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

Disclaimer: The views and opinions expressed in this presentation are those of the author and do not necessarily reflect the position of 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.

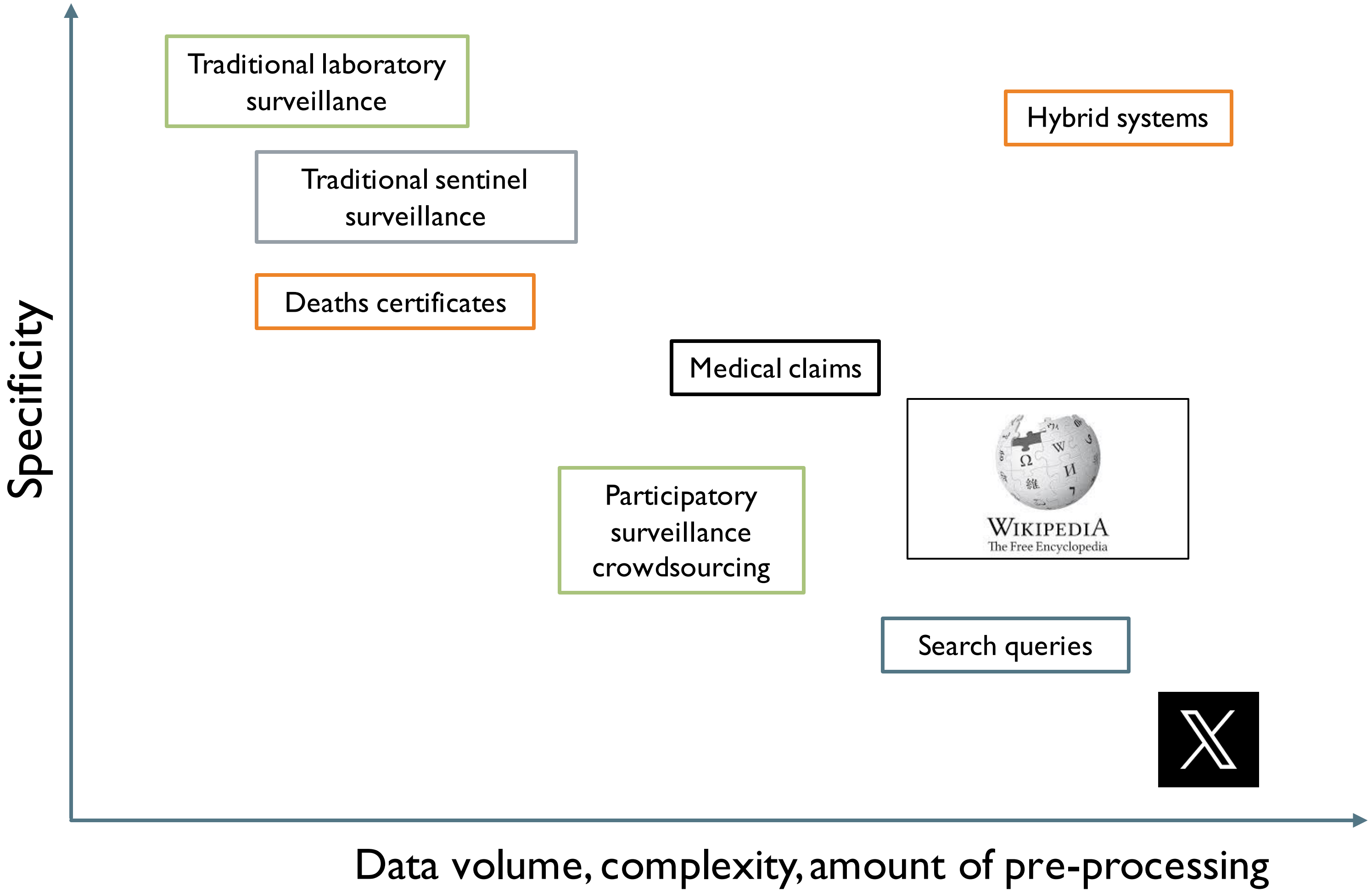
Traditional surveillance systems may find a limitation in the delay of data availability of up to two weeks compared to the reporting period due to the necessary collection and processing activities of the information sent by the network of sentinel physicians. Therefore, emerging methods, such as **monitoring 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.

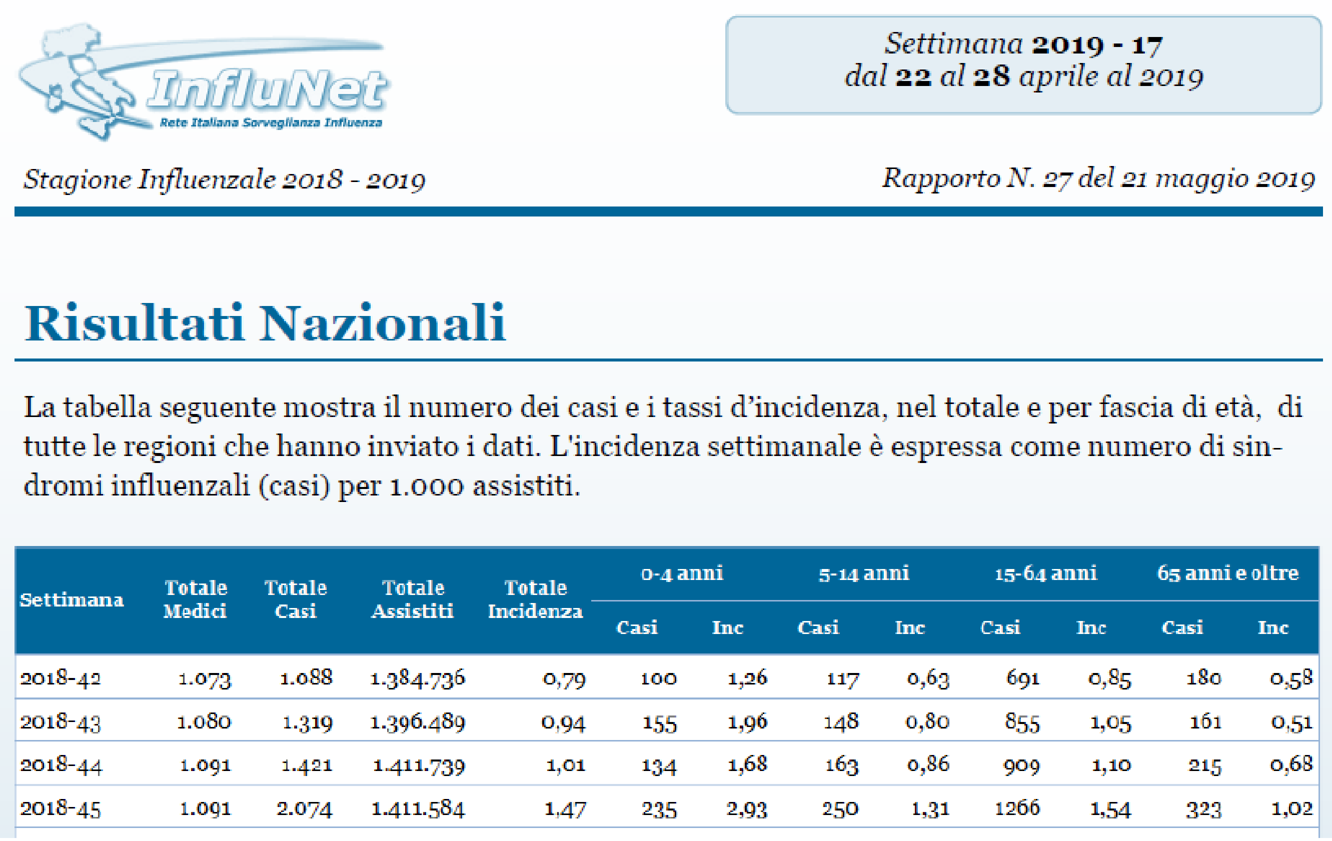

Specificity
Traditional laboratory surveillance
Hybrid systems
Traditional sentinel surveillance
Deaths certificates
Medical claims
Wikipedia The Free Encyclopedia
Participatory surveillance crowdsourcing
Search queries
X
Data volume, complexity, amount of pre-processing

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

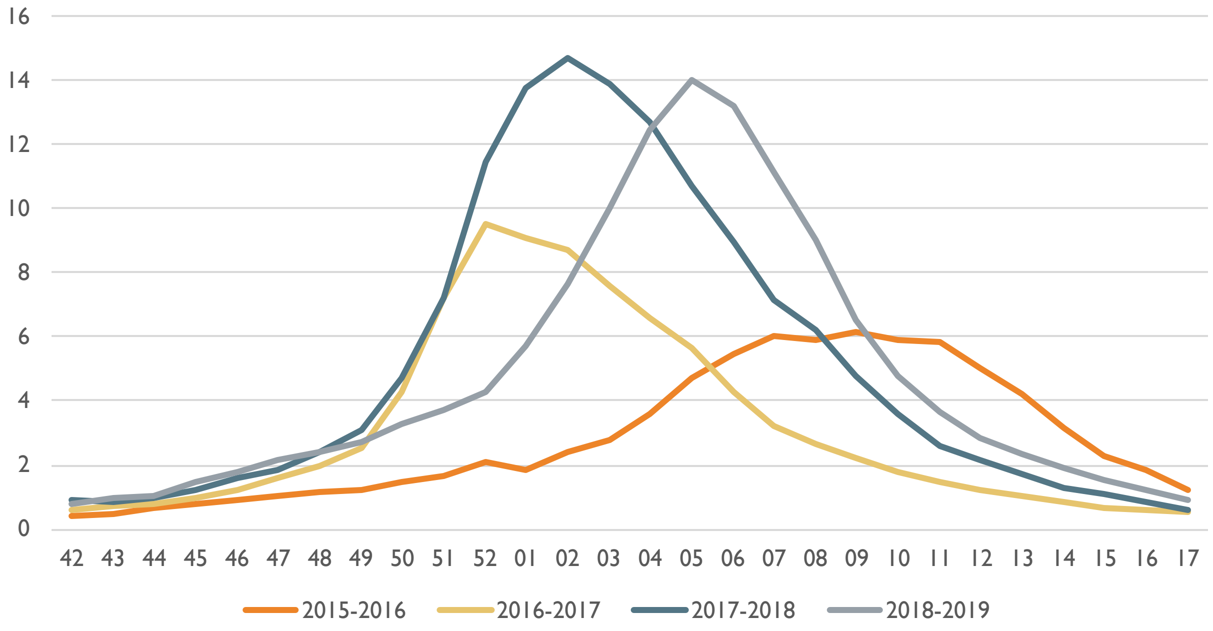
2. Data collection and preprocessing

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.

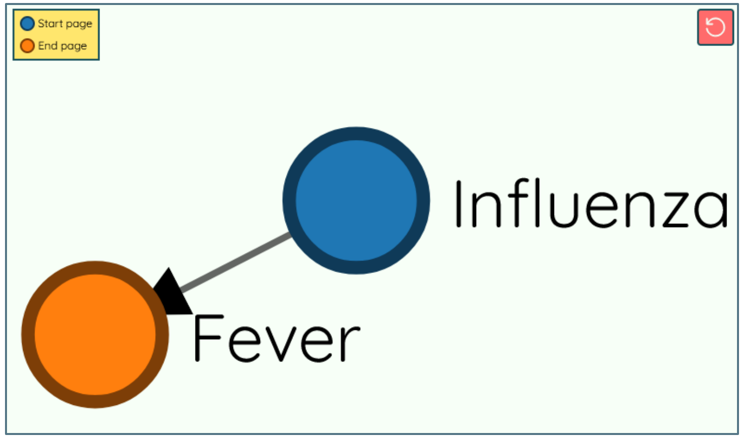
2. Data collection and preprocessing

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.

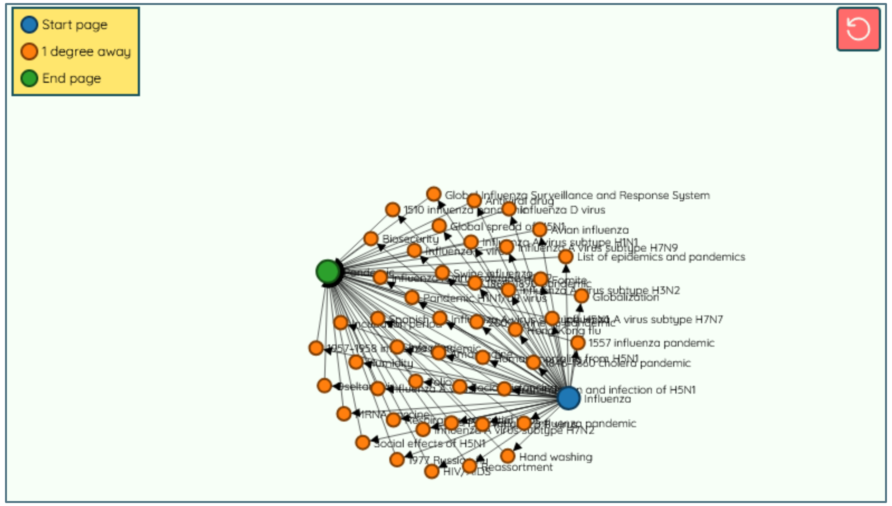
With regard to the selection of the most relevant Wikipedia articles, several approaches are available, which involve analyzing their relationships based on hyperlinks, categories, content similarity, or a combination of these factors.

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

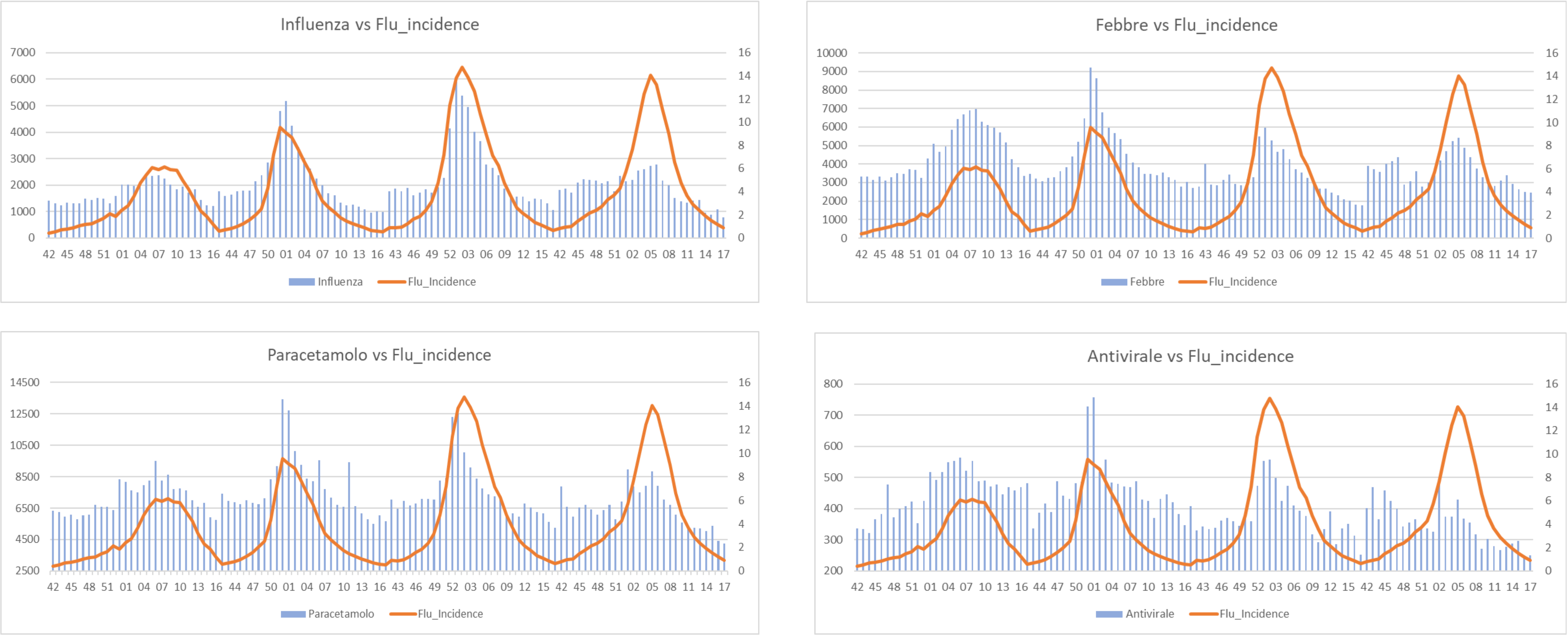
The selected pages were curated based on **thematic relevance**, with publications and prior studies on influence rate prediction reviewed to guide the selection of topics most likely to contribute to accurate modeling.
After this selection, distance metrics were applied as methods of verification and 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)



Influenza vs Flu_incidence

Febbre vs Flu_incidence

Paracetamolo vs Flu_incidence

Antivirale vs Flu_incidence

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:

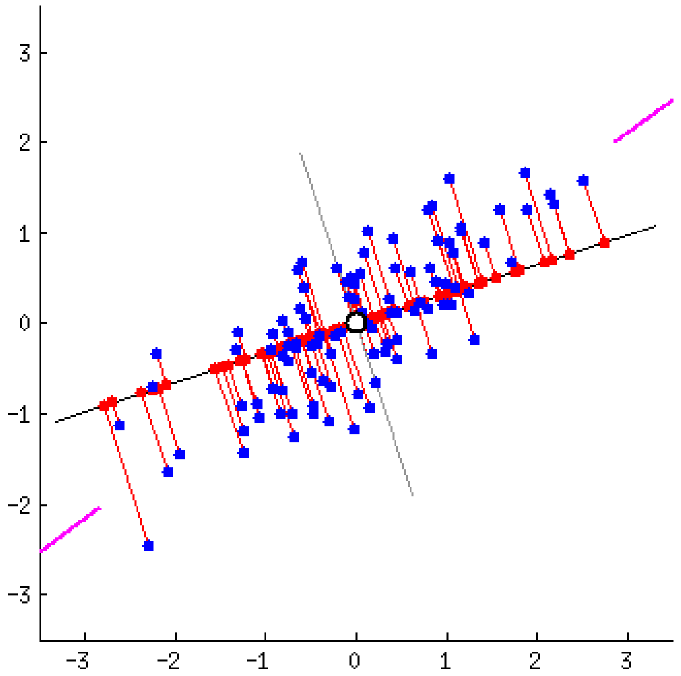
1. **loss of interpretability**: the principal components are often difficult to interpret in terms of the original variables;
2. **assumes linearity**: PCA assumes linear relationships between variables, which may not always be the case;
3. **sensitive to outliers**: PCA can be significantly affected 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



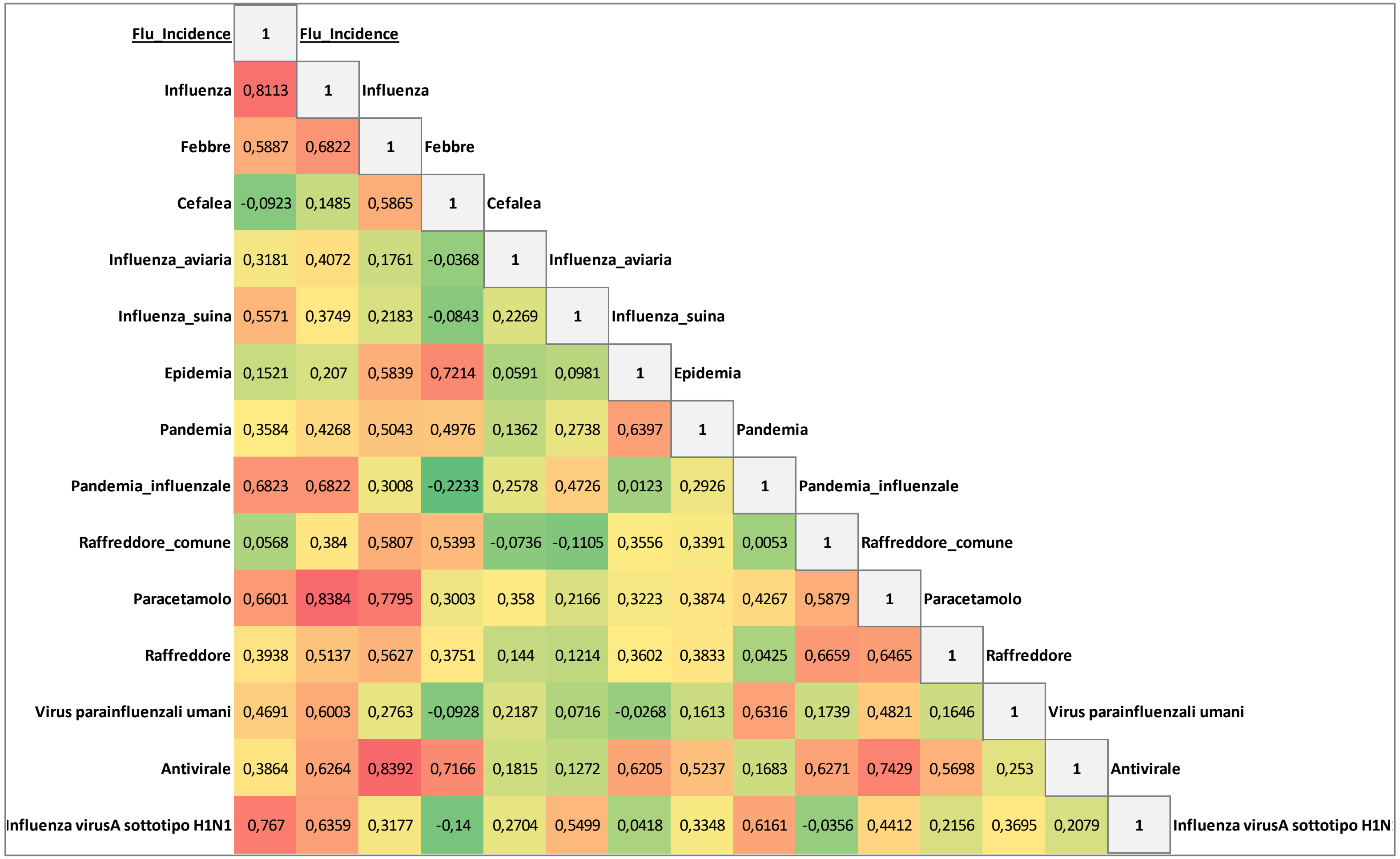
4. Model building: principal component regression (PCR)

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.

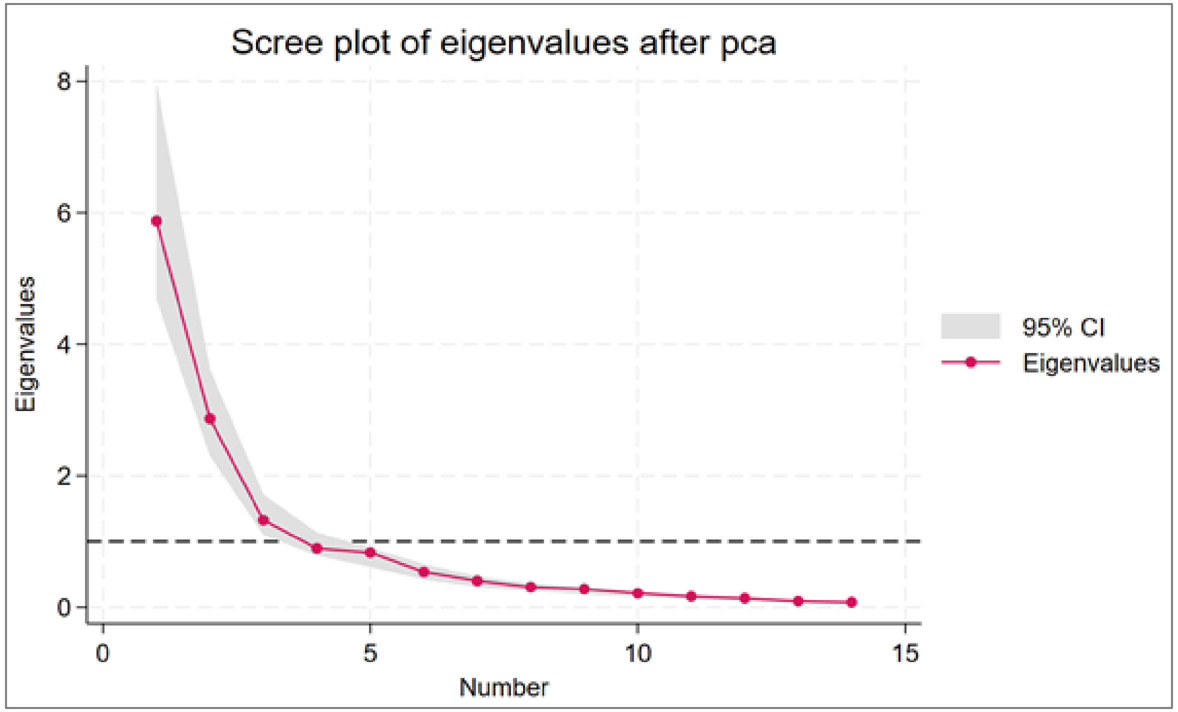


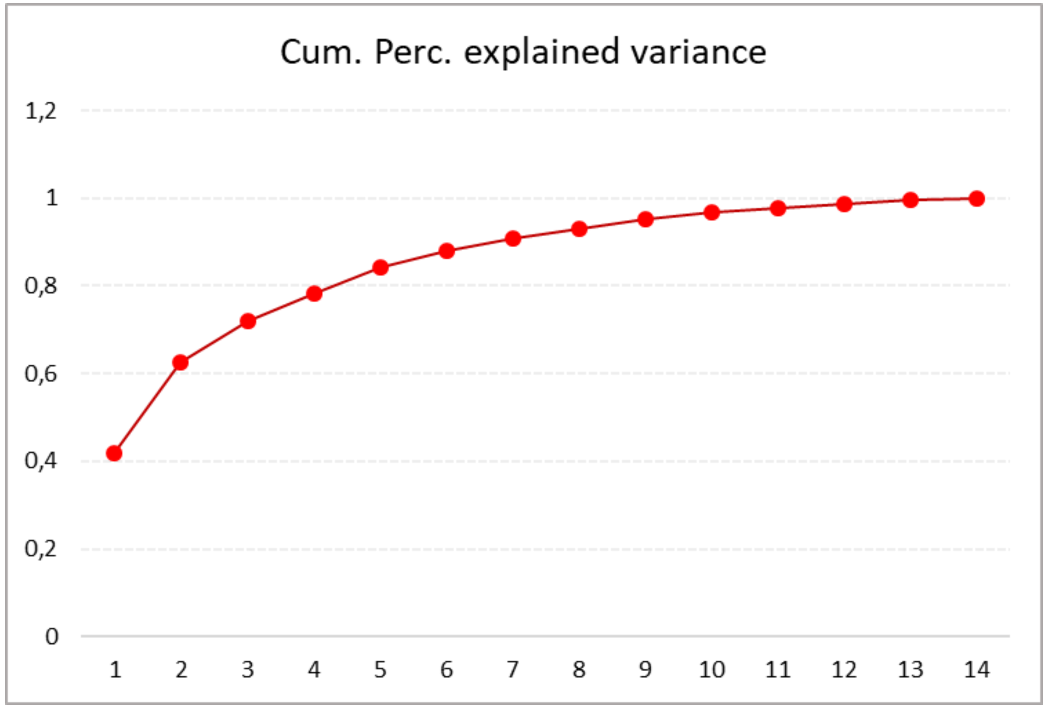
	Flu_Incidence	Influenza	Febbre	Cefalea	Influenza_aviaria	Influenza_suina	Epidemia	Pandemia	Pandemia_influenzale	Raffreddore_comune	Paracetamolo	Raffreddore	Virus parainfluenzali umani	Antivirale	Influenza virusA sottotipo H1N
Flu_Incidence	1														
Influenza	0,8113	1													
Febbre	0,5887	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.

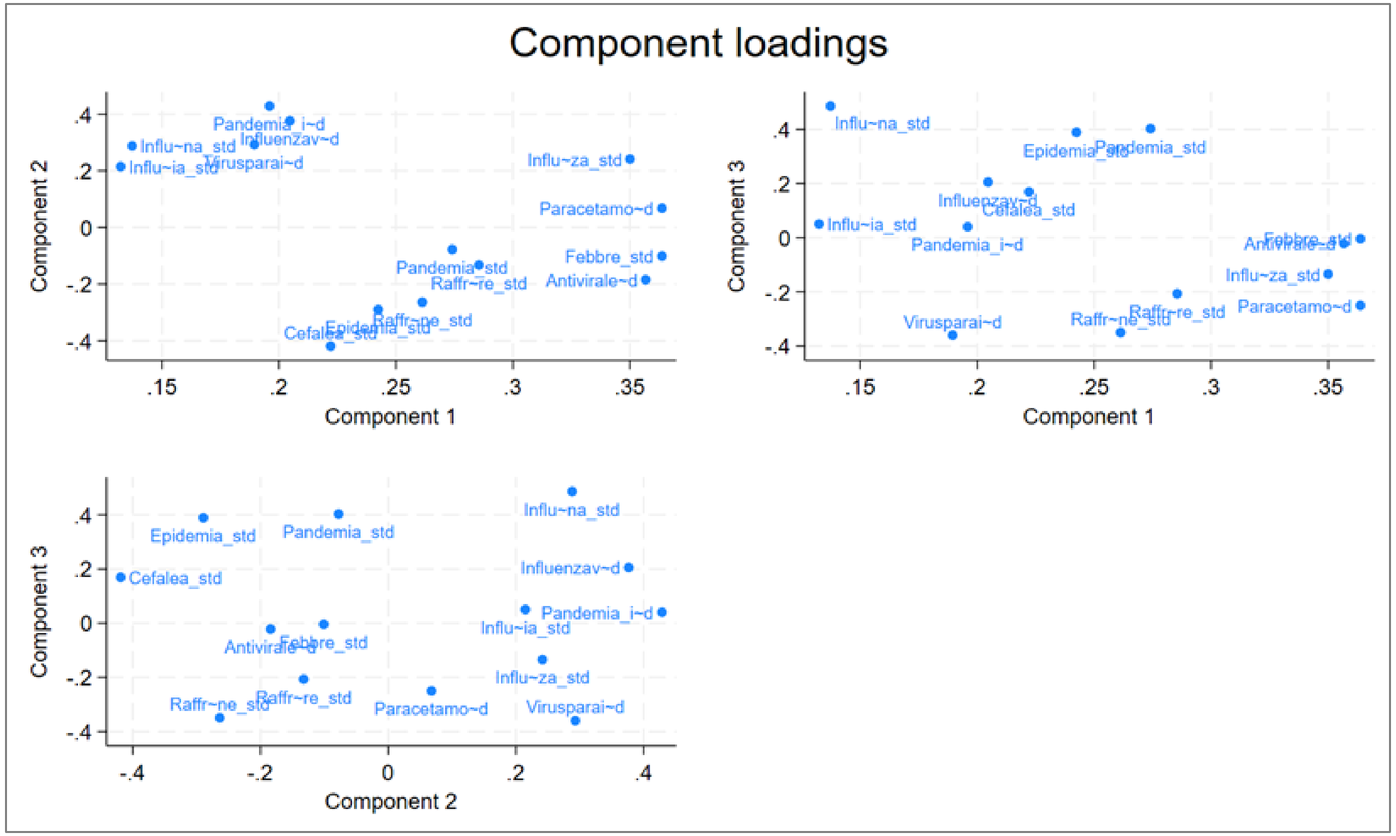
1. **Kaiser Criterion:** eigenvalue > 1;
2. **Cumulative Variance Explained:** select enough components to achieve a cumulative variance explained threshold, typically 70–90%;
3. **Scree Plot**: i.e. plotting the eigenvalues in descending order and look for the "elbow point" where the eigenvalues level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

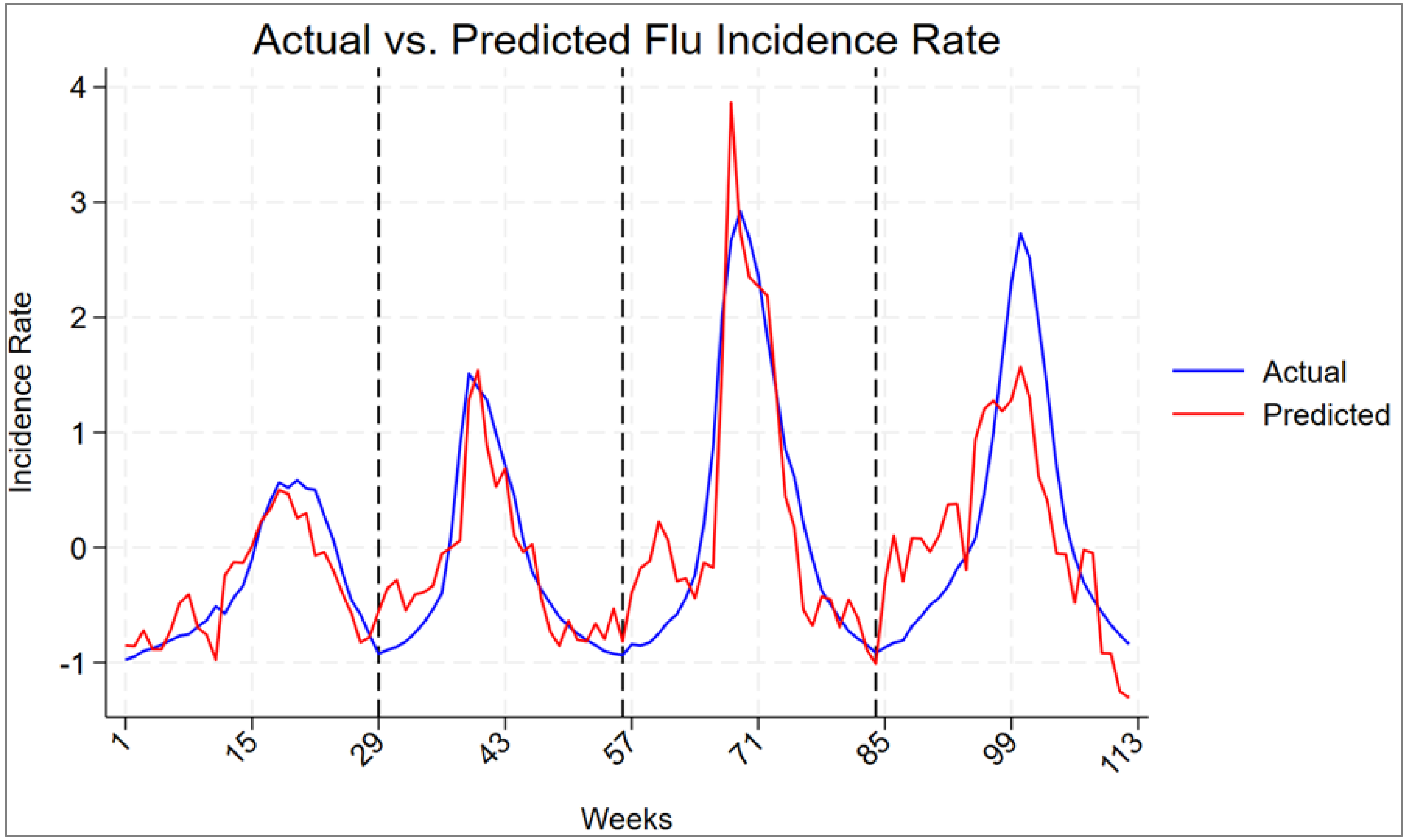


Variable	KMO
Influenza	0,8414
Febbre	0,9196
Cefalea	0,7868
Influenza_aviaria	0,7249
Influenza_suina	0,7365
Epidemia	0,7918
Pandemia	0,8715
Pandemia_influenzale	0,704
Raffreddore_comune	0,7933
Paracetamolo	0,8687
Raffreddore	0,8122
Antivirale	0,8888
Virus parainfluenzali umani	0,8155
Influenza virus A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

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



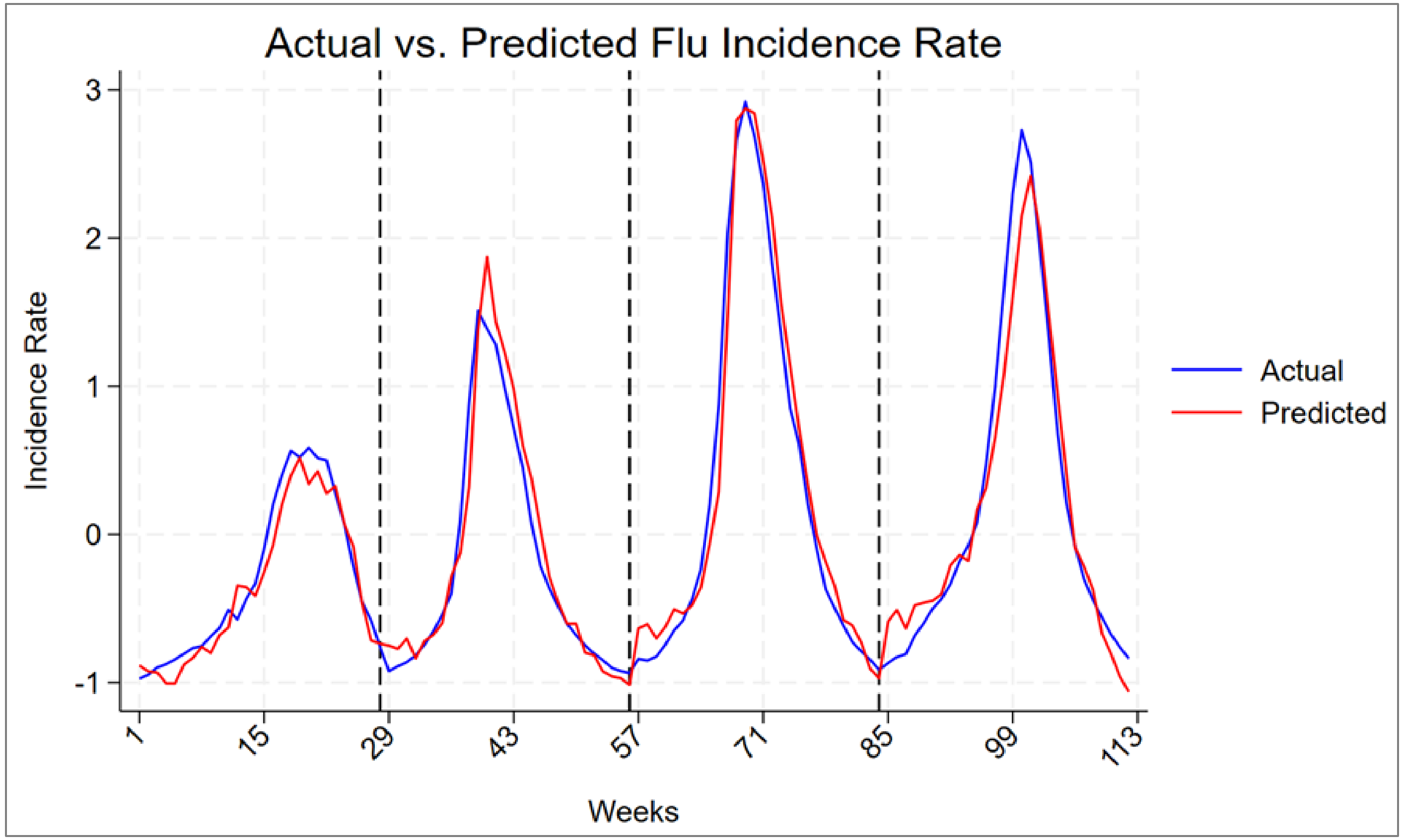
$\boldsymbol{r^2 = 0.7723}$

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