently under cold, dry conditions due to biological mechanisms—greater viral stability, rapid droplet evaporation, together with behavioural changes such as increased indoor gatherings during winter [37]. Early pandemic research, therefore, investigated whether similar mechanisms affected SARS-CoV-2.

Empirical studies from Europe and North America consistently report an inverse association between temperature and COVID-19 cases or mortality [14, 18]. Although Lithuania's climate is classified as continental, its winter–summer contrasts resemble the temperate regions examined in these studies. Global analyses also reinforce this relationship: Chen et al.[12] showed that countries located further from the equator—used as a proxy for lower long-term average temperature—experienced substantially higher case rates. Absolute humidity plays a parallel role - the literature distinguishes between absolute humidity (AH), which measures actual water vapour content, and relative humidity (RH), which varies strongly with temperature [19]. Foundational research in influenza demonstrates that low AH increases viral stability and transmission potential [46], and recent multicity analyses of COVID-19 replicate this pattern [38]. RH, by contrast, fluctuates too quickly with temperature to consistently represent moisture levels, particularly in cold climates where indoor AH drops during heating seasons. Long-term influenza studies similarly find AH to be the more robust predictor of viral activity [41]. Taken together, these results show temperature and absolute humidity as the most biologically and statistically informative meteorological variables. Their clear seasonal pattern and well-understood biological links to respiratory virus transmission justify including them as core predictors in the functional modelling framework of this thesis.

1.2 Vaccination and Epidemic Dynamics

COVID-19 vaccination campaigns substantially altered the trajectory of the pandemic. Clinical and observational evidence shows that vaccines reduce infection, transmission, and severe outcomes [20], and that regions with higher vaccination coverage experienced lower cases and mortality [48]. At the global level, modelling studies estimate that vaccination prevented between 14 and 20 million deaths in its first year [53]. However, the relationship between vaccination and cases is difficult to interpret because vaccination rates usually increase during or immediately after surges in cases, making it hard to separate the effects of one from the other. Vaccination rates frequently rose in response to worsening epidemiological conditions, and national rollouts were highly synchronised across municipalities. Consequently, simple correlations between vaccination and cases must be interpreted cautiously: shared national timing can produce strong associations even when municipality-level variation is minimal. The timing and speed of rollout are also important. Studies show that earlier campaigns provide greater epidemic control benefits [33], and modelling work has demonstrated nuanced interactions between vaccination schedules and natural epidemic cycles [10]. These findings underscore the temporal nature of vaccination effects—an aspect well suited to functional analysis, which emphasises trajectories rather than isolated time points.

Lithuania's 2021 vaccination campaign followed these general patterns. The introduction of the "Opportunity Passport" accelerated uptake across age groups and produced higher coverage than in neighbouring countries [51]. By early 2022, more than 72% of the population had received at least one dose [35]. Despite this high national uptake, municipal-level vaccination curves remained strongly synchronised, limiting their ability to explain spatial differences in cases—an issue that emerges clearly in the empirical results of this thesis.

1.3 Media Sentiment as Behavioural and Informational Signal

The COVID-19 pandemic unfolded alongside an unprecedented surge in news coverage. Public sentiment—reflecting collective fear, uncertainty, or reassurance, was closely linked to epidemic dynamics. Sharp declines in global sentiment followed major outbreak announcements [52], and news tone has been shown to influence behaviour: in the United Kingdom, evidence-based reporting decreased mobility in retail and recreational settings [11]. Sentiment also affects vaccination willingness, with positive vaccine-related sentiment preceding increased uptake [6]. Because sentiment is shaped by media attention, it often anticipates epidemiological changes. Coverage typically intensifies as early signs of deterioration emerge, producing negative sentiment weeks before case numbers peak. For this reason, sentiment functions more as an informational early-warning indicator than as a direct behavioural driver—particularly when derived from news headlines rather than social media posts.

From a methodological perspective, sentiment extraction can rely on either lexicon-based tools or supervised machine-learning models. Lexicon-based methods such as VADER are interpretable and computationally efficient, but they often misclassify pandemic-specific terminology and struggle with contextual meanings. Machine-learning approaches—including TF–IDF with linear classifiers and transformer-based models—provide higher accuracy and adapt better to domain-specific vocabulary [3, 15]. Additionally, because many advanced sentiment models are English-based, researchers commonly translate text prior to analysis. Machine translation has been shown to preserve sentiment polarity sufficiently for downstream classification tasks across multiple language pairs [2, 4, 5]. These findings support the translation-based pipeline employed in this thesis for Lithuanian news headlines.

1.4 Functional Data Approaches in Epidemic Research

Many pandemic-related processes evolve continuously in time: cases rise and fall in waves, vaccination uptake progresses through distinct phases, meteorological variables follow seasonal cycles, and sentiment fluctuates with changes in the information environment. Functional Data Analysis (FDA) treats such processes as smooth trajectories rather than as discrete observations, offering tools for identifying dominant temporal structures and modelling complex time-varying relationships.

Previous research demonstrates the value of FDA in epidemiological settings. Tang et al. [49] used functional time-series models to analyse COVID-19 case curves in the United States, finding improved predictive performance relative to traditional approaches. Functional clustering of SARS-CoV-2 case curves has revealed spatiotemporal patterns in European regions [44], and functional regression has been used to evaluate how time-varying covariates relate to mortality and transmission dynamics [9]. These studies support the modelling strategy used in this thesis: isolating dominant temporal modes using FPCA, analysing shared timing through FCCA and lagged correlations, and estimating joint influences using function-on-function regression.

2 Data Preparation

This chapter describes the data sources, pre-processing procedures, and initial exploratory analyzes undertaken prior to functional modeling.

2.1 Data Description

The empirical analysis integrates four independent datasets: epidemiological data (containing COVID-19 cases), vaccination data, meteorological measurements (temperature and absolute humidity), and media sentiment, derived from headlines from three major Lithuanian newspapers. All datasets are obtained daily for the period from January 2020 to the end of March 2022 and represent the period with the highest peak of coronavirus disease spread. As shown in Figure 1, data are later aggregated to a consistent weekly frequency as it reduces stochastic noise in cases and vaccination records, aligns with epidemiological reporting cycles, and ensures consistent temporal resolution across datasets.

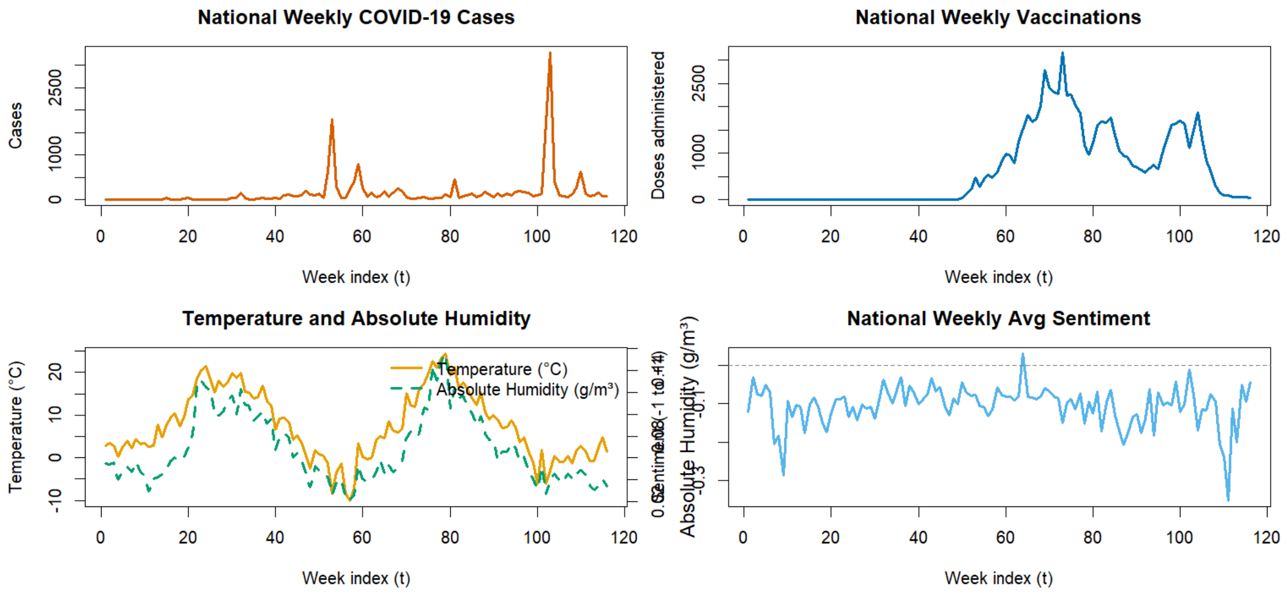

Figure 1: *National weekly aggregates for key variables used in the analysis: COVID-19 cases, vaccinations, meteorological conditions (temperature and absolute humidity), and average media sentiment.*

2.1.1 Epidemiological Data

Daily COVID-19 case counts were retrieved from the Lithuanian Ministry of Health Open Data Portal [30]. After removing two ambiguous municipality categories ("Nežinoma", "Nenustatyta"), the data set comprises 60 Lithuanian municipalities with complete daily data. Daily COVID-19 cases were aggregated into weekly totals, generating a weekly municipality level dataset. The national weekly case curve was calculated as the average in all municipalities, as shown in Figure 1. The raw and smoothed weekly case curves for all municipalities are shown in Figure 2. The data in the initial and smoothed plots emphasize two periods of sharper peaks during the major quarantine periods in Lithuania.

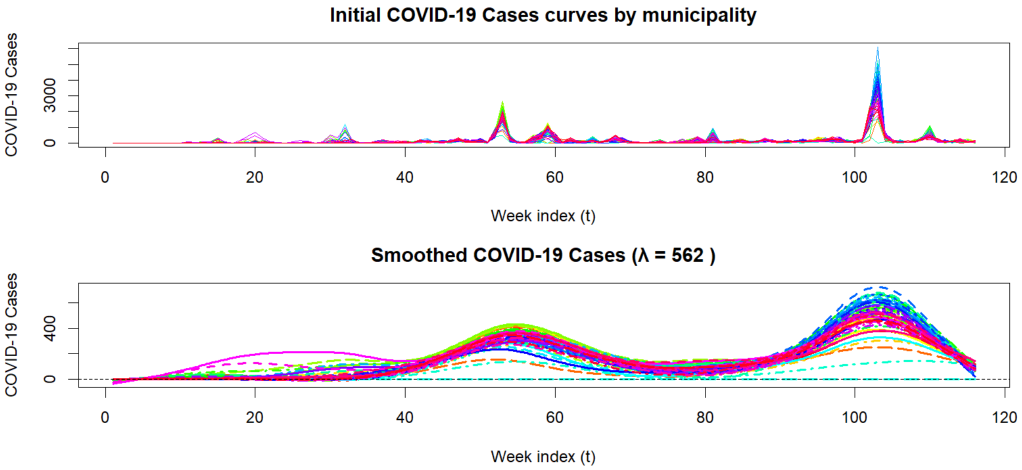

Figure 2: *Initial and smoothed weekly COVID-19 case curves for all 60 municipalities.*

2.1.2 Vaccination Data

Vaccination data were obtained from the Lithuanian national open data platform [34]. Individual vaccination records included dose type (first, second, or booster) and date of administration. Weekly municipality-level counts were constructed for variables *First dose*, *Second dose*, *Booster doses* and *Total vaccinations*. These variables reflect the dynamics of vaccination rollout in Lithuania. Although four variables were derived, first, second, and booster doses follow nearly identical temporal patterns nationally, justifying the use of the aggregated *total vaccinations* as the main variable in this thesis. Figure 3 presents the raw and smoothed total vaccination curves across municipalities, highlighting the heterogeneity in the timing and intensity, with more total vaccines administered in cities with higher population counts.

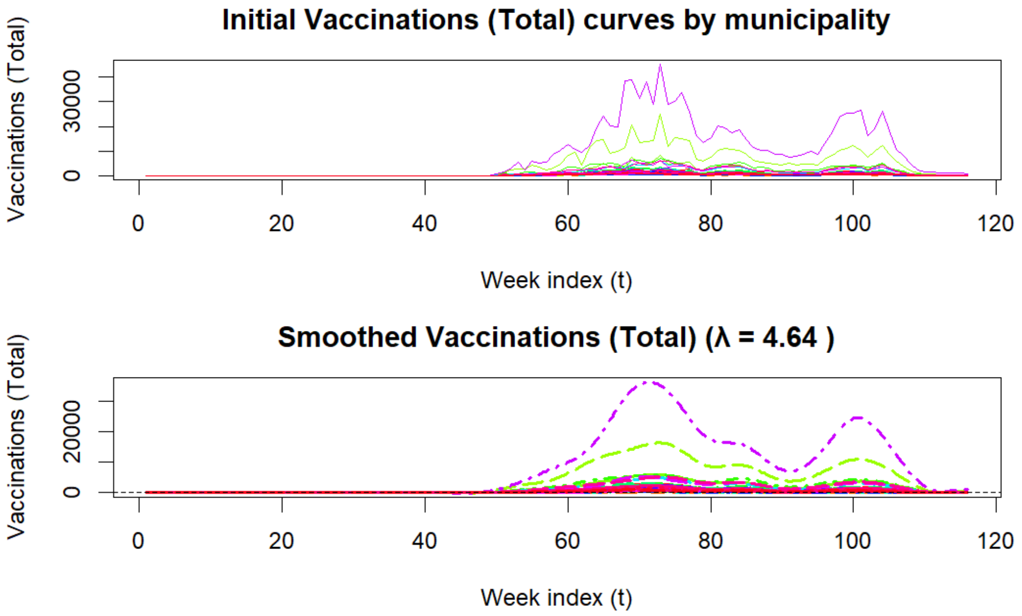

Figure 3: *Initial and smoothed weekly COVID-19 vaccination curves for all 60 municipalities.*

2.1.3 Weather Data

Daily air temperature °C and relative humidity (%) were retrieved from the Lithuanian Hydrometeorological Service API [31]. Absolute humidity (AH) was manually computed using the Magnus–Tetens approximation [1]. In all formulas below, T denotes air temperature in degrees Celsius, and the denominator $T + 273.15$ converts temperature to Kelvin for the calculation of absolute humidity. The daily meteorological values were aggregated to weekly means, and two variables - *Temperature* and *Absolute Humidity* were constructed.

Let T denote air temperature in °C and RH the relative humidity (in %). The saturation vapour pressure is approximated by

$$e_s(T) = 6.112 \,\text{exp}\left(\frac{17.67\, T}{T + 243.5} \right),$$

the actual vapour pressure by

$$e = \frac{\text{RH}}{100}\, e_s(T),$$

and absolute humidity by

$$\text{AH} = \frac{2.1674\, e}{T + 273.15}.$$

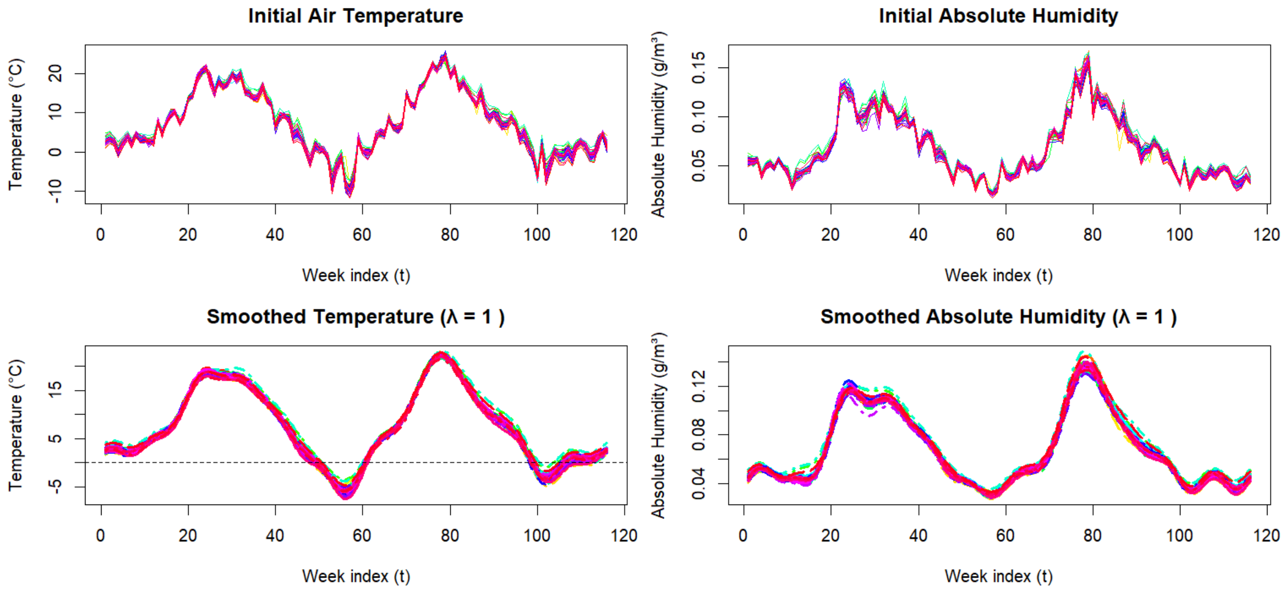

Figure 4: *Initial and smoothed weekly air temperature and absolute humidity (AH) curves for all 60 municipalities.*

Figure 4 shows the two weather variables before and after smoothing. The raw trajectories display substantial short-term variability, particularly during winter months, but a clear annual pattern is evident in both variables. It can be seen that temperature increases from early spring to midsummer and declines again toward winter, and absolute humidity showcases a closely aligned pattern. Although temperature and absolute humidity vary only modestly across municipalities, using municipality-specific curves preserves appropriate alignment with the spatial units used in the cases and vaccination datasets.

2.1.4 Media Sentiment Data

Approximately 33,000 COVID-related news headlines were scraped from Lithuania's three major news portals. Using Python, specific outlet queries searched *Delfi* (www.delfi.lt), *LRT* (www.lrt.lt), and *Lrytas* (www.lrytas.lt), and stored articles where headlines contained keywords "covid" or "korona". As articles were obtained in Lithuanian language, for further analysis purposes they were firstly translated to English and later, a supervised classifier, described in Section 3.2.4, assigned each headline a sentiment label (negative, neutral, positive) which was later aggregated to weekly means, denoted as *Sentiment Average* variable and used in further functional modelling. Municipality-level headline counts were insufficient for reliable sentiment curves, therefore, sentiment was modelled as a national-level covariate.

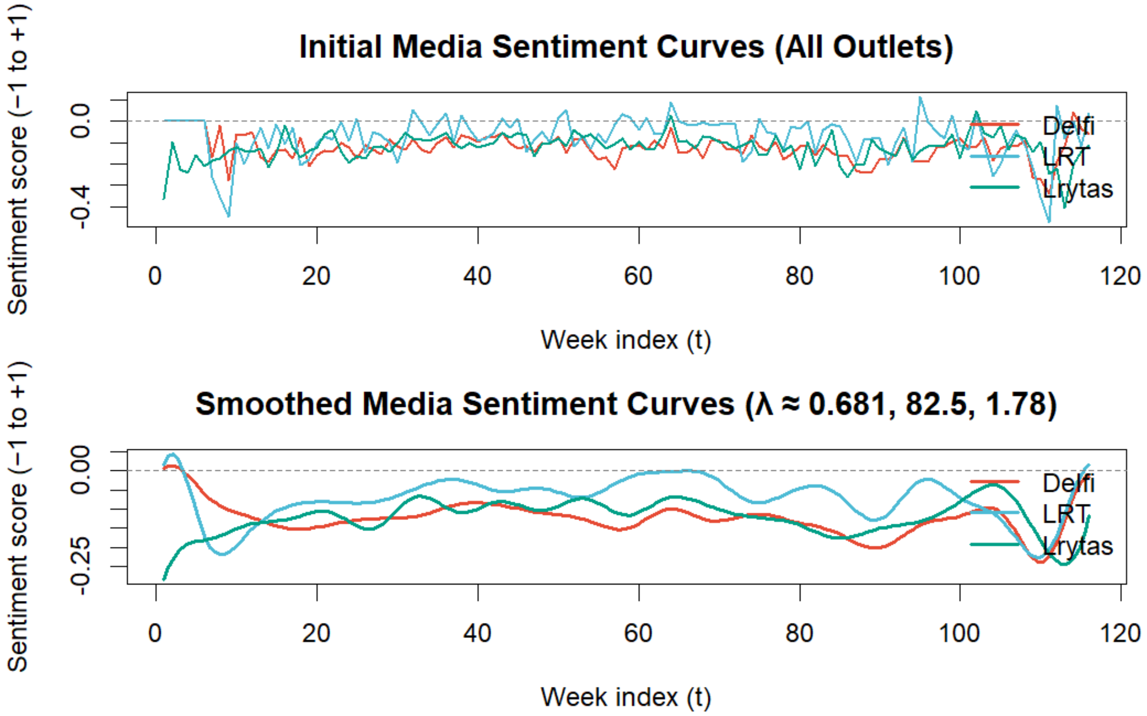

Figure 5: *Initial and smoothed weekly media sentiment curves across three major media outlets: Delfi, LRT, and Lrytas.*

Figure 5 shows that the initial sentiment curves exhibit substantial short-term fluctuations around zero, reflecting the noisiness of sentence-level sentiment scores and the variability of daily news coverage. After smoothing, a clearer temporal structure emerges: all outlets display predominantly negative sentiment throughout the period, generally remaining between -0.30 and 0.05. The trajectories differ in magnitude, with *Delfi* and *Lrytas* showing consistently more negative tone, while *LRT* (the national public broadcaster) maintains a more neutral stance. Notably, *LRT* is also the only outlet that occasionally reaches positive sentiment peaks, seen in the initial curves. Despite these differences, the overall shapes of the curves remain broadly similar across outlets. Mild oscillatory patterns coincide with major pandemic waves (e.g., during 2020–2021 and late 2021), suggesting sentiment becomes more negative during periods of epidemiological stress. Compared with meteorological variables, however, sentiment curves exhibit weaker seasonal regularity and more outlet-specific variation.

2.2 Preprocessing and Harmonization

All datasets were synchronised using the ISO 8601 week-date standard [25]. Data transformation and harmonization followed tidy data principles [54] and standard time-series pre-processing approaches [22].

2.2.1 ISO Week Alignment

Temporal alignment across all datasets followed the ISO 8601 standard, which defines a consistent system of week numbering with Monday as the first day of the week. Each observation was assigned an ISO week identifier of the form `YYYY-Www`, after which the weekly domain was indexed sequentially as $t = 1, \ldots, 116$. Only weeks present simultaneously in all data sources were retained, resulting in a harmonised panel of 116 aligned weeks covering the full analysis period.

2.2.2 Treatment of Missing Values

COVID-19 cases and vaccination variables contained no missing values after weekly aggregation. Some municipalities were missing weekly weather measurements - these were replaced with national same-week averages. The sentiment series contained a small number of missing days (4–5 per outlet) in the early days of the pandemic, when COVID-19 terminology was not yet widely used. These were imputed using linear interpolation with forward/backward fill. Manual inspection confirmed these gaps correspond to days without COVID coverage, not scraping errors, therefore assigning a neutral value is reasonable.

2.2.3 Construction of the Analytical Dataset

Following temporal harmonisation and standardisation of variable formats, all sources were merged into a unified weekly panel suitable for functional analysis. The resulting dataset consisted of 60 municipality-level case curves, dose-specific vaccination curves for the same 60 municipalities, and corresponding temperature and absolute humidity trajectories. In addition to these municipality-level variables, three national-level sentiment curves—derived from *Delfi*, *LRT*, and *Lrytas*—were incorporated as exogenous functional covariates. Together, these components form the analytical basis for all subsequent Functional Data Analysis procedures.

2.3 Smoothing and Functional Representation

Following the standard FDA formulation [43], each weekly time series was represented as a functional object using a basis expansion. Let $t \in [1,116]$ denote the weekly time domain, $X_i(t)$ the functional trajectory of municipality i, $\phi_j(t)$ the basis functions, and c_{ij} the corresponding expansion coefficients. The functional representation is

$$X_i(t) = \sum_{j=1}^{K} c_{ij}\,\phi_j(t),$$

where $\phi_j(t)$ are B-spline or Fourier basis functions, and the coefficients c_{ij} are estimated via penalized least squares.

2.3.1 Basis Selection

To construct functional representations of all weekly time series, appropriate basis systems were chosen according to the temporal characteristics of each variable. For epidemiological, vaccination, and sentiment data, a B-spline basis with $K = 25$ functions was used due to their non-periodic behaviour. In this setting, the basis functions take the form

$$\phi_j(t) = B_j(t),$$

allowing flexible modelling of irregular fluctuations while maintaining local support.

In contrast, temperature and absolute humidity exhibit clear seasonal periodicity. For these variables, a Fourier basis was employed:

$$\phi_j(t) \in \{\sin(\omega_j t), \cos(\omega_j t)\}.$$

The Fourier system efficiently captures cyclical behaviour and avoids edge effects associated with spline-based representations. The choice of $K = 25$ basis functions for all variables follows the practical recommendations of Ramsay and Silverman [43], providing adequate flexibility for reconstructing weekly dynamics without inducing overfitting.

2.3.2 Roughness Penalty

Functional smoothing was performed using a penalized least squares criterion. To prevent overfitting to short-term noise, the roughness penalty was applied to the integrated squared second derivative of the curve:

$$\text{Penalty}(X) = \lambda \int \left(X''(t)\right)^2 dt,$$

where $\lambda > 0$ controls the degree of smoothness. Larger values of λ enforce stronger regularization and yield smoother curves, whereas smaller values retain more local variation. This formulation follows the standard FDA framework proposed by Ramsay and Silverman [43].

2.3.3 Smoothing Parameter Selection

The optimal smoothing parameter λ for each variable was selected using Generalized Cross-Validation (GCV; [16]). For a given λ, the smoothing operation implemented in `smooth.basis()` defines a linear smoother

$$\hat{\mathbf{y}}_\lambda = S_\lambda \mathbf{y},$$

which allows the GCV score to be computed as

$$\text{GCV}(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}_\lambda\|^2}{\left(1 - \frac{\text{trace}(S_\lambda)}{n}\right)^2}.$$

Because epidemiological, vaccination, and meteorological datasets include a separate time series for each municipality, GCV was computed individually for every municipality and then averaged:

$$\overline{\text{GCV}}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} \text{GCV}_i(\lambda).$$

The value of λ minimizing this average was selected as the global smoothing parameter for that variable. Using a common λ across municipalities ensures that all functional trajectories are smoothed to a comparable degree, which is essential for downstream multivariate analyses such as FPCA and FCCA. Although the optimal λ may vary slightly between municipalities, imposing a shared smoothing parameter avoids distortions in covariance estimation and facilitates meaningful comparison of functional components. This approach is standard for replicated functional data [42]. For single national-level time series (e.g., media sentiment curves), GCV was computed directly using the observed vector. A custom grid search over λ values was implemented for each variable type. For B-spline bases, the grid ranged from 10^{-6} to 10^{4}, while Fourier bases used a more conservative range from 10^{-6} to 10^{0}, reflecting their differing regularity properties. These grids ensured that undersmoothing and oversmoothing regimes were thoroughly evaluated. Table 1 displays the optimal smoothing parameters selected by GCV for each variable.

Table 1: *Optimal smoothing parameters λ selected by generalized cross-validation.*

Variable	**Basis Type**	**Optimal λ**
Cases	B-spline	562.3
Vaccinations – Total	B-spline	4.642
Air Temperature	Fourier	1
Absolute Humidity	Fourier	1
Sentiment – Delfi	B-spline	0.6813
Sentiment – LRT	B-spline	82.54
Sentiment – Lrytas	B-spline	1.778
Sentiment - Average	B-spline	31.62

3 Functional Modelling

This chapter presents the methodological framework used to analyse the relationship between COVID-19 cases, vaccination uptake, meteorological conditions, and media sentiment. The analysis relies on Functional Data Analysis (FDA), which treats time series as smooth functions rather than discrete sequences, enabling the extraction of dominant temporal patterns and modelling complex time-varying dependencies.

3.1 Overview of Analytical Framework

The analytical objective is to examine how vaccination, weather conditions, and media sentiment relate to the temporal evolution of COVID-19 cases at national and municipal levels. The main methodological challenges are that all variables are observed weekly rather than continuously, that municipalities exhibit heterogeneous curve shapes—especially in vaccination uptake—and that potential effects may operate with temporal delays.

Functional Data Analysis (FDA) is used to address these issues. Weekly observations are first smoothed into continuous functional curves, after which Functional Principal Component Analysis (FPCA) is applied to extract dominant temporal patterns. Relationships between cases and each predictor are then analysed using reduced Functional Canonical Correlation Analysis (FCCA) and lagged correlation profiles. Finally, the function-on-function regression model quantifies how predictor trajectories are associated with COVID-19 cases over time.

3.2 Sentiment Classification Pipeline

Media sentiment was used as a functional predictor of COVID-19 cases. Empirical testing on 300 manually labelled headlines showed VADER misclassified pandemic-specific terms (e.g. ‘quarantine’, ‘self-isolation’), yielding inconsistent polarity scores, therefore, a supervised classification model was developed to assign polarity to headlines from Delfi, LRT, and Lrytas. This approach ensured more reliable sentiment estimates for COVID-19–specific language.

3.2.1 Text Preprocessing and Translation

To leverage established English Natural Language Processing (NLP) tools, all Lithuanian headlines were translated into English using a Google Cloud Translation API [21]. Machine translation has been shown to preserve sentiment polarity sufficiently for downstream classification tasks [4].

3.2.2 Manual Annotation and Label Construction

A set of 300 randomly sampled headlines was manually labelled into three classes: *negative*, *neutral*, and *positive*, following standard polarity guidelines [32]. These labels served as ground truth for supervised sentiment classification.

3.2.3 Feature Extraction: TF–IDF

Each headline was transformed into a Term Frequency–Inverse Document Frequency (TF–IDF) vector:

$$\mathsf{tfidf}_{w,d} = \mathsf{tf}_{w,d} \cdot \log\left(\frac{N}{\mathsf{df}_w}\right),$$

where $\mathsf{tf}_{w,d}$ is the frequency of word w in document d, df_w is the number of documents containing word w, and N is the total number of documents. TF–IDF weighting highlights discriminative terms and is widely used in sentiment classification due to its sparsity and interpretability [27].

3.2.4 Classification Model: Multinomial Logistic Regression

A multinomial logistic regression model was trained on the TF–IDF features:

$$P(Y = c \mid x) = \frac{\exp\left(\beta_c^\top x\right)}{\sum_k \exp\left(\beta_k^\top x\right)}, \qquad c \in \{\text{neg,neu,pos}\}.$$

Given the relatively small manually labelled dataset (300 headlines), linear models such as logistic regression outperform transformer-based approaches, which require substantially more labelled data. Logistic regression is well-suited for high-dimensional sparse inputs and has demonstrated robust performance in sentiment classification tasks [8]. The classifier achieved 91% accuracy on a held-out validation subset and outperformed the lexicon-based VADER tool, which produced inconsistent scores on pandemic-specific terminology.

3.2.5 Construction of Weekly Sentiment Curves

The trained model was applied to all $\sim$33,000 scraped headlines. Predicted labels were mapped to numerical sentiment values:

$$\text{negative} = -1, \qquad \text{neutral} = 0, \qquad \text{positive} = +1.$$

Daily averages were computed per outlet, then aggregated into weekly means to match the temporal resolution of epidemiological data. This produced four national-level sentiment curves:

$$\text{Sent}_{\text{Delfi}}(t), \quad \text{Sent}_{\text{LRT}}(t), \quad \text{Sent}_{\text{Lrytas}}(t), \quad \text{Sent}_{\text{avg}}(t).$$

These curves were smoothed using penalised B-splines. Because sentiment does not vary across municipalities, it contains no cross-sectional variation and cannot explain between-municipality differences for COVID-19 cases. It is therefore treated strictly as a national temporal covariate. In function-on-function regression, sentiment captures shared temporal alignment with cases rather than spatial heterogeneity.

3.3 Functional Principal Component Analysis (FPCA)

FPCA is used to identify the dominant patterns of variation across the 60 municipalities for cases, vaccination, temperature, and absolute humidity curves. Its role is dual: it reduces each functional trajectory to a small set of principal component scores, and it reveals interpretable modes of variation, such as shifts between early and late epidemic waves or differences between seasonal and non-seasonal behaviour. All functional variables in this thesis are treated as elements of $L^2[1,116]$, ensuring well-defined mean functions, covariance operators, and eigenfunction decompositions required for FPCA.

3.3.1 FPCA Model Formulation

Let $X_i(t)$ denote the smoothed trajectory for municipality i and $\mu(t)$ the mean function. FPCA decomposes each curve as:

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\, \phi_k(t),$$

where $\phi_k(t)$ are orthonormal eigenfunctions of the covariance operator, ξ_{ik} are PC scores, and λ_k are eigenvalues satisfying $\lambda_1 \geq \lambda_2 \geq \cdots$.

Before performing FPCA, all curves were centred:

$$X_i(t) \leftarrow X_i(t) - \hat{\mu}(t),$$

ensuring that PC scores reflect deviations from the temporal mean structure rather than absolute levels of cases.

Eigenfunctions satisfy:

$$\int \Sigma(s,t)\, \phi_k(s)\, ds = \lambda_k \phi_k(t),$$

where $\Sigma(s,t)$ denotes the covariance surface. In practice, only the first few PCs (typically 2–4) are retained because they explain most variability, consistent with standard FDA theory [23, 43]. FPCA is applied to all functional predictors, and their leading components are used to:

- compare temporal patterns across municipalities,
- construct reduced-dimensional inputs for FCCA,
- visualise heterogeneity across curves,
- stabilise multivariate analyses.

Despite limited spatial variation in some predictors (vaccination, weather), treating each municipality as a separate functional trajectory remains informative. The FDA allows the separation of a shared national temporal structure from residual municipal differences, ensuring that downstream analyses capture both synchronised and heterogeneous behaviours.

3.4 Multivariate Functional Relationships

Understanding how vaccination, weather, and sentiment relate to cases requires multivariate functional tools. Two approaches are used: reduced Functional Canonical Correlation Analysis (FCCA) and lagged FCCA.

3.4.1 Reduced FCCA

Classical FCCA seeks weight functions $a(t)$ and $b(t)$ that maximise the correlation between projections $\langle X, a\rangle$ and $\langle Y, b\rangle$ [29, 43]. This requires inversion of covariance operators, which is numerically unstable when data are noisy or sparsely observed.

To mitigate instability, a reduced-rank approach is used: FPCA is applied to each functional variable, and multivariate canonical correlation analysis is performed on the resulting score vectors [24, 28]. Let

$$\Xi = (\xi_{i1}, \ldots, \xi_{iK}), \qquad \Upsilon = (\upsilon_{i1}, \ldots, \upsilon_{iM})$$

denote FPCA score vectors for COVID-19 cases and a predictor (e.g. vaccination or temperature). Reduced CCA finds linear combinations:

$$U = a^\top \Xi, \qquad V = b^\top \Upsilon$$

that maximise:

$$\rho = \max_{a,b} \ \text{corr}(U, V).$$

The resulting canonical correlations $(\rho_1, \rho_2, \ldots)$ quantify the strength of association between dominant temporal modes of cases and the predictor. FCCA captures shared temporal structure rather than causal effects; synchronised national patterns (e.g. vaccination rollout aligning with epidemic waves) may yield high correlations even without a systematic relationship [7]. This is accounted for in interpreting results.

3.4.2 Lagged FCCA and Cross-Correlation Analysis

Epidemiological processes often contain delays: meteorological conditions may affect transmission with a several week lag, behavioural changes take time to manifest, and sentiment may shift before COVID-19 cases respond.

To incorporate potential delays, each functional predictor $X(t)$ was shifted by discrete lags ℓ using the standard functional lag operator [36]:

$$X_\ell(t) = X(t - \ell), \qquad \ell = -8, -7, \ldots, 8.$$

For each lag, FPCA was recomputed to maintain alignment over the effective domain [13]. Reduced FCCA was then applied to lagged score matrices, and the first canonical correlation $\rho_1(\ell)$ was extracted to produce a lag-correlation profile [40].

To assess whether the alignment between sentiment and cases could arise by chance, a circular-shift permutation test was applied, which preserves the autocorrelation structure of the sentiment curve while disrupting its temporal alignment with cases [50]. The test used 2000 circular permutations per sentiment series, generating phase-randomised surrogate curves while maintaining the aforementioned autocorrelation structure.

3.5 Municipality-Level Function-on-Function Regression

To fully utilise the spatio-temporal structure of the dataset, function-on-function (FoF) regression was used to model each municipality's case trajectory $Y_i(t)$ as a function of temperature, absolute humidity, vaccination, and sentiment curves. The model follows the general functional linear framework [26, 43, 45]:

$$\begin{aligned} Y_i(t) = {} & \beta_0(t) + \int_{s\leq t} \text{Temp}_i(s)\, \beta_{\text{Temp}}(s,t)\, ds + \int_{s\leq t} \text{AH}_i(s)\, \beta_{\text{AH}}(s,t)\, ds \\ & + \int_{s\leq t} \text{Vacc}_i(s)\, \beta_{\text{Vacc}}(s,t)\, ds + \int_{s\leq t} \text{Sent}(s)\, \beta_{\text{Sent}}(s,t)\, ds + \varepsilon_i(t), \end{aligned}$$

where $\beta_0(t)$ is the baseline mean function, and the bivariate coefficient surfaces $\beta_{\text{Temp}}(s,t)$, $\beta_{\text{AH}}(s,t)$, $\beta_{\text{Vacc}}(s,t)$, and $\beta_{\text{Sent}}(s,t)$ describe how predictor values observed at time s influence COVID-19 cases at time t.

To ensure temporal coherence, coefficient surfaces are interpreted under a historical framework, such that predictor values observed at time s influence COVID-19 cases at time t primarily when $s \leq t$. Although the function-on-function regression does not impose a hard constraint forcing $\beta(s,t) = 0$ for $s > t$, smoothness penalties strongly suppress implausible regions where future predictor values would affect past outcomes. Consequently, estimated effects are interpreted as predominantly historical rather than anticipatory. Coefficient surfaces were estimated using tensor-product spline bases with penalties on both dimensions [26, 45, 55]. Statistical significance was assessed using approximate F-tests implemented in the `pffr()` function of the `refund` package.

Because sentiment does not vary across municipalities, its effect is purely temporal and does not explain spatial variation for COVID-19 cases. Its coefficient surface therefore, reflects alignment with the national epidemic curve, consistent with functional regression theory for non-varying predictors [40].

3.6 Software and Implementation

All analyses were conducted using a combination of R and Python. In R, the `fda` package was used to construct functional data objects and perform smoothing and FPCA, while function-on-function regression was implemented through the `refund` package (`pffr`). Additional smoothing and basis-representation infrastructure relied on `mgcv`. Data management, temporal alignment, and general preprocessing were carried out using the `tidyverse` suite, together with `lubridate` for date handling and `zoo` for time-series interpolation.

Python was used primarily for sentiment classification. Text preprocessing and TF–IDF vectorisation were implemented in `scikit-learn`, and the sentiment classifier was trained using its multinomial logistic regression module. Data wrangling relied on `pandas` and `numpy`, and all Lithuanian headlines were translated to English via a Google Cloud Translation API [21] prior to modeling. Random seeds were fixed where applicable to ensure reproducibility of results.

4 Results

This chapter presents the empirical findings obtained from the functional analysis of COVID-19 cases, vaccination rollout, meteorological factors, and national media sentiment in Lithuania. The focus is on interpreting temporal structures revealed by functional smoothing, FPCA, FCCA, lagged analyses, and the functional regression model. Together, these results provide a full picture of the analysed drivers of the Lithuanian pandemic across 60 municipalities.

4.1 Descriptive Temporal Structure

Figure 6 presents the smoothed national trajectories for all variables, providing a functional overview before multivariate modelling. COVID-19 cases showcase two dominant epidemic waves, while vaccination follows with three national rollouts beginning in early 2021 (representing first, second and booster doses administered to the Lithuanian population). Temperature and absolute humidity display regular seasonal cycles, and sentiment remains predominantly negative with pronounced dips after major outbreaks.

These smoothed curves illustrate the key temporal features that FDA subsequently decomposes: multi-wave dynamics in cases, a reactive vaccination profile, strong meteorological seasonality, and a comparatively irregular sentiment signal. They provide the functional context for the FPCA, FCCA, and regression analyses that follow.

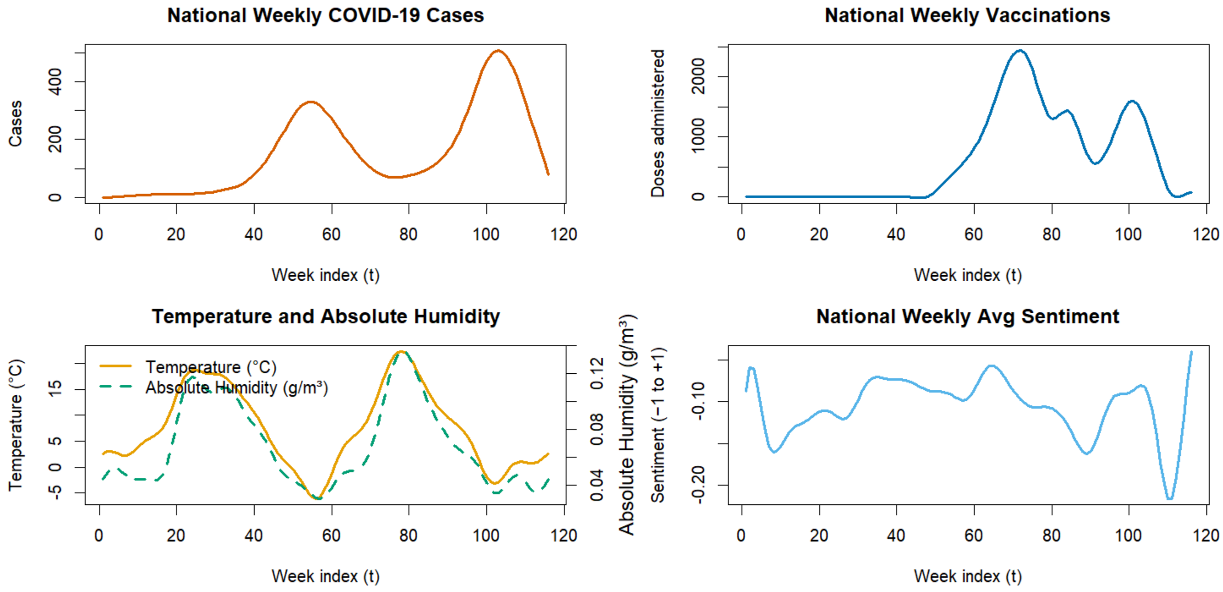

Figure 6: National smoothed weekly trajectories for cases, vaccination, meteorological variables, and sentiment.

4.2 Dominant Functional Modes

Functional Principal Component Analysis (FPCA) summarises the main temporal patterns across the 60 municipalities. The scree plots in Figure 7 show that each variable is driven by a small number of dominant components. For COVID-19 cases, the first component explains 63% of the total variance, reflecting the overall multi-wave structure of the epidemic, while PC2 and PC3 (19% and 10%) capture differences between early and late waves and the distinctive Omicron peak. Vaccination trajectories are almost entirely governed by a single component (PC1 99.6%), indicating near-perfect synchronisation in the national rollout.

Temperature and absolute humidity exhibit similarly concentrated structures: PC1 alone explains 78% and 68% of their variance respectively, reflecting the dominant annual seasonal cycle, with higher-order components accounting for minor deviations from this pattern.

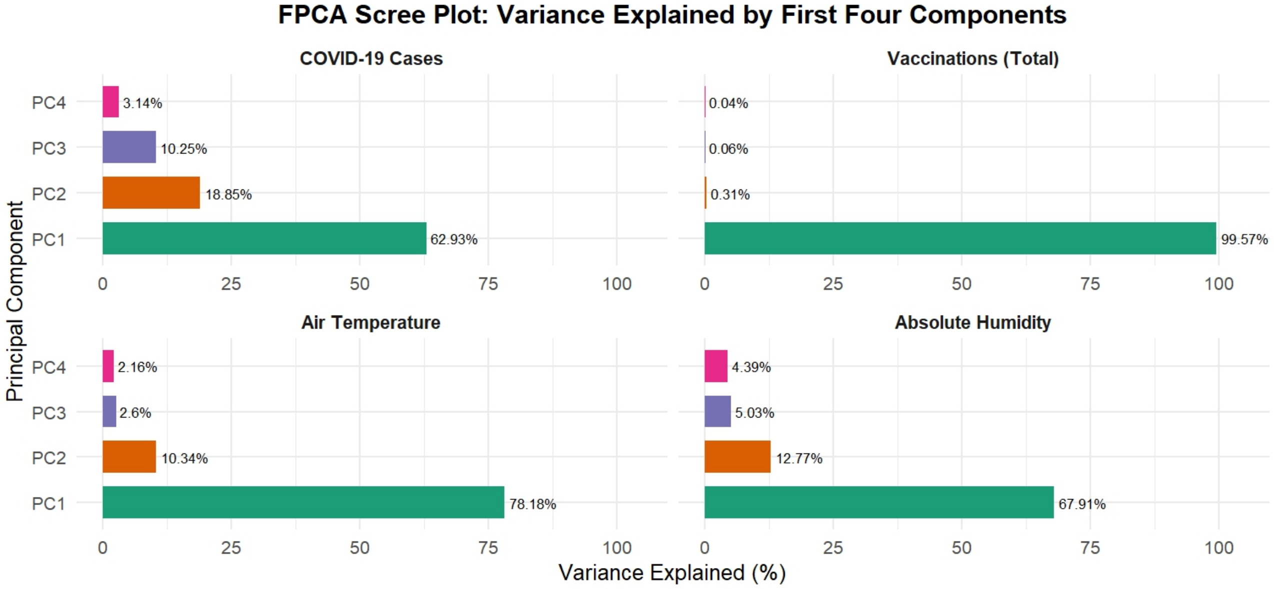

Figure 7: *FPCA scree panels showing the proportion of variance explained by the first four components for each variable.*

The corresponding eigenfunctions (Figure 8, displayed below) clarify the nature of these dominant modes. For COVID-19 cases, PC1 traces the broad multi-peak epidemic shape, PC2 contrasts early and late municipalities (shifting the weight between the first and second large waves), and PC3 isolates the sharp rise associated with the Omicron variant. The eigenfunctions for vaccination curves show that PC1 captures the timing and speed of the national rollout, while higher components contribute almost no meaningful structure. For temperature and absolute humidity, eigenfunctions exhibit smooth sinusoidal shapes characteristic of strong seasonality, with PC2 and PC3 adding small asymmetric or amplitude-adjustment features.

Taken together, these results confirm that shared national temporal structure—epidemic waves, vaccination rollout, and meteorological seasonality—dominates municipal variation. Local differences exist but contribute primarily to higher-order components, which explain only a small fraction of total variance.

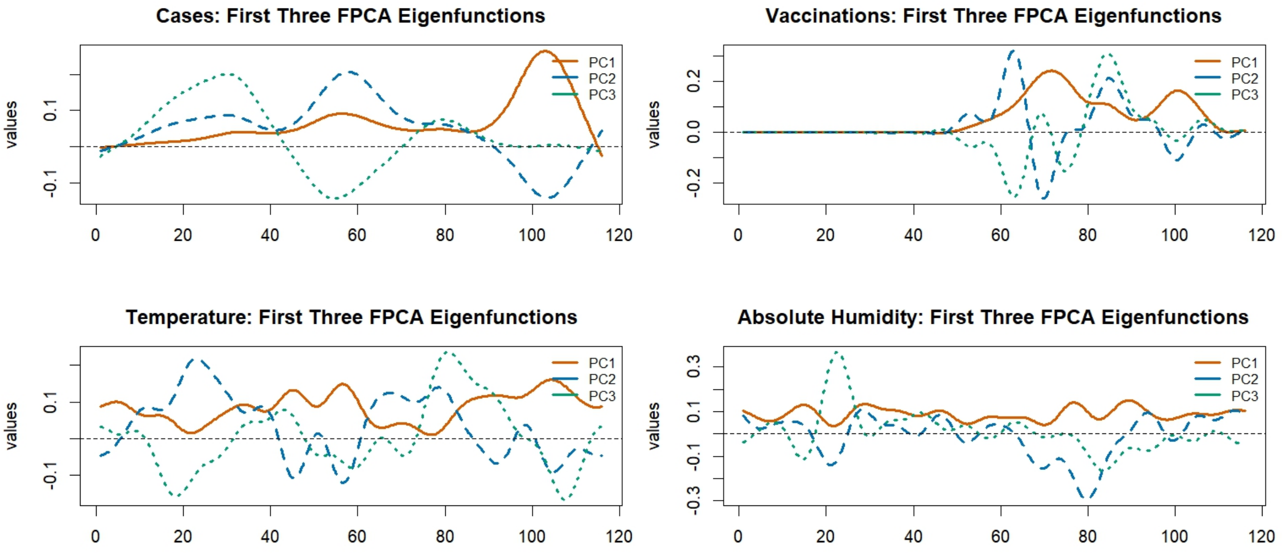

Figure 8: *First three eigenfunctions for COVID-19 cases, vaccination, temperature, and absolute humidity.*

4.3 Shared Temporal Structure: FCCA

To quantify how closely the dominant temporal modes of COVID-19 cases align with those of the predictor variables, reduced Functional Canonical Correlation Analysis (FCCA) was applied to the FPCA score matrices for all 60 municipalities. FCCA identifies linear combinations of the leading functional principal components that are maximally correlated, producing paired canonical variates for each municipality. Figure 9 presents the results for cases versus vaccination.

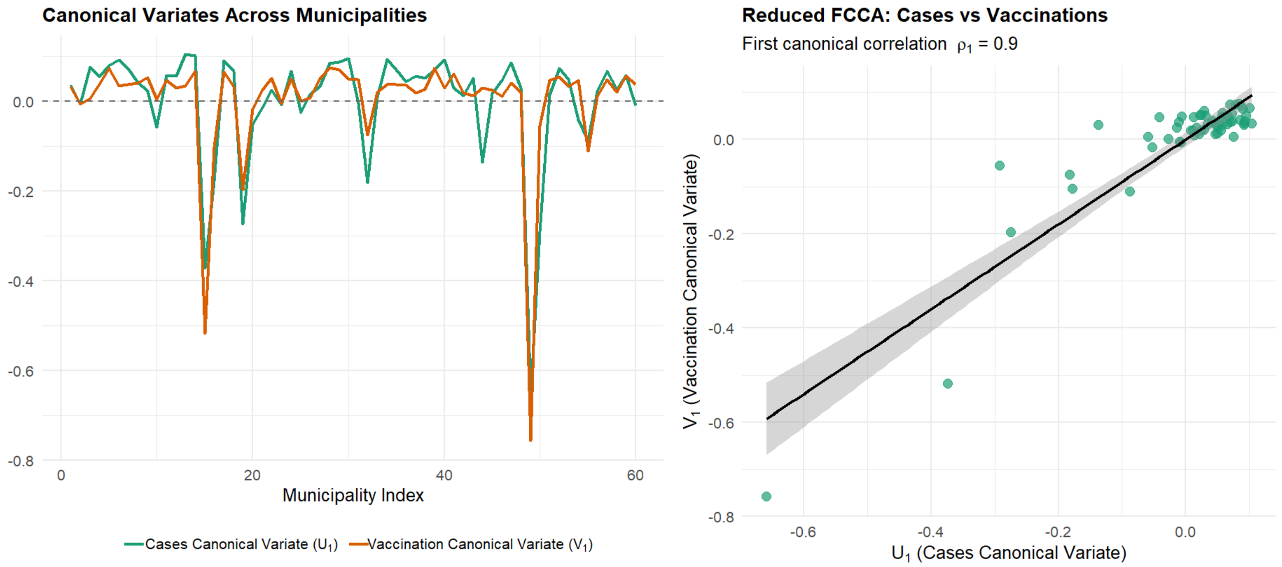

Figure 9: *Reduced FCCA for COVID-19 cases and vaccination.*

The left panel shows the first canonical variates U_1 (cases) and V_1 (vaccination) evaluated across municipalities. The two curves follow a nearly identical pattern: municipalities that score high on the dominant temporal mode of vaccination during the rollout period also score high on the dominant temporal mode of cases. The right panel displays the scatterplot of these canonical variates, yielding an extremely strong first canonical correlation ($\rho_1 \approx 0.90$).

Importantly, this association does not imply that COVID-19 cases increases vaccination. Rather, it reflects the fact that both processes share a highly synchronised national temporal structure. The first principal component of cases captures the timing and magnitude of epidemic waves, while the first principal component of vaccination captures the timing of the national rollout. FCCA therefore shows that municipalities participated in these national cycles in similar ways—municipalities that were relatively early or late in one process tended also to be early or late in the other. Associations with meteorological variables are weaker but systematic. Absolute humidity exhibits the strongest relationship with cases ($\rho_1 \approx 0.46$), consistent with mechanisms linking low moisture levels to increased respiratory virus transmission. Temperature shows a similar but slightly smaller canonical correlation, reflecting shared seasonal structure across municipalities.

Overall, FCCA reveals that:

1. the strongest shared temporal pattern occurs between cases and vaccination, driven by synchronised national timing rather than causal effects;
2. meteorological seasonality remains an important structural component of epidemic intensity; and
3. municipality-specific deviations exist but are secondary relative to the dominant national patterns governing all variables.

4.4 Lagged Relationships

Because the mechanisms linking weather, vaccination, and COVID-19 dynamics are inherently time-dependent, the temporal ordering of these processes was examined using lagged reduced FCCA at the municipality level. For each predictor, the FPCA score matrices were shifted forward and backward by up to eight weeks, and the first canonical correlation $\rho_1(\ell)$ was recomputed for each lag ℓ. Positive lags indicate that the predictor precedes changes in COVID-19 cases, while negative lags indicate that cases lead the predictor. The resulting profiles are shown in Figure 10 and Figure 11.

Figure 10 displays the lagged correlations for temperature, absolute humidity, and vaccination. For temperature, the highest canonical correlation occurs at a lag of eight weeks, with $\rho_1 \approx 0.50$. The gradual rise in correlation toward positive lags suggests that seasonal deterioration in meteorological conditions precedes increases in COVID-19 cases by several weeks. Given the use of weekly aggregation and functional representations, this lag should be interpreted as a phase shift between dominant seasonal patterns rather than a direct biological delay. Such timing is consistent with the gradual build-up of epidemic waves during colder months, mediated by behavioral and environmental factors. Absolute humidity reveals a similar lag structure, with a peak of $\rho_1 \approx 0.50$ at a lag of eight weeks, further supporting moisture-related mechanisms influencing respiratory spread. Although the magnitude of these associations is moderate, the temporal pattern is remarkably consistent across both meteorological variables: conditions characteristic of winter months precede the growth of epidemic waves by several weeks.

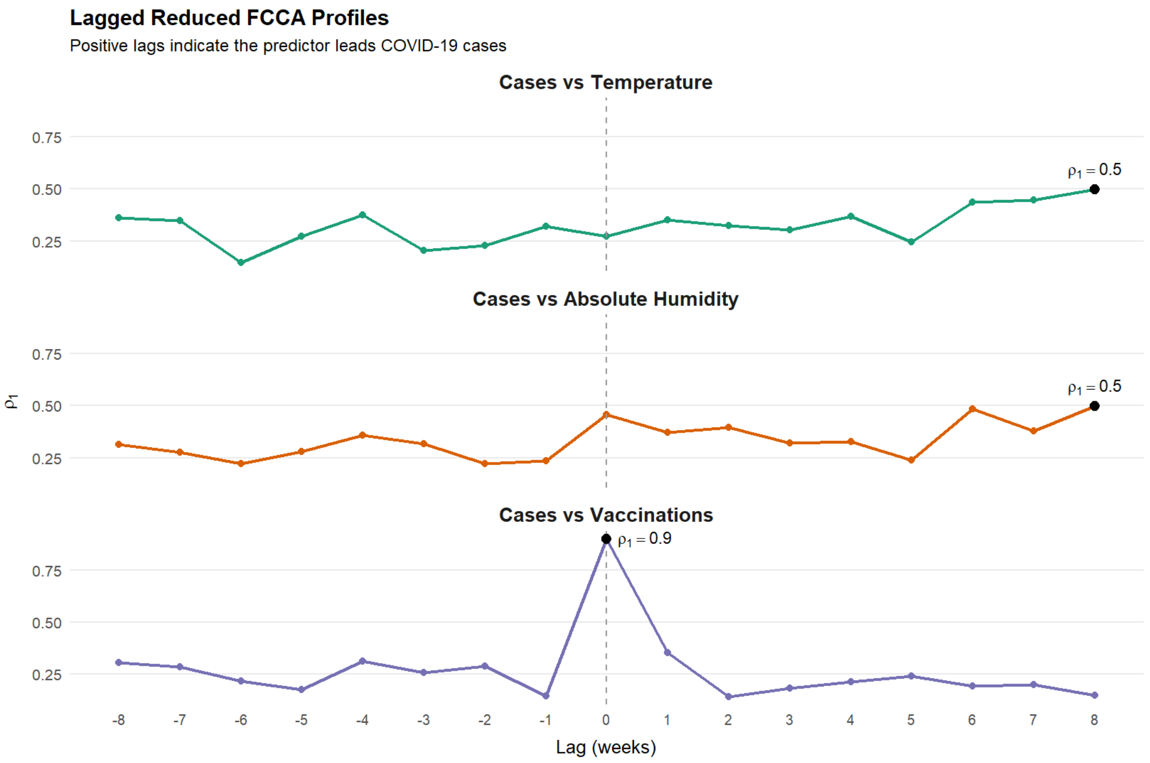

***Figure 10:** Lagged canonical correlation profiles for temperature, absolute humidity and COVID-19 cases.*

The vaccination profile is noticably different. Its lag curve shows a sharp maximum at lag zero, where $\rho_1 \approx 0.90$, but correlations fall rapidly when the predictor is shifted forward or backward. This pattern reflects the synchronised national vaccination rollout rather than a causal lead–lag relationship. Vaccination does not systematically precede changes in cases and instead the two processes share similar population-level timing because large-scale vaccination campaigns in Lithuania were initiated during or immediately following epidemic surges. The symmetric decay away from lag zero further indicates that this association arises from overlapping temporal structure, not anticipatory behavioural effects.

Figure 11 presents the national-level lagged correlation profiles for the three outlets and the combined sentiment index. Delfi sentiment leads cases by 6–8 weeks with strong significance ($p < 0.001$). LRT and Lrytas exhibit weaker lead effects; a combined sentiment index remains significant ($p \approx 0.041$). Sentiment curves display substantially more variability than meteorological variables but still exhibit interpretable temporal patterns. Delfi shows a clear lead effect, with the strongest positive correlation occurring approximately six to eight weeks before increases in cases, consistent with the tendency of major outlets to intensify negative coverage in the early phases of a worsening epidemic. The combined sentiment index showcases a similar though weaker pattern, reaching statistical significance at approximately the same lag range ($p \approx 0.04$). In contrast, LRT and Lrytas show noisier profiles with no statistically significant lead or lag structure. These differences reflect divergent editorial practices and audience roles: Delfi's faster responsiveness yields clearer alignment, while LRT—the public broadcaster—adopts a more neutral, less reactive tone.

Taken together, the lagged analyses clarify the temporal ordering among the examined processes. Meteorological effects precede changes in COVID-19 cases by several weeks, as expected from environmental transmission pathways. Media sentiment, particularly from commercial outlets, tends to deteriorate in advance of rising case numbers, acting as an informational early-warning signal rather than a behavioural driver. Vaccination, by contrast, aligns with epidemic waves rather than predicting them, reflecting policy timing rather than anticipatory public health effects. These results emphasise the importance of distinguishing shared temporal structure from genuine lead–lag relationships when interpreting functional correlations.

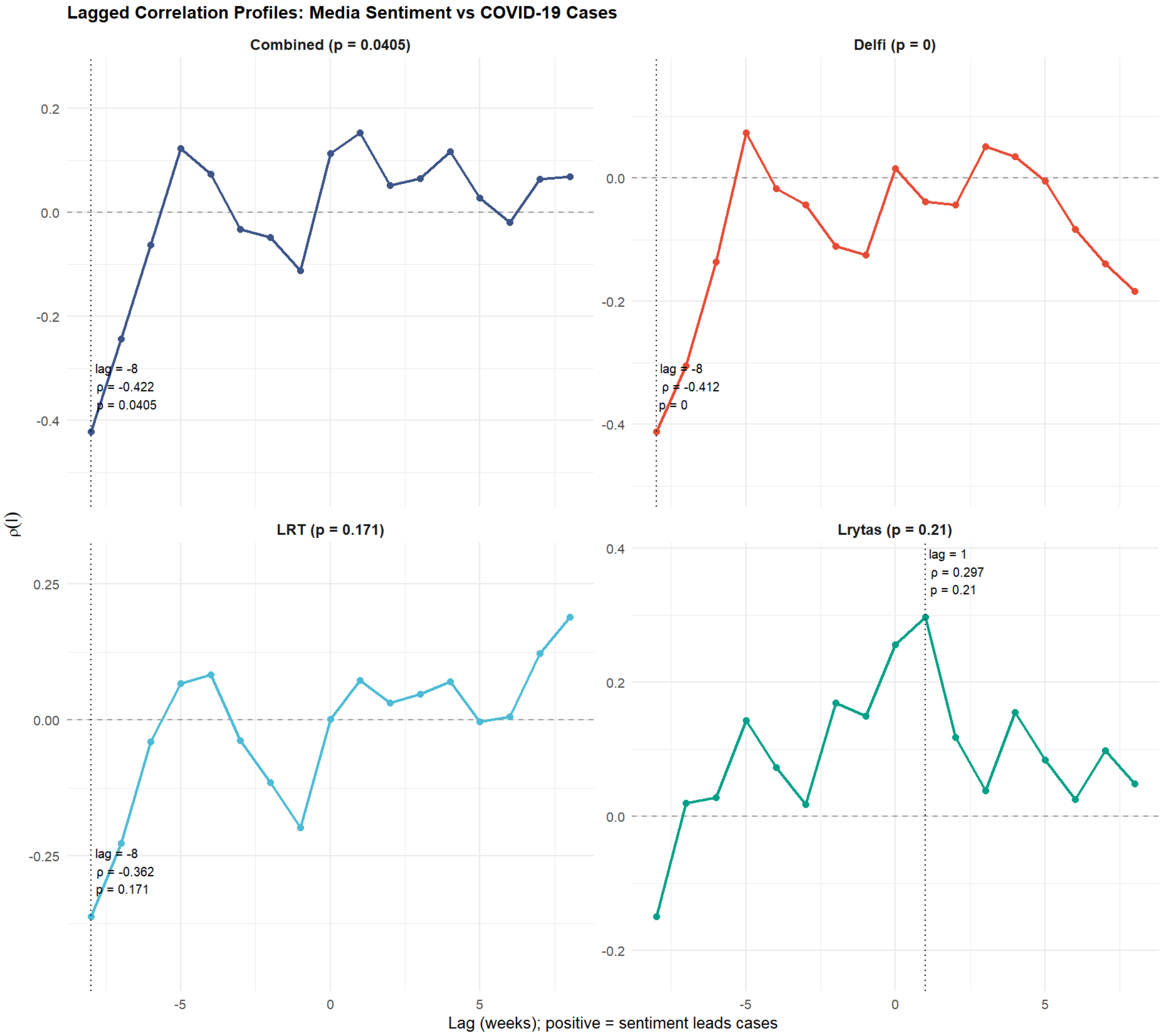

Figure 11: *Permutation-test lag profiles for media sentiment and COVID-19 cases.*

4.5 Regression-Based Evidence

To quantify the joint influence of meteorological conditions, vaccination rollout, and media sentiment on the temporal evolution of COVID-19 cases across municipalities, a function-on-function regression (FoFR) model was estimated using the `pffr` framework. Unlike concurrent regression models, which relate predictors and response at the same point in time, FoFR incorporates entire functional trajectories and allows the effect of past predictor values to vary smoothly across the cases curve.

The model achieves strong predictive accuracy: the adjusted $R^2 = 0.943$ and the functional coefficient of determination is $R^2_{\text{fun}} = 0.944$. All functional terms are highly significant ($p < 0.001$), indicating that temperature, absolute humidity, vaccination, and sentiment each contribute time-varying information conditional on the shared temporal structure of the epidemic. These values represent in-sample functional fit under penalized estimation, however, they do not imply predictive or causal performance.

To assess whether the high functional coefficient of determination reflects overfitting rather than substantive structure, a sequence of nested FoFR models was estimated. A model including only meteorological predictors already achieves a functional $R^2 = 0.931$, indicating that seasonal environmental variation accounts for the majority of temporal structure in COVID-19 cases. Adding vaccination trajectories increases the functional $R^2 = 0.943$, while inclusion of media sentiment yields a marginal increase to 0.944. The modest incremental gains demonstrate that the high overall fit is driven primarily by dominant seasonal patterns rather than excessive model flexibility, providing evidence against overfitting. Because meteorological variables primarily capture broad seasonal structure, an important question is whether the estimated effects of vaccination and media sentiment merely reflect residual seasonality. The nested regression results suggest this is not the case. While weather alone explains the majority of functional variation, the incremental contributions of vaccination and sentiment persist after accounting for seasonal effects. This indicates that vaccination and sentiment capture non-seasonal temporal information (such as policy timing and informational dynamics) rather than simply acting as proxies for meteorological seasonality.

The coefficient surfaces shown in Figure 12 provide a detailed decomposition of predictor effects across the two-dimensional (s,t) domain. Temperature and absolute humidity exhibit the strongest structured effects. Both surfaces show coherent regions of positive and negative influence, indicating that the effect of meteorological conditions varies across the epidemic timeline. Early in the epidemic waves, negative $\beta(s,t)$ regions dominate (blue), consistent with higher cases under cold, dry conditions. Later periods show positive $\beta(s,t)$ regions (red), reflecting nonlinear seasonal effects or changes in susceptibility and reporting over time. The vaccination surface exhibits only weak structure. This aligns with earlier FPCA and FCCA results: vaccination curves are almost perfectly synchronised across municipalities (PC1 explains 99.6% of their variation), leaving little spatial heterogeneity for the model to exploit. Weak effects reflect a lack of spatial variability, not the absence of epidemiological impact, thus, the FoFR model cannot isolate a strong independent vaccination signal, even though vaccines were effective.

The sentiment surface shows low-amplitude but systematic structure, with modest positive and negative regions that are mostly smooth in the time dimension. Unlike meteorological variables, the effect is not sharply localized along the diagonal. This pattern supports the interpretation from lagged FCCA that sentiment provides weak anticipatory or contemporaneous behavioural signals rather than a mechanistic effect on transmission.

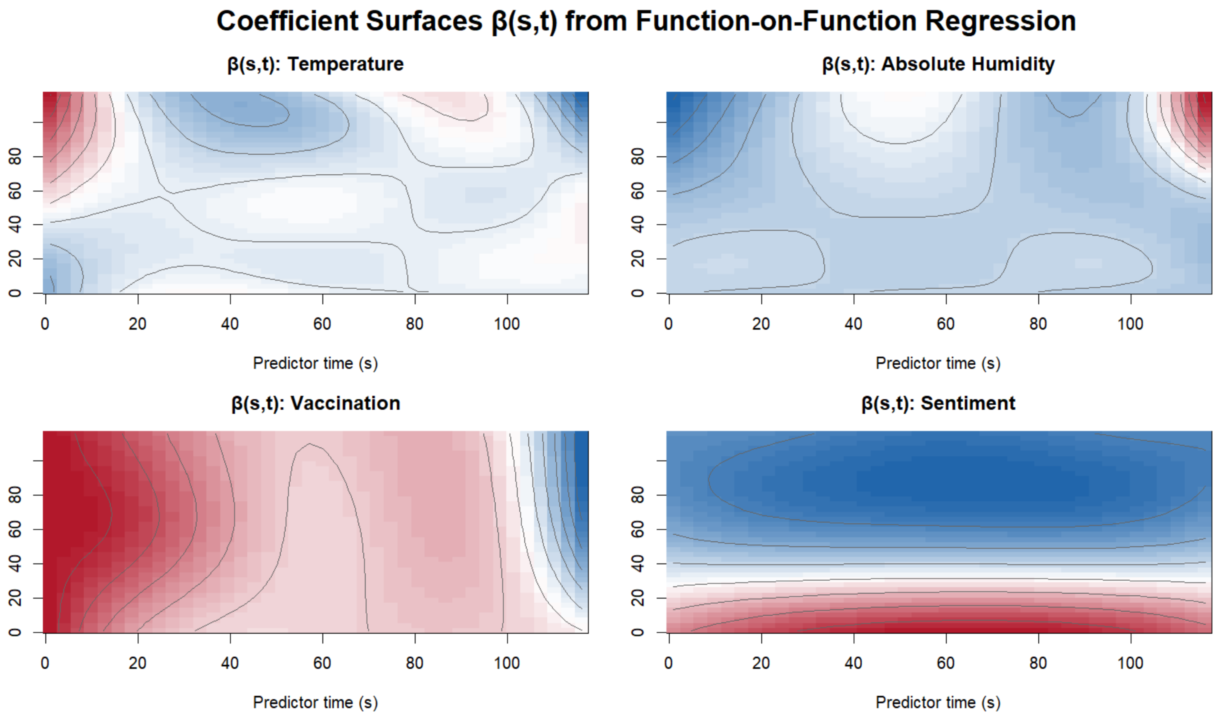

Figure 12: *Coefficient surfaces $\beta(s,t)$ for temperature, absolute humidity, vaccination, and sentiment from the function-on-function regression model.*

A formal statistical summary of the smooth and functional components is provided in Table 2. Overall, the FoFR analysis provides a coherent synthesis of the functional structure in all predictors. Weather conditions emerge as the dominant structural drivers of municipal epidemic variation, vaccination contributes minimal spatial signal and sentiment provides a small but systematic behavioural component. The high functional R^2 indicates that these combined effects explain nearly all between-municipality variation in COVID-19 cases once functional dependencies are incorporated.

Table 2: *Summary of smooth and functional terms in the function-on-function regression model.*

Term	**EDF**	**Ref.df**	**F-statistic**	p**-value**
Intercept(t)	17.318	19.000	3034.47	$< 2 \times 10^{-16}$
Temperature	20.219	22.032	12.67	$< 2 \times 10^{-16}$
Absolute Humidity	19.890	21.484	18.12	$< 2 \times 10^{-16}$
Vaccination	14.290	15.436	104.30	$< 2 \times 10^{-16}$
Sentiment	1.003	1.003	66.29	$< 2 \times 10^{-16}$
Adjusted $R^2 = 0.943$		Functional $R^2 = 0.944$		Deviance explained = 94.4%

4.6 Integrated Interpretation

The results obtained from descriptive smoothing, FPCA, FCCA, lagged FCCA, and the function-on-function regression converge toward a coherent explanation of COVID-19 dynamics in Lithuania. Across all methods, meteorological seasonality emerges as the dominant structural force: cold and low-humidity periods consistently precede COVID-19 cases, municipalities with similar seasonal profiles exhibit similar epidemic trajectories and the coefficient surfaces from the functional regression show strong negative effects of temperature and absolute humidity throughout the growth phase of each wave. Vaccination contributes mainly through national synchronisation rather than independent predictive influence, particularly after its introduction in 2021. Its temporal pattern is highly uniform across municipalities, and its association with cases is largely driven by shared timing of epidemic peaks. Lagged analyses show that vaccination typically reacts to, rather than leads, changes in cases, and its estimated effects in the functional regression remain weak once weather and sentiment are accounted for. Media sentiment provides a smaller but consistent behavioural signal. Lag profiles indicate that shifts in sentiment tend to precede increases in cases, and the corresponding regression surface shows short-term anticipatory effects concentrated near the diagonal.

Taken together, these methods reveal a stable temporal ordering: seasonal meteorological conditions set the structural baseline for transmission and observed cases reflect this seasonal modulation. Vaccination responds to epidemic pressure and sentiment captures early behavioural and informational responses. The coherence of findings across multiple functional techniques demonstrates the value of modelling COVID-19 not as isolated weekly measurements but as a system of connected temporal processes.

5 Conclusions and Recommendations

This thesis examined the temporal dynamics of COVID-19 in Lithuania by analysing cases, vaccination uptake, meteorological conditions, and media sentiment as smooth functional processes. Treating all variables as trajectories rather than discrete weekly observations made it possible to uncover the underlying structure of the epidemic and the ways in which environmental and informational patterns aligned with fluctuations in COVID-19 cases.

The analysis demonstrates that the evolution of SARS-CoV-2 in Lithuania was shaped predominantly by broad seasonal environmental patterns. Both temperature and absolute humidity exhibited strong periodicity, with their lowest values occurring during the winter months. Functional canonical correlations and lagged analyses showed that declines in temperature and absolute humidity generally preceded increases in cases by six to eight weeks. While these associations do not imply a direct causal effect (given that both weather and epidemic waves share seasonal rhythms) they provide a coherent and epidemiologically plausible explanation for the timing and intensity of national surges.

Vaccination dynamics, by contrast, were highly synchronised across municipalities. FPCA revealed that nearly all variation in vaccination curves was driven by a single temporal component representing the national rollout schedule. Apparent positive correlations between vaccination and COVID-19 cases at the aggregate level were largely attributable to temporal confounding, as vaccination was absent during the first epidemic wave and intensified during later surges. Once the common temporal structure was removed through functional modelling, vaccination contributed little unique explanatory value at the municipality level. This highlights the importance of avoiding naive ecological interpretations when processes share similar national timing.

Media sentiment, derived from supervised classification of more than 33,000 headlines, displayed more subtle but meaningful patterns. Sentiment, particularly in one of the commercial outlets Delfi, tended to deteriorate several weeks before observed increases in cases. Because sentiment reflects the information environment, including the intensity of media attention, its anticipatory behaviour likely captures early reporting and heightened public concern rather than behavioural changes that directly affect transmission. Permutation-based significance testing confirmed that these lead–lag patterns were unlikely to arise from chance alignment. Although sentiment cannot explain differences between municipalities, it appears to function as an early contextual signal of worsening epidemic conditions at the national level.

The function-on-function regression model integrated weather, vaccination and sentiment into a single predictive framework. After accounting for shared temporal structure, the model showed that cold temperatures and low absolute humidity were consistently associated with higher number of cases, and that sentiment contributed modest additional explanatory power by reflecting early shifts in the information landscape. Vaccination effects were comparatively small once national timing effects were attributed into the smooth functional components. Overall, the model captured over 94% of the functional variation in municipal case curves, indicating that the main temporal drivers of the epidemic were successfully represented.

Taken together, these findings suggest that Lithuania's COVID-19 trajectory was governed by an interplay of seasonal environmental patterns, centrally coordinated national interventions, and media-driven informational cycles. Local municipal differences played a secondary role compared with these broad national trends. The results underscore the importance of integrating meteorological indicators into epidemic surveillance systems, as they offer meaningful advance warning of periods conducive to transmission. Media sentiment, while indirect, may provide an additional early contextual signal that reflects emerging awareness of deteriorating conditions before they appear in official case counts.

Several limitations should be also acknowledged that have resulted out of this work. First, although functional methods capture dominant temporal structure effectively, they are less suited to isolating short-term causal effects or abrupt policy interventions. Second, the analysis relied on nationally aggregated media sentiment, which limits its ability to explain municipal-level variation. Third, vaccination effects were evaluated at the ecological level and may not reflect individual-level protection or heterogeneous uptake within municipalities. Finally, sentiment was derived from translated Lithuanian headlines using lexicon-based tools, which may introduce measurement error and attenuate nuanced emotional signals. These limitations do not undermine the core findings but clarify the scale and mechanisms to which the conclusions apply.

Future researches could extend this work by incorporating mobility data, socioeconomic factors, or local policy measures, which may capture spatial heterogeneity more directly than vaccination or sentiment. Improvements in sentiment analysis, especially models trained on Lithuanian-language text, may further refine the measurement of the information environment. The methodological framework presented here (integrating FPCA, reduced and lagged FCCA, permutation-based inference, and function-on-function regression) could be applied to other contexts, including state-level studies in countries like the United States, where regional contrasts in weather, media tone, vaccination uptake, and epidemic patterns are substantial. By bringing these functional tools together, the thesis highlights how FDA can offer a structured and interpretable way to analyse epidemics as systems of evolving, interdependent temporal processes.

References

[1] O. A. Alduchov, R. E. Eskridge. “Improved Magnus form approximation of saturation vapor pressure.” In: *Journal of Applied Meteorology* 35.4 (1996), pages 601–609. `https://doi.org/10.1175/1520-0450(1996)035<0601:IMFAOS>2.0.CO;2`.

[2] M. Araújo, A. Pereira, F. Benevenuto. “A comparative study of machine translation for multilingual sentence-level sentiment analysis.” In: *Information Sciences* 512 (2020), pages 1078–1102. `https://doi.org/10.1016/j.ins.2019.10.031`.

[3] V. Arya, A. K. Mishra, A. González-Briones. “Sentiments analysis of COVID-19 vaccine tweets using machine learning and VADER lexicon method.” In: *ADCAIJ* 11.4 (2023), pages 507–518. `https://doi.org/10.14201/adcaij.27349`.

[4] A. Balahur, M. Turchi. “Multilingual sentiment analysis using machine translation?” In: *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. 2012, pages 52–60. URL: `https://aclanthology.org/W12-3709/`.

[5] A. Barhoumi, C. Aloulou, N. Camelin, Y. Estève, L. Belguith. “Arabic sentiment analysis: An empirical study of machine translation’s impact.” In: *Language Processing and Knowledge Management International Conference (LPKM 2018)*. 2018. URL: `https://hal.science/hal-02042313`.

[6] A. Bari, M. Heymann, R. J. Cohen, R. Zhao, L. Szabo, S. Apas Vasandani, A. Khubchandani, M. DiLorenzo, M. Coffee. “Exploring coronavirus disease 2019 vaccine hesitancy on twitter using sentiment analysis and natural language processing algorithms.” In: *Clinical Infectious Diseases* 74 (2022), e4–e9. `https://doi.org/10.1093/cid/ciac141`.

[7] A. Bauer, F. Scheipl, H. Küchenhoff, A.-A. Gabriel. *Modeling spatio-temporal earthquake dynamics using generalized functional additive regression*. 2017. `https://doi.org/10.13140/RG.2.2.18420.04485`.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] T. Boschi, J. Di Iorio, L. Testa, et al. “Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy.” In: *Scientific Reports* 11 (2021), page 17054. `https://doi.org/10.1038/s41598-021-95866-y`.

[10] P. Castioni, S. Gómez, C. Granell, et al. “Rebound in epidemic control: How misaligned vaccination timing amplifies infection peaks.” In: *npj Complexity* 1 (2024), page 20. `https://doi.org/10.1038/s44260-024-00020-0`.

[11] H. Y. Chan, K. K. C. Cheung, S. Erduran. “Science communication in the media and human mobility during the COVID-19 pandemic: a time series and content analysis.” In: *Public Health* 218 (2023), pages 106–113. `https://doi.org/10.1016/j.puhe.2023.03.001`.

[12] S. Chen, K. Prettner, M. Kuhn, et al. “Climate and the spread of COVID-19.” In: *Scientific Reports* 11 (2021), page 9042. `https://doi.org/10.1038/s41598-021-87692-z`.

[13] J.-M. Chiou, H.-G. Müller, J.-L. Wang. "Functional response models." In: *Biometrika* 91.3 (2004), pages 605–619. URL: https://api.semanticscholar.org/CorpusID:19785308.

[14] C. A. Christophi, M. Sotos-Prieto, et al. "Ambient temperature and subsequent COVID-19 mortality in OECD countries and individual United States." In: *Scientific Reports* 11.1 (2021), page 8710. https://doi.org/10.1038/s41598-021-87803-w.

[15] A. Córdoba-Cabús, M. García-Borrego, Y. Ceballos. "Sentiment analysis toward COVID-19 vaccine in Latin American media on Twitter: the cases of Argentina, Chile, Colombia, Mexico, and Peru." In: *Vaccines* 11.10 (2023), page 1592. https://doi.org/10.3390/vaccines11101592.

[16] P. Craven, G. Wahba. "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation." In: *Numerische Mathematik* 31 (1979), pages 377–403. https://doi.org/10.1007/BF01404567.

[17] D. Cucinotta, M. Vanelli. "WHO declares COVID-19 a pandemic." In: *Acta Biomedica* 91.1 (2020), pages 157–160. https://doi.org/10.23750/abm.v91i1.9397.

[18] F. D'Amico, M. Marmiere, et al. "COVID-19 seasonality in temperate countries." In: *Environmental Research* 206 (2022), page 112614. https://doi.org/10.1016/j.envres.2021.112614.

[19] P. D. Davis, G. D. Parbrook, G. N. C. Kenny. "Humidification." In: *Basic Physics and Measurement in Anaesthesia*. 4th edition. Butterworth-Heinemann, 1995, pages 146–157. https://doi.org/10.1016/B978-0-7506-1713-0.50017-2.

[20] H. Du, S. Saiyed, L. Gardner. "Association between vaccination rates and COVID-19 outcomes in the United States: a population-level statistical analysis." In: *BMC Public Health* 24 (2024), page 220. https://doi.org/10.1186/s12889-024-17790-w.

[21] Google Cloud. *Google Cloud Translation API Documentation*. Accessed: 2025-01-10. 2024. URL: https://cloud.google.com/translate/docs/reference/rest.

[22] R. J. Hyndman, G. Athanasopoulos. *Forecasting: Principles and Practice*. 3rd edition. https://otexts.com/fpp3/. Melbourne: OTexts, 2021.

[23] L. Horváth, P. Kokoszka. *Inference for Functional Data with Applications*. New York: Springer, 2012. https://doi.org/10.1007/978-1-4614-3655-3.

[24] F. Yao, H.-G. Müller, J.-L. Wang. "Functional canonical analysis for sparse longitudinal data." In: *Journal of the American Statistical Association* 100.470 (2005), pages 577–590. https://doi.org/https://doi.org/10.1198/016214504000001745.

[25] *ISO 8601-1:2019 — Date and time — Representations for information interchange — Part 1: Basic rules*. Standard. Geneva: International Organization for Standardization, 2019.

[26] A. E. Ivanescu, A.-M. Staicu, F. Scheipl, S. Greven. "Penalized function-on-function regression." In: *Journal of Computational and Graphical Statistics* 24.2 (2015), pages 494–513. https://doi.org/10.1007/s00180-014-0548-4.

[27] D. Jurafsky, J. H. Martin. *Speech and Language Processing (3rd ed.)* 2023. URL: https://web.stanford.edu/~jurafsky/slp3/.

[28] P. Kokoszka, M. Reimherr. *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, 2017. https://doi.org/https://doi.org/10.1201/9781315117416.

[29] S. E. Leurgans, R. A. Moyeed, B. W. Silverman. "Canonical correlation analysis when the data are curves." In: *Journal of the Royal Statistical Society: Series B* 55.3 (1993), pages 725–740.

[30] Lithuanian Government Open Data Portal. *COVID-19 Cases and Deaths by Municipality (Atvejai ir Mirtys)*. https://data.gov.lt/datasets/1670/resource/9641/AtvejaiIrMirtys/csv. Data covering years 2020–2022. 2020.

[31] Lithuanian Hydrometeorological Service. *Meteo API – Operational meteorological observations*. 2022. URL: https://api.meteo.lt/.

[32] B. Liu. *Sentiment Analysis and Opinion Mining*. Springer Cham, 2012. https://doi.org/10.1007/978-3-031-02145-9.

[33] Y. Liu, S. R. Procter, et al. "Assessing the impacts of COVID-19 vaccination programme timing and speed on health benefits, cost-effectiveness, and relative affordability in 27 African countries." In: *BMC Medicine* 21 (2023), page 85. https://doi.org/10.1186/s12916-023-02784-z.

[34] LTData Open Data Portal. *COVID-19 Vaccination Dataset (COVID19-vakcinavimas)*. https://atviri-duomenys-ltdata.hub.arcgis.com/datasets/LTdata::covid19-vakcinavimas/explore. Data covering years 2020–2022. 2021.

[35] LTData Open Data Portal. *COVID-19 Vaccination Dataset (COVID19-vakcinavimas)*. 2022. URL: https://atviri-duomenys-ltdata.hub.arcgis.com/datasets/LTdata::covid19-vakcinavimas/explore.

[36] G. Mestre, J. Portela, G. Rice, A. Muñoz San Roque, E. Alonso. "Functional time series model identification and diagnosis by means of auto- and partial autocorrelation analysis." In: *Computational Statistics & Data Analysis* 155 (2021), page 107108. ISSN: 0167-9473. https://doi.org/10.1016/j.csda.2020.107108.

[37] M. Moriyama, W. J. Hugentobler, A. Iwasaki. "Seasonality of respiratory viral infections." In: *Annual Review of Virology* 7.1 (2020), pages 83–101. https://doi.org/https://doi.org/10.1146/annurev-virology-012420-022445.

[38] L. Nottmeyer, B. Armstrong, et al. "The association of COVID-19 incidence with temperature, humidity, and UV radiation – A global multi-city analysis." In: *Science of The Total Environment* 854 (2023), page 158636. https://doi.org/10.1016/j.scitotenv.2022.158636.

[39] OpenAI. *ChatGPT*. Version GPT-4.1, accessed free of charge, Internet access disabled. 2025. URL: https://chat.openai.com.

[40] V. M. Panaretos, S. Tavakoli. "Cramér–Karhunen–Loève representation and harmonic principal component analysis of functional time series." In: *Stochastic Processes and their Applications* 123.7 (2013), pages 2779–2807. https://doi.org/10.1016/j.spa.2013.03.015.

[41] A. Peci, A.-L. Winter, Y. Li, et al. “Effects of absolute humidity, relative humidity, temperature, and wind speed on influenza activity in Toronto, Ontario, Canada.” In: *Applied and Environmental Microbiology* 85 (2019), e02426–18. https://doi.org/10.1128/AEM.02426-18.

[42] J. O. Ramsay, G. Hooker, S. Graves. *Functional Data Analysis with R and MATLAB*. Springer, 2009. https://doi.org/10.1007/978-0-387-98185-7.

[43] J. O. Ramsay, B. W. Silverman. *Functional Data Analysis*. 2nd edition. Springer, 2005. https://doi.org/10.1007/b98888.

[44] M. Ribeiro, L. Azevedo, et al. “Understanding spatiotemporal patterns of COVID-19 incidence in Portugal: a functional data analysis from August 2020 to March 2022.” In: *PLOS ONE* 19.2 (2024), e0297772. https://doi.org/10.1371/journal.pone.0297772.

[45] F. Scheipl, A.-M. Staicu, S. Greven. “Functional additive mixed models.” In: *Journal of Computational and Graphical Statistics* 24.2 (2015), pages 477–501.

[46] J. Shaman, M. Kohn. “Absolute humidity modulates influenza survival, transmission, and seasonality.” In: *PNAS* 106.9 (2009), pages 3243–3248. https://doi.org/10.1073/pnas.0806852106.

[47] M. Siljander, R. Uusitalo, P. Pellikka, et al. “Spatiotemporal clustering patterns and sociodemographic determinants of COVID-19 (SARS-CoV-2) infections in Helsinki, Finland.” In: *Spatial and Spatio-Temporal Epidemiology* 41 (2022), page 100493. https://doi.org/10.1016/j.sste.2022.100493.

[48] A. Suthar, J. Wang, et al. “Public health impact of COVID-19 vaccines in the US: observational study.” In: *BMJ* 377 (2022), e069317. https://doi.org/10.1136/bmj-2021-069317.

[49] C. Tang, T. Wang, P. Zhang. “Functional data analysis: an application to COVID-19 data in the United States in 2020.” In: *Quantitative Biology* 10 (2022), pages 172–187. https://doi.org/10.15302/J-QB-022-0300.

[50] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J. D. Farmer. “Testing for nonlinearity in time series: the method of surrogate data.” In: *Physica D: Nonlinear Phenomena* 58.1-4 (1992), pages 77–94. URL: https://scispace.com/pdf/testing-for-nonlinearity-in-time-series-the-method-of-3obg021lq4.pdf.

[51] M. P. Walkowiak, J. B. Walkowiak, D. Walkowiak. “COVID-19 passport as a factor determining the success of national vaccination campaigns: does it work? The case of Lithuania vs. Poland.” In: *Vaccines* 9.12 (2021), page 1498. https://doi.org/10.3390/vaccines9121498.

[52] J. Wang, Y. Fan, et al. “Global evidence of expressed sentiment alterations during the COVID-19 pandemic.” In: *Nature Human Behaviour* 6 (2022), pages 349–358. https://doi.org/10.1038/s41562-022-01312-y.

[53] O. J. Watson, G. Barnsley, et al. “Global impact of the first year of COVID-19 vaccination: a mathematical modelling study.” In: *The Lancet Infectious Diseases* 22.9 (2022), pages 1293–1302. https://doi.org/10.1016/S1473-3099(22)00320-6.

[54] H. Wickham. “Tidy data.” In: *Journal of Statistical Software* 59.10 (2014), pages 1–23. https://doi.org/10.18637/jss.v059.i10.

[55] S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, 2006. https://doi.org/10.1201/9781315370279.

[56] World Health Organization. *WHO coronavirus (COVID-19) dashboard*. 2025. URL: https://data.who.int/dashboards/covid19/cases.

Appendix 1. Use of Artificial Intelligence Tools

Artificial intelligence tools were used for language editing purposes only, including improving grammar, sentence structure, and overall readability of the text. Specifically, the generative language model ChatGPT [39] was used to revise selected passages of text based on author-provided drafts. The prompts consisted of requests to improve clarity, grammar, and academic style while preserving the original meaning , and all AI-assisted revisions were subsequently reviewed and approved by the author. No artificial intelligence tools were used to generate scientific content, perform data analysis, derive mathematical results, or interpret any findings. The tool was accessed free of charge in 2025, with Internet access disabled.

Appendix 2. R code for Functional Data Analysis

The computational analyses were implemented in R. For transparency, the final analysis script is available in a public repository: `https://github.com/capemigle/Master-thesis`