scientific reports

Check for updates

OPEN

Explaining COVID-19 dynamics through user activity data from digital platforms with Yandex's self-isolation index as a case study

Piotr Żebrowski[1]✉, Grigory Boyarshinov[2], Anastasia Odintsova[2] & Elena Rovenskaya[1,3]

Social-distancing measures were among the very few available policy responses to the initial outbreak of COVID-19, and they remain an important tool for containing recurring wavers of this and possible future pandemics. However, policies aiming at limiting the intensity of people-to-people contacts incur substantial socio-economic costs while their effectiveness varies over time and across locations. Having a robust way of measuring the level of people-to-people contacts and monitoring compliance with social-distancing policies would greatly aid governments in better calibrating their responses to future pandemic outbreaks. In this paper we use the case example of the Yandex's self-isolation index to explore the potential of composite indices that aggregate multiple sources of activity data collected by digital platforms as proxies for evaluating the people-to-people contact intensity. To this end, we propose two error-corrected autoregressive distributed-lag models, inspired by the classical SIR model of infectious disease dynamics, and use them in testing for cointegration between the self-isolation index and the official data on the numbers of new COVID-19 cases and deaths, for the two largest cities in Russia, Moscow and St. Petersburg. We have found evidence for such cointegration, which confirms that the COVID-19 epidemic curve can be explained by the level of people-to-people contact intensity as measured by the self-isolation index. Our findings suggest that the self-isolation index is a useful real-time indicator of the level of compliance with social distancing measures in the population and thus can serve as a reliable tool for informing policymaking.

Keywords COVID-19 dynamics, Social distancing monitoring, Digital platforms, Cointegration, Autoregressive distributed lag models

The Coronavirus Disease (COVID-19)[1] pandemic has afflicted over 778 million people to date (May 2025), of whom approximately 7 million have died[2]. It also caused unprecedented disruptions in economic and political systems and devastated numerous communities across the world[3]. COVID-19 is caused by the SARS-CoV-2 virus, which mainly spreads through airborne transmission[4]. To reduce the virus transmission and contain initial surges in COVID-19 cases, most countries introduced a wide range of non-pharmaceutical interventions (NPIs), such as face-mask mandates and social distancing measures[5]. While vaccines (and later other pharmaceutical measures), became available as of December 2020, their roll-out across the world was uneven, with lower vaccination rates attained in low- and middle-income countries[6]. Global vaccination campaign succeeded in reducing COVID-19 deaths[7,8], but vaccine-induced immunity proved to decline within months from the primary vaccination cycle[9]. Due to immunity waning and appearance of more contagious variants of the SARS-CoV-2 virus, NPIs have continued to play an important role in managing the recurring waves of COVID-19 infections. They also remain our primary defense against outbreaks of emerging infectious diseases that may occur in the future.

The efficacy of NPIs has been assessed in different contexts using compartmental models of epidemic dynamics coupled with statistical modeling[10]. It is often evaluated in terms of the change in the effective reproduction number R_t, i.e., the average number of secondary cases generated by an individual case detected

[1]Advancing Systems Analysis Program, International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, 2361 Laxenburg, Austria. [2]Laboratory of Geoinformatics and the Arctic Big Data, Geophysical Center of the Russian Academy of Sciences (GC RAS), Moscow, Russia. [3]Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia. ✉email: zebrowsk@iiasa.ac.at

at time t, which serves as a key parameter to determine how the disease will spread in the population. Using data on the reported cases before and after the inception of a considered NPI, its impact can be estimated, provided no other significant factors interfere or the impact of these factors could be accounted for[11–18].

NPIs act on the reproduction number of the virus indirectly, e.g., by reducing individuals' exposure to the virus through changes in their behaviors, such as limiting the number of people-to-people contacts and altering context-specific mobility patterns. Thus, the efficacy of NPIs critically depends on citizens' compliance with the introduced measures. However, the level of compliance varies, both across and within countries, resulting in observable differences in the NPIs efficacy[13,17,19–21]. Socio-economic conditions[22,23], public believes and attitudes[24], risk perception[25], trust in the healthcare systems[26], and quality of institutions[21,27] have been found to influence compliance in different contexts. Many of these factors are obscure or unknown in real-time to authorities responding to developing disease outbreaks, making it difficult to anticipate population's compliance with, and thus effectiveness of, NPI measures. This uncertainty leaves governments facing high-stakes dilemmas when calibrating policy response: overly stringent measures are likely to contain the outbreak but possibly at the price of exceedingly high social and economic costs, while too lax (but still costly) policies may be ineffective. Ability to monitor citizen's compliance with NPI measures would reduce the level of uncertainty under which governments and public health authorities operate and thus would be a significant aid in formulating policy responses to disease outbreaks. A reliable way of monitoring population-level compliance with social distancing measures (e.g., travel restrictions, school and non-essential businesses closures, shelter-in-place orders) would be particularly useful for policymakers, as such measures, while proven effective in containing COVID-19 outbreaks, cause significant social and economic disruptions.

To help authorities make informed decisions on the scale and timing of social distancing measures, two elements are needed: (1) a reliable indicator of social distancing in the population; and (2) an understanding of the relationship between this indicator and the dynamics of the epidemic. The latter would enable experts to infer the level of social distancing necessary to slow down the virus spread, while the former would allow them to gauge whether the current compliance with already introduced measures ensures the required level of social distancing and whether social distancing policies should be strengthened or relaxed.

Monitoring all people-to-people interactions is not feasible in practice and may not be politically and socially desirable. Thus, in absence of direct measurements, a policy-relevant indicator of social distancing within a population must rely on aggregate and anonymized proxy data. Large digital platforms, which collect and integrate user activity data across multiple domains (including geolocation services, virtual shopping, online entertainment, ride-hailing services) have a high potential to serve as valuable sources of good-quality proxy data that allow for monitoring social distancing on aggregate levels and across different contexts (workplaces, public transport, etc.).

Aggregate population mobility patterns derived from anonymized geolocation data collected by digital platforms and mobile service providers were among the most widely used proxies for explaining the development of the COVID-19 pandemic in the short-run. Using the estimated flows of travelers based on Baidu geolocation services, Quilty et al.[28] studied the effects of introducing a cordon sanitaire around Wuhan in January 2020 on the spread of COVID-19 across major cities in China. Similarly, Moorley et al.[29] correlated the composite mobility grade based of Unacast mobile telephone tracking data with the daily estimates of the reproductive numbers R_t for eight counties in central State of New York. Gerlee et al.[30] employed data on the public transport usage and the Google mobility reports (GMR)[31] to predict hospital admissions due to COVID-19 infections in Sweden. GMR data was successfully incorporated into statistical models to improve accuracy of their predictions of COVID-19 spread[32,33] and changes in the reproductive number R_t[34], and were widely used for parameterizing and validating compartmental and metapopulation models of the COVID-19 dynamics[35]. Mobility data was also used as a proxy for the level compliance with social distancing measures. Vokó & Pitter[13] used GMR to calculate country-specific social distance indices for 28 European countries and showed that elevated levels of these indices coincided with breakpoints in the infection rates. Ilin et al.[36] leveraged GMR and aggregated mobility data from other platforms to assess how changes in the stringency of lockdown policies translate to changes in mobility behaviors and, ultimately, to changes in infection rates.

While mobility metrics are readily available and widely used in research, their usefulness for predicting COVID-19 dynamics and informing policies have proven to be limited. When evaluated over a range of diverse regions, mobility data does consistently carry statistically significant information on COVID-19 spread and the predictive power of mobility metrics is highly dependent on the level of spatial aggregation[37]. Moreover, as the COVID-19 pandemic developed, mobility information became a progressively worse proxy for frequency of risky in-person contacts which drive the dynamics of COVID-19 infections[38].

Estimates of in-person contact rates, either based on surveys or derived from high frequency positioning data from mobile devices, proved to be better predictors of COVID-19 spread compared to aggregate mobility metrics[39,40]. Yet, using such estimates for monitoring population compliance with social-distancing measures may be impractical, as it would involve conducting frequent surveys within a sufficiently large representative group of responders or require processing of large amounts of sensitive data to identify co-location events for mobile device users.

We posit that shifts in users' engagement with a wide range of services offered by digital platforms (such as Google, Baidou or Yandex), summarized by aggregate user activity metrics, reflect behavioral changes of population sufficiently well to serve as a useful proxy for the intensity of people-to-people contacts driving the spread of COVID-19. User activity metrics are based on data routinely collected by digital platforms and, in principle, could be made available to public and authorities just as the aggregate mobility metrics were. However, to our knowledge, the unique example of publicly available aggregate metric of user activity is the Self-isolation Index (SII) provided by Yandex[41], a major digital platform ecosystem operating in Russia and several other countries of the Commonwealth of Independent States, covering 50–60% of search sessions in Russian internet in years 2020–2021 (with 2025 market share of 65%)[42].

Yandex's Self-isolation Index gives insights into the changes in population's travel patterns as well as into shifts in types of activity people engage in, inferred from the usage of apps and web services in the Yandex ecosystem, such as search queries, delivery and taxi requests. SII was successfully used in machine-learning models to enhance the forecasting of the COVID-19 dynamics[43]. It could also serve as a readily available tool for real-time monitoring of population's compliance with the mandated social distancing measures, yet, to our knowledge, its suitability in this role has never been rigorously investigated. In this paper we address this knowledge gap.

To investigate the usefulness of Yandex's SII as a real-time indicator of the level of social distancing within a population, we statistically explore whether SII can explain the observed development of the COVID-19 epidemic in two largest urban centers of Russia: Moscow and St. Petersburg. To address this question, we propose two error-corrected autoregressive distributed-lag time series models and use them to test for statistically significant cointegration (interpreted as a stable long-run equilibrium) between the SII and the daily reported numbers of new infections or fatal cases associated with COVID-19. Whenever the available data provide sufficient evidence for such cointegration, we also estimate parameters of the short-run dynamics which govern the responses of these two quantities to changes of the self-isolation index.

Materials and methods

Data

Official data on daily reported COVID-19 infections and deaths in the Russian Federation were obtained from[44]. The data are available at the level of individual administrative units (oblasts) and has a daily resolution, starting from 12 March 2020.

The Yandex's self-isolation index was designed to reflect the average intensity of person-to-person contacts[45]. This was possible due to the platform's extensive user base. In years 2020 – 2022, for example, on average 80 to 100 million users in Russia (approximately 55–70% of the Russian population, with slight underrepresentation of adults of above 65 years of age – 10% of Yandex users vs. 27% of total adult population[46,47]) were visiting Yandex platform at least once a month. A typical user was spending about 13 h a month engaging with the Yandex ecosystem, including portal, mail service, movie and music streaming, blog, news, marketplace, and a range of other services[48]. Thus, the Yandex user base is a good cross-section of people living in Russian urban centers, which tend to have above-average access to computers and mobile devices.

The self-isolation index was derived from the anonymized user activity data from various services available through the Yandex ecosystem, including geolocation data from navigation and ride-hailing apps, delivery requests, movie streaming as well as search queries and use of blog platforms[49]. For example, higher values of self-isolation index were attained when fever than usual routing requests were registered for the navigator and subway route planning apps. Likewise, an elevated usage of movie streaming, delivery services and activity on internet forums resulted in higher values of the SII. The index was calibrated against a pre-pandemic baseline and takes values between 0 and 5, with 0 representing the level of activity during rush hours on a regular weekday, when most people are not at home, and 5 corresponding to virtually empty streets, with a vast majority of the population staying at home. The SII time series are available for individual cities at daily resolution for the period from 23 February 2020 to 22 September 2021.

The period for which the epidemic data and the SII data overlap spans between 12 March 2020 and 22 September 2021. There is a mismatch, however, in the level of spatial resolution, as the epidemic data is available for entire administrative units (federal subjects), while SII is calculated for individual cities. Therefore, in our analysis, we focus on the two largest cities in Russia, Moscow and St. Petersburg, which themselves are administrative units (federal cities). Both epidemic and SII data sets used in our study are publicly available, anonymized and highly aggregated. They are not considered personal data and thus their use does not require ethical approval.

The time series used in our analyses are displayed in Fig. 1. Points represent daily reported new cases of COVID-19 infections N_t, daily COVID-19 related deaths D_t, and daily values of self-isolation index SII_t for Moscow and St. Petersburg. Solid lines represent the smoothed versions of these time series, $RM_7(N_t)$, $RM_7\,(D_t)$ and $RM_7(SII_t)$, respectively, where $RM_7(\cdot_t)$ denotes a 7-day rolling mean (moving average) taken over the days from $(t-6)$ to t. The 7-day rolling mean filters out seasonal patterns at weekly time scales (e.g., lower than average numbers of new infections recorded on Tuesdays and Wednesdays and higher values of the SII on weekends).

In the initial stage of the pandemic unfolding in early spring of 2020, only a few new cases have been reported each day, with some days on which no new cases have been recorded at all. Numbers of new cases in double digits or higher (a signal sufficiently strong for the purpose of our analysis) have been consistently reported only after 23 March 2020 for Moscow and 6 April 2020 for St. Petersburg. We therefore take these dates as starting points for the time series of the daily new COVID-19 cases in the respective cities. As SII_t exhibit pronounced weekly cycles with index values higher on weekends, we limit our analysis to the period that ends on 19 September to ensure that the time series used for modelling include records of full weeks. The ranges of the time series included in the analysis are marked on Fig. 1A with a gray background.

The data on the daily reported deaths due to COVID-19 required a more substantial trimming to be used in our analysis. As some analysis of the excess mortality in Russia suggests, the number of the fatal cases due to COVID-19 might have been considerably underreported in 2020[50,51]. According to Kobak[50], on 28 December 2020, the Russian authorities admitted that most of the excess mortality between January and November 2020 was related to COVID-19, but the official statistics were not updated. Therefore, we discard the data points on deaths before 2 November 2020 which marks the beginning of the second wave of COVID-19 deaths. The period for which the data on COVID-19 deaths was included in our analysis is marked on Fig. 1B with the grey background.

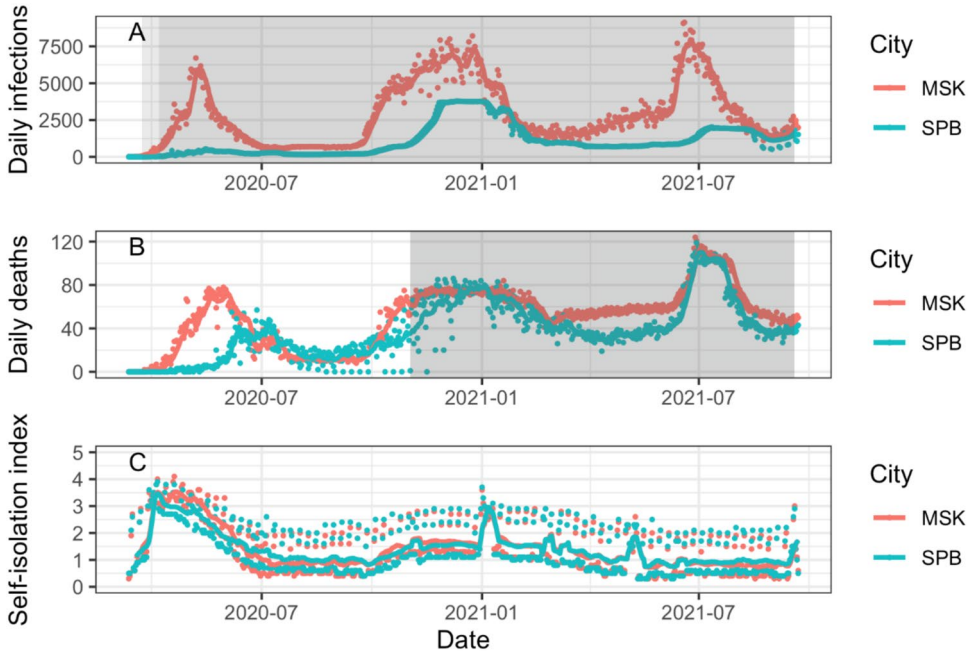

Fig. 1. Time series used in the study. (**A**) Daily reported new COVID-19 infections, (**B**) daily reported deaths, and (**C**) daily levels of the self-isolation index for Moscow (MSK) and St. Petersburg (SPB). Dots represent the original data while solid lines represent the rolling means over the previous 7 days. Shaded regions indicate the periods for which data was used in our analyses: 23 March 2020 – 19 September 2021 and 4 April 2020 – 19 September 2021 for the daily infections in Moscow and St. Petersburg, respectively, and 2 November 2020 – 19 September 2021 for the daily reported deaths (for both cities).

The SIR-inspired time series model

A visual inspection of Fig. 1 suggests the following relationship between the dynamics of COVID-19 and the self-isolation index: the waves of infections are preceded by the periods of low SII values, while higher SII levels appear to flatten the epidemic curve. If confirmed, such a relationship would justify SII_t as an informative indicator for explaining the observed dynamics of COVID-19, reflected by N_t or D_t. To explore this hypothesis, in this section we propose an autoregressive distributed-lag time series model inspired by the classical compartmental susceptible-infected-removed (SIR) model of infectious disease[52].

In the SIR model, at any given time t the number of new cases, N_t is proportional to the number of currently infected individuals I_t. In a discrete time, this relationship can be expressed as

$$N_t = \beta I_{t-1} = \beta \sum_{i=1}^{K} N_{t-i} \tag{1}$$

where K is an average period within which infected individuals remain contagious, and the proportionality coefficient β is interpreted as the average number of contacts per person times the probability of virus transmission (See Supplementary Information S1, where we show how the discrete-time Eq. (1) can be derived from the continuous-time SIR model). Applying the logarithmic transformation, Eq. (1) can be written in the additive form

$$\log N_t = \log \beta + \log \left(\sum_{i=1}^{K} N_{t-i} \right) \tag{2}$$

Equation (2) allows for the separation of the effects related to the intensity of people-to-people contacts and the probability of transmission, jointly captured by parameter β, and the effects related to the sheer number of infected individuals, captured by the sum of recent cases N_{t-i} over the period of the last K days.

It is important to note that Eq. (2) relies on the assumptions of a well-mixed and closed population, inherited from the SIR model. For large cities, such as Moscow and St. Petersburg, the assumption of a well-mixed population is plausible, since multiple public spaces and dense transport infrastructure enables uniform virus spread (i.e., without a tendency to develop spatially isolated clusters of infections). Populations of urban centers, however, cannot be considered closed. To represent the net flow of infected individuals to and from a city in question, we add a constant term μ to the right-hand side of Eq. (2).

To further improve the realism of model (2) we make two more adjustments. First, we observe that term $\log\left(\sum_{i=1}^{K} N_{t-i}\right)$ implies that the old cases N_{t-i} (with i close to K) contribute to the current number of cases N_t with the same weight as the more recent cases (with i close to 1). In other words, model (2) assumes that infected individuals that are close to recovery transmit the virus at the same rate as the newly infected persons. In the case of COVID-19, however, higher transmission rates are observed within the first several days of a COVID-19 infection, and then they gradually decline[53]. To reflect this reality in the model while maintaining its convenient linear form, we replace $\log\left(\sum_{i=1}^{K} N_{t-i}\right)$ with $\sum_{i=1}^{q} \gamma_i \log N_{t-i}$, where coefficients γ_i represent the effect strengths of the lagged values $\log N_{t-i}$ on the current value of $\log N_t$. Here $q \leq K$ is the time horizon within which these effects are not negligible. Importantly, coefficients γ_i should not be confused with transmission rates of individuals in the i-th day of infection.

Second, in Eq. (2), parameter β, i.e., the average number of contacts per person times the transmission probability, is constant. In reality, however, this parameter can vary from day to day, e.g., due to the social distancing measures that reduce the number of physical person-to-person contacts. While daily changes in the average number of contacts per person are impossible to track, we conjecture that they match the changes in the self-isolation index. Accordingly, term $\log\beta$ in Eq. (2), i.e., the effect of the logarithm of the average number of contacts per person (assuming constant transmission probability) on the current value of $\log N_t$, can be replaced with $\sum_{j=0}^{p} \beta_j \log SII_{t-j}$, where coefficients β_j represent the effect strengths of lagged values $\log SII_{t-j}$ on $\log N_t$. In other words, $\log\beta$ can be replaced with a linear model that regresses $\log N_t$ on past values of $\log SII_t$. The time horizon for detectable effects, p, should not exceed the length of the period within which newly infected persons can transmit the SARS-COV-2 virus, but it does not need to coincide with q.

Considering the abovementioned adjustments to Eq. (2), we propose the autoregressive distributed lag (ARDL) model $((M_1))$, given by formula (3), to explain the current level of $\log N_t$ with a linear combination of its past values (the autoregressive part) and past values of $\log SII_t$ (the distributed lag part) plus the white noise ε_t.

$$\log N_t = \mu + \sum_{i=1}^{q} \gamma_i \log N_{t-i} + \sum_{j=0}^{p} \beta_j \log SII_{t-j} + \varepsilon_t, \tag{3}$$

If reliable testing and/or registering the new COVID-19 cases is lacking, the daily reported new cases N_t may not reflect well the true epidemic curve. Yet, the true number of infections could be estimated using the reported numbers of COVID-19 fatalities[54]. While estimating the true numbers of infections based on the reported fatal cases of COVID-19 is beyond the scope of this paper, the abovementioned approach justifies the use of the daily reported deaths D_t in place of N_t in our analysis. Working with the time series of reported deaths D_t poses several problems, however. As visible on Fig. 1B, D_t exhibits considerable changes in volatility and on some days no deaths were reported, which precludes logarithmic transformation. Therefore, we smoothen the original data D_t using a 7-day rolling mean. This reduces the variance without obscuring the long-term patterns and allows us to apply logarithmic transformation to the time series (barring cases when no deaths were reported for 7 consecutive days). We also apply a 7-day rolling mean to $\log SII_t$. Thus, as an analog to model $((M_1))$, we propose model $((M_2))$ given by formula (4):

$$\log RM_7(D_t) = \mu + \sum_{i=1}^{q} \gamma_i \log RM_7(D_{t-i}) + \sum_{j=0}^{p} \beta_j \log RM_7(SII_{t-j}) + \varepsilon_t. \tag{4}$$

A reliable estimation of parameters of models $((M_1))$ and $((M_2))$ using the ordinary least squares (OLS) method requires that both processes $\log N_t$ and $\log SII_t$, and $\log RM_7(D_t)$ and $\log RM_7(SII_t)$, respectively, are stationary[55]. A suite of stationarity tests (cf. Supplementary Information S2, Supplementary Tables S2.1 and S2.2) compels us, however, to reject the hypotheses of stationarity of $\log N_t$, $\log SII_t$, $\log RM_7(D_t)$, and $\log RM_7(SII_t)$ – for both Moscow and St. Petersburg. Consequently, the OLS method cannot be used to estimate parameters of models (3) and (4) directly. However, the first differences of the abovementioned processes, $\Delta \log N_t$, $\Delta \log SII_t \Delta \log RM_7(D_t)$, and $\Delta \log RM_7(SII_t)$, do pass the stationary tests (see Supplementary Information S2, Supplementary Tables S2.3 and S2.4). Thus, for both cities, the analyzed time series are of type $I(1)$ (integrated of order one, i.e., stationary after applying difference operator once). This opens a possibility for employing the so-called error-corrected forms of ARDL models (M1) and (M2), for which OLS estimators of parameters are reliable.

Cointegration and the error-corrected form of models (M1) and (M2)

While both $\log N_t$ and $\log SII_t$ are non-stationary, there may exist a stable long-run relationship between them, which could be exploited for reliable OLS estimation of parameters. More specifically, there may exist vector $(\gamma, \theta) \in \mathbb{R}^2$ such that process $\gamma \log N_t + \theta \log SII_t$ is stationary – in which case processes $\log N_t$ and $\log SII_t$ are said to be cointegrated[56]. Notice that the mean of a stationary process is a constant, which can be included in the intercept term of the model, thus, without loss of generality, we can assume that the average of $\gamma \log N_t + \theta \log SII_t$ is zero. Consequently, cointegration implies that, on average, we expect to observe the long-run equilibrium relationship $\log N_t = -\frac{\theta}{\gamma} \log SII_t$.

If $\log N_t$ and $\log SII_t$ are cointegrated, then model $((M_1))$ given by formula (3) can be transformed into its so-called error-corrected form[57,58], defined by formula (5) below and denoted as $(M1_{EC})$:

$$\Delta \log N_t = \mu + \gamma \log N_{t-1} + \theta \log SII_{t-1} + \sum_{i=1}^{q-1} \alpha_i \Delta \log N_{t-i} + \sum_{j=0}^{p-1} \phi_j \Delta \log SII_{t-j} + e_t. \quad (5)$$

Cointegration of $\log N_t$ and $\log SII_t$ implies that both sides of the formula above consist of stationary processes, which offers a possibility for a reliable estimation of parameters of model $(M1_{EC})$. More specifically, OLS estimators of these parameters are consistent if: (A) errors e_t are serially independent with zero mean and a constant variance; and (B) errors e_t are uncorrelated with $\Delta \log SII_{t-h}$ for all lags $h \in \mathbb{Z}$ (i.e., $\Delta \log SII_t$ is exogenous)[55]. Parameters of the original model $(M1)$ can then be recovered using the following relationships:

$$\begin{aligned}
&\gamma_1 = \gamma + 1 + \alpha_1,\\
&\gamma_i = \alpha_i - \alpha_{i-1}, i = 2, \ldots, q-1,\\
&\gamma_q = -\alpha_{q-1},\\
&\beta_0 = \theta_0,\\
&\beta_1 = \theta + \phi_1 - \phi_0\\
&\beta_j = \phi_j - \phi_{j-1}, j = 2, \ldots, p-1,\\
&\beta_p = -\phi_{p-1}.
\end{aligned} \quad (6)$$

The error-corrected model $(M1_{EC})$ has an important practical interpretation. The ARDL term $\sum_{i=1}^{q-1} \alpha_i \Delta \log N_{t-i} + \sum_{j=0}^{p-1} \phi_j \Delta \log SII_{t-j}$ represents the short-term dynamics of $\Delta \log N_t$ driven by the past changes $\Delta \log N_t$ and $\Delta \log SII_t$. The short-term dynamics is not influenced by the level of $\log SII_t$, however, regardless of whether it is low or high. Yet, in the presence of a long-run equilibrium $\log N_t = -\frac{\theta}{\gamma} \log SII_t$, any given level of $\log SII_t$ will induce an adjustment of values of $\log N_t$ towards its new equilibrium value of $-\frac{\theta}{\gamma} \log SII_t$. Thus, to properly describe the evolution of $\Delta \log N_t$, the short term-dynamics needs to be corrected (hence the name of error correction model) with the term $\gamma \log N_{t-1} + \theta \log SII_{t-1} = -\gamma \left(-\frac{\theta}{\gamma} \log SII_{t-1} - \log N_{t-1}\right)$, where the expression in parentheses is the difference between the expected equilibrium value of $\log N_t$, equal to $-\frac{\theta}{\gamma} \log SII_{t-1}$, and the value of $\log N_{t-1}$ observed at the previous time step. The constant $-\gamma$ is interpreted as the rate of convergence to the new equilibrium.

In a similar fashion as above, if $\log RM_7(D_t)$ and $\log RM_7(SII_t)$ are cointegrated, then the model $(M2)$ can be transformed into its error-corrected form $(M1_{EC})$ given by formula (7) below:

$$\begin{aligned}
\Delta \log RM_7(D_t) = &\mu + \gamma \log RM_7(D_{t-1})\\
&+ \theta \log RM_7(SII_{t-1})\\
&+ \sum_{i=1}^{q-1} \alpha_i \Delta \log RM_7(D_{t-i})\\
&+ \sum_{j=0}^{p-1} \phi_j \Delta \log RM_7(SII_{t-j}) + e_t.
\end{aligned} \quad (7)$$

The conditions (A) and (B) for reliable estimation of parameters of model $(M2_{EC})$ are essentially the same as for the model $(M1_{EC})$, with $\Delta \log RM_7(SII_{t-h})$ replacing $\Delta \log SII_{t-h}$ in condition (B). Coefficients of the model $(M2)$ can be recovered using Eq.(6).

Estimation of parameters of error-corrected model and cointegration tests

An error-corrected ARDL model can be employed in testing for cointegration if its parameters can be reliably estimated using the OLS method, i.e., if conditions (A) and (B) above are satisfied[58]. The procedure for testing for cointegration involves: (i) estimation of parameters of the error-corrected ARDL model including the optimal orders q^* and p^* of the autoregressive and distributed lag parts of the model, respectively, and (ii) testing the null hypothesis $H_0 : \gamma = \theta = 0$[58,59].

In the Results section below, we present the results of cointegration tests between $\log N_t$ and $\log SII_t$, and $\log RM_7(D_t)$ and $\log RM_7(SII_t)$ obtained with use of the R package dLagM, version 1.1.8[60]. To carry out step (i), we use the ardlBoundOrders function. The function performs a search over combinations of orders q and p within the pre-specified limits $q_{\max}$ and $p_{\max}$. For each combination of q and p, the function estimates an error-correction ARDL model and it returns q^* and p^* for which the corresponding estimated model minimizes the selected goodness-of-fit criterion. As a goodness-of-fit criterion we use the Bayesian Information Criterion (BIC) since it strongly penalises the number of parameters in a model and thus reduces the risk of overfitting. To carry out step (ii), we use the obtained optimal orders q^* and p^* as parameters in the ardlBound function. This function fits an error-corrected ARDL model of the specified orders, performs the bounds test for cointegration as described in[59] and displays the model's diagnostics.

Bounds test for cointegration		
Critical value	$I(0)$ bound	$I(1)$ bound
10%	4.04	4.78
5%	4.94	5.73
1%	6.84	7.84

Table 1. Cointegration test based on the $(\mathrm{M1_{EC}})$ model for the Moscow data. F-statistic > $I(1)$ bound for critical value of 1%. F-statistic = 11.48. Conclusion: The null hypothesis of cointegration not rejected at 0.01 significance level.

Coefficients of model $(M1_{EC})$				
Coefficient	Estimate	Std. error	*t* statistic	*p* value
μ	0.148	0.060	2.460	0.014 (*)
γ	−0.019	0.008	−2.434	0.015 (*)
θ	−0.005	0.012	−0.423	0.673
ϕ_0	0.019	0.014	1.404	0.161
ϕ_1	−0.019	0.015	−1.254	0.210
ϕ_2	0.042	0.014	3.006	0.003 (**)
ϕ_3	−0.090	0.014	−6.531	1.56e−10 (***)
α_1	−0.433	0.042	−10.298	< 2e−16 (***)
α_2	−0.138	0.045	−3.082	0.002 (**)
α_3	−0.022	0.041	−0.544	0.587
α_4	0.017	0.040	0.428	0.669
α_5	−0.049	0.040	−1.211	0.226
α_6	0.171	0.040	4.218	2.90e−05 (***)
α_7	0.370	0.040	9.171	< 2e−16 (***)
α_8	0.285	0.040	7.202	2.09e−12 (***)
α_9	0.091	0.037	2.430	0.015 (*)

Table 2. Parameters of the $(\mathrm{M1_{EC}})$ model fitted to the Moscow data. Multiple R^2: 0.42, Adjusted R^2: 0.41. F-statistic: 22.42 on 15 and 520 DF, *p* value: < 2.2e−16. Significance codes used: $0 \leq$ (***) $\leq 0.001 \leq$ (**) $\leq 0.01 \leq$ (*) $\leq 0.05 \leq$ (.) ≤ 0.1.

Results

The relationship between daily reported new infections and the self-isolation index

To explore the relationship between $\log N_t$ and $\log SII_t$ we perform the cointegration tests based on model $(M1_{EC})$. In the testing procedure (i.e., steps (i) and (ii) described in the previous subsection), the upper limits on the orders of the autoregressive and distributed lag components are set to $q_{max} = p_{max} = 14$ days, which is a typical duration of a symptomatic COVID-19 infection[53].

The output of the ardlBound function, including the model diagnostics, is presented in full in Appendices S3 and S4 for the Moscow and St. Petersburg data, respectively. There, we discuss conditions (A) and (B) for a reliable OLS estimation of parameters and we assess the quality of the fitted models $(M1_{EC})$. Below, we present the results of cointegration test. If the presence of cointegration is concluded, we also present the estimated parameters of model $(M1_{EC})$, reconstruct the parameters of the model $(M1)$, and interpret the results.

Cointegration between new infections and the self-isolation index for Moscow

Table 1 summarizes the results of the bounds test for cointegration. The value of the F-statistic for the bounds test is above the $I(1)$ threshold for the critical value of 1%, which provides strong evidence for cointegration between $\log N_t$ and $\log SII_t$.

As discussed in Supplementary Information S3, conditions (A) and (B) for a reliable OLS estimation of parameters of model $(M1_{EC})$ are satisfied and we consider the quality of the fitted model to be good. The estimated coefficients of model $(M1_{EC})$ and the results of significance tests are presented in Table 2. From the estimated coefficients of model $(M1_{EC})$ (displayed on Fig. 2A and C), we recover the coefficients of the original ARDL model $(M1)$ using Eq. (3)–see Fig. 2B and D.

The estimated values of the error correction coefficients are $\gamma = -0.019$ and $\theta = -0.005$, implying the long-run equilibrium relationship $\log N_t = -\frac{\theta}{\gamma}\log SII_t = -0.26\log SII_t$. That is, an increase in $\log SII_t$ by 1 induces $\log N_t$ to eventually decrease by 0.26, and the rate of this adjustment (i.e., convergence to the new

equilibrium value) is equal to $-\gamma = 0.019$ per day. This agrees with the empirical observation that limiting people-to-people interactions leads to a reduction in the numbers of COVID-19 cases.

As the long-run equilibrium is restored rather slowly (adjustments of about 2% a day), it has a relatively small impact on the short-term dynamics of $\Delta \log N_t$. Indeed, the level of $\log SII_{t-1}$ has a negligible effect on $\Delta \log N_t$. However, $\Delta \log SII_{t-2}$ and $\Delta \log SII_{t-3}$ have statistically significant, even if small, effects on $\Delta \log N_t$. This suggests that the daily numbers of new infections react to changes in the level of self-isolation index with a delay of 2–3 days. This apparent delay coincides with the observation that people infected with SARS-CoV-2 can themselves become infectious one to two days before the onset of symptoms[61] (plus one day for results of positive tests to appear in the published statistics).

The autoregressive part of thestimated model $(M1_{EC})$ indicates that the short-run dynamics of $\Delta \log N_t$. is driven mainly by the past changes $\Delta \log N_{t-h}$ for $h = 1, \ldots 9$. For small h, the impact of $\Delta \log N_{t-h}$ on $\Delta \log N_t$ is negative, gradually becoming zero around $h = 4$, then further increasing and reaching a peak at $h = 7$, and then tapering off as h approaches 9, as presented on Fig. 2C. This agrees with the observed dynamics of COVID-19 infections, where the number of new infections is driven by the total number of active cases. Indeed, if the number of cases is on the rise for some time prior to $t-1$, then, even as $\Delta \log N_{t-1} < 0$, the number of new cases N_t is pushed up by the active cases detected more than 1 day ago. A continued decrease in the daily number of cases overcomes this momentum within 4–5 days and then reverses it, which manifests itself in $\Delta \log N_{t-h}$ becoming gradually positively impacting $\Delta \log N_t$ as h approaches 7. This impact decreases for h approaching 9. Lags $h \geq 10$ were not included in the model, which suggests that the cases older than 10 days do not contribute significantly to the increase in new cases. The contributions of the past infections, i.e., of active cases i days old, to the current number of daily infections are represented by coefficients γ_i in the autoregressive part of model $(M1)$ – cf. Fig. 2D.

Cointegration between the new infections and the self-isolation index for St. Petersburg

As discussed in Supplementary Information S4, condition (A) for a reliable OLS estimation of parameters is not satisfied and the quality of the fitted model $(M1_{EC})$ is poor. This undermines validity of the cointegration test, results of which (included in Supplementary Information S4) are thus not presented here. Available data for St. Petersburg do not support conclusive evidence for or against the hypothesis of cointegration between $\log N_t$ and $\log SII_t$. We highlight potential quality issues with data for St. Petersburg in the Discussion.

The relationship between the COVID-19 deaths and the self-isolation index

To investigate the existence of a long-run relationship between the self-isolation index and reported deaths related to COVID-19 we perform tests for cointegration between $\log RM_7(D_t)$ and $\log RM_7(SII_t)$ based on model $(M2_{EC})$. The testing procedure is essentially the same as the one employing model $(M1_{EC})$. However, to accommodate a considerable uncertainty in the time between contracting COVID-19 and death – ranging between a couple of days and a couple of weeks – we increase the upper limit for the orders of autoregressive and distributed-lag terms to $q_{max} = p_{max} = 21$ days. The full output of the ardlBound function, discussion of conditions for a reliable OLS estimation and assessment of the quality of models $(M2_{EC})$ fitted to the Moscow and St. Petersburg data are presented in Appendices S5 and S6, respectively.

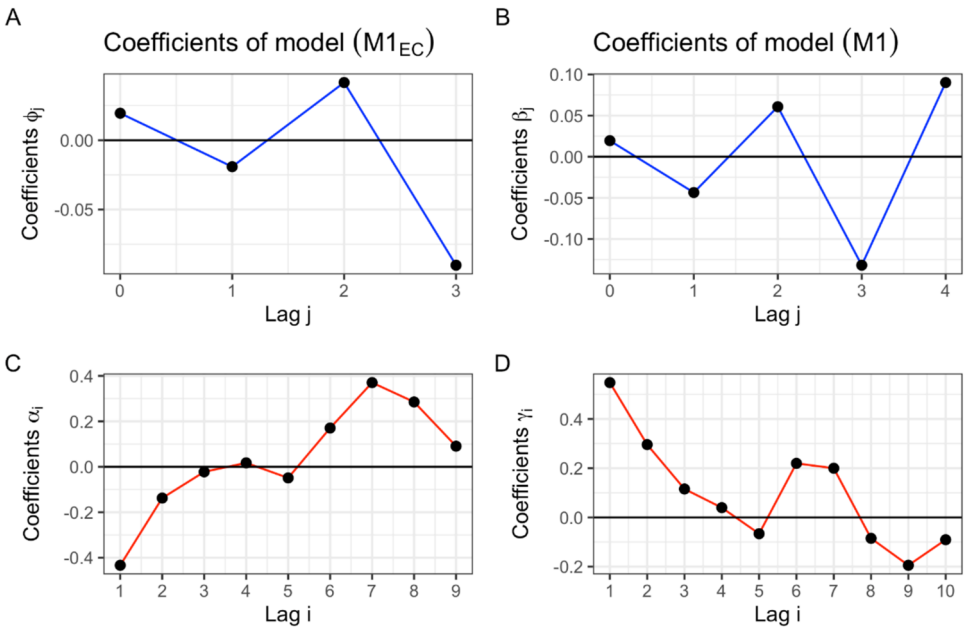

Fig. 2. Plots of model coefficients estimated for Moscow. (**A**) Distributed lag coefficients of model $(M1_{EC})$. (**B**) Distributed lag coefficients of model $(M1)$. (**C**) Coefficients of autoregressive part of model $(M1_{EC})$. (**D**) Coefficients of autoregressive part of model $(M1)$.

It is important to note that uncertainty in the time between the COVID-19 infection and death blurs any short-term responses of $\log D_t$ to changes in $\log SII_t$. This is further compounded by smoothening of these time series with 7-days rolling means. Consequently, coefficients of the model $(M2_{EC})$ do not lend themselves to easy interpretation. Hence, in this section, we will not attempt to draw conclusions from the values of short-term coefficients α_i and ϕ_j of model $(M2_{EC})$.

Cointegration between deaths and the self-isolation index for Moscow

The results of the bounds test for cointegration between $\log RM_7(D_t)$ and $\log RM_7(SII_t)$ applied to the Moscow data are presented in Table 3. The value of the F-statistic is above the $I(1)$ bound for the critical value of 5%, thus the hypothesis of cointegration is not rejected at the significance level of 0.05.

Despite slight violation of condition (A) for a reliable OLS estimation (light serial correlation in the residuals) we consider the fitted model $(M2_{EC})$ to have satisfactory quality – cf. Supplementary Information S5. Thus, we have good confidence in the results of the cointegration test and in the estimates of parameters of model $(M2_{EC})$, presented in Table 4.

The long-run equilibrium between $\log RM_7(D_t)$ and $\log RM_7(SII_t)$ is given by the equation $\log RM_7(D_t) = -\frac{\theta}{\gamma}\log RM_7(SII_t) = -0.055\log RM_7(SII_t)$. That is, on average, an increasing $\log RM_7(SII_t)$ by 1 is expected to decrease $\log RM_7(D_t)$ by 0.055. The rate of convergence to the equilibrium is equal to $-\gamma = 0.006$ per day. Qualitatively, this is in line with the observed dynamics of COVID-19 (social distancing measures result in fewer infections and thus in fewer fatal cases), as well as with the estimated equilibrium relationship between $\log N_t$ and $\log SII_t$ (although here both the equilibrium coefficient, and the rate of convergence are an order of magnitude smaller). This further strengthens our confidence in the soundness of conclusions drawn from the model $(M2_{EC})$ for Moscow.

Cointegration between deaths and the self-isolation index for St. Petersburg

Table 5 presents the results of the bounds test for cointegration between $\log RM_7(D_t)$ and $\log RM_7(SII_t)$ recorded for St. Petersburg. The value of the F-statistic is just above the $I(1)$ bound for the critical value of 5%. Thus, we conclude that the hypothesis of cointegration is not rejected at the significance level of 0.05.

As discussed in Supplementary Information S6, the conditions for a reliable OLS estimation of parameters are satisfied to a satisfactory degree and we consider the quality of the fitted model $(M2_{EC})$ to be good. We are, therefore, confident in the results of the cointegration test and in the estimated parameters of model $(M2_{EC})$, which are presented in Table 6.

The estimated coefficient $-\frac{\theta}{\gamma}$ in the long-run equilibrium $\log RM_7(D_t) = -\frac{\theta}{\gamma}\log RM_7(SII_t)$ for St. Petersburg is equal to 0.58, and the rate of convergence is equal to $-\gamma = 0.008$. Qualitatively, this is consistent with our findings for Moscow obtained with both models $(M1_{EC})$ and $(M2_{EC})$.

Discussion and conclusions

In this paper we proposed two error-corrected ARDL models, $(M1_{EC})$ and $(M2_{EC})$, which allow us to investigate the existence of a stable long-run relationship (cointegration) between the numbers of newly detected COVID-19 infections and the numbers of fatal cases, respectively, and the self-isolation index.

Applying model $(M1_{EC})$ to the data available for Moscow, we found very strong evidence for cointegration between $\log N_t$ and $\log SII_t$. For St. Petersburg, however, we were unable to conclude whether such cointegration relationship exists or not. In our opinion, this is due to a questionable quality of the official data on the daily numbers of new COVID-19 cases reported in this city. Indeed, the first wave of infections in spring 2020, clearly seen in Moscow's records, is barely visible in the St. Petersburg data (see Fig. 1A). Moreover, the two following waves (November–December 2020 and August–September 2021) have very unusual shapes, with an initial exponential-like growth sharply transiting to elevated but nearly constant levels of the new daily infections. This is in a stark contrast to the typically observed patterns of sharp rises in the numbers of new cases, followed in quick succession by comparably steep declines (as, e.g., in the case for Moscow). A very low volatility (i.e., deviations from the 7-days rolling mean) of the daily reported numbers of the new COVID-19 cases further rises our suspicions about reliability of the official infection statistics for St. Petersburg. Furthermore, these statistics also appear to be inconsistent with the number of COVID-19 related deaths reported in St. Petersburg, as the second and third wave of reported deaths slightly precede – rather than lag behind – the corresponding waves of the newly reported cases. Finally, the official statistics for St. Petersburg imply the case-fatality ratio that is 2–3 times higher than such ratio for Moscow, which also appears unlikely. Thus, we conjecture that insufficient

Bounds test for cointegration		
Critical value	$I(0)$ bound	$I(1)$ bound
10%	4.04	4.78
5%	4.94	5.73
1%	6.84	7.84

Table 3. Cointegration test based on the $(M2_{EC})$ model for the Moscow data. F-statistic > $I(1)$ bound for critical value of 5% F-statistic = 6.08 Conclusion: The null hypothesis of cointegration not rejected at 0.05 significance level.

Coefficients of model ($M2_{EC}$)				
Coefficient	**Estimate**	**Std. error**	***t* statistic**	***p* value**
μ	0.024	0.011	2.078	0.039 (*)
γ	−0.006	0.003	−2.092	0.037 (*)
θ	−3.15e−04	0.002	−0.172	0.864
ϕ_0	0.010	0.016	0.639	0.523
ϕ_1	−0.041	0.017	−2.397	0.017 (*)
ϕ_2	0.029	0.017	1.678	0.095 (.)
ϕ_3	0.022	0.017	1.249	0.213
ϕ_4	0.006	0.017	0.340	0.734
ϕ_5	−0.038	0.017	−2.185	0.030 (*)
ϕ_6	0.025	0.018	1.403	0.162
ϕ_7	0.005	0.020	0.272	0.786
ϕ_8	−0.015	0.019	−0.791	0.430
ϕ_9	0.021	0.019	1.102	0.271
ϕ_{10}	0.006	0.019	0.313	0.755
ϕ_{11}	0.015	0.019	0.780	0.436
ϕ_{12}	−0.049	0.019	−2.587	0.010 (*)
ϕ_{13}	−2.72e−05	0.019	−0.001	0.999
ϕ_{14}	0.068	0.019	3.502	0.001 (***)
ϕ_{15}	−0.024	0.017	−1.361	0.175
ϕ_{16}	0.016	0.017	0.941	0.348
ϕ_{17}	−0.010	0.017	−0.594	0.553
ϕ_{18}	0.017	0.017	0.965	0.335
ϕ_{19}	−0.021	0.017	−1.206	0.229
ϕ_{20}	−0.029	0.017	−1.690	0.092 (.)
ϕ_{21}	0.023	0.017	1.409	0.160
α_1	0.414	0.059	7.065	1.40e−11 (***)
α_2	0.222	0.058	3.814	1.70e−04 (***)
α_3	0.258	0.060	4.291	2.49e−05 (***)
α_4	0.041	0.062	0.664	0.507
α_5	0.058	0.063	0.916	0.361
α_6	0.031	0.061	0.509	0.611
α_7	−0.452	0.059	−7.722	2.32e−13 (***)
α_8	0.286	0.059	4.855	2.06e−06 (***)

Table 4. Parameters of the model $(\mathrm{M2_{EC}})$ fitted to the Moscow data. Multiple R^2: 0.66, Adjusted R^2: 0.62. F-statistic: 16.41 on 32 and 267 DF, *p* value: < 2.2e−16. Significance codes used: $0 \leq$ (***) $\leq 0.001 \leq$ (**) $\leq 0.01 \leq$ (*) $\leq 0.05 \leq$ (.) ≤ 0.1.

Bounds test for cointegration		
Critical value	$I(0)$ **bound**	$I(1)$ **bound**
10%	4.04	4.78
5%	4.94	5.73
1%	6.84	7.84

Table 5. Cointegration test based on the $(\mathrm{M2_{EC}})$ model for the St. Petersburg data. F-statistic > $I(1)$ bound for critical value of 5% F-statistic = 5.74 Conclusion: The null hypothesis of cointegration not rejected at 0.05 significance level.

Coefficients of model ($M2_{EC}$)				
Coefficient	**Estimate**	**Std. error**	***t* statistic**	***p* value**
μ	0.032	0.014	2.260	0.025 (*)
γ	−0.008	0.004	−2.220	0.027 (*)
θ	−0.005	0.004	−1.075	0.283
ϕ_0	−0.058	0.028	−2.078	0.039 (*)
ϕ_1	−0.062	0.030	−2.093	0.037 (*)
ϕ_2	0.085	0.028	3.038	0.003 (**)
α_1	0.406	0.056	7.275	3.15e−12 (***)
α_2	0.212	0.059	3.572	4.13e−04 (***)
α_3	0.048	0.060	0.806	0.421
α_4	0.027	0.054	0.504	0.615
α_5	0.057	0.054	1.058	0.291
α_6	0.041	0.054	0.759	0.448
α_7	−0.380	0.054	−7.098	9.44e−12 (***)
α_8	0.147	0.057	2.554	0.011 (*)
α_9	0.119	0.057	2.069	0.039 (*)
α_{10}	0.142	0.054	2.616	0.009 (**)

Table 6. Parameters of the model $(\mathrm{M2_{EC}})$ fitted to the St. Petersburg data. Multiple R^2: 0.48, Adjusted R^2: 0.45. F-statistic: 17.89 on 15 and 295 DF, *p* value: < 2.2e−16. Significance codes used: $0 \leq$ (***) $\leq 0.001 \leq$ (**) $\leq 0.01 \leq$ (*) $\leq 0.05 \leq$ (.) ≤ 0.1.

evidence for cointegration between number of COVID-19 cases and SII for St. Petersburg is more likely due to unreliability of data rather than due to inadequacy of model $(M1_{EC})$.

This conclusion is corroborated by our findings obtained with model $(M2_{EC})$, which uses the 7-day rolling mean of daily reported numbers of fatal cases of COVID-19 in place of the numbers of infections. With the help of this model, we have found strong evidence for cointegration between $\log RM_7(D_t)$ and $\log RM_7(SII_t)$ for both Moscow and St. Petersburg. This suggests that the numbers of COVID-19 related deaths may be a more robust, if less direct, proxy for the true dynamics of the pandemic in case the data on the numbers of new COVID-19 cases is unreliable.

The detected cointegration relationships between $\log SII_t$ and $\log N_t$, and between $\log RM_7(SII_t)$ and $\log RM_7(D_t)$ allows us to draw a conclusion that Yandex's self-isolation index is a useful metric for monitoring the level of intensity of people-to-people contacts within a population.

The long-run equilibrium between $\log N_t$ and $\log SII_t$ implied by cointegration has considerable practical importance. First, it informs us about the order of magnitude of changes to $\log N_t$ that could be brought about by limiting person-to-person contacts measured by the self-isolation index. Second, the rate at which $\log N_t$ adjust to changed levels of $\log SII_t$ gives us a measure of the expected time delays before social-distancing policies start having the desired effects on the dynamic of the pandemic. For Moscow, the approximate equilibrium is given by the equation $\log N_t = -0.26 \log SII_t$ and the rate of adjustments of $\log N_t$ to a new level of $\log SII_t$ of 2% per day. A qualitatively similar long-run relationship was detected between $\log RM_7(D_t)$ and $\log RM_7(SII_t)$, with the equilibrium coefficient ranging between -0.05 for Moscow and -0.58 for St. Petersburg, and the rate of convergence less than 1% a day. This could indicate that social distancing measures may be more effective in limiting the number of COVID-19 related deaths rather than the overall number of cases, but it takes longer for the results of such measures to make a detectable impact.

The error-corrected ARDL models $(M1_{EC})$ and $(M2_{EC})$ proposed in this paper were employed in a diagnostic mode, with the purpose of testing for cointegration between the self-isolation index and the time series reflecting COVID-19 dynamics. We advise caution, in employing them in a prognostic mode, especially for long-term projections. The structure of our models was chosen for the ease of interpretation, for reliability of parameter estimators, and for the possibility of relating them to the classical SIR model of epidemic dynamics. More advanced models, such as partially observed Markov processes used in[7], which feature better mechanistic representation of the COVID-19 dynamics, and which account for changing levels of vaccine-induced immunity as well as varying transmissibility of the virus (e.g., due to emergence of new variants of the SARS-Cov-2) may have better predictive performance. Moreover, the aggregation methodology of condensing diverse data into a single SII value may be suboptimal in terms of capturing the correlation between the patterns of user activity on the Yandex platform and actual people-to-people contact rates. Non-linear SII transformations could enhance this correlation and thus act as even better predictors of COVID-19 dynamics, however, at the cost of less straightforward interpretation of results. In this paper, we opted to work with the SII in its original scale.

In conclusion, our results obtained for the Yandex's self-isolation index in Moscow and St. Petersburg suggest that similar aggregated indices reflecting the intensity of people-to-people contacts based on anonymized user mobility and online activity data collected by Yandex and other digital platform may be useful for monitoring in real-time the level of population compliance with social distancing measures in other parts of the world beyond Russia. Caution is advised, however, until additional data covering other regions become available and allow

for more extensive testing of the performance of aggregate user activity indices as proxies for people-to-people contact rates. This testing is necessary to establish whether such indices suffer from inconsistent predictive power across different regions and scales, as was the case with pure mobility metrics (cf.[37]). Statistical models presented in this paper establish a method of testing whether aggregate user activity indices akin to SII can serve as reliable monitoring tools and support decision-making of public health authorities.

Data availability

The relevant data is available in the Supporting Information files, together with the R code replicating the presented analysis. The data set was composed from publicly available sources: Statistics of new COVID-19 cases and deaths due to COVID-19 in Russia ([https://datalens.yandex/7o7is1q6ikh23?tab=X1]) and the Yandex's Self-isolation index ([https://datalens.yandex/7o7is1q6ikh23?tab=q6]).

Received: 8 October 2024; Accepted: 13 October 2025
Published online: 19 November 2025

References

1. Coronavirus disease (COVID-19). *World Health Organization* https://www.who.int/health-topics/coronavirus (2024).
2. Coronavirus (COVID-19) Dashboard. *World Health Organization* https://covid19.who.int (2024).
3. Tooze, A. *Shutdown: How covid shook the world's economy*. (Viking, 2021).
4. Wang, C. C. et al. Airborne transmission of respiratory viruses. *Science* **373**, eabd9149. https://doi.org/10.1126/science.abd9149 (2021).
5. Prather, K. A., Wang, C. C. & Schooley, R. T. Reducing transmission of SARS-CoV-2. *Science* **368**, 1422–1424 (2020).
6. Mathieu, E. *et al*. Coronavirus pandemic (COVID-19). https://ourworldindata.org/coronavirus (2020).
7. He, D. et al. Evaluation of effectiveness of global COVID-19 vaccination campaign. *Emerg. Infect. Dis.* **28**(9), 1873–1876 (2022).
8. Horita, N. & Fukumoto, T. Global case fatality rate from COVID-19 has decreased by 968% during 25 years of the pandemic. *J. Med. Virol.* **95**, e28231. https://doi.org/10.1002/jmv.28231 (2022).
9. Menegale, F. et al. Evaluation of waning of SARS-CoV-2 vaccine–induced immunity. *JAMA Netw. Open.* **6**(5), e2310650. https://doi.org/10.1001/jamanetworkopen.2023.10650 (2023).
10. Perra, N. Non-pharmaceutical interventions during the COVID-19 pandemic: A review. *Phys. Rep.* **913**, 1–52. https://doi.org/10.1016/j.physrep.2021.02.001 (2021).
11. Flaxman, S. et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
12. Auger, K. A. et al. Association between statewide school closure and COVID-19 incidence and mortality in the US. *JAMA* **324**(9), 859–870 (2020).
13. Vokó, Z. & Pitter, J. G. The effect of social distance measures on COVID-19 epidemics in Europe: an interrupted time series analysis. *GeroScience.* **42**(4), 1075–1082 (2020).
14. Leung, K., Wu, J. T., Liu, D. & Leung, G. M. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet* **395**(10233), 1382–1393 (2020).
15. Haug, N. et al. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* **4**(12), 1303–1312 (2020).
16. Chung, H. W. et al. Effects of government policies on the spread of COVID-19 worldwide. *Sci. Rep.* **11**, 20495. https://doi.org/10.1038/s41598-021-99368-9 (2021).
17. Brauner, J. M. et al. Inferring the effectiveness of government interventions against COVID-19. *Science* **371**(6531), eabd9338. https://doi.org/10.1126/science.abd9338 (2021).
18. Wilk, A. M., Łakomiec, K., Psiuk-Maksymowicz, K. & Fujarewicz, K. Impact of government policies on the COVID-19 pandemic unraveled by mathematical modelling. *Sci. Rep.* **12**, 16987. https://doi.org/10.1038/s41598-022-21126-2 (2022).
19. Hsiang, S. et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**(7820), 262–267 (2020).
20. Schröder, M. et al. COVID-19 in South Africa: outbreak despite interventions. *Sci. Rep.* **11**, 4956. https://doi.org/10.1038/s41598-021-84487-0 (2021).
21. Alfano, V. & Ercolano, S. The efficacy of lockdown against COVID-19: A cross-country panel analysis. *Appl. Health Econ. Health Policy* **18**, 509–517 (2020).
22. Maloney, W. F., Taskin, T. Determinants of social distancing and economic activity during COVID-19: A global view. World bank policy research working paper No. 9242. http://documents.worldbank.org/curated/en/325021589288466494 (2020).
23. Gauvin, L. et al. Socio-economic determinants of mobility responses during the first wave of COVID-19 in Italy: from provinces to neighbourhoods. *J. R. Soc. Interface.* **18**, 2021009220210092. https://doi.org/10.1098/rsif.2021.0092 (2021).
24. Kantor, B. & Kantor, J. Non-pharmaceutical interventions for pandemic COVID-19: A cross-sectional investigation of US general public beliefs, attitudes, and actions. *Front. Med.* **7**, 384. https://doi.org/10.3389/fmed.2020.00384 (2020).
25. Xu, H. et al. Relationship between COVID-19 infection and risk perception, knowledge, attitude, and four nonpharmaceutical interventions during the late period of the COVID-19 epidemic in China: online cross-sectional survey of 8158 adults. *J. Med. Internet Res.* **22**(11), e21372. https://doi.org/10.2196/21372 (2020).
26. Chan, H. F., Skali, A., Savage, D. A., Stadelmann, D. & Torgler, B. Risk attitudes and human mobility during the COVID-19 pandemic. *Sci. Rep.* **10**, 19931. https://doi.org/10.1038/s41598-020-76763-2 (2020).
27. Alfano, V. & Ercolano, S. Social capital, quality of institutions and lockdown. Evidence from Italian provinces. *Struct. Chang. Econ. Dyn.* **59**, 31–41 (2021).
28. Quilty, B. J. et al. The effect of travel restrictions on the geographical spread of COVID-19 between large cities in China: A modelling study. *BMC Med.* **18**, 259. https://doi.org/10.1186/s12916-020-01712-9 (2020).
29. Morley, C. P. et al. Social distancing metrics and estimates of SARS-CoV-2 transmission rates: Associations between mobile telephone data tracking and R. *J. Public Health Manag. Pract.* **26**(6), 606–612 (2020).
30. Gerlee, P. et al. Predicting regional COVID-19 hospital admissions in Sweden using mobility data. *Sci. Rep.* **11**, 24171. https://doi.org/10.1038/s41598-021-03499-y (2021).
31. COVID-19 Community mobility reports. *Google*: https://www.google.com/covid19/mobility/ (2022).
32. Sulyok, M. & Walker, M. Community movement and COVID-19: a global study using Google's community mobility reports. *Epidemiol Infect.* **148**, e284. https://doi.org/10.1017/S0950268820002757 (2020).
33. García-Cremades, S. et al. Improving prediction of COVID-19 evolution by fusing epidemiological and mobility data. *Sci. Rep.* **11**, 15173. https://doi.org/10.1038/s41598-021-94696-2 (2021).
34. Bryant, P. & Elofsson, A. Estimating the impact of mobility patterns on COVID-19 infection rates in 11 European countries. *PeerJ* **8**, 9879 (2020).

35. Deng, Y., Lin, H., He, D. & Zhao, Y. Trending on the use of Google mobility data in COVID-19 mathematical models. *Adv. Cont. Discr. Mod.* **2024**, 21. https://doi.org/10.1186/s13662-024-03816-5 (2024).
36. Ilin, C. et al. Public mobility data enables COVID-19 forecasting and management at local and global scales. *Sci. Rep.* **11**, 13531. https://doi.org/10.1038/s41598-021-92892-8 (2021).
37. Delussu, F., Tizzoni, M. & Gauvin, L. The limits of human mobility traces to predict the spread of COVID-19: A transfer entropy approach. *PNAS Nexus* **2**, 10. https://doi.org/10.1093/pnasnexus/pgad302 (2023).
38. Jewell, S. et al. It's complicated: characterizing the time-varying relationship between cell phone mobility and COVID-19 spread in the US. *NPJ Digit. Med.* **4**, 152. https://doi.org/10.1038/s41746-021-00523-3 (2021).
39. Koher, A., Jørgensen, F., Petersen, M. B. & Lehmann, S. Epidemic modelling of monitoring public behavior using surveys during pandemic-induced lockdowns. *Commun. Med.* **3**, 80. https://doi.org/10.1038/s43856-023-00310-z (2023).
40. Crawford, F. W. et al. Impact of close interpersonal contact on COVID-19 incidence: Evidence from 1 year of mobile device data. *Sci. Adv.* **8**, eabi5499. https://doi.org/10.1126/sciadv.abi5499 (2022).
41. Self-isolation Index. *Yandex:* https://yandex.ru/company/researches/2020/podomam (2021).
42. Search engines in Russia. *Yandex Radar*: https://radar.yandex.ru/search?period=all&device-category=5, (2025).
43. Krivorotko, O., & Zyatkov, N. Modeling of the COVID-19 epidemic in the Russian regions based on deep learning. In *2023 5th international conference on problems of cybernetics and informatics (PCI), Baku, Azerbaijan*; https://doi.org/10.1109/PCI60110.2023.10325993 (2023).
44. Coronavirus: Dashboard (Statistics: Russian Federation). *Yandex DataLens* https://datalens.yandex/7o7is1q6ikh23?tab=X1 (2021).
45. Coronavirus: Dashboard (Self-isolation). *Yandex DataLens* https://datalens.yandex/7o7is1q6ikh23?tab=q6 (2021).
46. Aleksey Begin. How many users are there in Yandex? (Сколько пользователей в Яндексе?). *Inclient*: https://inclient.ru/yandex-stats/ (2025).
47. Population ages 65 and above (% of total population)—Russian Federation. *World Bank*: https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?locations=RU (2025).
48. Yandex statistics for 2022 (Статистика Яндекса в 2022 году). *ER10:* https://er10.kz/read/texnologii/obzory/statistika-jandeksa-v-2022-godu/
49. Yandex provides Self-Isolation Index. *Kommersant* https://www.kommersant.ru/doc/4308505 (2020).
50. Kobak, D. Excess mortality reveals Covid's true toll in Russia. *Significance* **18**(1), 16–19 (2021).
51. Scherbov, S., Gietel-Basten, S., Ediev, D., Shulgin, S. & Sanderson, W. COVID-19 and excess mortality in Russia: Regional estimates of life expectancy losses in 2020 and excess deaths in 2021. *PLoS ONE* **17**(11), e0275967. https://doi.org/10.1371/journal.pone.0275967 (2022).
52. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**(772), 700–721 (1927).
53. Puhach, O., Meyer, B. & Eckerle, I. SARS-CoV-2 viral load and shedding kinetics. *Nat. Rev. Microbiol.* **21**(3), 147–161 (2022).
54. Brännström, Å., Sjödin, H. & Rocklöv, J. A method for estimating the number of infections from the reported number of deaths. *Front. Public Health*. https://doi.org/10.3389/fpubh.2021.648545 (2022).
55. Wooldridge, J. M. *Introductory econometrics: A modern approach* (ed. 5th) (Cengage Learning, 2012).
56. Lutkepohl, H. *New introduction to multiple time series analysis* 1st edn. (Springer, Berlin, 2005).
57. Engle, R. F. & Granger, C. W. J. Co-integration and error correction: representation, estimation, and testing. *Econometrica* **55**(2), 251–276 (1987).
58. Hassler, U. & Wolters, J. Autoregressive distributed lag models and cointegration. *Allgemeines Statistisches Arch.* **90**, 59–74 (2006).
59. Pesaran, M. H., Shin, Y. & Smith, R. J. Bounds testing approaches to the analysis of level relationships. *J. Appl. Econometric.* **16**, 289–326 (2001).
60. Demirhan, H. dLagM: An R package for distributed lag models and ARDL bounds testing. *PLoS ONE* **15**(2), e0228812. https://doi.org/10.1371/journal.pone.0228812 (2020).
61. Pollock, A. M. & Lancaster, J. Asymptomatic transmission of covid-19. *BMJ* **371**, 4851. https://doi.org/10.1136/bmj.m4851 (2020).

Acknowledgements
The authors thank Prof. Alexey Gvishiani and Dr. Alena Rybkina for their contributions to the design of the study and for their advice and support in its realization.

Author contributions
P.Ż.: conceptualization, investigation, methodology, data analysis, software, interpretation, visualization, original draft preparation, and manuscript preparation. G.B.: data acquisition and curation, investigation, validation, and manuscript preparation. A.O.: funding acquisition, supervision and administration, conceptualization, investigation, and manuscript preparation. E.R.: funding acquisition, supervision and administration, conceptualization, investigation, data analysis, interpretation, original draft preparation, and manuscript preparation.

Funding
International Institute for Applied Systems Analysis, Ministry of Science and Higher Education of the Russian Federation, 075-00439-25-01

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-24240-z.

Correspondence and requests for materials should be addressed to P.Ż.