Regional differences in modelling Covid-19 infections using Google Trends data: evidence from Hungary

László Kovács
Corvinus University of Budapest,
Department of Statistics,
Budapest, Hungary
Email:
laszlo.kovacs2@uni-corvinus.hu

Fanni Büki
Corvinus University of Budapest,
Department of Statistics,
Budapest, Hungary
Email: fanni.buki@uni-corvinus.hu

This study models the spread of Covid-19 based on internet search data at Hungary's NUTS 2 regional level. Using a modified version of the search index composition method proposed by Li et al. (2020), adapted to Hungarian data, we explore the regional predictability of Covid-19 infections and examine its correlation with socio-economic variables. Our findings indicate significant regional variations in the effectiveness of internet search data for predicting coronavirus case numbers.

We identify key Covid-19-related Google search terms relevant to the Hungarian context and assess their predictive power using ordinary least squares and Bayesian vector autoregression models, with the susceptible–infectious–recovered model serving as a benchmark.

Our findings suggest that internet search data can be a valuable tool for predicting the regional spread of Covid-19, though predictability varies. Forecasting infection numbers based on internet search data is more successful in regions with high health awareness influencing internet search behaviour. The case of Northern Hungary demonstrates that internet search-based forecasting can also be effective in areas where the population relies on the internet as an alternative source for health-related information.

This study highlights the importance of regional analysis in pandemic forecasting, offering new insights for public health strategies and contributing to the literature on the applicability of internet search data.

Keywords:
Covid-19,
Google Trends,
Bayesian VAR model,
SIR model,
NUTS 2

Online first publication date: 30 October 2025

Introduction

This study aims to predict the spread of the Covid-19 pandemic based on internet search data at the NUTS 2 regional level in Hungary. Our forecasting method is derived from a version of the search index composition created from internet search terms, as proposed by Li et al. (2020), and adapted to Hungarian data. The primary objective of this study is to examine the significance of regional differences in the predictability of Covid-19 spread. The study demonstrates that the effectiveness of using internet search data to predict the spread of the coronavirus can vary significantly across regions. These regional differences in predictability correlate with territorial patterns of healthcare coverage (such as the number of people per pharmacy, average length of hospital stays and number of doctors per capita), endemic disease incidences (most notably diabetes) and specific demographical properties (such as the ageing index and the proportion of people with reduced working capacity).

Although we currently live without strict restrictions, future measures similar to those implemented during the Covid-19 pandemic may become necessary due to the emergence of a new Covid-19 variant or a virus strain with similar behaviour. In such cases, a tool capable of estimating virus spread at a regional level could be invaluable for determining the timing and severity of necessary interventions.

In recent years, especially since the outbreak of the Covid-19 pandemic, Google Trends and other search engine search volume data received increasing attention and have been utilised in various studies to predict the spread of different infectious viruses (Ben et al. 2022, Fantazzini 2020). Our study uses internet search volume data from Hungarian Google Trends for forecasting. Unlike official statistics, internet search data are available almost in real-time, which is an advantage for forecasting purposes. Based on our review of Hungarian and international literature, this analysis using Hungarian data has not been conducted before, making this research unique domestically. Furthermore, the regional differences identified in the study could provide new insights into the international literature regarding the applicability of internet search data in predicting the spread of pandemics.

In this study, the authors review the literature on internet search-based pandemic spread forecasting, focusing on the findings obtained in the case of the Covid-19 pandemic. Studies examining regional inequalities in Covid-19 are also presented. Then, a detailed description of the data sources used in the research is provided. After this, the authors introduce their methods for statistical analysis, including the data transformations performed and the forecasting models applied, followed by a presentation of the findings of the study and the conclusions drawn from them. Finally, the most important findings of the research are summarized, highlighting the limitation of the analysis, and proposing further study directions based on the findings.

Literature review

The application of internet search data to forecast the spread of various infectious diseases is well-documented in international literature. McGough et al. (2017) used internet search data to predict the spread of the Zika virus in Latin America, highlighting its potential as a supplementary real-time tracking tool. The study found that the effectiveness of search data varies by country, emphasising the need for region-specific evaluations in epidemic forecasting.

Xu et al. (2019) observed that both the approach and method can impact the interpretability and usability of the findings. Their study on the relationship between Chinese citizens' internet searches related to cancer and the number of cancer-related deaths did not apply lagged variables, limiting the reliability of the findings.

Austys–Burneikaitė (2019) forecasted flu epidemics in Lithuania using Google Trends search volume indexes. They translated commonly used influenza-related search terms from studies in other languages into Lithuanian, determined cross-correlations between search volumes and flu cases and built an ARIMA model incorporating lagged usage frequency of search terms as exogenous variables. Technically, weekly new infection numbers were used, and a SARIMAX(1,0,0)(0,1,0)[52] model was fitted to the latest infection numbers, with the exogenous variable being the search term frequency with the strongest cross-correlation. Thus, the seasonal unit root arising from weekly seasonality in the number of infection cases was handled by seasonal differencing. Their findings suggest that Google Trends data could help predict the progression of flu epidemics, underscoring its potential utility for Hungary given its similar socio-economic background (HCSO 2022, Mihályi 2014). Furthermore, Hungary's larger population and land area allows for a sub-national regional examination of the forecasting accuracy of internet search data for epidemic spread, as more sub-national differences can be observed in socio-economic status and Covid-19 infections (Oroszi et al. 2022a, 2022b, Uzzoli et al. 2021).

Huang et al. (2020) used the Baidu search index to predict cases of HIV/AIDS, syphilis and gonorrhoea in China. A search index composition was created and applied in a vector autoregression (VAR) model. The forecasts indicated that Baidu search data could predict new infections. This search index composition forms the basis for our approach to Google Trends data and is introduced in detail in chapter Methods. In this study, the VAR model is fitted over a more extended time period, from January 2011 to October 2016. Monthly Baidu search frequencies for keywords related to HIV/AIDS, syphilis and gonorrhoea are considered for the search index composition, alongside the monthly number of new cases. The fitted VAR model is then used to forecast the number of new cases from November 2016 to October 2017. The out-of-sample R^2 were all above 85%, demonstrating the strong fit and predictive performance of the Baidu-augmented VAR model. Similarly, Li et al. (2019) developed an ARIMAX model using Baidu search data to predict HIV/AIDS cases. The model is

fitted to the monthly number of new HIV/AIDS cases, with the exogenous variable being the search index composition derived from the search volume of HIV/AIDS-related terms on Baidu. Technically, a SARIMAX(0,1,2(1,0,0)[12] model is fitted; thus, the unit root in the number of monthly new cases is accounted for by modelling the first differences. Their findings showed that incorporating search data improved in-sample accuracy, measured by the Akaike information criterion (AIC), compared to traditional ARIMA models.

Regarding the Covid-19 pandemic, studies on regional differences in internet search-based epidemic forecasting have also emerged. Saegner–Austys (2022) provides a detailed overview of research on Google Trends-based forecasting of Covid-19 spread. Based on 54 studies, they concluded that most found a positive correlation between internet interest in the coronavirus and the number of new infections. Specifically, 45 of the 54 reported a positive correlation between search frequencies for keywords such as 'coronavirus' and 'Covid-19' (with variations in keyword selection across studies), and the number of daily/weekly new coronavirus cases (with differences in time series frequencies across studies). The methods applied in these studies ranged from simple cross-correlations to machine-learning models (like random forest regressions). However, all employed some form of time series modelling, whether linear (ARIMAX, VAR) or non-linear (random forest regressions, long short-term memory neural networks). Notably, the positive findings were predominantly observed in high-income countries with high internet penetration. Negative findings were often attributed to modelling on very short time series during the early stages of the pandemic. At that time, Google searches related to the coronavirus focused more on media coverage than substantive information about the virus, which may explain the negative findings in the literature.

Both Saegner–Austys (2022) and our literature review indicate that no study has substantially examined the applicability of Google Trends for predicting the spread of the coronavirus at a sub-country level using time series methods. For instance, using simple static correlation analysis, Mavragani–Gkillas (2020) assessed the applicability of Google Trends for predicting Covid-19 spread across different U.S. states. Moreover, their quantile regression analysis did not examine residual autocorrelation, which could lead to spurious regression. Rovetta (2021) demonstrated the time dependence of contemporaneous Pearson and Spearman correlation coefficients between Google search volumes for coronavirus-related terms and the number of infections using regional data from Italy. The study investigated the differences in contemporaneous correlation coefficients between search volumes and the number of new infections from 1 February–4 December 2020 (period 1) and 20 February–18 May 2020 (period 2). The findings indicated that correlations changed significantly, at least at the 10% level, for regional search term volumes between the two periods but did not change significantly when worldwide search volumes of the same terms were examined.

Ben et al. (2022) analysed gastrointestinal symptom-related search terms to forecast Covid-19 outbreaks in the U.S., U.K., Australia and China, identifying varying optimal lags for forecasting across countries. The two symptoms for which volumes had the strongest correlation with daily newly confirmed Covid-19 cases – at lags 9 and 12 days at the global level – were 'fever' and 'cough'. Meanwhile, 'diarrhoea' and 'loss of taste' displayed the highest correlation with daily newly confirmed cases at lags of 12 and 5 days, respectively, at the global level. These symptoms exhibited similar lag correlation patterns with daily newly confirmed Covid-19 cases at the country level. However, the specific lag days varied across countries (see Table 2 of the study for country-specific lags). The authors also noted that their study did not account for the potential within-country heterogeneity of their findings. Our study aims to address this gap by examining within-country heterogeneity in Hungary.

Fantazzini (2020) used data from 158 countries to forecast Covid-19 cases with a two-week forecasting lag, demonstrating the utility of Google Trends data. However, the study did not examine the shared characteristics of countries where Google Trends data improved forecasting accuracy. The authors developed Google Trends-augmented ARIMAX and VAR models and simple ARMA, ETS and SIR models for benchmarking. Daily data from January to March 2020 were used as the training sample for model estimation, while April to May 2020 was reserved for out-of-sample forecasting. The mean squared errors show that Google-augmented time series models performed best for 86 out of 158 countries. Our study identifies the common socio-economic characteristics of Hungarian NUTS 2 regions, described using the variables introduced in chapter Data, where Google Trends data are effective for out-of-sample forecasting of coronavirus spread.

Most studies on regional differences in the coronavirus pandemic focus on mortality. Using cluster analysis, Bucci et al. (2023) examined daily mortality time series for the coronavirus across European NUTS 2 regions. Their study confirms the importance of analysing coronavirus mortality patterns at a sub-country level, as these patterns can vary significantly within a country. Igari (2023) highlighted differences in excess mortality across European NUTS 3 regions during various pandemic waves.

Numerous studies note that in Central and Eastern European regions, the population's health status is generally poorer, and the healthcare system is less accessible, increasing the risk of excess mortality due to the coronavirus (Kovács–Uzzoli 2020, Uzzoli et al. 2020, 2021, Kovalcsik et al. 2021). Generally, regional differences are primarily influenced by the phenomenon that regions with lower socio-economic status tend to have less developed healthcare systems (Bambra et al. 2020, McGowen–Bambra 2022, Munford et al. 2022).

In Hungary, coronavirus mortality patterns resemble the territorial trends observed in other Central and Eastern European countries (Karlinsky–Kobak 2021,

Tóth 2022). Studies have primarily focused on demographic and comorbidity factors (Bíró et al. 2021, Elek et al. 2022, Gombos et al. 2020). Kovács–Vántus (2022) highlighted the role of healthcare capacity, while Oroszi et al. (2022a, 2022b) and Uzzoli et al. (2021) noted that lower socio-economic regions face challenges in testing capacity and experience higher mortality risks. Páger et al. (2024) identified spatial autocorrelation of mortality with hotspots in less urbanised northeastern and central Hungarian regions.

Almost all reviewed studies examining the territorial patterns of coronavirus mortality suggest regional level, localised rather than national epidemic management measures due to significant within-country differences in coronavirus mortality. Our study can effectively support such measures by conducting a regional examination of the predictability of case numbers based on internet search data. Our literature review shows no such regional study has been conducted in Central–Eastern Europe. Furthermore, in the broader international literature, epidemic spread forecasting based on related internet searches primarily relies on static (cross-sectional), correlation-based methods that do not account for lagged effects or unit roots. Consequently, the regional characteristics of internet search-based epidemic spread forecasts revealed by this study can serve as a reference for further similar research in other countries.

Data

The data used in this study come from two primary sources. County-level internet search term volume data were obtained from the Google Trends website, while county-level daily new coronavirus case numbers were sourced from *atlatszo.hu*'s *Koronamonitor* application. The analysis focuses on the pandemic's second, third and fourth waves, covering the period from 1 October 2020 to 28 February 2022. Following weekly aggregation, the datasets comprise 73 observations [1], [2]. The first wave was excluded from our analysis because strict epidemiological measures kept case numbers and mortality in Hungary significantly lower than in subsequent waves [1] (Igari 2023, Oroszi 2022b). Thus, including it would have introduced a structural break in the time series, necessitating a separate model. However, the weekly aggregated time series for the first wave would contain only 26 observations (from 31 March 2020, to 30 September 2020), which is insufficient for reliable parameter estimation in a VAR model (Kirchgässner et al. 2013).

Weekly aggregation was primarily necessary because, after 11 June 2021, daily new case numbers were unavailable on weekends [1]. Imputing the weekend data using an imputation method would have introduced the assumption that the number of new cases behaves similarly on weekdays and weekends or that these missing weekends behave equally to the weekend data observed in the first part of the time series. To avoid these assumptions, we opted to analyse the data every week. Furthermore,

sub-country Google Trends data can become quite noisy daily (e.g. consecutive 0 periods), which can also be mitigated using weekly frequency. Alternatively, mixed-frequency methods such as mixed-frequency data sampling regression (MIDAS) can be applied to retain a daily frequency for Google Trends data and/or a weekend-imputed daily new infections data (Ghysels et al. 2016). However, this is not the approach adopted by the international literature on epidemic forecasting using Google Trends data. The studies reviewed all applied weekly or monthly aggregated time series for periods exceeding one year (McGough et al. 2017, Austys–Burneikaitė 2019, Li et al. 2019, Huang et al. 2020, Fantazzini 2020, Saegner–Austys 2022, Ben et al. 2022, Llewellyn et al. 2023). Our study examines the differences in out-of-sample forecasting accuracy of these established methods at a sub-country level. That is why we have chosen to apply weekly aggregation.

The socio-economic status of the regions was characterised using economic, demographic and health variables for the year 2021 from the Hungarian Central Statistical Office (HCSO) website [3]. The specific range of variables was determined based on literature examining the socio-economic factors influencing the spread and mortality of coronavirus in Hungary (Tóth 2022, Bíró et al. 2021, Elek et al. 2022, Páger et al. 2024). The variables include:

- demographic variables
  - population density as of 31 December 2021 (people/km²)
  - ageing index (number of elderly individuals aged 65 years and over per 100 individuals younger than 14 years) as of 31 December 2021 (%)
  - age-standardised mortality rate per 100,000 inhabitants in 2021 (per mille)
  - life expectancy at birth for women in 2021 (years)
  - life expectancy at birth for men in 2021 (years)
- economic variables and internet penetration
  - gross domestic product per capita in 2021 (thousand HUF)
  - average monthly old-age pension per capita in 2021 (HUF/month)
  - internet subscriptions per 1,000 inhabitants in 2021 (units)
- health status related variables
  - percentage of the population receiving benefits for reduced working capacity in 2021 (%)
  - proportion of hypertension patients registered with general practitioners per 10,000 inhabitants aged 19 and over in 2021 (cases/10,000 inhabitants)
  - proportion of ischaemic heart disease (I20–I25) patients registered with general practitioners per 10,000 inhabitants aged 19 and over in 2021 (cases/10,000 inhabitants)
  - proportion of diabetes patients registered with general practitioners per 10,000 inhabitants aged 19 and over in 2021 (cases/10,000 inhabitants)
- healthcare system related variables
  - number of working doctors per 10,000 inhabitants in 2021 (persons)

- number of operational hospital beds per 10,000 inhabitants in 2021 (units/person)
- average length of hospital care in 2021 (days)
- number of inhabitants per pharmacy in 2021 (persons/pharmacy).

For Google search volume data, we used the search terms identified by Li et al. (2020) as a starting point. However, our findings indicated that only a fraction of the search terms from Li et al. (2020) generated interest among Hungarian internet users. These included coronavirus symptoms and Covid-19 symptoms. For other terms, only Budapest had sufficient data for analysis. Therefore, we used alternative Covid-19-related terms that were more frequently searched across Hungary [2].

The most frequent Hungarian search terms in Google Trends for the studied period are as follows:

- coronavirus symptoms/covid symptoms (hereinafter: symptom)
- coronavirus testing/covid testing (hereinafter: test)
- coronavirus incubation period/covid incubation period (hereinafter: incubation)
- coronavirus news/covid news (hereinafter: news)
- coronavirus medicine/covid medicine (hereinafter: medicine)
- delta variant (hereinafter: delta)
- post covid (hereinafter: post-covid)
- PCR test (hereinafter: PCR).

In the Hungarian language, 'coronavirus' and 'covid' are primarily used interchangeably; therefore, we treated such cases as identical terms. This means that when someone searches for 'coronavirus symptoms' and 'covid symptoms', they generally intend the same meaning. Consequently, the search volumes of these two separate search terms were combined. The search volume indices for each term, grouped by region, are presented in the appendices (see in Appendix Figure A1).

County-level (NUTS 3) observations are available on Google Trends, *Koronamonitor* and the HCSO websites. For some search terms, certain counties either lacked sufficient search data for Google Trends to display them or had highly incomplete data at several time points. To avoid consistently zero search volume time series over long periods, we aggregated the search volume indices of the search terms to the NUTS 2 region level, treating Budapest as a separate unit. Mixed-frequency models for panel data, as proposed by Yang et al. (2023), can also be considered to address regional data observed at different frequencies. A decision to aggregate, such as in the case of weekly aggregation, aligns with the approach taken by most research using Google Trends data for epidemic forecasting (McGough et al. 2017, Austys–Burneikaitė 2019, Li et al. 2019, Huang et al. 2020, Fantazzini 2020, Saegner–Austys 2022, Ben et al. 2022, Llewellyn et al. 2023).

Methods

In our study, we apply the search index composition proposed by Li et al. (2020) and Huang et al. (2020) to quantify Google search volume for the coronavirus topic across Hungarian regions. Li et al. (2020) examined the out-of-sample R^2 of the search index composition concerning the Covid-19 pandemic using an ARIMAX(0,1,2) model. The method itself is mainly consistent with those used in studies forecasting the number of HIV/AIDS, syphilis and gonorrhoea cases (Huang et al. 2020, Li et al. 2019). Thus, the method is robustly applicable to forecasting the spread of various epidemics based on internet searches.

When compiling the composition, the goal is to include the most relevant search terms for the topic, in this case, Covid-19. In the study by Li et al. (2020), a combination of several methods was used to determine the appropriate terms. Primarily, potential keywords were collected from a Chinese website, with additional terms derived from a semantic correlation analysis. Examples of webpages used in the study include Baidu (the most popular search engine in China), Microblog, Wikipedia and WeChat. This approach enabled the collection of 14 terms reflecting search behaviour related to Covid-19. Since not all terms were directly related to the Covid-19 virus, the sample was further refined to relevant keywords in two steps.

First, the Spearman rank correlation coefficient between the Baidu search index (as Baidu is the largest search engine in China, not Google) and the presumed case numbers were examined for each term. Based on the findings, terms with a rank correlation coefficient smaller than 0.4 were removed from the collection. Then, the cross-correlation between the different lags of the search terms and the coronavirus case numbers was analysed. The lag of each term that exhibited the highest correlation with the case numbers was selected (Huang et al. 2020).

We examine the Google Trends search volume indices for the terms specified in chapter Data at the NUTS 2 level in Hungary, using the two filtering methods described above, from 1 October 2020, to 28 February 2022.

In the study by Li et al. (2020), all the lags of the five selected search terms were used to construct the search index composition. The weight of each term was determined based on the highest absolute value of the correlation coefficient (ϱ) for each term, measuring the effect size of a given term.

Equation (1) shows the calculation of the weights:

$$weight_{ki} = \frac{\rho_{ki}}{\sum_{i=1}^{n} \rho_{ki}} \quad (1)$$

where k is the potential lag, n is the number of keywords and ρ_{ki} is the correlation coefficient of the kth lag of the ith keyword (Huang et al. 2020). Equation (2) shows the search index composition:

$$SearchIndex_k = \sum_{i=0}^{n} weight_{ki} \times keyword_{ki} \quad (2)$$

where k is the potential lag, n is the number of keywords, $weight_{ki}$ is the weight of the kth lag of the ith keyword and $keyword_{ki}$ is the search volume index of the kth lag of the ith keyword for a given period (Li et al. 2020).

A bivariate VAR model describes the bivariate relationship between the internet search index and the weekly new coronavirus case numbers. This model is based on the concept that both time series are endogenous, meaning each has its own equation within the framework. In both equations, the lagged values of the explanatory variables are included up to a predetermined lag k (Woodward et al. 2017). Thus, equations (3) and (4) show the bivariate VAR model:

$$Y_t = \beta_0 + \beta_{11}Y_{t-1} + \beta_{12}Y_{t-2} + \ldots + \beta_{1k}Y_{t-k} + \beta_{21}X_{t-1} + \beta_{22}X_{t-2} + \ldots + \beta_{2k}X_{t-k} + u_t \quad (3)$$

$$X_t = \alpha_0 + \alpha_{11}Y_{t-1} + \alpha_{12}Y_{t-2} + \ldots + \alpha_{1k}Y_{t-k} + \alpha_{21}X_{t-1} + \alpha_{22}X_{t-2} + \ldots + \alpha_{2k}X_{t-k} + v_t \quad (4)$$

where Y_t and X_t are the internet search index and the time series of weekly new coronavirus cases, respectively, u_t and v_t are the residual or error terms, t is the current time point and k is the number of lags (Woodward et al. 2017).

The model coefficients are first estimated using the ordinary least squares (OLS) method. According to the fundamental assumption of OLS estimation, the error terms (u_t and v_t) jointly behave as white noise. The Portmanteau test is appropriate for assessing this assumption, with the null hypothesis (H_0) stating that the error terms can be considered white noise, which means that an error term is neither correlated with its lags nor the other error terms. The second assumption of the OLS-based VAR model is that the input time series must be stationary. This condition is crucial for avoiding spurious correlations (Kirchgässner et al. 2013). The stationarity of the input time series is tested using the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) and augmented Dickey–Fuller (ADF) tests. If a time series is non-stationary, it can often be rendered stationary through differencing. Additionally, the stability of the VAR models is assessed by examining the roots of the characteristic polynomials of the equations. The model is considered stable if all roots of the characteristic polynomial for each VAR equation lie within the unit circle (Hamilton 2020).

The maximum lag k of the model is determined using information criteria. The Akaike, Hannan–Quinn and Bayes–Schwarz information criteria are applied jointly and align with standard practice. During modelling, a lag is considered optimal if recommended by at least two criteria. In cases where all three criteria suggest different lags, the Hannan–Quinn criterion is adopted as a compromise, as it provides a balanced choice between the two extreme cases.

To assess the robustness of the findings, model equations (3) and (4) with the optimal k lags identified by the information criteria, are also estimated using a Bayesian approach. Weakly informative priors are employed to introduce regularisation, following the method proposed by Gelman–Hill (2007). Assuming that the error terms $u_t \sim N(0, \sigma_u)$ and $v_t \sim N(0, \sigma_v)$ are independent (jointly white noise), the posterior distribution of the model parameters, $p(\alpha_0, \alpha, \beta_0, \beta, \sigma | Y_t, X_t)$, s derived using Bayes' theorem (5).

$$p(\alpha_0, \alpha, \beta_0, \beta, \sigma | Y_t, X_t) = \frac{p(Y_t, X_t | \alpha_0, \alpha, \beta_0, \beta, \sigma) \cdot p(\alpha_0, \alpha, \beta_0, \beta, \sigma)}{p(Y_t, X_t)} \quad (5)$$

where $\alpha = [\alpha_{11}, \dots, \alpha_{2k}]$, $\beta = [\beta_{11}, \dots, \beta_{2k}]$ and $\sigma = [\sigma_u, \sigma_v]$. $p(Y_t, X_t | \alpha_0, \alpha, \beta_0, \beta, \sigma)$ is the model likelihood function, $p(Y_t, X_t)$ is the empirical distribution of our time series, and $p(\alpha_0, \alpha, \beta_0, \beta, \sigma)$ is the prior distribution of parameters. These latter prior distributions are defined as weakly informative, following Muth et al. (2018).

$$\sigma_u \sim Exp(s_y) \quad (6)$$

$$\sigma_v \sim Exp(s_x) \quad (7)$$

$$\alpha_{ij}, \sim N(0, 2.5 \cdot s_x / s_y) \quad (8)$$

$$\beta_{ij}, \sim N(0, 2.5 \cdot s_y / s_x) \quad (9)$$

$$\alpha_0 \sim N(\bar{x}, 2.5 \cdot s_x) \quad (10)$$

$$\beta_0 \sim N(\bar{y}, 2.5 \cdot s_y) \quad (11)$$

where s_y and s_x are the empirical standard deviations of the time series, applied to scale the priors for the observed data's distribution. Definitions (6)–(11) are weekly informative, as the expected value of the model parameters is 0, with a moderately large standardised standard deviation of 2.5. Consequently, they provide some regularisation of the model effects reflected in the α, β parameters. The posterior distribution of (5) is determined numerically using Markov Chain Monte Carlo simulation. The empirical mean of the simulated posterior distribution can be used as the point estimate for the model parameters when predicting the number of infections. If the observed data do not support the hypothesis that the internet search index can help predict the number of infections, then the empirical mean of the simulated posterior parameter distribution will not differ substantially from 0.

As a benchmark model for predicting the number of weekly new Covid-19 infections without the internet search index, the susceptible–infectious–recovered (SIR) model is applied (Brauer–Castillo-Chavez 2012). The fundamental principle of the SIR model is that susceptible individuals (S) become infected with a rate of β as they encounter already infected individuals (I). Thus, β can be interpreted as the infectious contact rate. Additionally, infected individuals recover (R) at a rate of γ. This system is a set of differential equations, as given in equation (12).

$$\frac{dS}{dt} = -\beta \cdot I \cdot S \quad (12)$$

$$\frac{dI}{dt} = \beta \cdot I \cdot S - \gamma \cdot I \quad (13)$$

$$\frac{dR}{dt} = \gamma \cdot I \quad (14)$$

Once equations (12)–(14) are solved numerically, we can predict the number of new infections for any given t time, denoted as I_t, which can serve as a benchmark estimation for the number of new infections without using internet search data. However, this method assumes that the parameters β, γ are known. We can estimate these parameters with OLS by solving equation (15) on the observed time series of length T.

$$\min_{\beta,\gamma} \sum_{t=1}^{T} \left(X_t - I_t(\beta,\gamma)\right)^2 \quad (15)$$

where $I_t(\beta,\gamma)$ is the estimated number of new infections from the solved SIR system of equations (12)–(14) with given β,γ parameters (Harko et al. 2014).

In our statistical analysis process, after preparing the data and examining basic assumptions such as stationarity, we first analyse cross-correlations between the Google search volume indices of individual terms and weekly new coronavirus cases for each region. We allow for a maximum of seven lags, while the number of observations permits a maximum of $73/5 = 14.6 \sim 14$ variables in the VAR model. Beyond this, the model would become overfitted (too many parameters relative to the number of observations) and thus unusable for forecasting (Hamilton 2020). For example, in a VAR model with two endogenous variables and seven lags, each equation contains 14 explanatory variables. The search terms corresponding to the strongest cross-correlations (absolute value greater than 0.3) for a given region were included in the region's search index composition. The search index composition and the weights of the individual terms were calculated for each region according to equations (1) and (2).

After constructing the search index composition, we built models by region. In each case, we fitted our three models using the first 65 observations and used the last eight observations (two months) to assess forecasting accuracy with an expanding estimation window. For example, when forecasting the number of new infections on week 70, weeks 1–69 are used to estimate model parameters in all cases. Forecasting accuracy for the number of weekly new infections is measured using the R^2 measure, defined as the squared correlation between the estimated and actual number of infections in the last eight weeks (Hastie et al. 2011).

Diagnostic information for the OLS VAR models is reported using the entire time series.

Results

First, the stationarity of each time series is examined. The p-values of the ADF and KPSS tests for the levels and first differences of the time series included in our analysis for Budapest are presented in Table 1. The findings for other regions are provided in the appendices, which also include graphical representations of the time series for each region.

Based on Table 1, the first difference transformed every time series to exhibit stationary behaviour according to the KPSS test at all common significance levels. However, the ADF test indicates that the time series of case numbers remains non-stationary at every common significance level, even after first differencing. This finding may represent a type II error, given the KPSS test findings. Therefore, following the principle of parsimony, we do not apply further transformations to the time series. If these ADF test findings pose a problem for the stability of the VAR

models, it will be evident from the roots of the characteristic polynomials. Similar conclusions can be drawn for other regions based on the comparable tables in the appendices.

Table 1

The p-values of stationarity tests for the levels and first differences of the examined time series in Budapest

Variable	Levels		First differences	
	ADF p-value	KPSS p-value	ADF p-value	KPSS p-value
Cases	0.626	<0.01	0.367	>0.1
Medicine	0.117	>0.1	<0.01	>0.1
News	0.230	<0.01	<0.01	>0.1
Incubation	0.024	>0.1	<0.01	>0.1
Test	0.073	0.067	<0.01	>0.1
Symptom	0.063	0.064	<0.01	>0.1
Delta	0.523	0.026	<0.01	>0.1
PCR	0.022	0.022	<0.01	>0.1
Post	0.418	0.084	<0.01	>0.1

After transforming the time series, the composition of the internet search index is determined, and the optimal OLS VAR and Bayesian VAR models for each region are constructed, along with the SIR model. Diagnostics for the OLS VAR models are summarised in Table 2.

Table 2

Model diagnostics of the regional OLS VAR models

Region	Optimal lag	Portmanteau multivariate white noise test p-value	Range of characteristic polynomial roots
Budapest	3	0.517	0.017–0.778
Pest	2	0.489	0.392–0.741
Western Transdanubia	1	0.935	0.019–0.319
Southern Transdanubia	2	0.824	0.317–0.635
Central Transdanubia	2	0.681	0.480–0.705
Northern Hungary	1	0.972	0.038–0.274
Northern Great Plain	2	0.934	0.368–0.667
Southern Great Plain	4	0.977	0.179–0.816

Based on Table 2, the VAR model appears to be diagnostically adequate across all regions: the residuals can be considered white noise at all common significance levels, and the roots of the characteristic polynomials lie within the unit circle. Consequently, no second differencing in the case numbers is necessary.

Out-of-sample R^2 measures for the three models over the last eight weeks are reported for each region in Table 3.

Table 3

Out-of-sample R-squared measures for all three models

(%)

Region	OLS VAR	Bayesian VAR	SIR	Gain of the best-performing VAR compared to SIR
Budapest	22.86	24.00	1.33	+22.67
Pest	8.03	9.19	79.99	−70.8
Western Transdanubia	5.02	0.17	71.78	−66.76
Southern Transdanubia	21.83	13.76	0.03	+21.8
Central Transdanubia	5.29	12.91	59.83	−46.92
Northern Hungary	14.53	11.79	8.48	+6.05
Northern Great Plain	11.05	1.78	0.01	+11.04
Southern Great Plain	33.26	11.14	0.47	+32.79

Note: regions where the out-of-sample forecasting accuracy is higher for the VAR models are highlighted with grey background.

The findings in Table 3 indicate that the VAR models enhanced with the internet search index outperform the classic SIR model in Budapest, Southern Transdanubia, Northern Hungary, Northern Plain and Southern Plain regions. Thus, in these regions, the volume of internet searches can be used to predict the number of infections more accurately. These findings are robust across estimation methods, as both OLS and Bayesian VAR models surpass the SIR model in these regions. Furthermore, we conclude that the SIR models perform generally poor in regions where VAR models overperform. One notable exception is Northern Hungary, where the out-of-sample performance of the SIR and VAR models is more similar (R^2 difference is within 10 percentage points). This suggests that the internet search index in Northern Hungary provides only a marginal predictive advantage for weekly new infections, or its effect may be non-linear, making this region an outlier.

To identify common characteristics of the regions where the internet search index enhances infection prediction, we examined the Spearman rank correlation between the percentage-point gain in R^2 (compared to the SIR model) of the best-performing VAR and 16 variables describing the socio-economic status of the regions. The rank correlation coefficient was used to detect potential non-linear relationships, whereas the classic Pearson correlation cannot (Zhang et al. 2016). Since Budapest exhibits extreme outliers in most socio-economic variables, we conducted the correlation calculations on a logarithmic scale. Our findings are presented in Table 4.

Based on Table 4, the R^2 gain for VAR models is strongly correlated (with an absolute correlation value of at least 0.7) with four variables: the number of people per pharmacy, the average length of hospital care and the number of doctors and diabetics per ten thousand people. Based on the signs of these correlations, it can be inferred that regions where internet search volume is a strong predictor of Covid-19 infections are characterised by a low number of people per pharmacy, a short average

length of hospital care and a high proportion of doctors and diabetics. These findings are further supported by the regional distribution of these three variables, as shown in Appendix Figure A2.

Table 4

Rank correlations of Granger causality test p-values with socio-economic variables

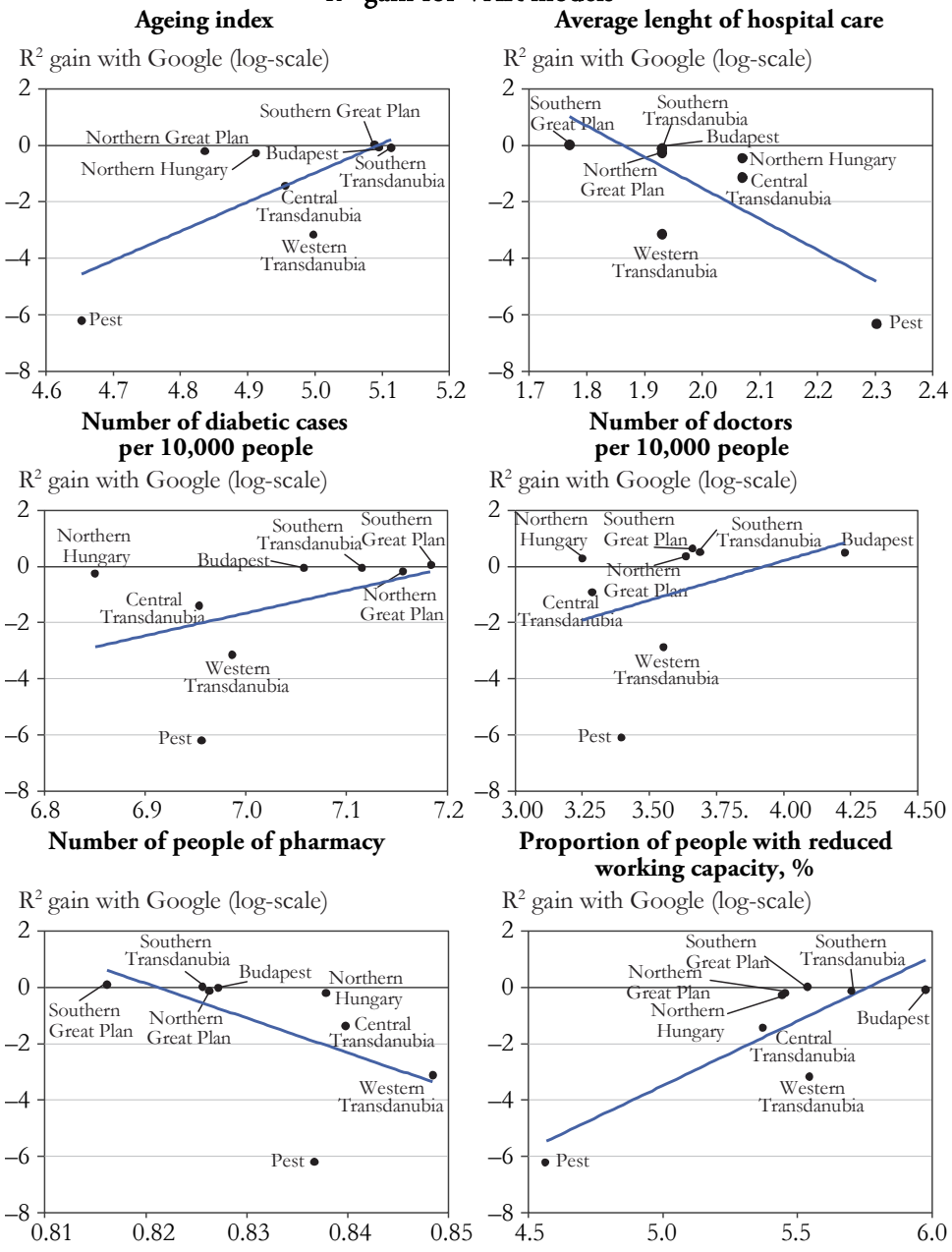
Variable	Rank correlation
Number of inhabitants per pharmacy in 2021 (persons/pharmacy)	−0.786
Average length of hospital care in 2021 (days)	−0.779
Number of working doctors per 10,000 inhabitants in 2021 (persons)	0.714
Proportion of diabetes patients registered with general practitioners per 10,000 inhabitants aged 19 and over in 2021 (cases/10,000 inhabitants)	0.701
Ageing index (number of the elderly aged 65 years and over per 100 individuals younger than 14 years) as of 31 December 2021 (%)	0.667
Percentage of the population receiving benefits for reduced working capacity in 2021 (%)	0.643
Proportion of hypertension patients registered with general practitioners per 10,000 inhabitants aged 19 and over in 2021 (cases/10,000 inhabitants)	0.571
Proportion of ischaemic heart disease (I20–I25) patients registered with general practitioners per 10,000 inhabitants aged 19 and over in 2021 (cases/10,000 inhabitants)	0.571
Number of operational hospital beds per 10,000 inhabitants in 2021 (units/person)	0.524
Average monthly old-age pension per capita in 2021 (HUF/month)	−0.452
Population density as of 31 December 2021 (people/km²)	−0.429
Age-standardised mortality rate per 100,000 inhabitants in 2021 (per mille)	0.192
Internet subscriptions per 1,000 inhabitants in 2021 (units)	−0.180
Gross domestic product per capita in 2021 (thousand HUF)	−0.167
Life expectancy at birth for women in 2021 (years)	−0.140
Life expectancy at birth for men in 2021 (years)	0.001

Note: arranged in descending order according to the absolute value of the rank correlation coefficient.

Based on the maps in Appendix Figure A2, Budapest and the Southern Great Plain show the most significant improvement in infection forecasting when utilising the internet search index. These regions also have the lowest number of people per pharmacy (indicating a relatively low pharmacy workload) and the shortest average length of hospital care (indicating effective inpatient healthcare). The Southern Great Plain also has the highest number of diabetic cases, while Budapest has the highest number of doctors per 10,000 people. Southern Transdanubia and the Northern Great Plain exhibit similar trends, characterised by a low pharmacy workload, short hospital stays, high diabetes incidence and a relatively high number of doctors. Northern Hungary is an outlier, showing R^2 gains when applying the internet search index despite having the lowest diabetes incidence, high pharmacy workload, extended hospital stays and few doctors. This anomaly aligns with Table 3, where Northern Hungary shows a significantly lower R^2 gain than other regions. Depicting the top correlations in the scatter plot of Figure 1 leads to similar conclusions.

Figure 1

R^2 gain for VAR models*



* As a function of the ageing index, the average length of hospital care, the number of doctors and diabetics per ten thousand people the number of people per pharmacy, and the percentage of people with reduced working capacity on a logarithmic scale.

Figure 1 exhibits tendencies similar to those in Appendix Figure A2, complemented by correlations between the R^2 gain with the internet search index and variables such as the ageing index and the percentage of people with reduced working capacity (Table 4). Pest is a notable outlier in three scatter plots for healthcare-related variables (people per pharmacy, diabetes incidence and doctors per 10,000 people), showing a significantly smaller R^2 gain than expected. This can be attributed to its younger population and the lowest percentage of people with reduced working capacity, suggesting a healthier population is less likely to search for disease symptoms online, leading to poorer VAR model performance than SIR.

A similar outlier tendency can be observed in Western Transdanubia, albeit on a much smaller scale. The region generally exhibits a lower R^2 gain with the internet search index, as indicated by most of the healthcare and demographic variables in Figure 1. Notably, Western Transdanubia has by far the highest number of people per pharmacy – the variable most strongly correlated with the R^2 gain – suggesting that regional pharmacies with lower average workload may be particularly effective in raising health awareness, thereby increasing internet interest in diseases.

While it is not as significant an outlier than Pest or Western Transdanubia, Northern Hungary exhibits a larger R^2 gain than the healthcare and demographic variables in Figure 1 would suggest. This outlier behaviour can be attributed to the region having the country's highest age-standardised mortality and lowest life expectancy. Moreover, given that it has some of the worst figures for the number of people per pharmacy and doctors per 10,000 people, it can be inferred that the region has a population with generally poor health and low healthcare efficiency. This may drive individuals to rely more on the internet for health information.

Discussion and conclusions

Overall, the findings presented in Tables 3 and 4, along with Figure 1 and Figure A2 in Appendix, suggest that internet search data can effectively forecast infection numbers in regions where health awareness is prevalent and reflected in internet search behaviours. The four variables that are strongly correlated with the R^2 gain of the VAR models underscore the importance of health awareness in enabling internet-based epidemic forecasting.

The strong correlation between the R^2 gain, the internet search index, and the number of people per pharmacy aligns with the literature, which highlights the crucial role of pharmacies in raising health awareness (Agomo–Ogunleye 2014, Hermawatiningsih et al. 2024, Eastwood et al. 2022, Agomo et al. 2018). A lower workload (fewer people per pharmacy) allows pharmacies to fulfil this role more effectively.

While it may seem counterintuitive, a higher incidence of diabetes can indicate an effective healthcare system. Adequate healthcare resources facilitate the early

detection of diseases such as diabetes, as supported by multiple studies (Howard et al. 2010, Heuclin et al. 2009). This perspective is reinforced by the case of Northern Hungary, which has the lowest proportion of diagnosed diabetics despite its low socio-economic indicators. The literature suggests that this discrepancy may result from underdetection due to the region's limited number of medical professionals.

Additionally, managing diabetes requires active patient and environmental engagement (American Diabetes Association 2003, Nathan 2015), which likely translates into higher internet search volumes for health-related topics. Therefore, regions with a high proportion of diagnosed diabetics often demonstrate greater health awareness, contributing to the effective use of internet searches for subjects such as the coronavirus.

In conclusion, regions where internet searches effectively predict new Covid-19 cases tend to be characterised by high health awareness, supported by an effective healthcare system and an older population together with a significant proportion of people with reduced working capacity. In such populations, existing health issues and accessible healthcare encourage internet searches for information on emerging diseases like Covid-19, making these searches a reliable predictor.

However, Northern Hungary presents an exception, where internet searches remain an effective predictor despite inadequate healthcare infrastructure and high age-standardised mortality. This suggests that, in the serious absence of a robust healthcare system, the region's population turns to the internet as a crucial alternative source of health-related information, compensating for the lack of reliable medical guidance.

These findings represent a novelty compared to the literature summarized in chapter Literature review, as those studies either did not examine the effectiveness of internet search-based epidemic forecasting in countries with lower socio-economic status (Saegner–Austys 2022) or only investigated the performance of internet search data in infection forecasting without exploring its correlation with socio-economic variables (Fantazzini 2020, Ben et al. 2022). Moreover, the studies reviewed that investigated intra-country heterogeneity typically employed static, contemporaneous correlation analyses rather than time series methods (Mavragani–Gkillas 2020, Rovetta 2021).

Naturally, all these conclusions should be interpreted with notable limitations, as the raw number of new coronavirus infections and the internet search volumes of coronavirus-related terms needed to be aggregated on a weekly and NUTS 2 level to fit time series models. On the one hand, this approach aligns with the international literature using Google Trends data in epidemic forecasting on a cross-country level, demonstrating in which sub-country regions these methods can be effective (in terms of out-of-sample R^2). On the other hand, these simple aggregations lead to a loss of within-weekly and within-NUTS 2 level variance in the data. This may explain why the Google-augmented VAR models perform poorly in some regions compared to

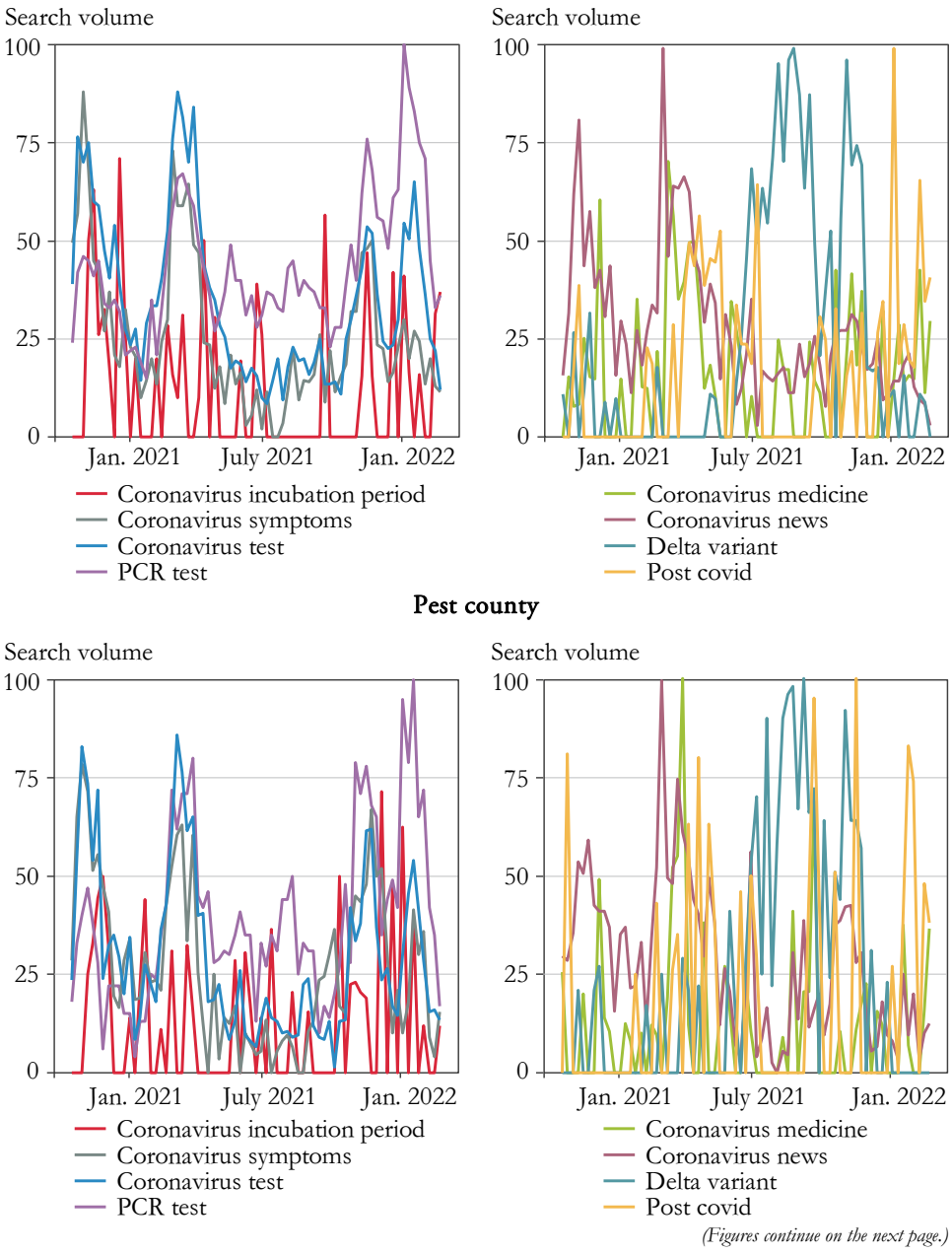
the SIR model in out-of-sample forecasting. Ghysels et al. (2016) and Yang et al. (2023) proposed that mixed-frequency models could address this issue, presenting a fascinating new avenue for research in sub-country level epidemic forecasting using internet search data.

Furthermore, our study focuses only on the eight Hungarian regions. Therefore, the correlations identified should not be generalised to other countries, as they are descriptive of the Hungarian context. Increasing the number of observations could be achieved by constructing VAR models at the county or district level rather than at the NUTS 2 regional level. However, our time series data (Google search volumes and weekly new coronavirus case numbers) already contain multiple contiguous periods with zero values, even at the county level, creating numerical parameter estimation issues. Consequently, the level of geographical aggregation cannot be reduced. A promising avenue for expanding the number of observations in correlation analysis would be to develop VAR models for all European NUTS 2 regions.

Appendix

Figure A1

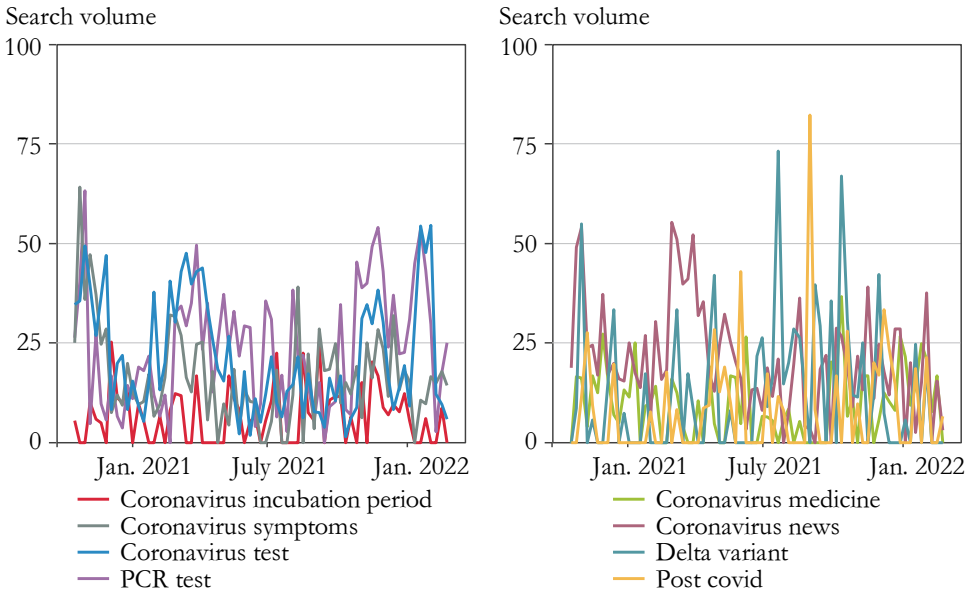
Google Trends search volumes for our keywords in each Hungarian region



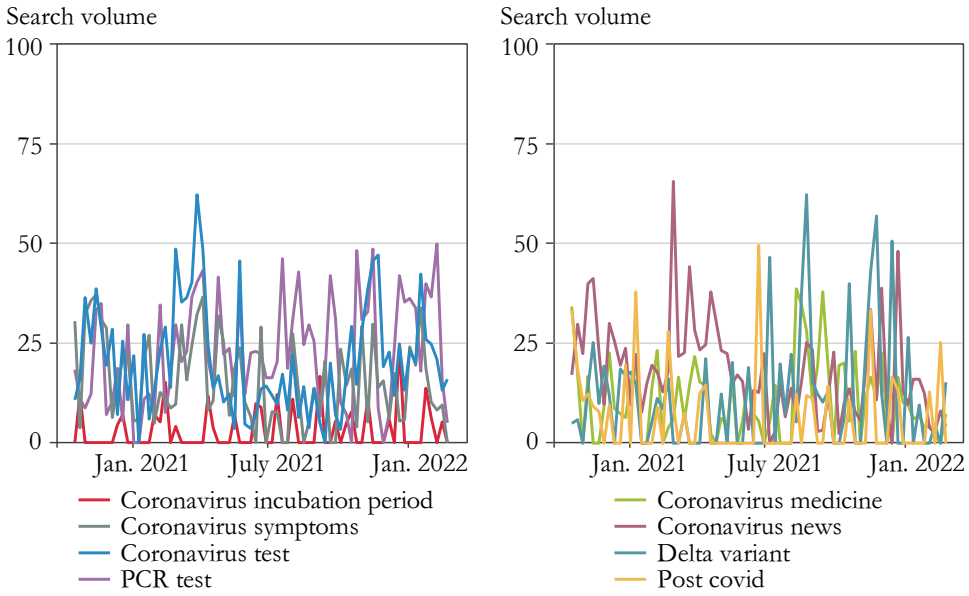
(Figures continue on the next page.)

(Continued.)

Western Transdanubia



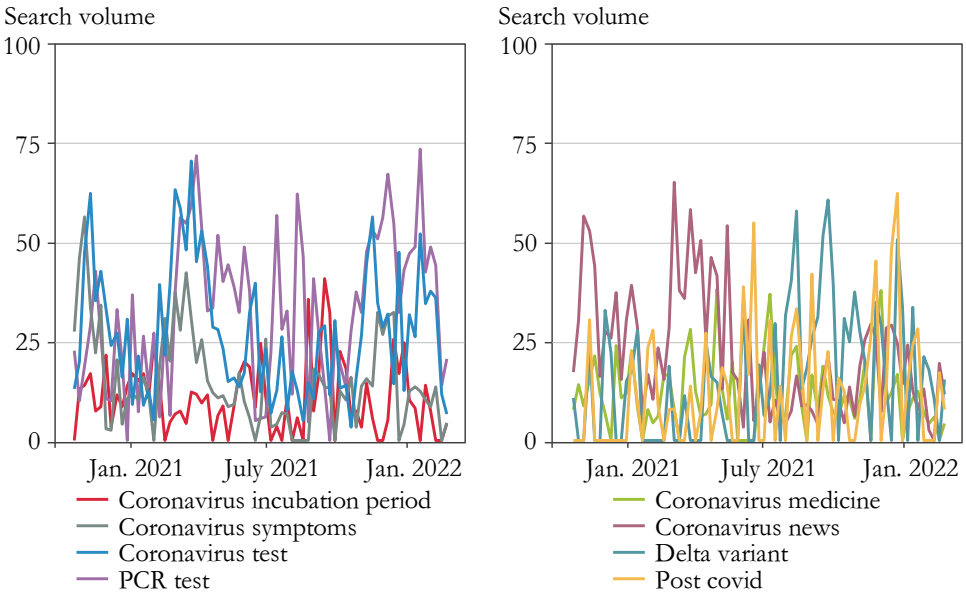
Southern Transdanubia



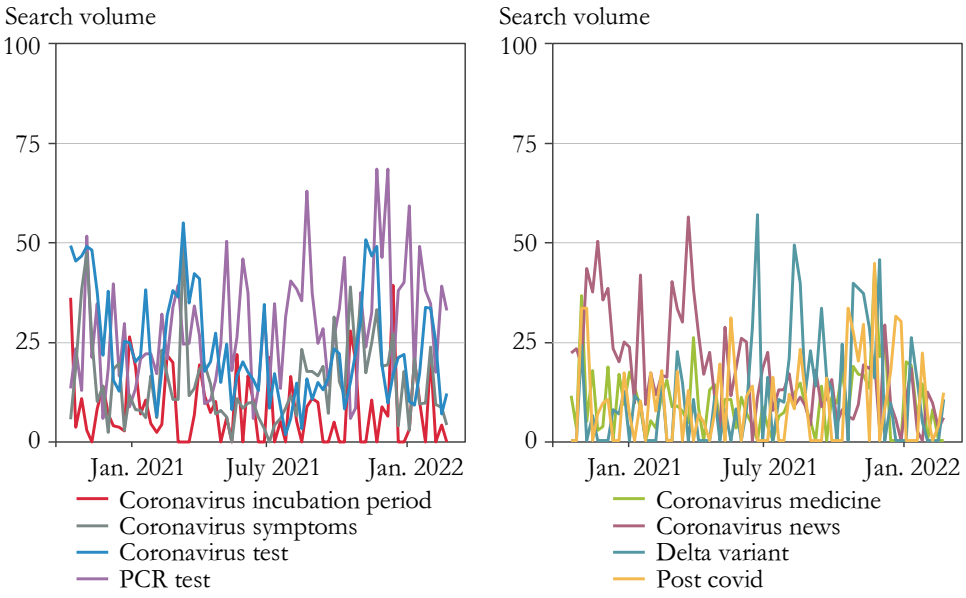
(Figures continue on the next page.)

(Continued.)

Central Transdanubia



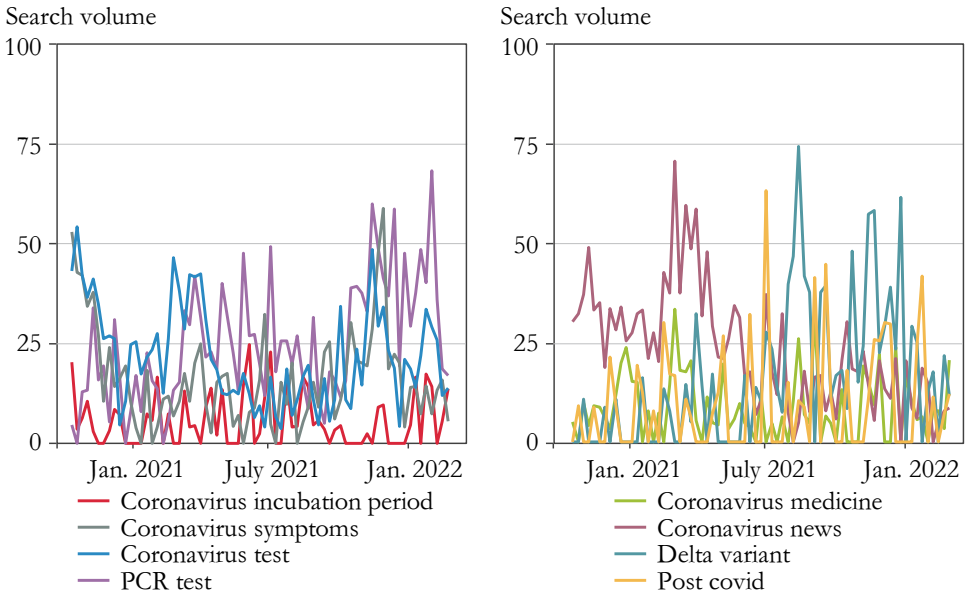
Southern Great Plain



(Figures continue on the next page.)

(Continued.)

Northern Great Plain



Northern Hungary

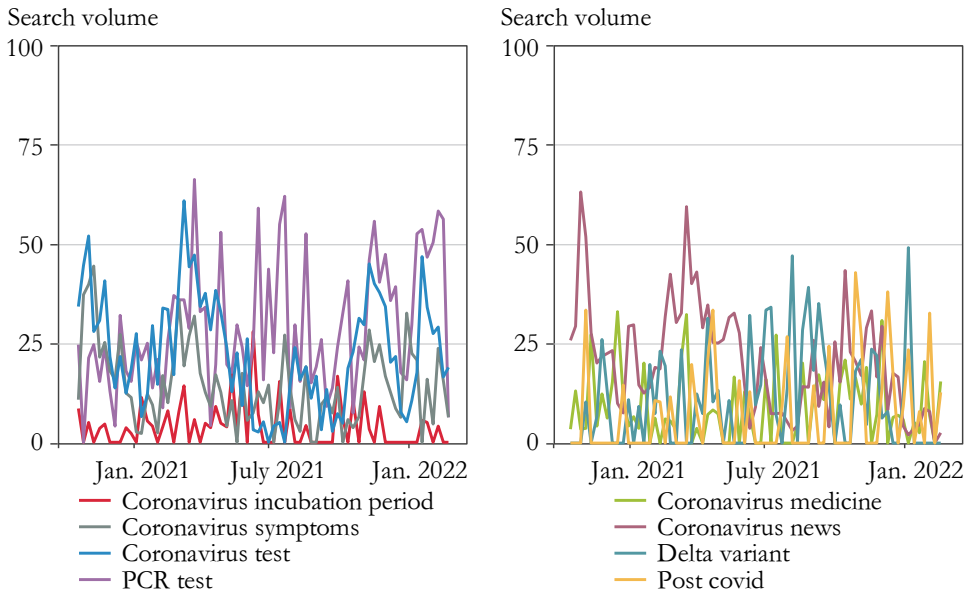
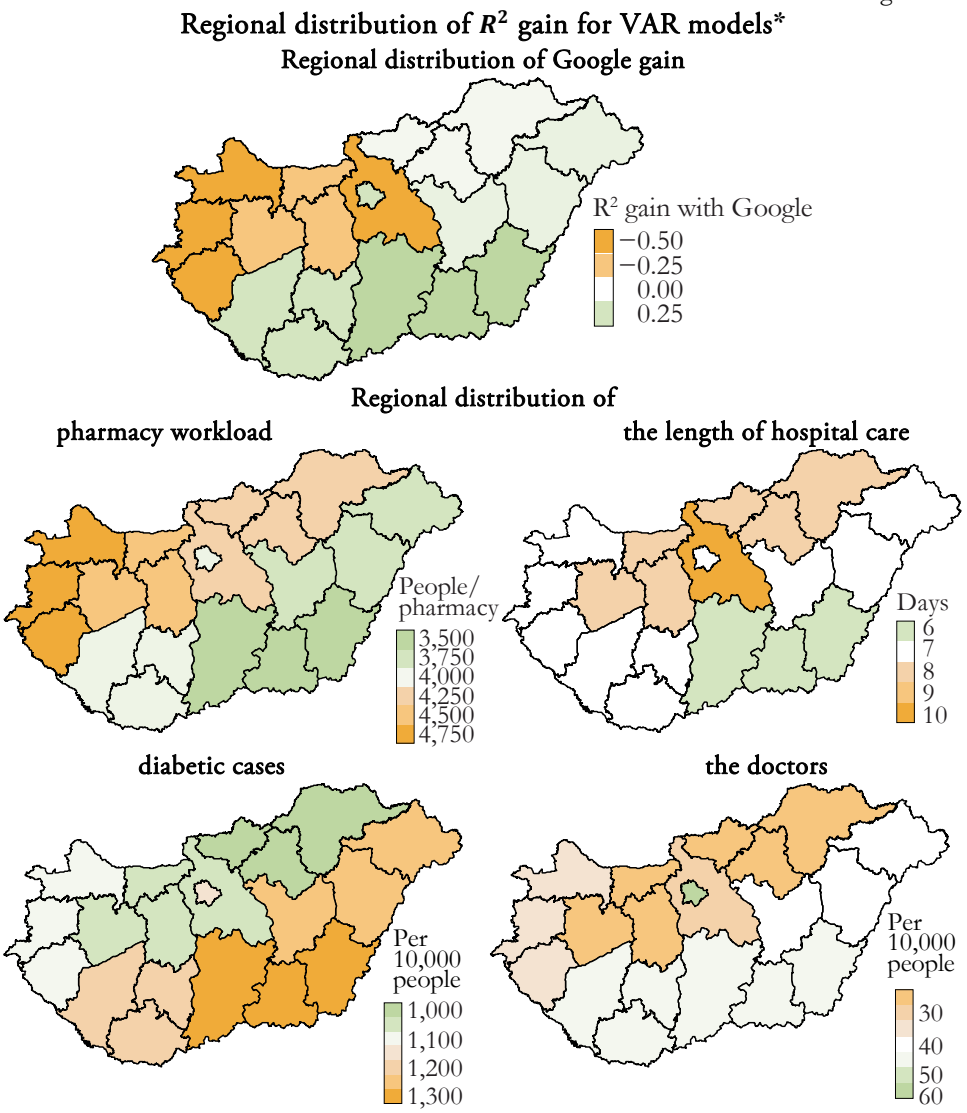


Figure A2



* As well as the number of people per pharmacy, average length of hospital care and the number of doctors and diabetics per ten thousand people in Hungary.

Table A1

Tests of stationarity for each Hungarian region

Variable	Levels		First differences	
	ADF p-value	KPSS p-value	variable	ADF p-value
	Pest county			
Cases	0.590	<0.01	0.388	>0.1
Medicine	0.052	>0.1	<0.01	>0.1
News	0.093	<0.01	<0.01	>0.1
Incubation	0.021	>0.1	<0.01	>0.1
Test	0.022	>0.1	<0.01	>0.1
Symptom	0.093	0.088	<0.01	>0.1
Delta	0.756	0.040	<0.01	>0.1
PCR	0.020	0.046	<0.01	>0.1
Post	0.025	>0.1	<0.01	>0.1
	Southern Great Plain			
Cases	0.706	<0.01	0.399	>0.1
Medicine	0.097	>0.1	<0.01	>0.1
News	<0.01	<0.01	<0.01	>0.1
Incubation	<0.01	>0.1	<0.01	>0.1
Test	0.093	0.054	<0.01	>0.1
Symptom	0.344	>0.1	<0.01	>0.1
Delta	0.083	0.055	<0.01	>0.1
PCR	<0.01	<0.01	<0.01	>0.1
Post	0.204	>0.1	<0.01	>0.1
	Southern Transdanubia			
Cases	0.664	<0.01	0.529	>0.1
Medicine	0.334	>0.1	<0.01	>0.1
News	0.414	0.012	<0.01	>0.1
Incubation	<0.01	>0.1	<0.01	>0.1
Test	0.096	>0.1	<0.01	>0.1
Symptom	0.056	0.045	<0.01	>0.1
Delta	0.261	>0.1	<0.01	>0.1
PCR	0.025	0.027	<0.01	>0.1
Post	<0.01	>0.1	<0.01	>0.1
	Western Transdanubia			
Cases	0.670	<0.01	0.493	>0.1
Medicine	0.233	>0.1	<0.01	>0.1
News	0.367	0.033	<0.01	>0.1
Incubation	0.017	>0.1	<0.01	>0.1
Test	0.050	>0.1	<0.01	>0.1
Symptom	0.090	0.074	<0.01	>0.1
Delta	0.231	>0.1	<0.01	>0.1
PCR	0.028	>0.1	<0.01	>0.1
Post	0.016	>0.1	<0.01	>0.1

(Table continues on the next page.)

(Continued.)

Variable	Levels		First differences	
	ADF p-value	KPSS p-value	variable	ADF p-value
	North Great Plain			
Cases	0.750	<0.01	0.465	>0.1
Medicine	<0.01	>0.1	<0.01	>0.1
News	0.190	<0.01	<0.01	>0.1
Incubation	0.084	>0.1	<0.01	>0.1
Test	0.160	0.047	<0.01	>0.1
Symptom	0.130	>0.1	<0.01	>0.1
Delta	0.517	<0.01	<0.01	>0.1
PCR	0.324	<0.01	<0.01	>0.1
Post	<0.01	>0.1	<0.01	>0.1
	North Hungary			
Cases	0.672	<0.01	0.426	>0.1
Medicine	0.091	>0.1	<0.01	>0.1
News	0.338	0.013	<0.01	>0.1
Incubation	0.392	>0.1	<0.01	>0.1
Test	0.132	>0.1	<0.01	>0.1
Symptom	0.165	0.093	<0.01	>0.1
Delta	0.052	>0.1	<0.01	>0.1
PCR	0.104	<0.01	<0.01	>0.1
Post	<0.01	0.040	<0.01	>0.1
	Central Transdanubia			
Cases	0.588	<0.01	0.489	>0.1
Medicine	0.018	>0.1	<0.01	>0.1
News	0.358	<0.01	<0.01	>0.1
Incubation	0.024	>0.1	<0.01	>0.1
Test	0.069	>0.1	<0.01	>0.1
Symptom	0.031	0.079	<0.01	>0.1
Delta	0.486	<0.01	<0.01	>0.1
PCR	0.184	0.062	<0.01	>0.1
Post	0.172	0.026	<0.01	>0.1

REFERENCES

AMERICAN DIABETES ASSOCIATION (2003): Treatment of hypertension in adults with diabetes *Clinical Diabetes* 21 (3): 122–127. https://doi.org/10.2337/diaclin.21.3.122

AGOMO, C. O.–OGUNLEYE, J. (2014): An investigation of strategies enhancing the public health role of community pharmacists: a review of knowledge and information *Journal of Pharmaceutical Health Services Research* 5 (2): 135–145. https://doi.org/10.1111/jphs.12056

AGOMO, C.–UDOH, A.–KPOKIRI, E.–OSUKU-OPIO, J. (2018): Community pharmacists' contribution to public health: assessing the global evidence base *The Pharmaceutical Journal* 10 (4). https://doi.org/10.1211/cp.2018.20204556

AUSTYS, D.–BURNEIKAITĖ, M. (2019): Google Trends and forecasting of influenza epidemics in Lithuania *European Journal of Public Health* 29 (Supplement_4): ckz185.403. https://doi.org/10.1093/eurpub/ckz185.403

BAMBRA, C.–RIORDAN, R.–FORD, J.–MATTHEWS, F. (2020): The Covid-19 pandemic and health inequalities *Journal of Epidemiology and Community Health* 74 (11): 964–968. https://doi.org/10.1136/jech-2020-214401

BEN, S.–XIN, J.–CHEN, S.–JIANG, Y.–YUAN, Q.–SU, L.–CHRISTIANI, D. C.–ZHANG, Z.–DU, M.–WANG, M. (2022): Global internet search trends related to gastrointestinal symptoms predict regional Covid-19 outbreaks *Journal of Infection* 84 (1): 56–63. https://doi.org/10.1016/j.jinf.2021.11.003

BÍRÓ, A.–BRANYICZKI, R.–ELEK, P. (2021): Time patterns of precautionary health behaviours during an easing phase of the Covid-19 pandemic in Europe *European Journal of Ageing* 19: 837–848. https://doi.org/10.1007/s10433-021-00636-4

BUCCI, A.–IPPOLITI, L.–VALENTINI, P. (2023): Analysing spatiotemporal patterns of Covid-19 confirmed deaths at the NUTS-2 regional level *Regional Statistics* 13 (2): 214–239. https://doi.org/10.15196/rs130202

BRAUER, F.–CASTILLO-CHAVEZ, C. (2012): *Mathematical models in population biology and epidemiology* Springer.

EASTWOOD, K. A.–ALLEN-WALKER, V. A.–MAXWELL, M.–MCKINLEY, M. C. (2022): Raising awareness of pre-conception care in community pharmacies: a feasibility study *Pilot and Feasibility Studies* 8 (1): 44. https://doi.org/10.1186/s40814-022-01001-7

ELEK, P.–FADGYAS-FREYLER, P.–VÁRADI, B.–MAYER, B.–ZEMPLÉNYI, A. (2022): Effects of lower screening activity during the Covid-19 pandemic on breast cancer patient pathways: evidence from the age cut-off of organized screening *Health Policy* 126 (8): 763–769. https://doi.org/10.1016/j.healthpol.2022.05.013

FANTAZZINI, D. (2020): Short-term forecasting of the Covid-19 pandemic using Google Trends data: evidence from 158 Countries *Applied Econometrics* 59: 1–26. https://doi.org/10.2139/ssrn.3671005

GELMAN, A.–HILL, J. (2007): *Data analysis using regression and multilevel/hierarchical models* Cambridge University Press.

GHYSELS, E.–KVEDARAS, V.–ZEMLYS, V. (2016): Mixed frequency data sampling regression models: the R package midasr *Journal of Statistical Software* 72: 1–35. https://doi.org/10.18637/jss.v072.i04

GOMBOS, K.–HERCZEG, R.–ERŐSS, B.–KOVÁCS, S. Z.–UZZOLI, A.–NAGY, T.–KISS, S.–SZAKÁCS, Z.–IMREI, M.–SZENTESI, A.–NAGY, A.–FABIAN, A.–HEGYI, P.–GYENESEI, A. (2020): Translating scientific knowledge to government decision makers has crucial importance in the management of the Covid-19 pandemic *Population Health Management* 24 (1): 35–45. https://doi.org/10.1089/pop.2020.0159

HAMILTON, J. D. (2020): *Time series analysis* Princeton University Press.

HARKO, T.–LOBO, F. S.–MAK, M. K. (2014): Exact analytical solutions of the susceptible-infected-recovered (SIR) epidemic model and of the SIR model with equal death and birth rates *Applied Mathematics and Computation* 236: 184–194. https://doi.org/10.1016/j.amc.2014.03.030

HASTIE, T. J.–TIBSHIRANI, R.–FRIEDMAN, J. (2011): *The elements of statistical learning: data mining, inference, and prediction (2nd ed.)*. Springer.

HEUCLIN, T.–DUBOS, F.–HUE, V.–GODART, F.–FRANCART, C.–VINCENT, P.–HOSPITAL NETWORK FOR EVALUATING THE MANAGEMENT OF COMMON CHILDHOOD DISEASES–MARTINOT, A. (2009): Increased detection rate of Kawasaki disease using new diagnostic algorithm, including early use of echocardiography *The Journal of Pediatrics* 155 (5): 695–699. https://doi.org/10.1016/j.jpeds.2009.04.058

HERMAWATININGSIH, O. D.–RAISING, R.–LA BASY, L.–MARITHA, V. (2024): Health awareness education through blood checking in pharmacies *Jurnal IPMAS* 4 (3): 149–158. https://doi.org/10.54065/ipmas.4.3.2024.480

HOWARD, D. H.–THORPE, K. E.–BUSCH, S. H. (2010): Understanding recent increases in chronic disease treatment rates: more disease or more detection? *Health Economics, Policy and Law* 5 (4): 411–435. https://doi.org/10.1017/S1744133110000149

HUANG, R.–LUO, G.–DUAN, Q.–ZHANG, L.–ZHANG, Q.–TANG, W.–SMITH, M. K.–LI, J.–ZOU, H. (2020): Using Baidu search index to monitor and predict newly diagnosed cases of HIV/AIDS, syphilis and gonorrhea in China: estimates from a vector autoregressive (VAR) model *BMJ Open* 10 (3): e036098. https://doi.org/10.1136/bmjopen-2019-036098

IGARI, A. (2023): Spatiotemporal inequalities of excess mortality in Europe during the first two years of the Covid-19 pandemic *Regional Statistics* 13 (03): 510–535. https://doi.org/10.15196/rs130306-

KARLINSKY, A.–KOBAK, D. (2021): Tracking excess mortality across countries during the Covid-19 pandemic with the world mortality dataset *ELife* 10: e71974. https://doi.org/10.7554/eLife.69336

KIRCHGÄSSNER, G.–WOLTERS, J.–HASSLER, U. (2013): *Introduction to modern time series analysis* Springer Science & Business Media. https://doi.org/10.1007/978-3-642-33436-8

KOVÁCS, S. Z.–UZZOLI, A. (2020): A koronavírus-járvány jelenlegi és várható egészségkockázatainak területi különbségei Magyarországon *Tér és Társadalom* 34 (2): 155–170. https://doi.org/10.17649/TET.34.2.3265

KOVÁCS, L.–VÁNTUS, K. (2022): A hazai koronavírus-halálozás járási különbségeinek összefüggései az egészségügyi ellátással Területi Statisztika 62 (3): 253–289. https://doi.org/10.15196/TS620301

KOVALCSIK, T.–BOROS, L.–PÁL, V. (2021): A Covid-19-járvány első két hullámának területisége Közép-Európában *Területi Statisztika* 61 (3): 263–290. https://doi.org/10.15196/TS610301

LI, K.–LIANG, Y.–LI, J.–LIU, M.–FENG, Y.–SHAO, Y. (2020): Internet search data could be used as novel indicator for assessing Covid-19 epidemic *Infectious Disease Modelling* 5: 848–854). https://doi.org/10.1016/j.idm.2020.10.001

LI, K.–LIU, M.–FENG, Y.–NING, C.–OU, W.–SUN, J.–WEI, W.–LIANG, H.–SHAO, Y. (2019): Using Baidu search engine to monitor AIDS epidemics inform for targeted intervention of HIV/AIDS in China *Scientific Reports* 9 (1): 1–12. https://doi.org/10.1038/s41598-018-35685-w

LLEWELLYN, M.–ROSS, G.–RYAN-SAHA, J. (2023): Covid-era forecasting: Google Trends and window and model averaging *Annals of Tourism Research* 103: 103660. https://doi.org/10.1016/j.annals.2023.103660

MAVRAGANI, A.–GKILLAS, K. (2020): Covid-19 predictability in the United States using Google Trends time series *Scientific Reports* 10 (1): 20693. https://doi.org/10.1038/s41598-020-77275-9

MCGOUGH, S. F.–BROWNSTEIN, J. S.–HAWKINS, J. B.–SANTILLANA, M. (2017): Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data *PLOS Neglected Tropical Diseases* 11 (1): e0005295. https://doi.org/10.1371/journal.pntd.0005295

MCGOWAN, V.–BAMBRA, C. (2022): Covid-19 mortality and deprivation: pandemic, syndemic and endemic health inequalities *Lancet Public Health* 7 (11): e966-e975. https://doi.org/10.1016/S2468-2667(22)00223-7

MIHÁLYI, P. (2014): Post-socialist transition in a 25-year perspective *Acta Oeconomica* 64 (1): 1–24. https://doi.org/10.1556/aoecon.64.2014.s1.1

MUNFORD, L.–KHAVANDI, S.–BAMBRA, C. (2022): Covid-19 and deprivation amplification: an ecological study of geographical inequalities in mortality in England *Health and Place* 78: 102933. https://doi.org/10.1016/j.healthplace.2022.102933

MUTH, C.–ORAVECZ, Z.–GABRY, J. (2018): User-friendly Bayesian regression modeling: a tutorial with rstanarm and shinystan *The Quantitative Methods for Psychology* 14 (2): 99–119. https://doi.org/10.20982/tqmp.14.2.p099

NATHAN, D. M. (2015): Diabetes: advances in diagnosis and treatment *Jama* 314 (10): 1052–1062. https://doi.org/10.1001/jama.2015.9536

OROSZI, B.–JUHÁSZ, A.–NAGY, C.–HORVÁTH, J. K.–KOMLÓS, K. E.–TÚRI, G.–MCKEE, M.–ÁDÁNY, R. (2022a): Characteristics of the third Covid-19 pandemic wave with special focus on socioeconomic inequalities in morbidity, mortality and the uptake of Covid-19 vaccination in Hungary *Journal of Personal Medicine* 12: 388. https://doi.org/10.3390/jpm12030388

OROSZI, B.–JUHÁSZ, A.–NAGY, C.–HORVÁTH, J. K.–KOMLÓS, K. E.–TÚRI, G.–MCKEE, M.–ÁDÁNY, R. (2022b): A Covid-19 járvánnyal összefüggő megbetegedések és halálozások egyenlőtlen terhei, valamint összefüggésük a társadalmi-gazdasági helyzettel Magyarországon a második járványhullám alatt *Népegészségügy* 99 (1): 76–91.

PÁGER, B.–TÓTH, G. C.–UZZOLI, A. (2024): The role of socioeconomic variables in the regional inequalities of Covid-19 mortality in Hungary *Eastern Journal of European Studies* 15 (1): 272–297. https://doi.org/10.47743/ejes-2024-0112

ROVETTA, A. (2021): Reliability of Google Trends: analysis of the limits and potential of web infoveillance during Covid-19 pandemic and for future research *Frontiers in Research Metrics and Analytics* 6: 670226. https://doi.org/10.3389/frma.2021.670226

SAEGNER, T.–AUSTYS, D. (2022): Forecasting and surveillance of Covid-19 spread using Google Trends: literature review *International Journal of Environmental Research and Public Health* 19 (19): 12394. https://doi.org/10.3390/ijerph191912394

TÓTH, C. G. (2022): Narrowing the gap in regional and age-specific excess mortality during the Covid-19 in Hungary *Eastern Journal of European Studies* 13 (1): 185–207. https://doi.org/10.47743/ejes-2022-0109

UZZOLI, A.–EGRI, Z.–SZILÁGYI, D.–PÁL, V. (2020): Does better availability mean better accessibility? Spatial inequalities in the care of acute myocardial infarction in Hungary *Hungarian Geographical Bulletin* 69 (4): 401–418. https://doi.org/10.15201/hungeobull.69.4.5

UZZOLI, A.–KOVÁCS, S. ZS.–FÁBIÁN, A.–PÁGER, B.–SZABÓ, T. (2021): Spatial analysis of the Covid-19 pandemic in Hungary – changing epidemic waves in time and space *Region* 8 (2): 147–165. https://doi.org/10.18335/region.v8i2.343

XU, C.–WANG, Y.–YANG, H.–HOU, J.–SUN, L.–ZHANG, X.–CAO, X.–HOU, Y.–WANG, L.–CAI, Q.–WANG, Y. (2019): Association between cancer incidence and mortality in web-based data in China: infodemiology study *Journal of Medical Internet Research* 21 (1): e10677. https://doi.org/10.2196/10677

WOODWARD, W. A.–GRAY, H. L.–ELLIOTT, A. C. (2017): *Applied time series analysis with R* CRC press.

YANG, Y.–JIA, F.–LI, H. (2023): Estimation of panel data models with mixed sampling frequencies *Oxford Bulletin of Economics and Statistics* 85 (3): 514–544. https://doi.org/10.1111/obes.12536

ZHANG, W. Y.–WEI, Z. W.–WANG, B. H.–HAN, X. P. (2016): Measuring mixing patterns in complex networks by Spearman rank correlation coefficient *Physica A: Statistical Mechanics and its Applications* 451: 440–450. https://doi.org/10.1016/j.physa.2016.01.056

INTERNET SOURCE

HUNGARIAN CENTRAL STATISTICAL OFFICE [HCSO] (2022): *Population number of Hungary by sex and age, 1 January*. https://www.ksh.hu/interaktiv/korfak/orszag_en.html (downloaded: May 2024)

DATA SOURCES

[1] ÁTLÁTSZÓ (atlo): *Coronavirus map of Hungary*. https://atlo.team/koronaterkep/ (downloaded: April 2024)

[2] GOOGLE: *Google Trends*. https://trends.google.com/trends/explore?q=covid&geo=HU (downloaded: April 2024)

[3] HUNGARIAN CENTRAL STATISTICAL OFFICE [HCSO]: *Interactive map display application*. https://www.ksh.hu/teruletiatlasz_timea (downloaded: July 2024)