Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Forecasting influenza-like illness in Italy using Wikipedia: a principal components regression approach[1]

Gianluca Mura,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Forecasting influenza-like illness in Italy using Wikipedia: a principal components regression approach

Gianluca Mura[1]

Abstract

In recent years, there has been growing academic interest in leveraging Wikipedia data to understand the epidemiology of infectious diseases. Studies on Influenza-Like Illness (ILI) have shown that Wikipedia can serve as a valid surveillance system, addressing gaps in traditional surveillance methods. Influenza, being a highly prevalent and impactful disease, presents an opportunity to explore the potential of using Wikipedia data for accurate and timely predictions. In this paper, I propose a novel approach to predict Influenza-Like Illness rates in Italy by analysing the number of consultations of a selected set of Italian language Wikipedia pages. I employ Principal Components Regression (PCR) as my predictive model, utilizing the reduced-dimensional representations of the Wikipedia data. The performance of the model is evaluated by comparing the predicted rates with the official data published by InfluNet, the Italian epidemiological influenza surveillance system, for four consecutive influenza seasons, spanning from 2015 to 2019. This study contributes to the growing body of literature on utilizing alternative data sources for disease surveillance and highlights the potential of machine learning models for real-time forecasting in public health emergencies.

Keywords: Influenza-Like Illness (ILI), Wikipedia Page Consultations, Machine Learning, Principal Components Regression, Surveillance System.

JEL classification: C38, C53, I18

[1] Bank of Italy (Gianluca.Mura@bancaditalia.it). The views and opinions expressed in this paper are those of the author and do not necessarily reflect the position of Bank of Italy.

1. Introduction

Seasonal influenza remains a significant public health challenge, causing substantial morbidity and mortality worldwide, with an estimated 290,000 to 650,000 respiratory deaths annually (De Toni et al., 2021). Beyond its health implications, influenza leads to economic losses through work absenteeism and decreased productivity, with up to 88% of the economic burden attributable to indirect costs (de Courville et al., 2022). Traditional surveillance systems, such as sentinel physician networks and laboratory confirmations, play a crucial role in monitoring influenza outbreaks but are limited by reporting delays of up to two weeks and are expensive (Simonsen et al., 2016; Gianfredi et al., 2021). In recent years, the emergence of digital epidemiology has introduced novel data streams with potential to supplement traditional surveillance methods by capturing health-seeking behaviours in the population through near real-time signals of disease activity from internet-based data sources. (McClymont et al., 2024). The recent COVID-19 pandemic has further demonstrated the critical importance of timely disease surveillance and forecasting capabilities. While traditional surveillance systems were overwhelmed during the pandemic, digital surveillance methods gained prominence, offering potential complementary tools for tracking disease spread (Santangelo et al., 2022).

Against this background, Wikipedia, as a widely used online resource for health information, represents a particularly valuable data source for tracking influenza trends, since its access logs contain information about public interest in health topics across different languages and geographic regions (Smith, 2020). Moreover, prior research has demonstrated strong correlations between Wikipedia page views and official influenza incidence data, positioning it as a promising resource for timely and accurate disease surveillance and forecasting (McIver & Brownstein, 2014; Generous et al., 2014; Zimmer et al., 2018). However, effective forecasting systems must account for inherent challenges in surveillance data, including data instability, surveillance coverage variations, and media-influenced attention spikes that can skew internet-based signal detection (Chakraborty et al., 2018). To overcome these problems, internet-based surveillance systems should ideally complement rather than replace traditional methods, creating hybrid approaches that leverage the strengths of each data source (Hammond et al., 2022).

In Italy, RespiVirNet, coordinated by the Istituto Superiore di Sanità, is the integrated national surveillance system for influenza-like illnesses and respiratory viruses, combining epidemiological and virological monitoring to enable timely vaccine updates and identify circulating viruses such as influenza. Traditional surveillance systems of this kind, as previously highlighted, while extremely valuable, inherently suffer from delays associated with data collection and reporting. To address this limitation, this study investigates whether Italian Wikipedia access logs can effectively predict influenza-like illness (ILI) rates in Italy.

Specifically, I propose a forecasting model based on Principal Components Regression (PCR) that utilizes page-view data from relevant Italian-language Wikipedia articles, which provides methodological advantages by addressing the multicollinearity inherent in such data while preserving the variance structure critical for predictive accuracy. By employing dimension reduction techniques, PCR enables efficient handling of high-dimensional data without sacrificing explanatory power.

The model's performance is assessed against official data from InfluNet, Italy's epidemiological influenza surveillance system, across four consecutive influenza seasons (2015-2019). Ultimately, this research – while being aware of the limitations inherent in this type of data – aims to contribute to the growing field of digital disease surveillance (Santangelo et al., 2022) by showing the practical utility of Wikipedia data for enhancing real-time monitoring capabilities in public health systems and exploring the potential integration of these alternative data streams into existing surveillance frameworks.

2. Essential review of related literature

Several studies have investigated the possibility of exploiting Wikipedia access log data to monitor and predict rates of influenza syndromes and other infectious diseases. In a work that can be considered pioneering, Generous et al. (2014) conducted a comprehensive evaluation of Wikipedia as a global disease monitoring tool, examining the correlation between article access logs and official disease incidence data across multiple diseases and countries. Their findings revealed that Wikipedia data could successfully estimate influenza-like illness (ILI) levels in certain contexts, while acknowledging limitations in others.

Contemporary and related work is that of McIver and Brownstein (2014), who developed a Poisson model using selected Wikipedia articles to estimate ILI prevalence in the United States. Their approach demonstrated that Wikipedia usage could provide near real-time estimates of ILI activity with accuracy, outperforming Google Flu Trends particularly during abnormal influenza seasons. They identified specific health-related articles (including "Influenza," "Fever," and "Oseltamivir") that consistently correlated with official CDC data.

Hickmann et al. (2015) combined Wikipedia access data with a mathematical epidemiological model and data assimilation methods. Their research, specifically focused on forecasting the 2013-2014 influenza season in the United States, highlighted the value of combining Wikipedia data with traditional epidemiological models (i.e. integration of mechanistic disease modelling with digital surveillance), achieving predictive accuracy for both the timing and magnitude of peak influenza activity during the study period.

Priedhorsky et al. (2017) implemented a systematic article selection process based on semantic relatedness showing how filtering inputs through Wikipedia's category links significantly improved prediction quality. However, they also identified limitations in forecasting beyond the immediate present, finding "very little forecasting value, even just one time step ahead". This study underscores the importance of precise article selection and the limited predictive horizon of purely observational data.

Moreover, few studies have compared Wikipedia's forecasting capabilities with other data sources. Sharpe et al. (2016) conducted a comparative analysis of Google, Twitter, and Wikipedia using Bayesian change point analysis, reporting varying levels of sensitivity and positive predictive value (PPV) across platforms.

De Toni et al. (2021) propose a language-agnostic approach for using Wikipedia to monitor influenza-like illness across multiple European countries. Their method employs two algorithms, Personalized PageRank and CycleRank, to automatically select relevant Wikipedia pages without requiring expert input. Their results showed strong correlations between Wikipedia data and official epidemiological records across the studied countries, demonstrating the potential of Wikipedia as an effective ILI surveillance tool in diverse linguistic contexts.

Finally, with reference to Wikipedia pages in Italian, Gianfredi et al. (2021) conducted a cross-sectional study examining the correlation between Wikipedia page views and official influenza data in Italy between October 2015 and May 2019. Using Pearson correlation coefficients, they identified strong temporal correlations between Wikipedia search trends and official ISS influenza bulletins. These findings indicate that Wikipedia access patterns mirror actual influenza incidence in Italy, suggesting potential value as a complementary real-time surveillance tool that could address the delay in traditional systems.

3. Data and methods

Epidemiological data

As noted earlier, the RespiVirNet (formerly InfluNet) surveillance system monitors influenza-like illnesses (ILI) during the seasonal period through a network of sentinel physicians. The system relies on a network of participating general practitioners (GPs) and paediatricians (PLSs) and is coordinated by the Istituto Superiore di Sanità (ISS). Epidemiological surveillance is conducted via weekly reports of ILI cases from a sample of GPs and PLSs during the period from October to April.

To conduct this analysis, the figures on the reference week and the weekly incidence data, expressed as the number of influenza syndromes (cases) per 1,000 patients (total incidence), were extrapolated. Each influenza season, and thus the relevant sample survey, has an average duration of 28 weeks, beginning on the 42nd week of one year and ending on the 17th of the following year. The decision to limit this analysis to four influenza seasons (equivalent 112 weekly observations) prior to the COVID-19 pandemic (2015-2019) was based on methodological considerations regarding data stability. Including pandemic-era data might have introduced significant complications to the analysis, as information-seeking behaviours likely changed dramatically during this period (McClymont et al., 2024). Research suggests that the pandemic created an unprecedented "information crisis" that altered online engagement patterns. In fact, during early 2020, Wikipedia experienced extraordinary spikes in page creation and viewership related to coronavirus content, potentially distorting typical seasonal patterns of influenza-related information seeking. The pandemic appears to have shifted public attention toward specific topics like epidemiology and public health, creating anomalous conditions that would require specialized analytical approaches (Colavizza, 2020). By focusing on pre-pandemic seasons, this study may provide a more stable baseline for understanding the relationship between Wikipedia access data and influenza trends under more typical circumstances, though future research could explore how these patterns evolved during the pandemic period.

Influenza seasons considered in the analysis

Table 1

Influenza Seasons	From	To	No. of weeks
2015-2016	2015-42	2016-16	28
2016-2017	2016-42	2017-17	28
2017-2018	2017-42	2018-17	28
2018-2019	2018-42	2018-17	28

[1] Note: the year 2015 ended with the 53rd week, the others with the 52nd

Source: author's elaboration

Wikipedia data

Wikipedia provides various tools and APIs to analyze its data, including access logs, page content, hyperlinks, and editor activity. There are also numerous free third-party solutions and open-source projects have been developed by the community, enhancing the functionality and accessibility of Wikipedia data. I utilized data from Meta-Wiki, specifically from the Pageviews Analysis tool, which provides detailed statistics on Wikipedia page access. This tool aggregates user visits to Wikipedia articles, allowing for the analysis of search trends over time: information on daily accesses as of 1 July 2015 was identified by selecting all types of platforms (desktop, mobile and mobile application) and discarding accesses generated by crawlers and other automated programs). The number of accesses was then aggregated by ISO week and purged of irrelevant observations because they were not included in the annual flu season survey. Data preprocessing included the identification and treatment of extreme values in the Wikipedia page view time series. Examination of the dataset revealed few abnormal values that fell substantially outside expected ranges for typical page view patterns. Specifically, four outliers were identified and treated: two adjacent values for the 'Epidemia' page during the first influenza season (2015-2016), and two individual values for the 'Influenza suina' page occurring during the third (2017-2018) and fourth (2018-2019) influenza seasons respectively. To maintain the integrity of the time series while minimizing distortion, these outliers were replaced with imputed values calculated as the arithmetic mean of the immediately preceding and following observations. This approach preserved the underlying temporal trends while preventing extreme values from disproportionately influencing the principal component structure and subsequent regression analysis.

Regarding the selection of the most relevant Wikipedia articles, several approaches are viable, which involve analyzing their relationships based on hyperlinks, categories, content similarity, or a combination of these factors. For this study, the selected pages were manually picked based on thematic relevance, with publications and prior studies on influence rate prediction reviewed to guide the selection of topics most likely to contribute to accurate modeling. After this selection, distance metrics – in terms of shortest path distance between "Influenza" article and the other articles – were applied as methods of verification and confirmation, ensuring the dataset represented a cohesive and meaningful subset of Wikipedia content. At the end of this process, 14 articles were selected. Finally, as previous studies have shown (e.g. Generous et al., 2014), since geo-localized data are not freely available, the language of Wikipedia articles was used as a proxy to delimit the target country. Given the limited diffusion of the Italian language abroad, it is the author's opinion that this does not have any relevant consequences for the purposes of this study.

Selected Wikipedia articles

Table 2

Article in Italian	Corresponding article in English	Total number of views over the study period (4 flu seasons)	No. of weeks
Influenza	Influenza	227.775	-
Febbre	Fever	446.206	1
Cefalea	Headache	230.886	1
Influenza aviaria	Avian influenza	91.820	1
Influenza suina	Swine influenza	49.771	1
Epidemia	Epidemic	62.671	2
Pandemia	Pandemic	96.235	2
Pandemia influenzale	Influenza pandemic	21.314	1
Raffreddore comune	Common cold	225.309	1
Paracetamolo	Paracetamol	806.148	1
Raffreddore	Cold	12.419	2
Virus parainfluenzali umani	Human parainfluenza viruses	11.355	1
Antivirale	Antiviral drug	45.896	1
Influenza virus A sottotipo H1N1	Influenza A virus subtype H1N1	93.237	1

Note: Shortest Path Distance is defined as the minimum number of hyperlink clicks (edges) required to navigate from one Wikipedia page (node) to another through a directed network of internal links

Source: author's elaboration based on pageviews.toolforge.org/

4. Methodology

Overview

This methodology combines dimension reduction via Principal Component Analysis (PCA) with linear regression through a multi-step process. Before briefly mentioning the algebraic formulation of PCR, it is useful to recall the pros and cons of PCA with reference to the objective of this study. When the goal is timely influenza surveillance, this technique offers several complementary advantages.

First, it enhances signal extraction: weekly access counts inevitably include some noise, i.e. spikes in page views may mirror media attention or automated activity rather than genuine changes in health-related concern. Because PCA performs a mathematical transformation of the original variables into new orthogonal (uncorrelated) variables that maximize explained variance, it effectively consolidates what multiple pages have in common while eliminating redundancies. As a result, a small set of components can dampen such random fluctuations while preserving the persistent temporal pattern that tends to accompany an epidemic wave. Second, PCA contributes to statistical stability: the raw page-view series suffer from some degree of collinearity, a circumstance that inflates variance and undermines predictive precision. Regressing influenza-like-illness (ILI) rates on the uncorrelated component scores alleviates this difficulty and produces more reliable coefficients even when the number of original predictors exceeds the number of weekly observations. Third, given that InfluNet provides only about thirty weekly data points per season, using just the leading principal components rather than individual Wikipedia pages helps prevent over-fitting to limited data.

This approach promotes statistical parsimony while still incorporating information from all Wikipedia pages in the initial analysis before dimension reduction. This lean structure mirrors the "sloppiness" seen in complex biological and physical models: just a few "stiff" parameter combinations drive the predictions, while all other directions have negligible influence (Transtrum et al., 2015). Acknowledging this fact supports the choice to analyse principal-component signals instead of individual page-view coefficients.

On the other hand, these strengths, coexist with limitations that deserve explicit acknowledgment. PCA orders components strictly by the variance they explain (descending order), yet components accounting for only a modest fraction of total variance can still hold valuable predictive information, excluding them solely on variance grounds can therefore hinder forecast accuracy (Hadi et al., 2012). Moreover, a single component may blend traffic to biomedical articles with visits triggered by unrelated popular events, complicating substantive interpretation. Such peculiarities do not negate the method's effectiveness, but they remind that PCA is an exploratory method and therefore translating patterns of online behaviour into actionable public-health signals still requires careful judgement.

Principal component regression

Principal Component Regression provides an effective approach for addressing multicollinearity Wikipedia page view data while preserving the variance structure essential for accurate influenza prediction. In this section, I provide a concise overview of the key mathematical steps involved in performing principal component regression. For consistency and clarity, all mathematical notations follow standard linear algebra conventions, though alternative representations exist.

To model influenza incidence rates, as a first step, the covariates – daily view counts for 14 influenza-related Wikipedia pages – were standardized to zero mean and unit variance employing $\boldsymbol{Z} = (\boldsymbol{X} - \mu)\boldsymbol{D}^{-1/2}$, where $\mu \in R^{14}$ is the vector containing the means and **D** is the diagonal matrix of variances. This results in a standardized matrix $\boldsymbol{Z} \in R^{112\,x\,14}$ where each element z_{ij} represents the standardized page views for Wikipedia article j during week i. It should be noted that, while standardization is not an absolute requirement for PCA, I standardized all variables prior to my analysis to ensure that pages with higher overall consultation volumes would not disproportionately influence the component structure, thereby allowing each Wikipedia page an equal opportunity to contribute to the extracted patterns.

In the second step, it is performed the Principal Component Analysis on $\boldsymbol{Z}$, which can be efficiently computed using Singular Value Decomposition (SVD). To this end, let
$\boldsymbol{\Sigma} \in R^{14\,x\,14}$ be the empirical covariance matrix of the standardised predictors that can be computed as:

$$\boldsymbol{\Sigma} = \frac{1}{n-1}\boldsymbol{Z}^T\boldsymbol{Z}$$

Diagonalising the covariance matrix produces the spectral (eigen) decomposition, a fundamental transformation that separates the variance structure into its principal directions:

$$\boldsymbol{\Sigma} = \boldsymbol{V} \wedge \boldsymbol{V}^T$$

where $\boldsymbol{V}$ is the orthogonal matrix of eigenvectors (i.e. the loading directions), $\boldsymbol{V}^T$ is its transpose, and $\wedge$ is a diagonal matrix containing the singular values (square roots of eigenvalues). From this decomposition, is it possible to obtain the principal component scores:

$$\mathbf{W} = \mathbf{ZV}$$

where $\boldsymbol{W} \in R^{112\,x\,14}$ is the matrix of principal component scores. Each column $w_k$of $\boldsymbol{W}$ represents the $w-th$ principal component, with variance λ_k.

The final step involves regress the weekly influenza-like illness rates $y \in R^{112}$ on a subset of m principal components:

$$y = \boldsymbol{W}_m\,\gamma_m + \varepsilon$$

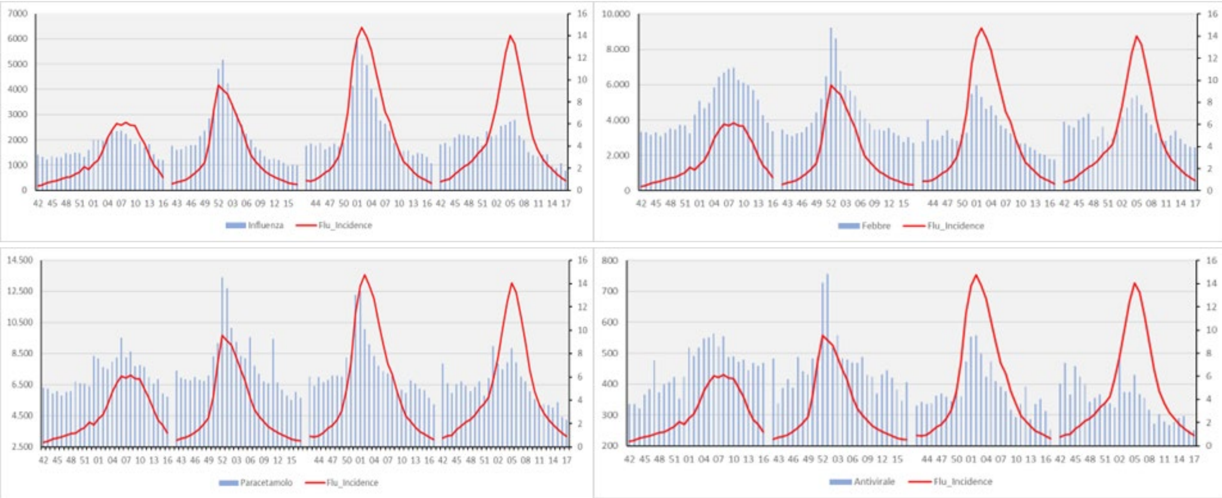
where $\boldsymbol{W}_m \in R^{112\,x\,m}$ contains the first m columns of $\boldsymbol{W}$ (representing the principal components that explain the most variance), $\gamma_m \in R^m$ is the vector of regression coefficients, and $\varepsilon \in R^{112}$ represents the error term.

5. Results and discussion

Exploratory data analysis

The exploration of temporal patterns in both Wikipedia page views and official ILI rates can reveal some insights into their relationship. Descriptive statistics indicated substantial variability in both datasets, with Wikipedia page views demonstrating cyclical patterns that often corresponded to official influenza seasons. Graph 1 illustrate the time-series comparison between access frequencies for four selected Wikipedia pages ("Influenza," "Febbre," "Paracetamolo," and "Antivirale") and official ILI rates over the four observed influenza seasons (2015-2019).

Access logs for selected Wikipedia pages plotted alongside official ILI rates Graph 1



Sources: ILI rates: respivirnet.iss.it/; Wikipedia: pageviews.toolforge.org/

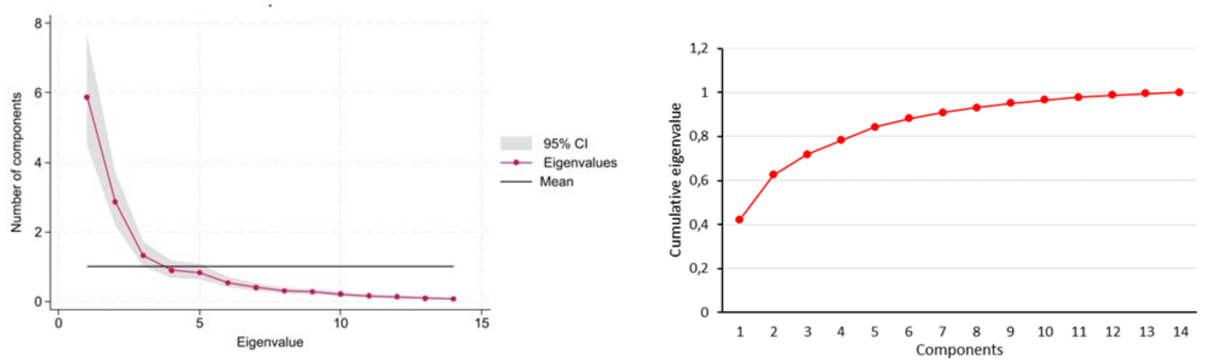
Visual inspection of these time-series plots suggested some degree of temporal alignment between Wikipedia page views and official ILI rates. Increases in page views frequently preceded or coincided with rises in reported influenza cases, indicating possible predictive utility. The "Influenza" page demonstrated the strongest correlation (r = 0.81), followed by "Paracetamolo" (r = 0.66), while "Febbre" and "Antivirale" showed moderate correlations (r = 0.59 and r = 0.39, respectively). These varying correlation strengths, with *p-values* confirming statistical significance, might suggest different information-seeking behaviours associated with disease progression, from initial symptom investigation to treatment options.

The comparative analysis across the four influenza seasons revealed significant epidemiological heterogeneity that presents notable forecasting challenges. The 2017-2018 season exhibited the most severe epidemic profile with an earlier start and substantially higher peak (14.74 cases per 1,000 patients) compared to other seasons. The 2018-2019 season showed a delayed peak that occurred in February rather than January. Meanwhile, the 2016-2017 season had a relatively moderate intensity but demonstrated the sharpest decline post-peak, while the 2015-2016 season exhibited a more prolonged epidemic period with multiple smaller peaks. Additionally, examining the magnitude of page view fluctuations suggested interesting patterns: the "Paracetamolo" page consistently received the highest absolute number of consultations, potentially reflecting its relevance beyond influenza cases, while the "Antivirale" page showed more pronounced relative increases during peak periods despite lower baseline consultation rates. These observations suggest that different Wikipedia pages may provide complementary signals that together enhance forecasting capabilities.

Principal component analysis results

After identifying the correlation patterns identified in the exploratory analysis, it was then proceeded with PCA. Graph 2 presents the scree plot of eigenvalues after PCA, which exhibits a characteristic elbow shape with a steep decline for the first three components, followed by a gradual levelling off thereafter. The first component demonstrated a substantially higher eigenvalue (approximately 5.9), indicating its dominance in explaining dataset variance.

Scree plot of eigenvalues (left) and explained cumulative variance (right) — Graph 2



Source: author's elaboration

Several established criteria exist for determining the optimal number of principal components to retain (Mooi et al., 2018): the Kaiser criterion, which suggests retaining components with eigenvalues greater than 1; the scree plot method, which identifies the "elbow point" in the plot of eigenvalues where additional components contribute minimally to explained variance; parallel analysis, which compares observed eigenvalues to those from randomly generated data; the cumulative percentage of variance explained, typically targeting 70-90%; and information criteria such as AIC and BIC, which balance model fit with complexity. Each approach offers unique advantages, often leading researchers to apply multiple criteria for increased confidence in their selection.

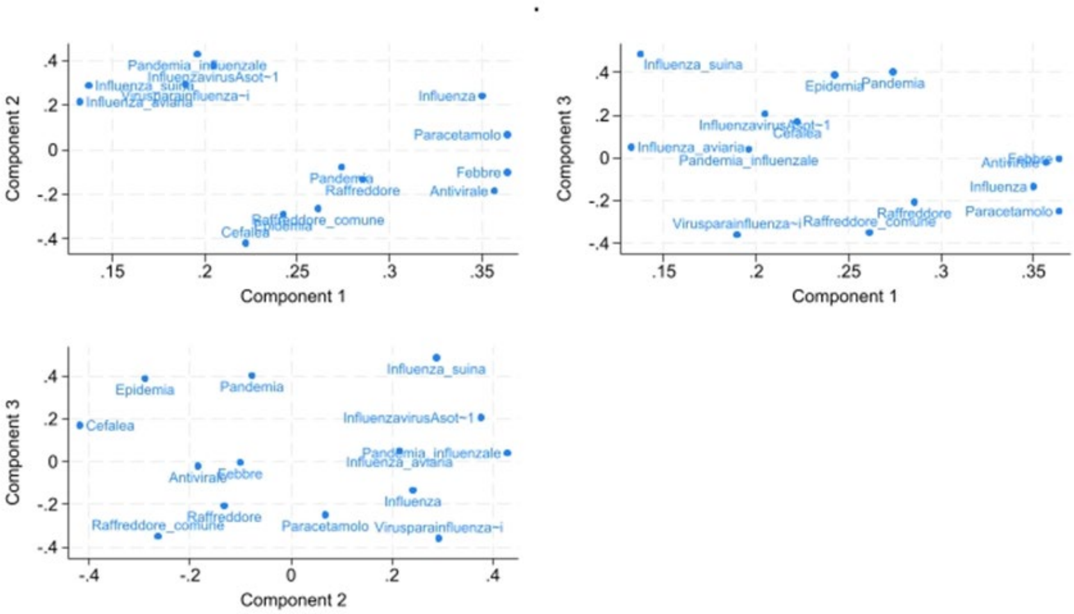
For this analysis, I primarily employed the Kaiser criterion, which suggested retaining three components, as the eigenvalues for the first three components were 5.9, 2.9, and 1.3, respectively, while the fourth component's eigenvalue fell below the threshold at 0.9. The cumulative percentage of variance explained supported this decision, with the first three components collectively accounting for about 72% of the total variance, a proportion that aligns with statistical best practices. The three-component solution thus balanced dimensionality reduction with information preservation, avoiding both under-specification and over-parameterization.

To better understand the underlying structure of the principal components, I utilized component visualization diagrams to examine the relationships between components. Graph 4 presents a three-panel graph displaying the pairwise relationships between the first three principal components, allowing for visual inspection of how the original Wikipedia variables contribute to each dimension.

These component projection graphs provide an intuitive representation of the multidimensional relationships, with variables positioned according to their correlation with each component pair. Distinct clustering patterns among the Wikipedia pages are showed, suggesting meaningful latent dimensions in users' information-seeking behaviour. Pages related to general influenza symptoms (such as "Influenza" and "Febbre") appear grouped differently from pandemic-specific pages (such as "Influenza virus A sottotipo H1N1" and "Pandemia influenzale"), visually confirming the differentiation of components by thematic content. This graphical representation supports a more detailed quantitative analysis of the component structure.

Component loadings

Graph 3



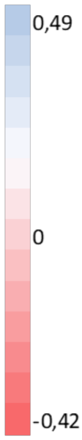
Source: author's elaboration

For this purpose, it is useful to check the exact measure of component loadings, which reveal distinct patterns of variable contributions to each principal component. The first principal component (PC1), explaining 42.1% of variance, exhibited strong positive loadings from "Influenza" (0.35), "Febbre" (0.36), "Paracetamolo" (0.36), and "Antivirale" (0.36), suggesting it represents general information-seeking about influenza symptoms and common treatments. This component captures the core influenza-related search behaviour that aligns with typical symptom investigation and treatment options. The second principal component (PC2), accounting for 20.0% of variance, showed distinct loading patterns with strong positive contributions from "Pandemia influenzale" (0.43), "Influenza virus A sottotipo H1N1" (0.38), and "Influenza_suina" (0.29), while displaying negative loadings for "Cefalea" (-0.42) and "Raffreddore comune" (-0.26). This component appears to differentiate between searches related to severe or pandemic influenza strains versus those focused on more common, less severe symptoms. It potentially reflects a dimension of concern about serious influenza outcomes as opposed to routine seasonal symptoms. The third principal component (PC3), explaining 10.0% of variance, demonstrated a more complex pattern with high positive loadings for "Influenza suina" (0.49), "Pandemia" (0.40), and "Epidemia" (0.39), contrasted with negative loadings for "Virus parainfluenzali umani" (-0.36), "Raffreddore comune" (-0.35), and "Paracetamolo" (-0.25). This component might represent a distinction between epidemic/pandemic-related information-seeking and searches for differential diagnosis or treatment of common respiratory conditions. It potentially captures public concern about wider disease spread versus management of individual symptoms.

Loadings of Wikipedia pages along the first three components

Graph 4

	PC1	PC2	PC3
Influenza	0,35	0,24	-0,13
Febbre	0,36	-0,10	0,00
Cefalea	0,22	-0,42	0,17
Influenza_aviaria	0,13	0,21	0,05
Influenza_suina	0,14	0,29	0,49
Epidemia	0,24	-0,29	0,39
Pandemia	0,27	-0,08	0,40
Pandemia_influenzale	0,20	0,43	0,04
Raffreddo_comune	0,26	-0,26	-0,35
Paracetamolo	0,36	0,07	-0,25
Raffreddore	0,29	-0,13	-0,21
Virus_parainfluenzali_umani	0,19	0,29	-0,36
Antivirale	0,36	-0,18	-0,02
Influenza_virus_A_sottotipo_H1N1	0,20	0,38	0,21



Source: author's elaboration

The adequacy of the dataset for PCA was confirmed through the Kaiser-Meyer-Olkin test, which returned an overall measure of 0.8308, exceeding the 0.80 threshold that indicates good sampling adequacy. Individual KMO values for variables ranged from 0.7040 ("Pandemia influenzale") to 0.9196 ("Febbre"), further validating the appropriateness of the PCA approach. This also suggests strong intercorrelations among variables that are effectively addressed through the dimensionality reduction provided by PCA.

These findings indicate that the principal components extracted represent meaningful latent dimensions of information-seeking behaviour related to influenza, with a good degree of interpretability in terms of the original Wikipedia variables. The meaningfulness of these dimensions is evidenced by their distinct loading patterns, which align with theoretically coherent aspects of health information-seeking: general symptom and treatment information (PC1), pandemic-specific concerns (PC2), and epidemic spread versus routine illness management (PC3). The KMO measure of sampling adequacy further validates that these components capture genuine underlying structures rather than artifacts of random variation (Mooi et al., 2018).

Principal components regression forecasting results

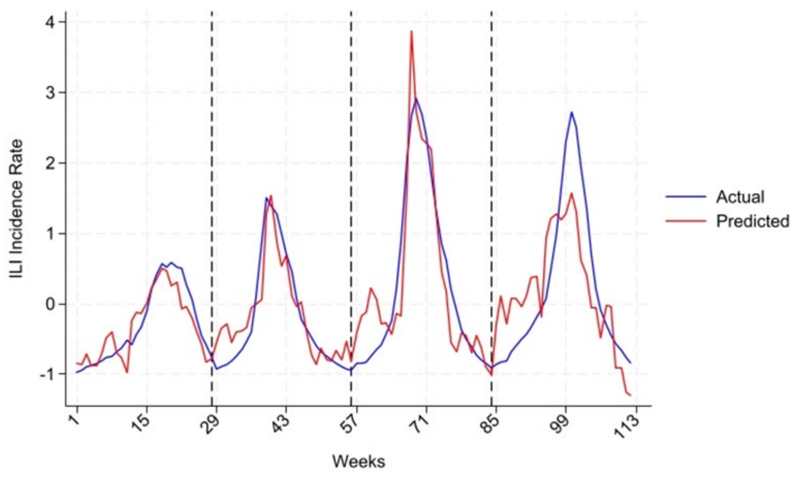
The retention of three components provided a parsimonious yet informative basis for the subsequent regression analysis, achieving an optimal balance between dimensionality reduction and information preservation. As mentioned above, this parsimony is valuable for predictive modelling because it reduces potential multicollinearity among the original variables while retaining a high percentage of the original variance, enough to capture the essential patterns in Wikipedia page views related to influenza trends without over-fitting to noise.

The graph illustrates the relationship between observed ILI rates and those predicted by the PCR over (non-consecutive) 112 weeks. The model demonstrates reasonable accuracy in capturing the cyclical nature of influenza activity, successfully identifying most seasonal peaks and troughs, particularly during the first two seasonal cycles.

The observed discrepancies between actual and predicted values highlight well known limitations inherent to Wikipedia access data as a surveillance source. In fact, while Wikipedia article views exhibit lower signal-to-noise ratios than general internet searches, they remain susceptible to external influences unrelated to disease prevalence. The model's overestimation around week 71, for instance, may reflect increased public interest driven by media coverage rather than actual disease activity. Conversely, the significant underestimation for the last period suggests that during unusually severe outbreaks, information-seeking behaviour on Wikipedia may not scale proportionally with disease prevalence.

These variations likely stem from complex factors affecting online information-seeking behaviour, including media attention that drives article views independently of illness rates, changes in public awareness or concern that don't directly correlate with epidemiological patterns, background seasonal variations in internet usage patterns, external events that may trigger interest in health-related content.

Actual versus predicted Flu incidence rates | Graph 5



Note: standardized values.

Source: author's elaboration

Despite these limitations, as shown in Table 5.1, the model demonstrates good explanatory power, with an R-squared value of 0.7723. While this level of explanatory power appears substantial for social science data and epidemiological modelling, it is important to interpret these results with appropriate caution, recognizing the inherent limitations of statistical models in capturing complex disease dynamics. The adjusted R-squared of 0.7660 is only marginally lower than the unadjusted value, which may indicate that the penalty for including three regressors is minimal. Statistical significance of the overall regression is excellent, confirming that the principal components collectively provide substantial explanatory power to the model; nevertheless, statistical significance alone does not guarantee practical significance or real-world applicability, and these results should be considered preliminary until further validation across additional influenza seasons and geographical contexts.

Overall goodness-of-fit of the regression

Table 3

Influenza Seasons	Value	Degrees of freedom/Units
R-squared (R^2)	0.7723	0.7723
Adjusted R^2	0.7660	0.7660
F-statistic	122.11	122.11
Prob > F (overall p-value)	< 0.0001	< 0.0001

Sources: TableNote

The prediction-error metrics provide support for the model's performance. With an RMSE of 0.4750 and an MAE of 0.3668 (both in standard deviation units), the model demonstrates reasonably strong predictive accuracy, with typical errors of approximately half a standard deviation. While encouraging, these error levels may still be too large for certain public health applications, particularly when precise estimates are required for resource allocation decisions. The RMSE represents about 12% of the total observed range of the standardized dependent variable, which, though relatively modest, still indicates a non-negligible level of uncertainty in the predictions.

Model performance

Table 4

Influenza Seasons	Value	Scale/Notes
RMSE	0.4750	Standard-deviation units (variables were z-scored)
MAE	0.3668	Sd units
RMSE / Range	12%	% of (max – min) of outcome

Source: author's elaboration

Ultimately, the model's ability to maintain this level of accuracy across different influenza seasons varies and would benefit from further evaluation across a broader range of epidemic scenarios. Despite its promise, the Wikipedia-based approach may still be susceptible to biases related to internet access disparities, changes in information-seeking behaviour over time, and platform-specific trends that could affect long-term reliability. When compared to other digital surveillance approaches, this Wikipedia-based PCR model shows potential advantages, but its relative effectiveness requires critical assessment considering systematic comparisons. Prior comparative evaluation of digital surveillance tools has been made (Sharpe et al., 2016). However, the rapid evolution of artificial intelligence capabilities and epidemic intelligence systems creates a pressing need for updated comparative assessments that account for technological advancements (MacIntyre et al., 2023). Claims of superiority over search query-based systems should therefore be considered tentative until more recent comparative analyses are conducted using identical evaluation criteria and time periods. In any case, models that draw on large-scale Internet data offer a promising complement to conventional surveillance, yet they should be treated as supplementary decision-support tools rather than outright replacements for established epidemiological methods (McIver et al, 2014)

6. Conclusions

This study provides evidence that Wikipedia access logs offer valuable data for enhancing influenza surveillance systems in Italy. The findings demonstrate that using Principal Components Regression on Italian-language Wikipedia page views produces reasonably accurate predictions of influenza-like illness rates when compared with official InfluNet data. The model's R-squared value of 0.7723 indicates substantial explanatory power, consistent with previous research showing strong correlations between Wikipedia data and official epidemiological records (De Toni et al., 2021; Gianfredi et al., 2021).

The methodological approach using PCR proved effective in addressing multicollinearity concerns while preserving the variance structure essential for prediction. By extracting three principal components that collectively accounted for approximately 72% of the total variance, the model achieved a balance between dimensionality reduction and information preservation. The loading patterns of these components revealed meaningful dimensions in user information-seeking behaviour, consistent with theoretical expectations about how individuals search for health information during disease outbreaks.

Nevertheless, this approach faces some important limitations. As demonstrated by the prediction errors and visual analysis of time series data, Wikipedia-based surveillance remains susceptible to external influences unrelated to disease prevalence, including media-driven attention spikes that may skew predictions. Additionally, the model showed variable performance across different epidemic scenarios, suggesting that its reliability may depend on the specific characteristics of each influenza season.

These findings align with the broader literature suggesting that internet-based surveillance systems should complement rather than replace traditional methods (Hammond et al., 2022). The Wikipedia-based approach offers advantages in terms of near real-time availability and cost-effectiveness, potentially addressing the reporting delays inherent in traditional surveillance systems. However, as previous research has emphasized, the combination of traditional and digital surveillance sources appears to provide the most robust framework for disease monitoring (McClymont et al., 2024).

Future research should explore more sophisticated techniques for filtering noise from Wikipedia signals, integration with other digital data streams, and extension to additional infectious diseases beyond influenza. Longer time series spanning more influenza seasons would also enhance model validation and improve understanding of how Wikipedia-based surveillance performs across varied epidemic profiles. Overall, this study contributes to the growing field of digital disease surveillance by demonstrating the practical utility of Wikipedia data for enhancing real-time monitoring capabilities while acknowledging the continued importance of established epidemiological methods.

Acknowledgements

I thank Dr. Stefano Andreozzi for valuable suggestions and comments.

References

Chakraborty P, Lewis B, Eubank S, Brownstein JS, Marathe M, Ramakrishnan N. (2018) What to know before forecasting the flu. PLoS Comput Biol 14(10): e1005964. https://doi.org/10.1371/journal.pcbi.1005964.

Colavizza, G. (2020). COVID-19 research in Wikipedia. Quantitative Science Studies, 1(4), 1349–1380. https://doi.org/10.1162/qss_a_00080.

de Courville C, Cadarette SM, Wissinger E, Alvarez FP. The economic burden of influenza among adults aged 18 to 64: A systematic literature review. Influenza Other Respi Viruses. 2022;16(3):376-385. doi:10.1111/irv.12963.

De Toni G, Consonni C, Montresor A (2021). A general method for estimating the prevalence of influenza-like-symptoms with Wikipedia data. PLoS ONE 16(8): e0256858. https://doi.org/10.1371/journal.pone.0256858.

Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R (2014) Global Disease Monitoring and Forecasting with Wikipedia. PLoS Comput Biol 10(11): e1003892. doi: 10.1371/journal.pcbi.1003892.

Gianfredi, V., Santangelo, O. E., & Provenzano, S. (2021). Correlation between flu and Wikipedia's pages visualization. *Acta Bio Medica: Atenei Parmensis*, *92*(1), e2021056.

Hadi, A. S., & Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, *52*(1), 15-19.

Hammond, A., Kim, J. J., Sadler, H., & Vandemaele, K. (2022). Influenza surveillance systems using traditional and alternative sources of data: a scoping review. *Influenza and Other Respiratory Viruses*, *16*(6), 965-974.

Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. (2015). Forecasting the 2013–2014 Influenza Season Using Wikipedia. PLoS Comput Biol 11(5): e1004239. doi: 10.1371/journal.pcbi.1004239.

Istituto Superiore di Sanità. (2025). RespiVirNet: National epidemiological and virological surveillance system for influenza and other respiratory viruses (https://www.iss.it/en/respivirnet).

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

McClymont, H., Lambert, S.B., Barr, I. *et al.* Internet-based Surveillance Systems and Infectious Diseases Prediction: An Updated Review of the Last 10 Years and Lessons from the COVID-19 Pandemic. *J Epidemiol Glob Health* 14, 645–657 (2024). https://doi.org/10.1007/s44197-024-00272-y

MacIntyre, C. R., Chen, X., Kunasekaran, M., Quigley, A., Lim, S., Stone, H., ... & Gurdasani, D. (2023). Artificial intelligence in public health: the potential of epidemic early warning systems. *Journal of International Medical Research*, *51*(3), 03000605231159335.

McIver DJ, Brownstein JS (2014) Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Comput Biol 10(4): e1003581. doi:10.1371/journal.pcbi.1003581.

Mooi, E., Sarstedt, M., & Mooi-Reci, I. (2018). *Market research*. Springer Nature Singapore Pte Ltd.

O'Leary, D. E., & Storey, V. C. (2020). A Google–Wikipedia–Twitter model as a leading indicator of the numbers of coronavirus deaths. *Intelligent Systems in Accounting, Finance and Management*, *27*(3), 151-158.

Priedhorsky, R., Osthus, D., Daughton, A. R., Moran, K. R., Generous, N., Fairchild, G., ... & Del Valle, S. Y. (2017, February). Measuring global disease with Wikipedia: Success, failure, and a research agenda. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1812-1834).

Santangelo, O. E., Gianfredi, V., & Provenzano, S. (2022). Wikipedia searches and the epidemiology of infectious diseases: A systematic review. *Data & Knowledge Engineering*, *142*, 102093.

Sharpe, J. D., Hopkins, R. S., Cook, R. L., & Striley, C. W. (2016). Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. *JMIR public health and surveillance*, *2*(2), e5901.

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Simonsen, L., Gog, J., Olson, D., & Viboud, C. (2016). Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems. Journal of Infectious Diseases, 214(suppl.4). https://doi.org/10.1093/infdis/jiw376.

Smith DA (2020) Situating Wikipedia as a health information resource in various contexts: A scoping review. PLoS ONE 15(2): e0228786. https://doi.org/10.1371/journal.pone.02287.

Transtrum, M. K., Machta, B. B., Brown, K. S., Daniels, B. C., Myers, C. R., & Sethna, J. P. (2015). Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of chemical physics*, *143*(1).

Wikimedia Foundation. (2025). Pageviews Analysis. Meta-Wiki. https://pageviews.toolforge.org/

Zimmer C, Leuba SI, Yaesoubi R, Cohen T. 2018 Use of daily Internet search query data improves real-time projections of influenza epidemics. J. R. Soc. Interface 15: 20180220. http://dx.doi.org/10.1098/rsif.2018.0220.

4th IFC and Bank of Italy Workshop on "Data Science in Central Banking"

18-20 February 2025, Rome, Italy

Forecasting influenza-like illness in Italy using Wikipedia: a principal component regression approach

Gianluca Mura

Bank of Italy

AGENDA

1. Context and motivation
2. Data collection and preprocessing
3. Principal component analysis (PCA): an overview
4. Model building: principal component regression (PCR)
5. Model performance evaluation
6. Conclusions

1. Context and motivation

Influenza remains a significant public health challenge, with substantial epidemiological, clinical, and economic impacts. Factors such as its high transmissibility, antigenic variability, seasonal epidemics (and occasional pandemics), and severe complications in vulnerable populations (e.g., children, elderly) contribute to its impact on society.

According to recent estimates, the European Centre for Disease Prevention and Control (ECDC) reports that seasonal influenza causes approximately 40 million symptomatic cases and 28,000 influenza-related deaths annually in Europe.

Additionally, **influenza-like-illness** (ILI) imposes substantial social costs due to lost working days and decreased productivity among working adults and students, particularly those who are unvaccinated, under 65 years of age, or experiencing severe disease. Considerable work time and productivity loss is also attributable to caregiver burden.

These figures underline the critical role of surveillance systems in guiding vaccine composition, healthcare planning, and understanding influenza-related complications.

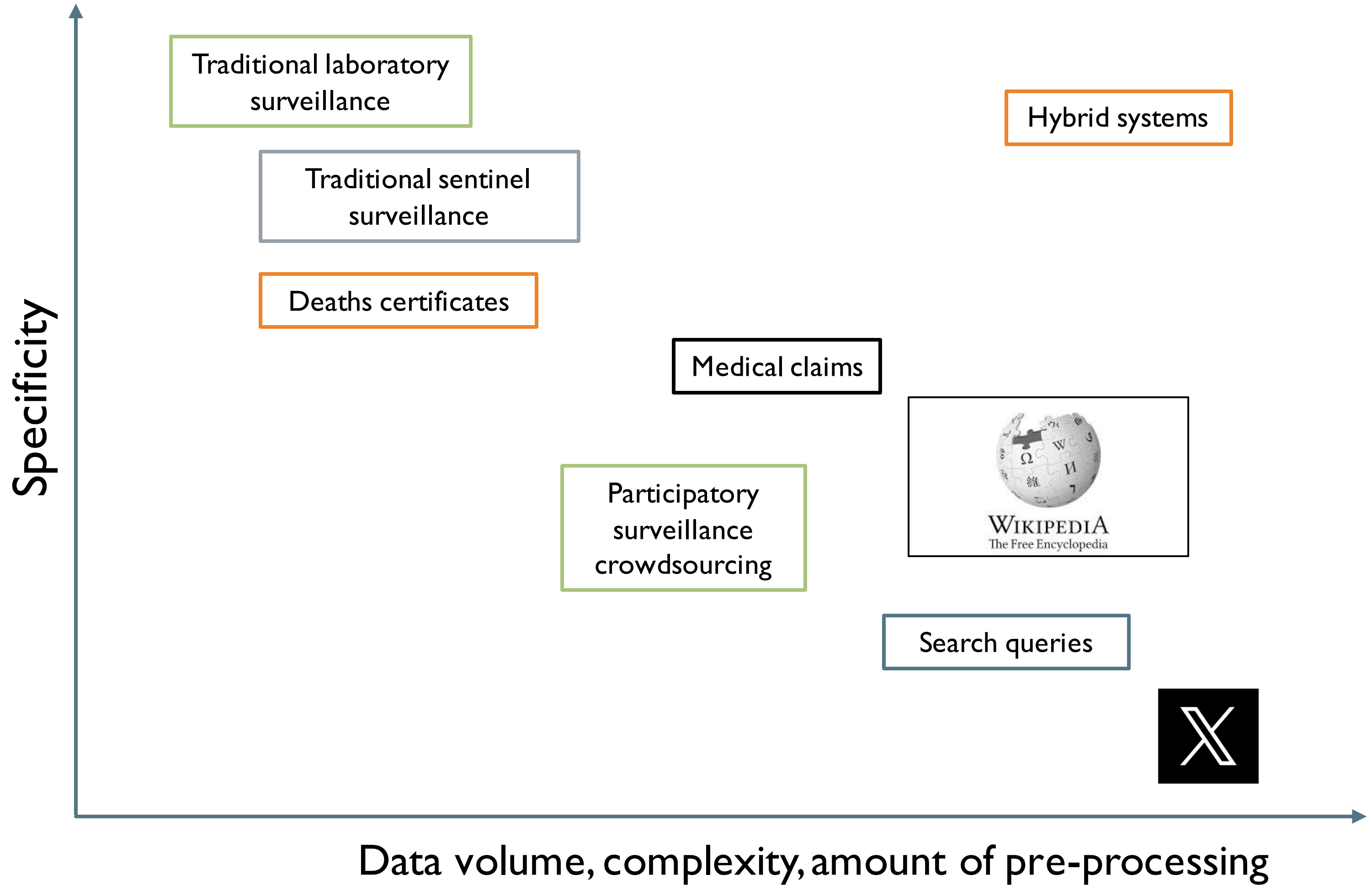
Traditional surveillance systems may find a limitation in the delay of data availability of up to two weeks compared to the reporting period due to the necessary collection and processing activities of the information sent by the network of sentinel physicians. Therefore, emerging methods, such as **monitoring Internet activity and social media trends**, have gained attention for their potential to enhance near real-time influenza surveillance (better situational awareness).

In this context, several initiatives have been promoted by major health organizations (WHO, ECDC, CDC) to develop alternative systems based on big data for forecasting influenza and other viral diseases.

1. Context and motivation

Surveillance systems have evolved significantly, incorporating both classical and big data-driven approaches. Classical methods now include **electronic death certificates, patient-level hospital discharge records, and medical claims data**,

Alongside these, **big data-driven systems** have emerged, relying on streams from internet search queries, social media, and crowdsourcing platforms to accelerate the detection and monitoring of syndromic trends.



Source: author's elaboration based on "Simonsen L. et al. - Infectious disease surveillance in the big data era: towards faster and locally relevant systems. J Infect Dis 2016;214:S380-5".

2. Data collection and preprocessing

In Italy, the RespiVirNet (formerly **InfluNet**) surveillance system monitors influenza-like illnesses (ILI) during the seasonal period through a network of so-called sentinel physicians.

The system relies on a network of voluntarily participating general practitioners (GPs) and pediatricians (PLS) and is coordinated by the Istituto Superiore di Sanità (ISS). Epidemiological surveillance is conducted via weekly reports of ILI cases from a sample of GPs and PLS during the period from October to April.

InfluNet
Rete Italiana Sorveglianza Influenza

Settimana **2019 - 17**
dal **22** *al* **28** *aprile al 2019*

Stagione Influenzale 2018 - 2019

Rapporto N. 27 del 21 maggio 2019

Risultati Nazionali

La tabella seguente mostra il numero dei casi e i tassi d'incidenza, nel totale e per fascia di età, di tutte le regioni che hanno inviato i dati. L'incidenza settimanale è espressa come numero di sindromi influenzali (casi) per 1.000 assistiti.

Settimana	Totale Medici	Totale Casi	Totale Assistiti	Totale Incidenza	0-4 anni		5-14 anni		15-64 anni		65 anni e oltre	
					Casi	Inc	Casi	Inc	Casi	Inc	Casi	Inc
2018-42	1.073	1.088	1.384.736	0,79	100	1,26	117	0,63	691	0,85	180	0,58
2018-43	1.080	1.319	1.396.489	0,94	155	1,96	148	0,80	855	1,05	161	0,51
2018-44	1.091	1.421	1.411.739	1,01	134	1,68	163	0,86	909	1,10	215	0,68
2018-45	1.091	2.074	1.411.584	1,47	235	2,93	250	1,31	1266	1,54	323	1,02

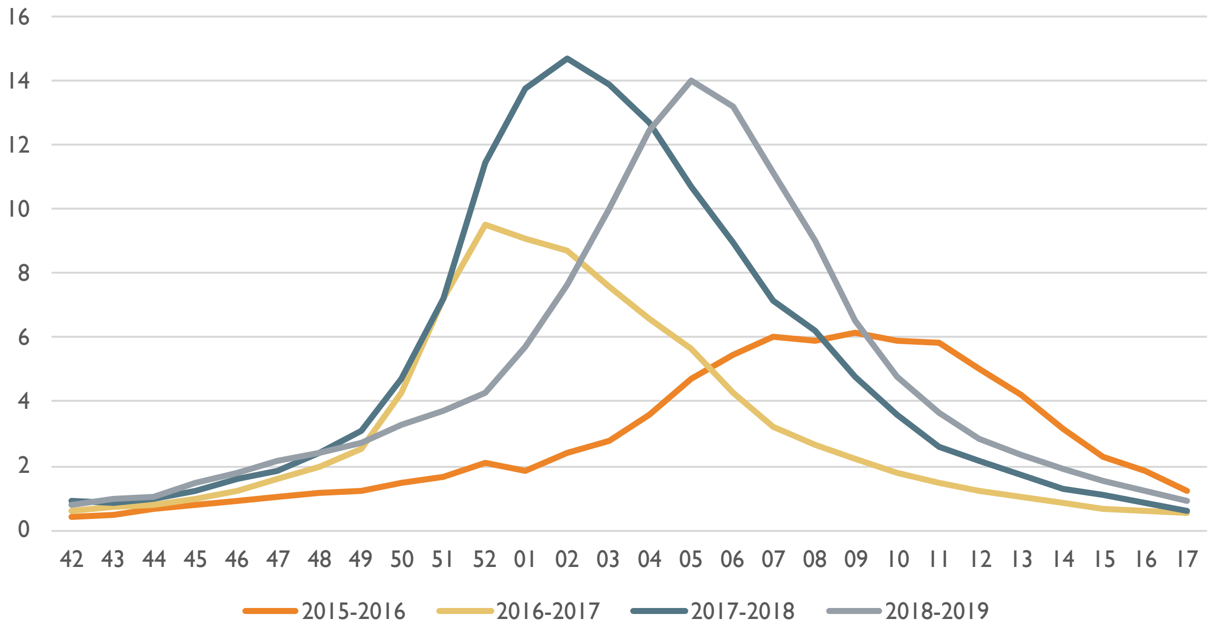
2. Data collection and preprocessing

The data on the reference week and the weekly incidence data, expressed as the number of influenza syndromes (cases) per 1,000 patients (total incidence), were extrapolated. Each influenza season, and thus the relevant sample survey, has an average duration of 28 weeks, beginning on the 42nd week of one year and ending on the 17th of the following year.

Due to the limited data available on Wikipedia access logs, four influenza seasons were selected for equivalent 112 weekly observations. As is usually the case, the four influenza seasons examined had different characteristics with different intensities and epidemic peaks.

Influenza Seasons	From	to	n. of weeks
2015-2016	2015-42*	2016-16	28
2016-2017	2016-42	2017-17	28
2017-2018	2017-42	2018-17	28
2018-2019	2018-42	2018-17	28

Note: the year 2015 ended with the 53rd week, the others with the 52nd



Source: InfluNet reports

2. Data collection and preprocessing

Key arguments in favor of using Wikipedia for predicting Flu incidence rates.

- **Widespread usage and accessibility**: i)Wikipedia is one of the most accessed online resources for health information, surpassing established platforms like WebMD and the World Health Organization in terms of web traffic; ii) Its content is frequently updated in real time, making it a dynamic and responsive source for health trend.
- **Correlations with epidemiological data**: several studies have shown moderate to strong correlations between Wikipedia page views and official epidemiological data, such as influenza incidence tracked by health institutes
- **Global relevance and multilingual content**: Wikipedia offers health content in over 275 languages, providing a globally representative dataset that complements regionally limited surveillance systems (Note: the language of Wikipedia articles is not always a good proxy for the geographical scope of interest).
- **Real-time surveillance**: Wikipedia page views can provide near real-time surveillance of public interest in specific diseases, which is faster than traditional health monitoring systems that may face delays in reporting. This rapid data acquisition makes it suitable for short-term forecasting and early detection of disease outbreak.

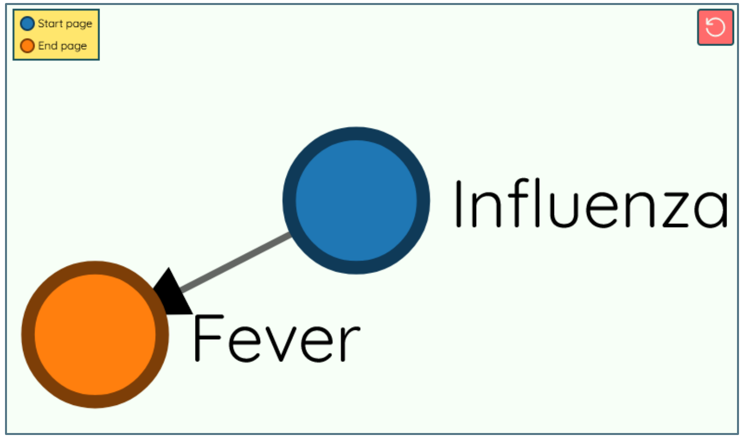
2. Data collection and preprocessing

Wikipedia provides **various tools and APIs** to analyze its data, including access logs, page content, hyperlinks, and editor activity. There are also numerous free third-party solutions and open-source projects have been developed by the community, enhancing the functionality and accessibility of Wikipedia data.

With reference to the number of Wikipedia **access logs**, information on daily accesses as of 1 July 2015 was identified by selecting all types of platforms (desktop, mobile and mobile application) and discarding accesses generated by crawlers and other automated programs) . The number of accesses was then aggregated by ISO week and purged of irrelevant observations because they were not included in the annual flu season survey.

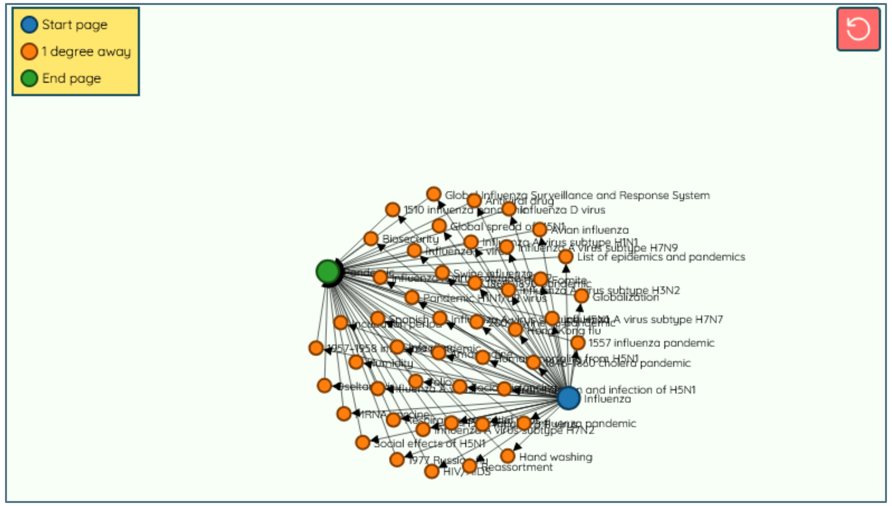
With regard to the selection of the most relevant Wikipedia articles, several approaches are available, which involve analyzing their relationships based on hyperlinks, categories, content similarity, or a combination of these factors.

1 path with 1 degree of separation from **Influenza** to **Fever**



Source: www.sixdegreesofwikipedia.com

47 paths with 2 degrees of separation from **Influenza** to **Pandemic**



2. Data collection and preprocessing

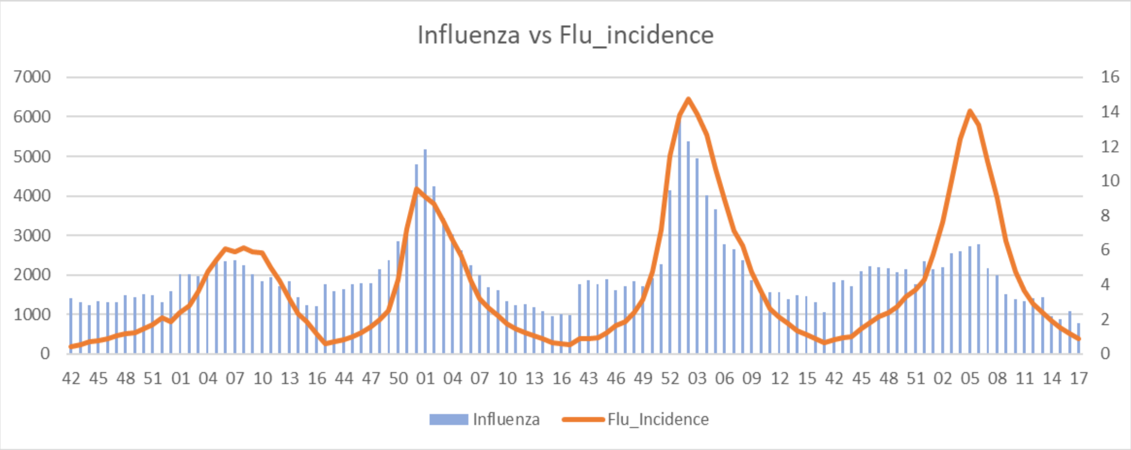
The selected pages were curated based on **thematic relevance**, with publications and prior studies on influence rate prediction reviewed to guide the selection of topics most likely to contribute to accurate modeling.
After this selection, distance metrics were applied as methods of verification and confirmation, ensuring the dataset represented a cohesive and meaningful subset of Wikipedia content. At the end of this process, 14 articles were selected.

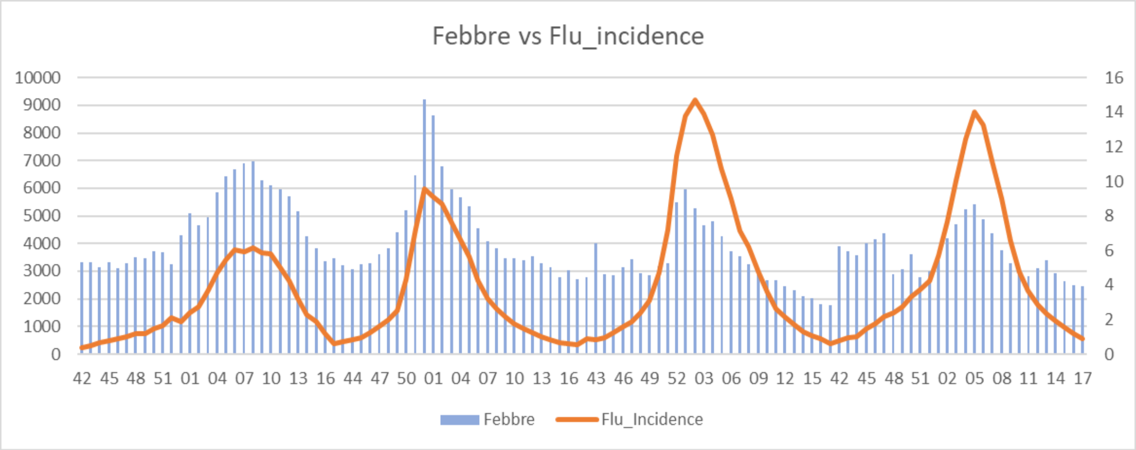
Wikipedia articles in Italian	Corresponding article in English	Total number of views over the study period (4 flu seasons)	Shortest Path Distance
Influenza	Influenza	227.775	-
Febbre	Fever	446.206	1
Cefalea	Headache	230.886	1
Influenza_aviaria	Avian influenza	91.820	1
Influenza_suina	Swine influenza	49.771	1
Epidemia	Epidemic	62.671	2
Pandemia	Pandemic	96.235	2
Pandemia_influenzale	Influenza pandemic	21.314	1
Raffreddore_comune	Common cold	225.309	1
Paracetamolo	Paracetamol	806.148	1
Raffreddore	Cold	12.419	2
Virus parainfluenzali umani	Human parainfluenza viruses	11.355	1
Antivirale	Antiviral drug	45.896	1
Influenza virus A sottotipo H1N1	Influenza A virus subtype H1N1	93.237	1

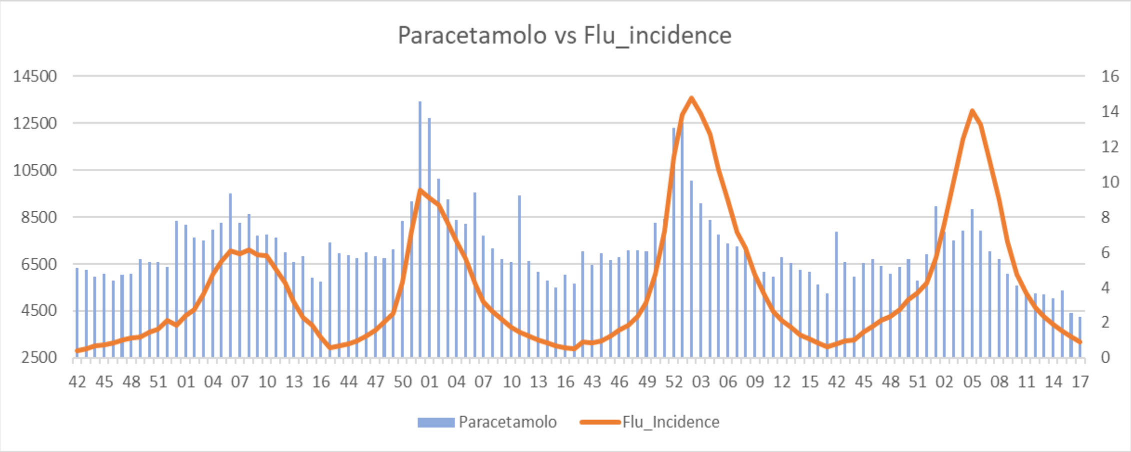
Source: https://meta.wikimedia.org/

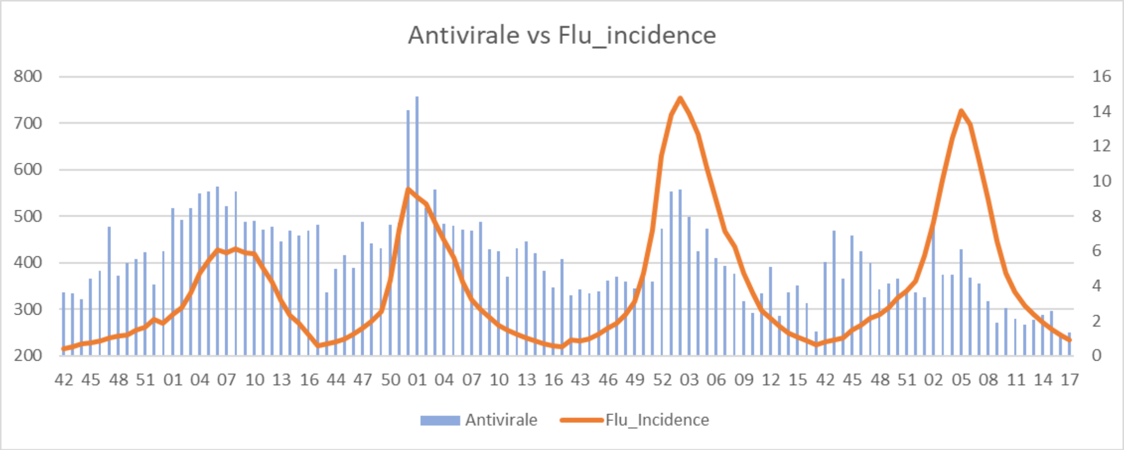
2. Data collection and preprocessing

Access logs for selected Wikipedia pages (left axis) plotted alongside official ILI rates (right axis)









3. Principal component analysis (PCA): an overview

PCA is a statistical technique used to **reduce the dimensionality** of a dataset while retaining as much variability as possible. It transforms correlated variables into a **new set of uncorrelated variables**, called principal components, ordered by the amount of variance they capture. PCA is widely used for **simplifying data structures, identifying patterns, and preparing data for predictive modeling** or visualization.

Pros of PCA:

1. **dimensionality reduction**: PCA can reduce the number of variables while retaining most of the important information;
2. **removes multicollinearity** between variables by creating orthogonal principal components;
3. **noise reduction**: PCA can help filter out noise in the data by focusing on the components that explain the most variance.
4. **better visualization**: PCA allows for easier visualization of high-dimensional data by projecting it onto lower dimensions.
5. **improved model performance**: using principal components as inputs can improve the performance of some machine learning models.

Cons of PCA:

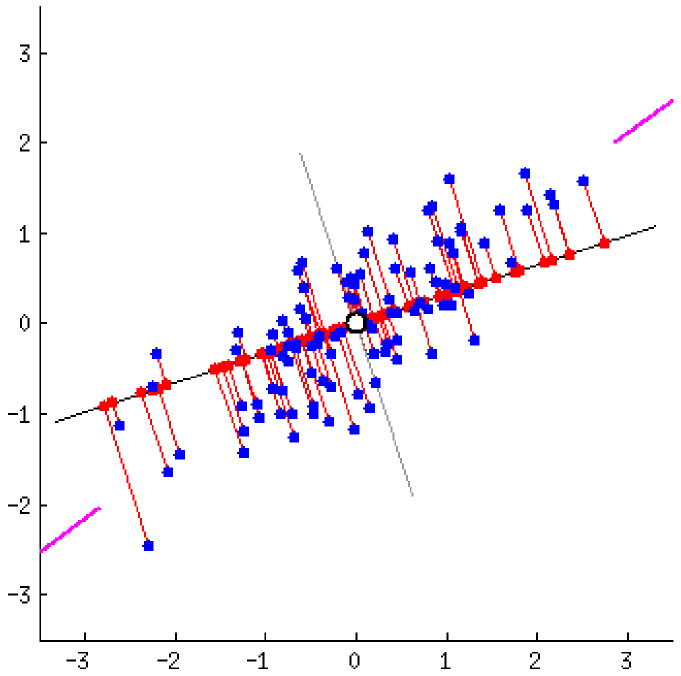
1. **loss of interpretability**: the principal components are often difficult to interpret in terms of the original variables;
2. **assumes linearity**: PCA assumes linear relationships between variables, which may not always be the case;
3. **sensitive to outliers**: PCA can be significantly affected by outliers in the dataset;
4. **may lose some information**: while PCA aims to retain the most important information, some potentially valuable details may be lost in the process of dimensionality reduction;
5. **requires standardization**: variables need to be standardized before applying PCA, which can be an additional preprocessing step.

3. Principal component analysis (PCA): an overview

The graph below is intended to explain the transformation of 2D data into 1D using PCA.

It shows the new vector, which is determined by rotating around the mean of the 2D distribution to find the direction that minimizes the projection error (i.e. maximizes the variance captured).

The original data points are projected onto this new line, representing their positions along the principal component, effectively reducing the dimensionality while preserving as much information as possible



4. Model building: principal component regression (PCR)

Principal Component Regression (PCR) combines PCA and linear regression to address issues like multicollinearity and dimensionality reduction. The consequential steps followed to perform the PCR were:

1. **Data Preparation**: check for missing values and outliers; assess normality of variables; standardize the data to ensure all variables are on the same scale.
2. **Correlation analysis**: create a correlation matrix to identify highly correlated variables.
3. **PCA computation** on standardized predictor variables, examination of the scree plot to determine the number of components to retain; cumulative explained variance plot analysis.
4. **Choice of the number of components** to be considered (different methods available).
5. **PCR model building**: selected principal components were used as predictors in a linear regression model; the model is fitted using ordinary least squares.
6. **Model Evaluation**: assess model performance using appropriate metrics.
7. **Interpretation**: analyze the contribution of original variables to principal components; interpret PCR coefficients in terms of the original variables.
8. **Prediction:** generation of predictions using the PCR model; comparison of actual vs. predicted values.
9. **Refinement**: evaluation of possible variations to the initial model to improve performance.

4. Model building: principal component regression (PCR)

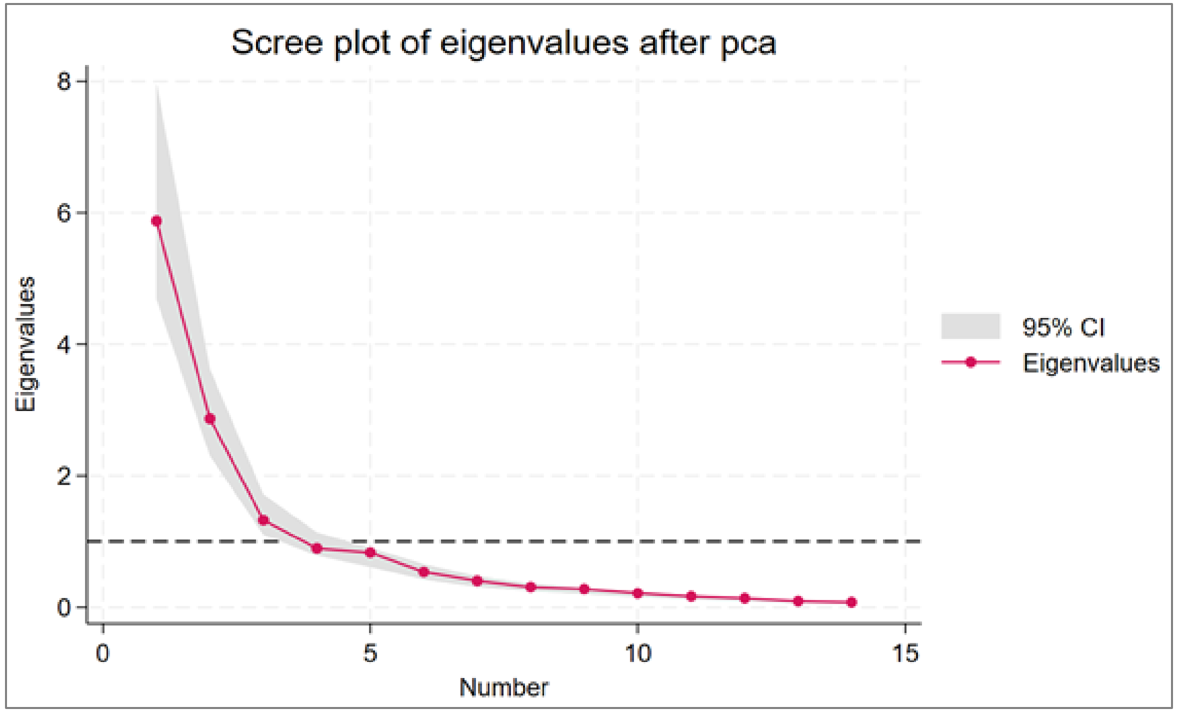
The correlation matrix revealed several medium and high correlations between the independent variables, as hypothesized, justifying the application of PCA to address multicollinearity. Additionally, the statistical significance of these correlations was assessed, with *p-values* confirming their reliability.

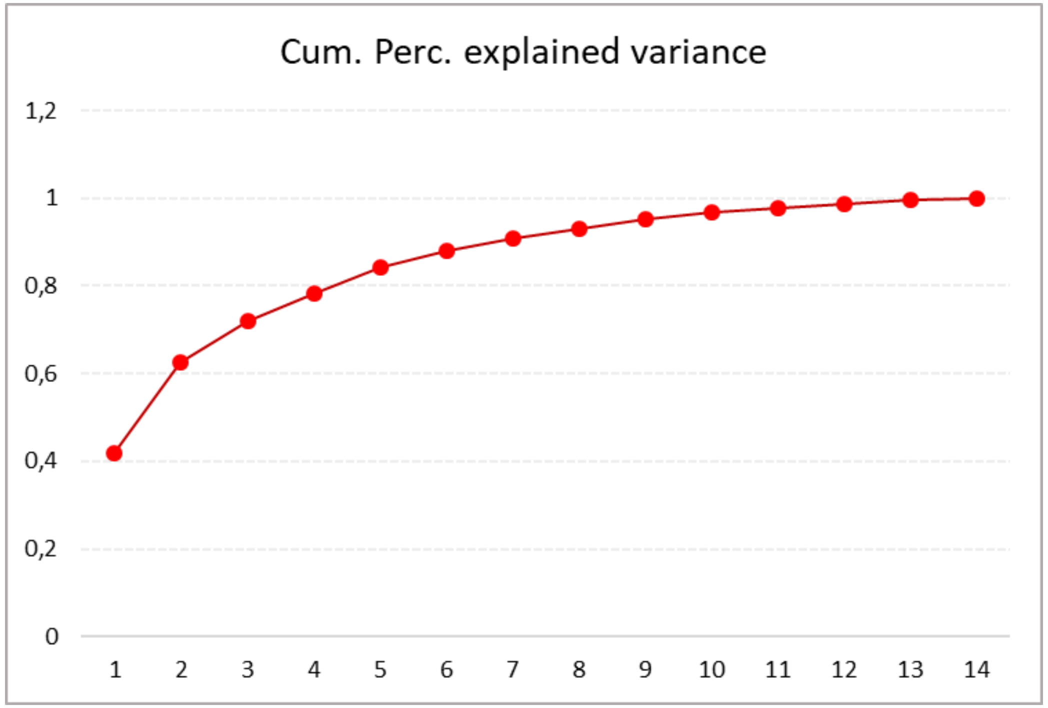
	Flu_Incidence	Influenza	Febbre	Cefalea	Influenza_aviaria	Influenza_suina	Epidemia	Pandemia	Pandemia_influenzale	Raffreddore_comune	Paracetamolo	Raffreddore	Virus parainfluenzali umani	Antivirale	Influenza virusA sottotipo H1N
Flu_Incidence	1														
Influenza	0,8113	1													
Febbre	0,5887	0,6822	1												
Cefalea	-0,0923	0,1485	0,5865	1											
Influenza_aviaria	0,3181	0,4072	0,1761	-0,0368	1										
Influenza_suina	0,5571	0,3749	0,2183	-0,0843	0,2269	1									
Epidemia	0,1521	0,207	0,5839	0,7214	0,0591	0,0981	1								
Pandemia	0,3584	0,4268	0,5043	0,4976	0,1362	0,2738	0,6397	1							
Pandemia_influenzale	0,6823	0,6822	0,3008	-0,2233	0,2578	0,4726	0,0123	0,2926	1						
Raffreddore_comune	0,0568	0,384	0,5807	0,5393	-0,0736	-0,1105	0,3556	0,3391	0,0053	1					
Paracetamolo	0,6601	0,8384	0,7795	0,3003	0,358	0,2166	0,3223	0,3874	0,4267	0,5879	1				
Raffreddore	0,3938	0,5137	0,5627	0,3751	0,144	0,1214	0,3602	0,3833	0,0425	0,6659	0,6465	1			
Virus parainfluenzali umani	0,4691	0,6003	0,2763	-0,0928	0,2187	0,0716	-0,0268	0,1613	0,6316	0,1739	0,4821	0,1646	1		
Antivirale	0,3864	0,6264	0,8392	0,7166	0,1815	0,1272	0,6205	0,5237	0,1683	0,6271	0,7429	0,5698	0,253	1	
Influenza virusA sottotipo H1N1	0,767	0,6359	0,3177	-0,14	0,2704	0,5499	0,0418	0,3348	0,6161	-0,0356	0,4412	0,2156	0,3695	0,2079	1

4. Model building: principal component regression (PCR)

Selecting the optimal number of factors (or principal components) is crucial in PCA to balance dimensionality reduction and information retention. Several approaches offer support for determining the number of factors, and it is common to use a combination of methods rather than relying on a single criterion.

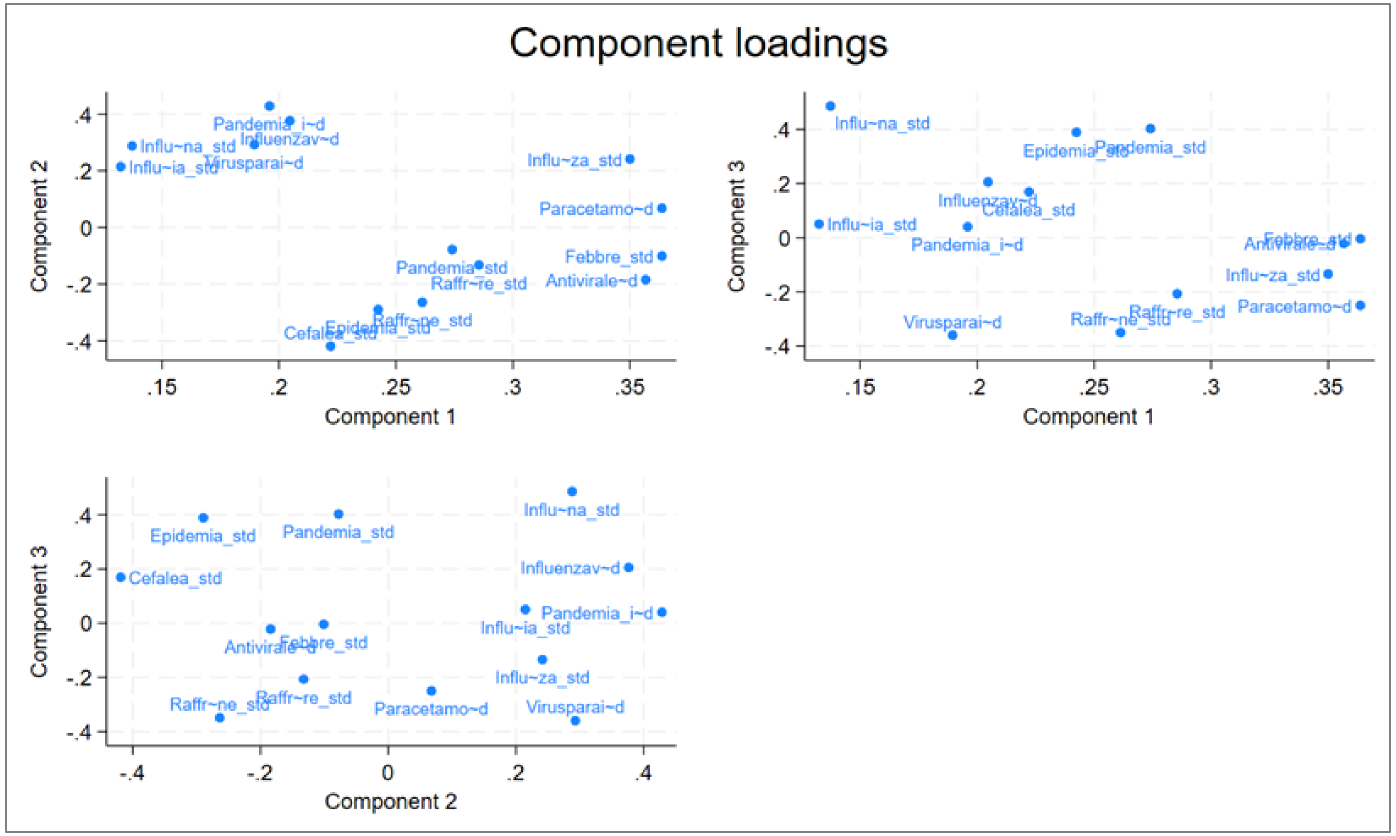
1. **Kaiser Criterion:** eigenvalue > 1;
2. **Cumulative Variance Explained:** select enough components to achieve a cumulative variance explained threshold, typically 70–90%;
3. **Scree Plot**: i.e. plotting the eigenvalues in descending order and look for the "elbow point" where the eigenvalues level off. Components are normally retained before this point.





4. Model building: principal component regression (PCR)

Loading plots (left side) visually represent the relationships between variables and components, helping to identify key contributors, variable groupings, and the underlying structure of the data in dimensionality reduction. I also assessed the adequacy of the dataset for PCA by performing the **Kaiser-Meyer-Olkin (KMO) test**. A KMO value greater than 0.80 indicates good sampling adequacy, suggesting that the variables share sufficient common variance to justify the use of PCA. This also suggest that the dataset is well-suited for dimensionality reduction, and the extracted components are likely to be meaningful.

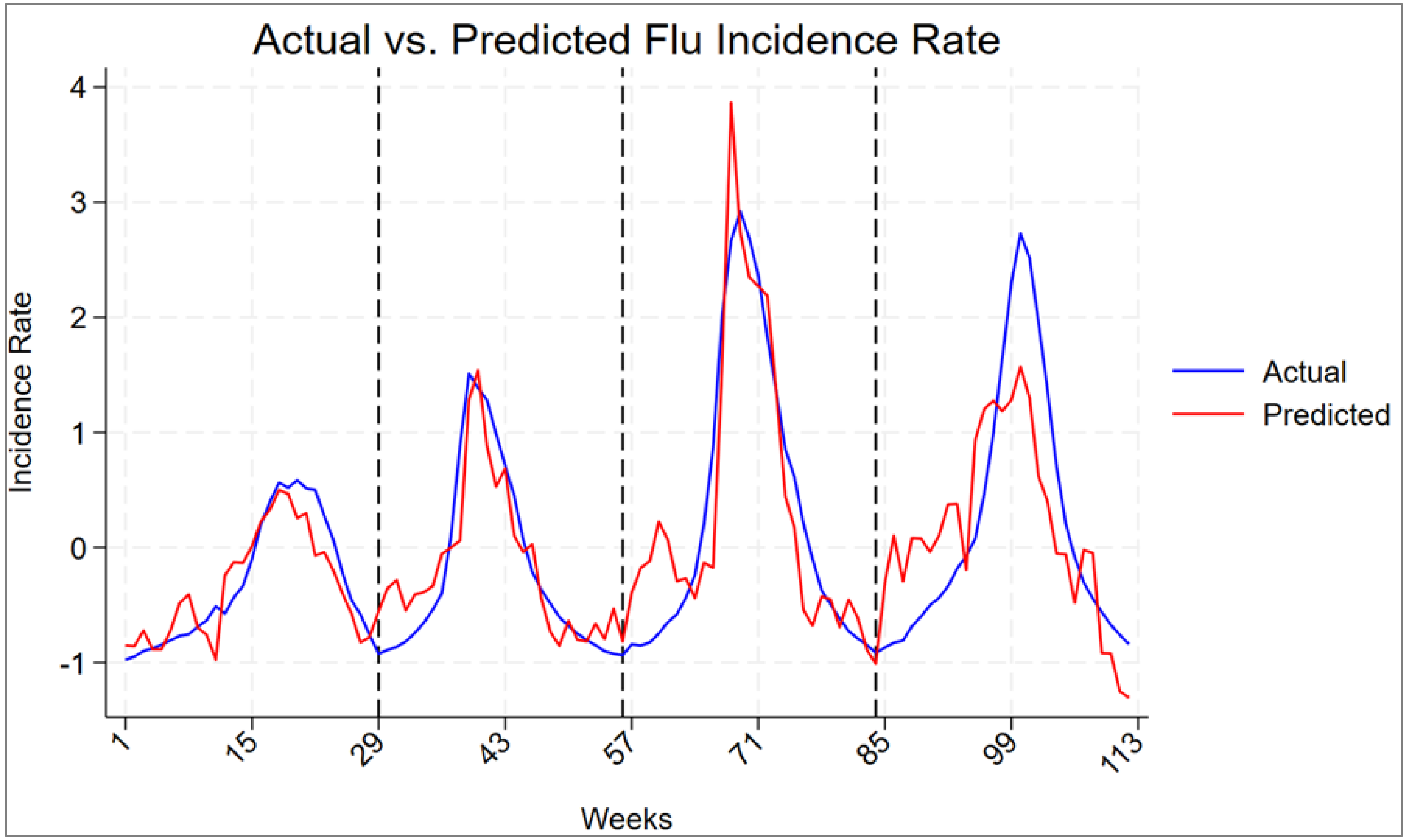


Variable	KMO
Influenza	0,8414
Febbre	0,9196
Cefalea	0,7868
Influenza_aviaria	0,7249
Influenza_suina	0,7365
Epidemia	0,7918
Pandemia	0,8715
Pandemia_influenzale	0,704
Raffreddore_comune	0,7933
Paracetamolo	0,8687
Raffreddore	0,8122
Antivirale	0,8888
Virus parainfluenzali umani	0,8155
Influenza virus A sottotipo H1N1	0,8657
Overall	**0,8308**

5. Model performance evaluation

The model performs well in explaining the variability of the dependent variable (explain 77.23% of the variance in the dependent variable) and shows good predictive accuracy. Additionally, the overall model is highly statistically significant.

$$Flu_{incidence} = -2.62e - 09 - 09 + 0.2750 \cdot pc1 + 0.3337 \cdot pc2 + 0.0794 \cdot pc3 + \varepsilon$$



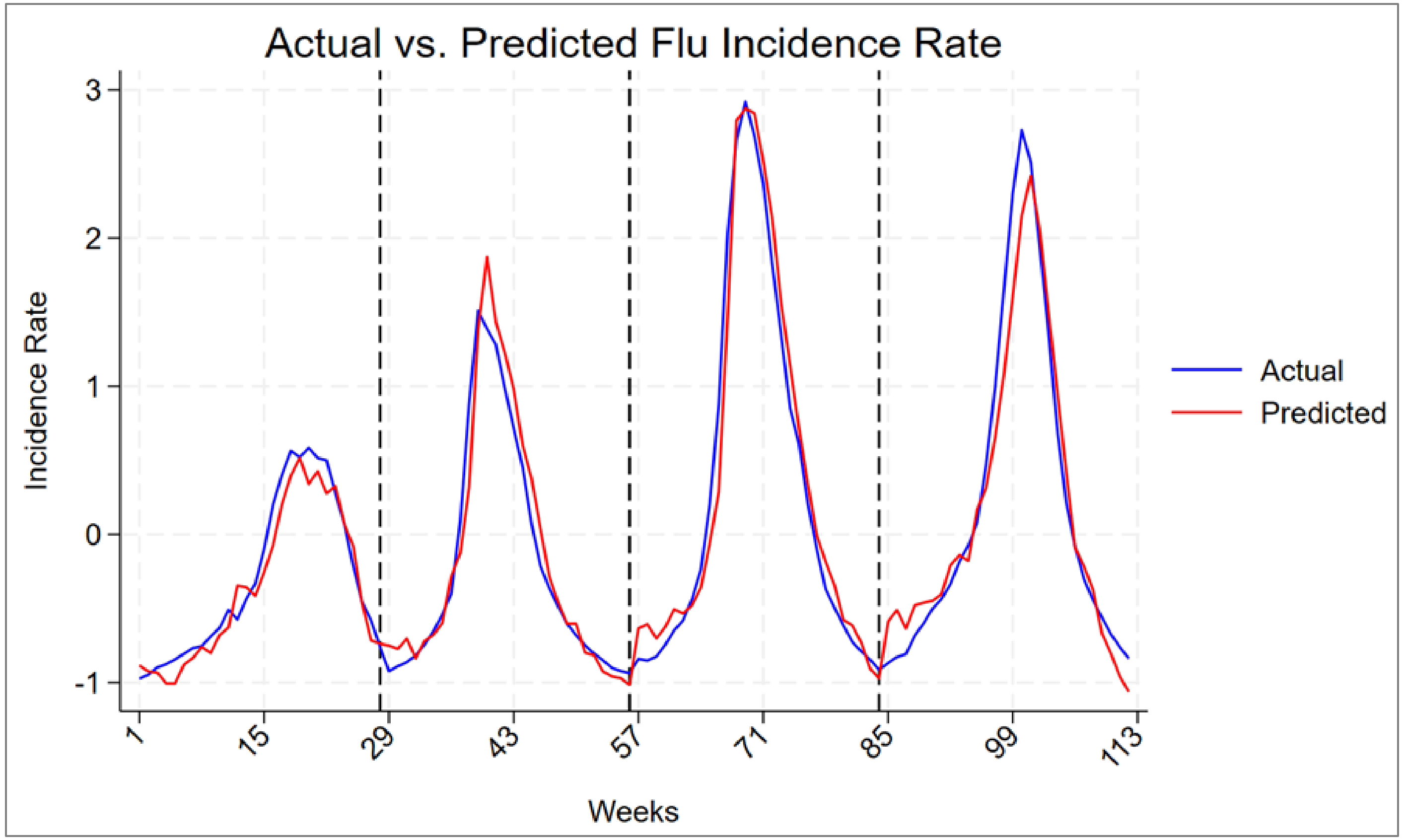
$\boldsymbol{r^2 = 0.7723}$

RMSE = 0.4837 ~12% of variable range

5. Model performance evaluation

To enhance the predictive performance of the model, a new explanatory variable was introduced: the **flu spread rate of the previous week**. This variable, representing known data, leverages temporal dependencies inherent in influenza dynamics. The rationale for this choice is supported by findings that historical data – such as previous week influenza levels – offer crucial insights into current trends. These temporal dependencies allow the model to dynamically adjust predictions.

$$Flu\_incidence = -3.42e - 09 + 0.0942 \cdot pc1 + 0.0932 \cdot pc2 - 0.0607 \cdot pc3 + 0.7542 \cdot Flu\_incidence_{t-1} + \varepsilon$$



$\boldsymbol{r^2 = 0.9572}$

RMSE = 0.2106

5. Conclusions

With this work, drawing on experiments and analyses previously conducted in the US and Europe, I aimed to assess the feasibility of developing a statistical model capable of estimating ILI rates in Italy using access log data from specific Wikipedia page consultations.

The Principal Component Regression-based approach proved particularly well-suited to the explanatory variables, as it efficiently manages the high dimensionality and correlations present in the dataset.

Moreover, the selection of Wikipedia articles can be expanded and continuously adjusted to account for contingent situations, such as changes in public interest during health emergencies.

The encouraging initial results confirm the potential for defining a hybrid model that combines traditional surveillance systems with real-time data sources like Wikipedia, providing enhanced forecasting capabilities and improving public health monitoring in Italy.

This study contributes to the broader goal of leveraging alternative data streams to complement traditional methods, addressing timeliness and adaptability in disease surveillance.

THANK YOU

gianluca.mura@bancaditalia.it