RESEARCH PAPER

SPATIAL MODELLING OF KEY REGIONAL-LEVEL FACTORS OF COVID-19 MORTALITY IN RUSSIA

Egor A. Kotov[1*], Ruslan V. Goncharov[1], Yuri V. Kulchitsky[1], Varvara A. Molodtsova[1], Boris V. Nikitin[2,3]
[1]Faculty of Urban and Regional Development, HSE University, Myasnitskaya str. 13-4, Moscow 101000, Russia
[2]Institute of Regional Consulting, office 903, Nakhimovsky prosp. 32, Moscow 117218, Russia
[3]Faculty of Geography, Moscow State University, Leninskie Gory 1, Moscow 119899, Russia
***Corresponding author:** kotov.egor@gmail.com

Received: June 30th, 2021 / Accepted: April 24th, 2022 / Published: June 30th, 2022
https://DOI-10.24057/2071-9388-2021-076

ABSTRACT. Intensive socio-economic interactions are a prerequisite for the innovative development of the economy, but at the same time, they may lead to increased epidemiological risks. Persistent migration patterns, the socio-demographic composition of the population, income level, and employment structure by type of economic activity determine the intensity of socio-economic interactions and, therefore, the spread of COVID-19.

We used the excess mortality (mortality from April 2020 to February 2021 compared to the five-year mean) as an indicator of deaths caused directly and indirectly by COVID-19. Similar to some other countries, due to irregularities and discrepancies in the reported infection rates, excess mortality is currently the only available and reliable indicator of the impact of the COVID-19 pandemic in Russia.

We used the regional level data and fit regression models to identify the socio-economic factors that determined the impact of the pandemic. We used ordinary least squares as a baseline model and a selection of spatial models to account for spatial autocorrelation of dependent and independent variables as well as the error terms.

Based on the comparison of AICc (corrected Akaike information criterion) and standard error values, it was found that SEM (spatial error model) is the best option with reliably significant coefficients. Our results show that the most critical factors that increase the excess mortality are the share of the elderly population and the employment structure represented by the share of employees in manufacturing (C economic activity according to European Skills, Competences, and Occupations (ESCO) v1 classification). High humidity as a proxy for temperature and a high number of retail locations per capita reduce the excess mortality. Except for the share of the elderly, most identified factors influence the opportunities and necessities of human interaction and the associated excess mortality.

KEYWORDS: COVID-19, spatial models, socio-economic factors, climatic factors, excess mortality, Russian regions

CITATION: Kotov E.A., Goncharov R.V., Kulchitsky Y.V., Molodsova V.A., Nikitin B.V. (2022). Spatial Modelling of Key Regional-Level Factors of Covid-19 Mortality In Russia. Geography, Environment, Sustainability, 2(15), p. 71-83
https://DOI-10.24057/2071-9388-2021-076

ACKNOWLEDGEMENTS: The reported study was funded by RFBR according to the research project № 20-04-60490 «Ensuring balanced regional development during a pandemic with spatially differentiated regulation of socio-economic interaction, sectoral composition of the economy and local labour markets».

Conflict of interests: The authors reported no potential conflict of interest.

INTRODUCTION

Intensive socio-economic interactions are a prerequisite for the innovative development of the economy, but at the same time, they may lead to increased epidemiological risks. Persistent migration patterns, socio-demographic composition of the population, income level, and employment structure by type of economic activity determine the intensity of socio-economic interactions and, therefore, the spread of COVID-19.

Most research on COVID-19 focuses on factors affecting COVID-19 infection rates and the resulting mortality. Many papers employ spatial regression models to achieve a better model fit and more trustworthy estimates of the effects. With this paper we aim to add to the existing body of research by revealing various factors for the case of Russian regions with a specific focus on physical human interaction using models that could be compared between countries. Below we provide an in-depth review of previous research along with the variable selection process.

MATERIALS AND METHODS

Data

The full data set and analysis code for this paper is available on GitHub, so the findings are fully reproducible and auditable: https://github.com/e-kotov/ru-covid19-regional-excess-mortality (doi: 10.5281/zenodo.6515455).

The dependent variable

A meta-analysis of 63 research papers (Franch-Pardo et al. 2020) showed that the most frequently used indicators for COVID-19 analysis are COVID-19 infection and mortality rates.

However, the use of these parameters relies heavily on the quality of data collection and reporting. When there is little trust in the collected data, it cannot be used, which is why the data on infection rates should be avoided even when it is available. Therefore, in this study, we used excess mortality as our target variable. The downside of using excess mortality is that this data becomes available much later than COVID-19 infection rates and reported deaths. However, recently published infection and death rates seem to correlate well with the excess mortality, so analysis for more recent periods can be performed on the data similar to what most researchers use.

Another reason to use excess mortality is that apart from deaths caused directly by the COVID-19 infection, it also takes into account deaths caused by the interruption of the regular healthcare provision. Excess mortality is also helpful for comparing data between different counties as it compensates for the possible differences in the mortality statistics collection (Rodríguez-Pose and Burlina 2021; Yarmol-Matusiak et al. 2021).

Our excess mortality variable is the ratio of per capita mortality for April 2020 - February 2021 to the mean over the previous five years.

The independent variables

We used the existing research to guide the selection of variables. A review of recent studies allowed us to divide the variables into several groups. Additionally, we also used our own Human Interaction group, which was a primary focus of this research. The final list of examined variables is presented in Appendix A.

Human Interaction

Under this group, we summarised multiple variables that fall under different groups in other studies but indicate how much physical human contact is required (or is possible, if there is a choice) for day-to-day activities, even during lockdowns.

Although the type of economic activity suggests a certain income level, we think it is an excellent indicator of how much physical human contact with clients or co-workers a particular job requires. For example, the share of the population working in retail influences the number of physical contacts. A higher number of retail outlets per capita may indicate a larger number of people working in retail, leading to more opportunities for violating the lockdown or distancing measures for both workers and consumers.

However, a higher area of retail per capita may allow better social distancing. Similarly, the mobility-related variables such as airport density, road and rail-road density, and the number of buses per capita may be regarded as indicators of how many people may be in direct contact and at what distance.

Some researchers (Andersen et al. 2021; Chakraborti et al. 2021; Desmet and Wacziarg 2021; Hass and Jokar Arsanjani 2021; Henning et al. 2021; Mollalo et al. 2020; Rahman et al. 2020; Scarpone et al. 2020) include very similar variables (retail outlets provision, big retail provision and road densities) as so-called environment factors.

Demographic

This group includes age structure with a specific focus on the share of population past working age, urbanisation and ethnic mix (Agnoletti et al. 2020; Amdaoud et al. 2021; Andersen et al. 2021; Ascani et al. 2021; Bański et al. 2021; Chakraborti et al. 2021; Desmet and Wacziarg 2021; Ehlert 2021; Hass and Jokar Arsanjani 2021; Henning et al. 2021; Konstantinoudis et al. 2021; Luo et al. 2021; Maiti et al. 2021; Mogi et al. 2020; Mollalo et al. 2020; Oto-Peralías 2020; Perone 2021; Rahman et al. 2020; Raymundo et al. 2021; Rodríguez-Pose and Burlina 2021; Sannigrahi et al. 2020; Scarpone et al. 2020; Sun et al. 2020; Zemtsov and Baburin 2020). We were primarily concerned with the age structure due to the higher COVID-19 fatality risks for the older population, using the share of post-, under- and working-age population in the analysis.

This group also includes migration flows at intra- and inter-regional levels, as well as international level (Chakraborti et al. 2021; Chen et al. 2021; Maiti et al. 2021; Wang et al. 2021). Even though international travel was heavily restricted at the beginning of the pandemic, it was not restricted early enough. Therefore, past international migration flows might be indicative of the international travel at the beginning of 2020, which influenced the spread of the virus and the excess mortality early on. The inter- and intra-regional travel within Russia were not as restricted and were even encouraged at some point to stimulate internal tourism.

Socio-economic

These indicators include unemployment rate, poverty rate, real income, salary, and employment across different economic activities (Agnoletti et al. 2020; Amdaoud et al. 2021; Andersen et al. 2021; Ascani et al. 2021; Bański et al. 2021; Chakraborti et al. 2021; Desmet and Wacziarg 2021; Ehlert 2021; Konstantinoudis et al. 2021; Luo et al. 2021; Maiti et al. 2021; Mogi et al. 2020; Mollalo et al. 2020; Oto-Peralías 2020 p.; Rahman et al. 2020; Raymundo et al. 2021; Rodríguez-Pose and Burlina 2021; Sannigrahi et al. 2020; Scarpone et al. 2020; Sun et al. 2020; Zemtsov and Baburin 2020). We included income-related variables in the analysis (see Appendix A) as we expected them to reveal the regions where the population cannot afford to obey the lockdowns and cease work or cannot afford extra medical care due to low income. However, employment by economic activities was regarded as part of a different group of variables - human interaction.

Mobility

These indicators include mobility patterns, passenger flows on public transport, mean travel time and more (Andersen et al. 2021; Ascani et al. 2021; Luo et al. 2021; Maiti et al. 2021; Rodríguez-Pose and Burlina 2021; Zemtsov and Baburin 2020). We included some mobility-related variables in the human interaction group above.

Healthcare provision and population health

In this category, other researchers note healthcare expenses per capita, the number of ventilators per capita, medical personnel per capita (Amdaoud et al. 2021; Bański et al. 2021; Konstantinoudis et al. 2021; Luo et al. 2021; Maiti et al. 2021; Mollalo et al. 2020; Perone 2021; Rahman et al. 2020; Raymundo et al. 2021; Rodríguez-Pose and Burlina 2021; Sannigrahi et al. 2020; Scarpone et al. 2020;

Sun et al. 2020; Zemtsov and Baburin 2020). Others also include indicators of public health, such as the number of smokers, the number of people with diabetes, the share of the overweight population (Andersen et al. 2021; Desmet and Wacziarg 2021; Ehlert 2021; Konstantinoudis et al. 2021; Luo et al. 2021; Mogi et al. 2020; Mollalo et al. 2020; Zemtsov and Baburin 2020). Even though a meta-analysis by Kolosov et. al (2021) suggests that there is no significant influence of the level of healthcare provision on mortality, we still tested this hypothesis for Russia on a regional level. Since healthcare indicators are usually highly correlated, we used the number of doctors per capita as an indicator of the current level of healthcare and its variation over five years as an indicator of how the healthcare provision had changed recently.

Climate and Environment

Many researchers also considered climate factors. Most of them used mean temperature and humidity, as well as precipitation and UV exposure (Hass and Jokar Arsanjani 2021; Konstantinoudis et al. 2021; Luo et al. 2021; Maiti et al. 2021; Oto-Peralías 2020; Perone 2021; Qi et al. 2020; Rahman et al. 2020; Wang et al. 2021). Some papers also used data on droughts and floods (Luo et al. 2021) as well as air quality via CO2 levels and other emissions (Agnoletti et al. 2020; Chakraborti et al. 2021; Hass and Jokar Arsanjani 2021; Luo et al. 2021; Maiti et al. 2021; Oto-Peralías 2020; Perone 2021; Rodríguez-Pose and Burlina 2021; Wang et al. 2021). We argue that temperature and humidity, apart from possibly affecting the survival of the virus, may also influence the willingness and opportunities of the population for outdoor vs indoor social gatherings.

Indices

Various indices may be regarded as a separate group, as they usually combine multiple indicators. A self-isolation index published by Yandex was used by Russian researchers (Zemtsov and Baburin 2020). Some indices are more focused on a particular topic, such as the healthcare quality index (Perone 2021), social trust index (Amdaoud et al. 2021), and economic diversity index (Ascani et al. 2021). Some indices are more comprehensive, for example, Community Need Index which covers income, culture, education, living conditions and healthcare (Henning et al. 2021) and the infection risk index (only available as a pre-print at the moment[1]). Due to the underlying data and methodology, it is often hard to calculate similar indices for different countries. We considered using the Herfindahl-Hirschman Index (HHI) for employment structure, which was applied in the study of Ascani et al. (2021). However, it was found that the shares of employment across different economic activities are a much better predictor of the excess mortality.

As we demonstrate in Fig. 1 below, most variables are subject to spatial autocorrelation. Therefore, we used spatial regression models to achieve the best results.

Clusters of the Excess mortality variable in Fig. 1 seem to provide limited support for the hypothesis that a pandemic should follow the pattern of the spatial diffusion of innovations (Hägerstrand 1973). During the first year, we can see that high excess mortality clustered in the regions of the Central Federal District while low mortality was observed in relatively remote regions that do not have intensive communication with the Central Federal District. However, we do not see high-value clustering in

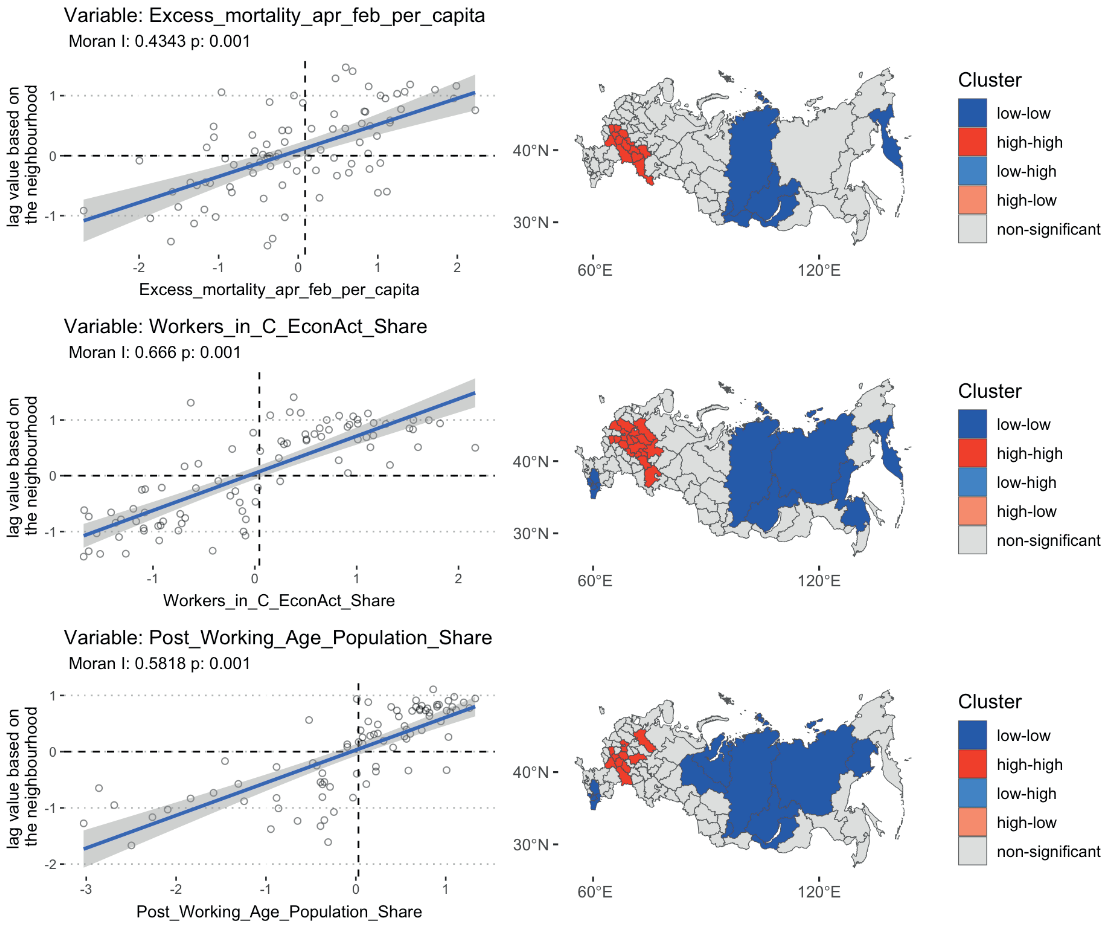

Fig. 1. Spatial autocorrelation tests for excess mortality and some explanatory variables

[1] Baum C.F. and Henry M. (2020). Socioeconomic Factors influencing the Spatial Spread of COVID-19 in the United States [SSRN Scholarly Paper]. DOI: 10.2139/ssrn.3614877

the Far Eastern Federal District and around The Republic of Tatarstan. The absence of such clustering may be due to the large size of the regions. We would expect the clustering to be present at the municipal level, confirming that the virus spread from the most populated cities to the least populated ones, which is in line with the spatial diffusion of innovations theory.

Clustering of other variables is given in Fig. 1 for illustrative purposes only. Clustering is different for different variables which justifies the incorporation of spatial effects into the regression analysis to compensate for these variations. Since the spatial clustering of excess mortality has diverging patterns, we can assume that it cannot be explained only by spatial autocorrelation of explanatory variables. Therefore, a model that compensates for the unobserved spatially correlated model errors should be used.

Method

We performed exploratory data analysis for all variables listed in Appendix A. Some variables were log-transformed for a better fit in linear models.

The selection of variables for the models was performed in the following way. For all the dependent variables we calculated Pearson correlation coefficients. We also fitted an ordinary least squares model for every independent variable against the excess mortality and calculated the R^2 and the p-value of the model (see the LM R2 and LM p-value columns in Table 1). After that, independent variables were ranked by descending R^2 and correlation (see Table 1). Using the list of top-ranked independent variables we eliminated the ones with the highest correlation (with a correlation coefficient of more than 0.7) to avoid potential multicollinearity in the models.

Then we constructed a series of baseline ordinary least squares (OLS) regression models following the basic equation:

$$y_i = \beta_0 + X_i\beta + \varepsilon_i \tag{1}$$

where y_i is excess mortality in the region *i*, β_0 is the intercept, X_i is a vector of selected explanatory variables, β is a vector of regression coefficients, and ε_i is a random error term.

We tried various combinations of factors in OLS regressions based on exploratory data analysis and corresponding model interpretation. After obtaining the best OLS model (1) we tested the independent variable, explanatory variables, and the OLS model residuals for spatial autocorrelation. The matrix of spatial neighbours for the spatial autocorrelation test and the resulting spatial models were created based on region boundary polygons from OpenStreetMap (OpenStreetMap contributors 2017) with GeoDa software[2] (Anselin et. al 2006) using first-order queen contiguity. Regions without neighbours (such as Kaliningrad Region and Sakhalin Region) were manually connected to 2-3 closest regions[3].

Based on the results of spatial autocorrelation tests we applied a selection of spatial models (LeSage and Pace 2009).

Spatially Lagged-X Model (SLX) was used to compensate for spatial autocorrelation of the explanatory variables:

$$y_i = \beta_0 + X_i\beta + W_iX_i\theta + \varepsilon_i \tag{2}$$

where, in addition to the OLS (1), W_i is a vector of spatial weights (a corresponding row of the spatial weights matrix), θ is the *k×1* coefficient vector for the exogenous spatially lagged independent variables.

Spatial Lag Model (SLM, also referred to as SAR – spatial autoregressive model) was used to compensate for spatial autocorrelation of the dependent variable:

$$y_i = \beta_0 + X_i\beta + \rho W_i y_i + \varepsilon_i \tag{3}$$

where *p* is the spatial lag parameter.

Spatial Error Model (SEM) was used to compensate for spatial autocorrelation of the error terms:

Table 1. Explanatory variables ranked by the highest correlation with excess mortality

Variable Name	Correlation	LM R2	LM p-value
Workers_in_C_EconAct_Share	0.5820	0.3387	0.0000000
Workers_in_G_EconAct_Share	0.5387	0.2902	0.0000001
Population_log	0.5344	0.2856	0.0000001
Floor_Area_per_capita	0.5007	0.2507	0.0000011
Workers_in_O_EconAct_Share	-0.4971	0.2471	0.0000013
Road_Density_log	0.4732	0.2239	0.0000048
Workers_in_P_EconAct_Share	-0.4722	0.2230	0.0000051
Population_Density_log	0.4573	0.2091	0.0000108
Workers_in_B_EconAct_Share_log	-0.4223	0.1783	0.0000569
Population_Below_Living_Wage_Share	-0.4081	0.1666	0.0001056
SME_in_GRDP_Share	0.3890	0.1513	0.0002335
Workers_in_R_EconAct_Share	-0.3839	0.1473	0.0002873
GRDP_in_GDP_Share_log	0.3780	0.1429	0.0003618
Migr_Outflow_InterReg_3Y_mean_per_capita_x10000	-0.3592	0.1290	0.0007364
Buses_per_capita_log	-0.3502	0.1226	0.0010187

[2] https://geodacenter.github.io

[3] The specific neighbours for those regions can be viewed by downloading the provided data and code.

$$y_i = \beta_0 + X_i\beta + u_i,\ u_i = \lambda W_i u_i + \varepsilon_i \tag{4}$$

where λ is a spatial lag parameter for the spatially correlated errors, u_i is a spatial component of the error, and ε_i is a spatially uncorrelated error.

Spatial Durbin Model (SDM) was used to compensate for spatial autocorrelation of the explanatory and the dependent variables:

$$y_i = \beta_0 + \rho W_i y_i + X_i\beta + W_i X_i\theta + \varepsilon_i \tag{5}$$

Spatial Durbin Error Model (SDEM) was used to compensate for spatial autocorrelation of the explanatory variables and the error terms:

$$y_i = \beta_0 + X_i\beta + W_i X_i\theta + u_i,\ u_i = \lambda W_i u_i + \varepsilon_i \tag{6}$$

SARAR (spatial autoregressive model with spatially autocorrelated disturbances, also referred to as SAC – spatial autoregressive combined model) was used to compensate for spatial autocorrelation of the independent variable and the error terms (Kelejian and Prucha 1998):

$$y_i = \beta_0 + X_i\beta + \rho W_i y_i + u_i,\ u_i = \lambda W_i u_i + \varepsilon_i \tag{7}$$

General Nesting Spatial Model (GNS, also referred to as mixed spatial autoregressive combined model) was applied as the final model that combines all the models above and tries to compensate for spatial autocorrelation of all components:

$$y_i = \beta_0 + X_i\beta + \rho W_i y_i + W_i X_i\theta + u_i,\ u_i = \lambda W_i u_i + \varepsilon_i \tag{8}$$

The code for all plots and tables was written in R language. The data, code and weights matrix are available in the supplementary materials on GitHub (Kotov 2022).

RESULTS AND DISCUSSION

Baseline OLS selection

We analysed a series of OLS models (see Fig. 2 and Fig. 3 below) by looking at the coefficients of the variables and their 95% confidence intervals. All coefficients are robust and z-standardised, which makes their effect on excess mortality comparable regardless of the unit size of any individual variable. If the confidence interval of a coefficient is entirely located to the right or to the left of the center at the 0 mark in Fig. 2 and Fig. 3, it means that there is a statistically significant negative or positive effect on the excess mortality.

We started with the model M0, which takes into account population density (using population density and residential floor per capita variables), employment in economic sectors that require close human interaction (B – mining, C – manufacturing, G – retail and services, P – education, and in small and medium enterprises in general), and local transportation opportunities and constraints (number of buses and cars per capita). M0 clearly showed that population density and transportation constraints have no effect on mortality. The only two significant variables are the share of employees in manufacturing (C) and retail & services (G), as these are the only variables with confidence intervals that do not cross the zero-line. The proximity of the confidence interval to zero may be due to the inclusion of insignificant variables in the model, therefore we removed some of these variables starting with the model M4 below.

M1 is an extension of M0 with climate variables (temperature and humidity). It was found that temperature does not affect mortality, while the effect of humidity is uncertain and should be tried in further models.

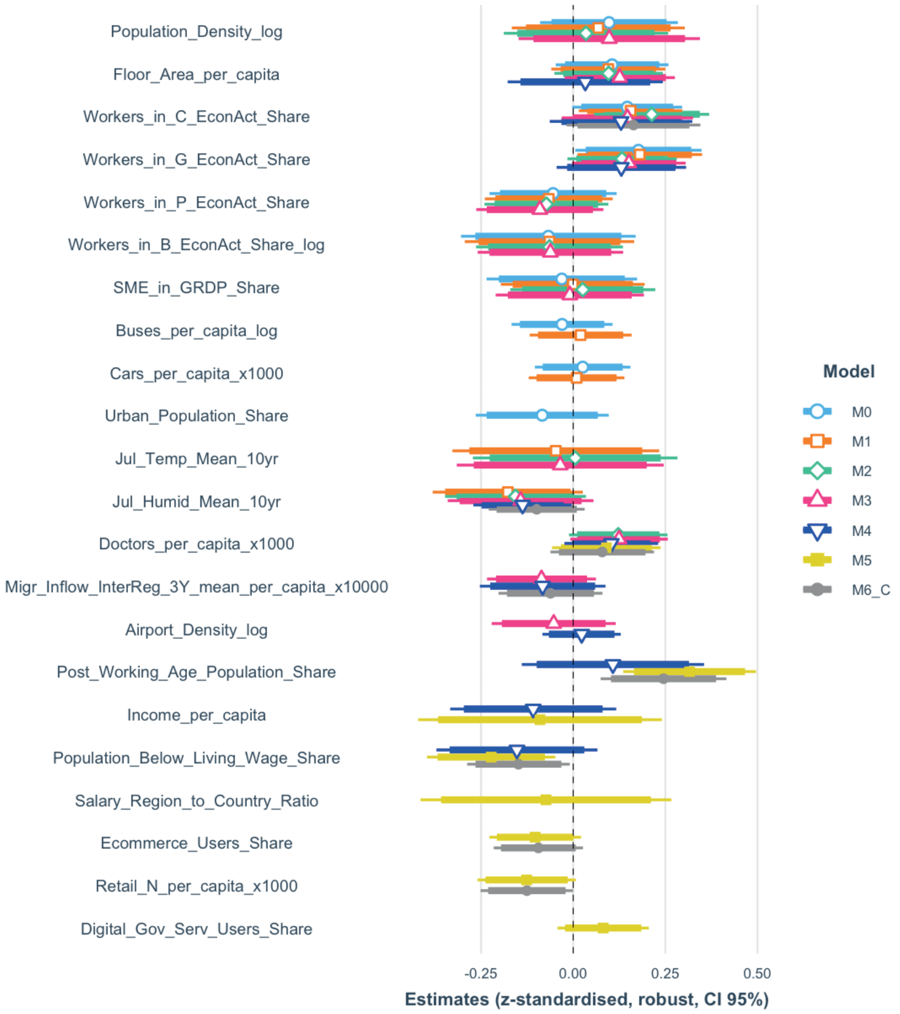

Fig. 2. Comparison of the coefficients of the models M0-M6_C

M2 adds the healthcare component (number of doctors per capita). It was found that, depending on the model, the confidence interval may touch the zero mark, but this factor is still worth considering in further models.

M3 adds interregional migration flow and opportunity for the spread of COVID-19 following the Hägerstrand's (1973) spatial diffusion of innovation (via the airport density). It was found that airports have no detectable effect, which suggests that air travel between regions was likely not a significant factor in the COVID-19 spread across Russia, while inter-regional migration is worth considering.

With the next model M4, we eliminated the non-significant variables from previous models and added age (as the elderly are the most affected by both the virus and the deterioration of regular medical care) and income (following the hypothesis that in poorer regions the population will ignore the restrictions more frequently as they must provide money for their families). It was found that on its own M4 has almost no significant coefficients, however, it provides information on the potential of individual variables. Population density expressed as residential floor area per capita proved to be insignificant, as its coefficient in M4 falls almost to zero. The coefficients for the share of workers in manufacturing (C) and retail & services (G), the number of doctors per capita, and humidity in M4 and previous models vary slightly but mostly remain significant, suggesting that these two economic domains with intensive and close human interaction are important negative factors of the excess mortality.

M5 builds on M4 by adding digital skills, the share of e-commerce users and the overall provision of retail businesses. In M5 we can see that income expressed as the share of the population with income below the living wage has a high negative impact on excess mortality. This is counterintuitive but may suggest that those households interacted less, as they had no money to spend. The higher share of e-commerce users, as well as the higher number of retail locations per capita, also had a negative effect, as the reliance on face-to-face contact was lower in regions with high values for these variables. Interestingly, M5 also suggests that a higher share of the population using government services over the Internet somehow negatively affects mortality.

Finally, models M6_C, M6_G and M6_CG are the ultimate models with the most significant variables that demonstrate a noticeable and explainable effect. The difference is that M6_C uses the share of employees in manufacturing (C), while M6_G replaces it with the share in retail and services (G). M6_CG uses the shares in both C and G economic activities. We can see the comparison of these M6 models in Fig. 3. Clearly, the share of employees in C and G is almost equally important, both according to the models and the logic behind the variables, however with the M6_C model we are able to capture the effect of retail with the number of retail locations per capita and e-commerce. M6_G and M6_CG, despite their overall similarity to M6_C, do not reproduce the same effects reliably. M6 models also suggest that the number of doctors is irrelevant, which makes sense compared to the previous models as this variable had a positive effect on excess mortality, which could only be explained by assuming that contacts through doctors were stimulating additional infections. The insignificance of the healthcare provision is also in line with previous findings (Kolosov et al. 2021).

A statistical summary of all OLS models is provided in Fig. 4 below. It shows that models M6_CG and M6_C are the best according to most model quality metrics. They have the lowest AICc (corrected Akaike information criterion), highest R-squared and adjusted R-squared, and lowest RMSE (root-mean-square error). Therefore, we used these models and their variables as the best baseline for the spatial extension of the model.

Extension of the best OLS with a spatial component

We used M6_CG as the baseline OLS model and extended it with spatial specifications as described in the methodology in equations (2) through (8). As we can see from Fig. 5, the best models are SEM (Spatial Error Model), LAG (Spatial Lag Model) and OLS. These models have the lowest corrected Akaike Information Criterion (AICc), but the values are very close and not significantly different. However, SEM helps to compensate for the spatial autocorrelation of some unobserved and unaccounted spatially autocorrelated factors. The LAG model corrects for the spatial autocorrelation of the excess mortality (as seen at the top of Fig. 1) and the fact that the spread of COVID-19 is indeed quite likely to occur between the neighbouring regions, which is also observed on the global scale. Other models (all models below OLS in Fig. 5) do produce lower model errors, however, they add very little in terms of interpretability of the results in general and the model coefficients.

Fig. 6 provides a comparison of the OLS model coefficients with and without its spatial extensions. The graph suggests that compensating for spatial autocorrelation increases confidence in the significance of several variables, including the number of retail locations per capita, the share of post-working age population and

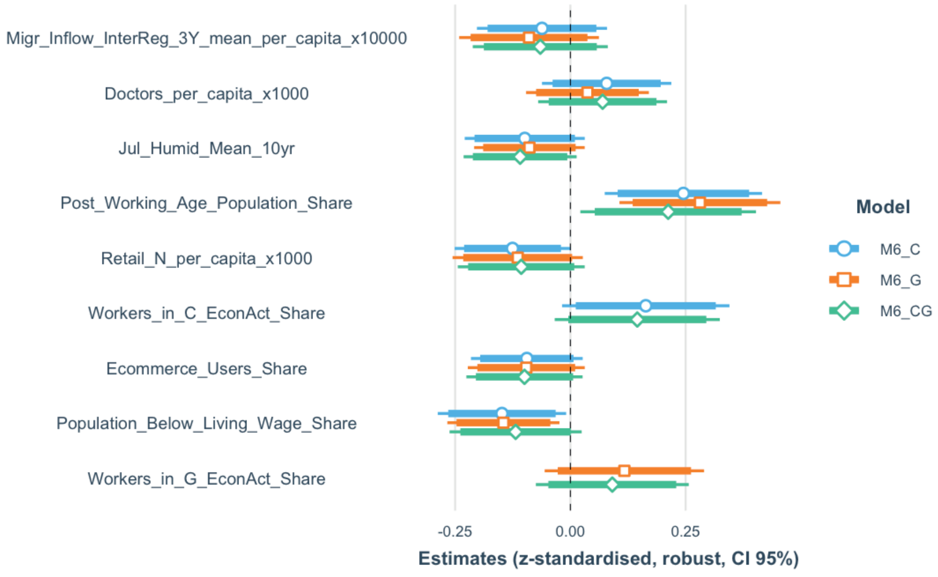

Fig. 3. Comparison of the coefficients of the model M6 variants

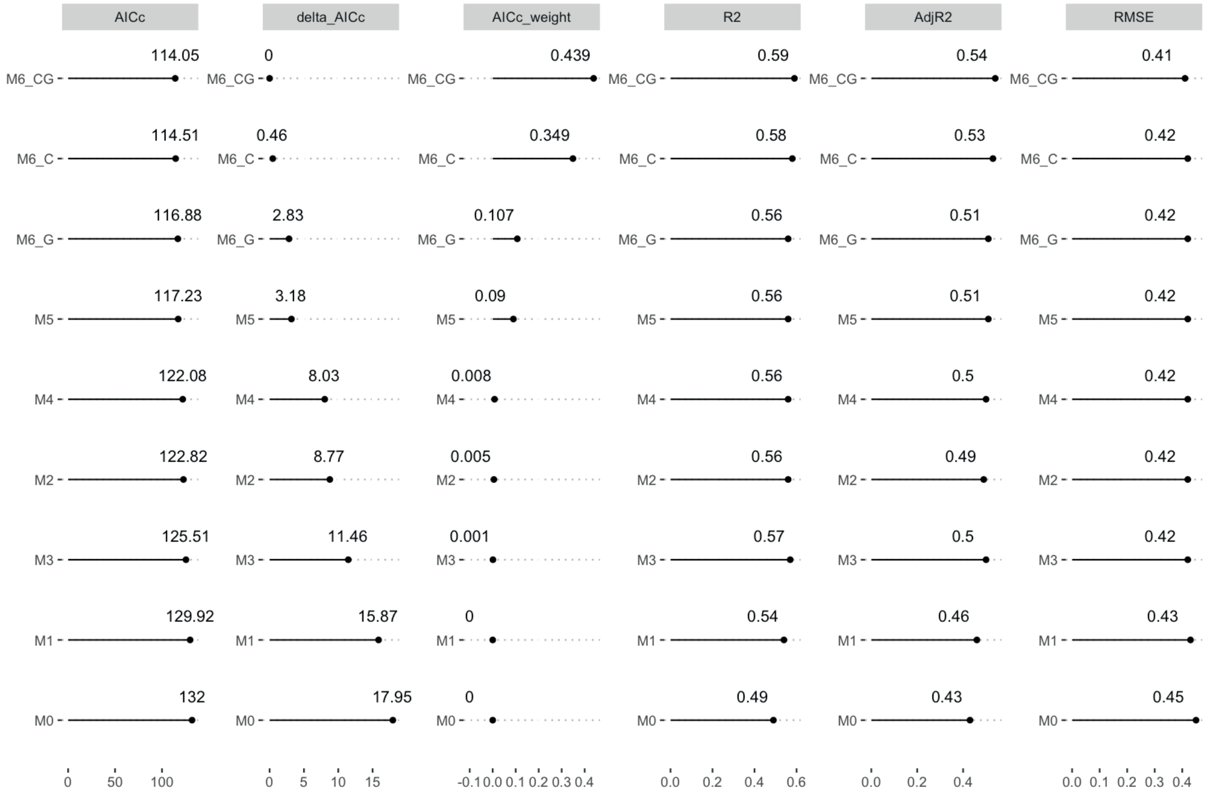

Fig. 4. Summary of baseline OLS models

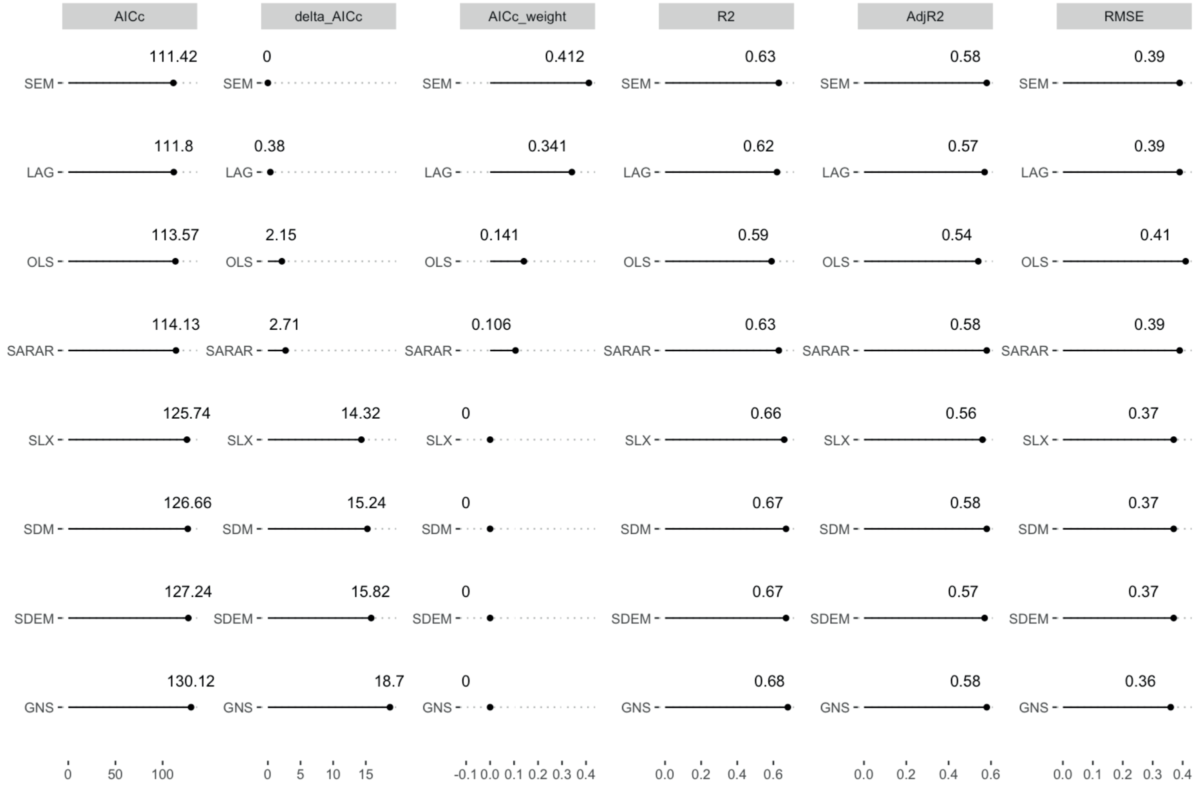

Fig. 5. Comparison of the OLS model performance with and without spatial extensions for the final set of variables

the share of employees in manufacturing (C), but not the share of workers in retail (G) and the number of doctors per capita. The effect of the share of e-commerce users on excess mortality is still insignificant, which might be due to the use of old data as Rosstat has not yet published the 2020 data, and pre-2020 the share of consumers who shop online was lower.

We can also observe that most variables do not show any external effects on neighbouring regions («lag.» variables at the bottom of Fig. 6). That is, the values of these variables in the neighbours of any given region do not affect the excess mortality in the region of interest. The only exception is the number of doctors per capita, which increases the excess mortality in neighbouring regions without any logical explanation.

Lambda and rho are significant in the corresponding SEM and LAG models, but not in their derivatives. SEM and LAG models are almost equivalent in all other aspects (model quality based on AICc, R-squared and RMSE, as well as model coefficients). This reinforces the statement

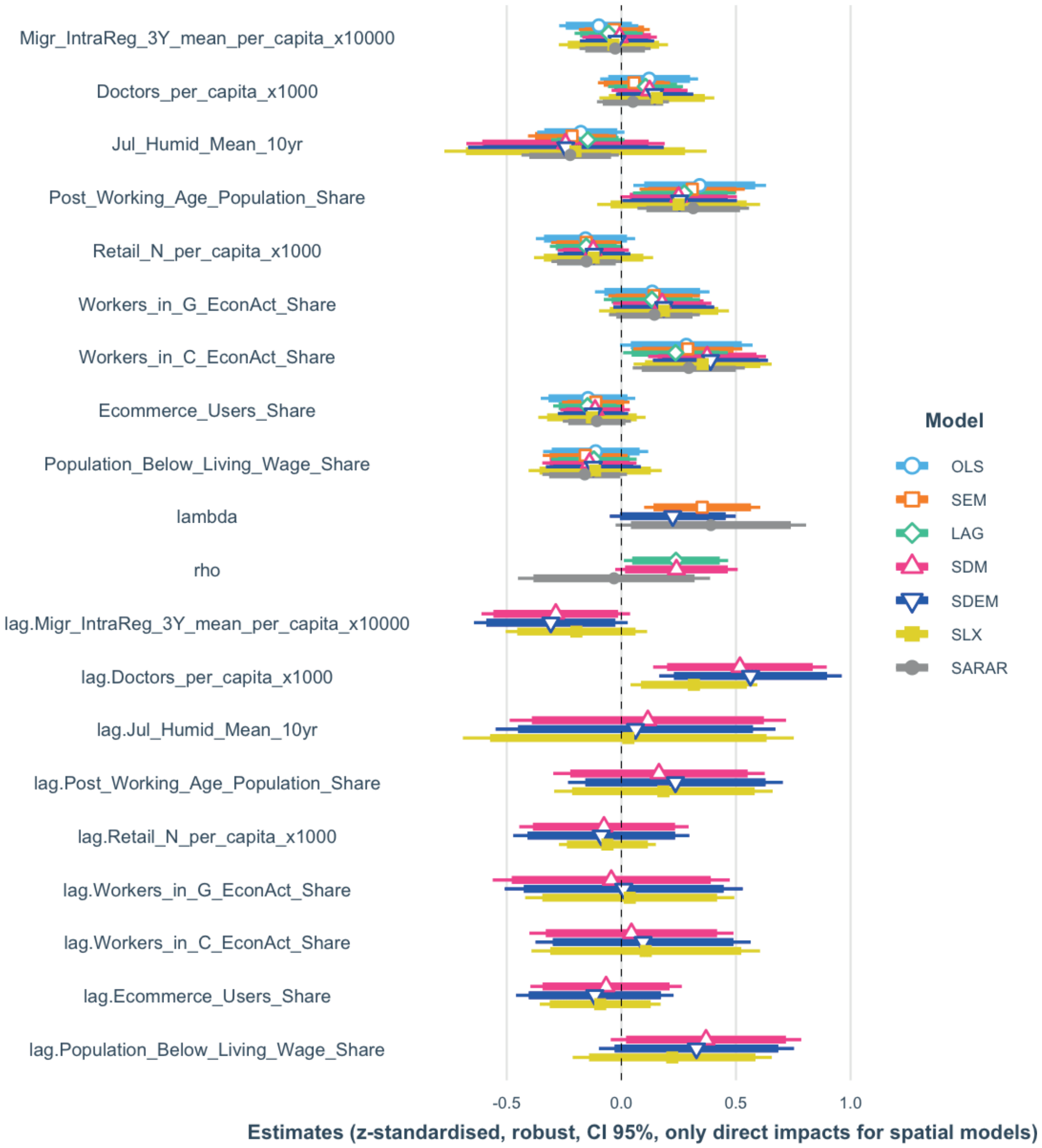

Fig. 6. Comparison of the OLS model coefficients with and without spatial extensions

that there are some unobserved spatially autocorrelated components missing from the model. However, their absence is partially compensated by the spatial extension of the OLS. Given that in the SARAR model both lambda and rho are insignificant and demonstrate an opposite effect, we can conclude that only LAG (compensates for the autocorrelation of excess mortality) or SEM (compensates for unobserved spatially autocorrelated variables) models can be considered as the best fit.

In Table 2 we directly compare the z-standardised coefficients and the number of significant variables in all models. It can be seen that the LAG extension of the OLS model results in more robust estimates for a larger number of variables.

As we can see from Table 2, the most important factor for excess mortality according to the best SEM model is the share of the post-working age population. It has the highest value of the standardised coefficient (0.31), which confirms the well-known fact that the early COVID-19 wave largely affected the elderly. Almost equally important (0.29) is the share of workers in manufacturing, where very close contact is common and social distancing is sometimes impossible. What is more, in manufacturing jobs workers often reside together in communal accommodation.

High humidity has a negative (-0.22) effect on excess mortality. In fact, humidity is highly correlated with temperature, and even though the temperature did not make it into the final model, we can assume humidity as a proxy for temperature. Climate conditions might explain not only the specifics of the COVID-19 virus related to humidity and temperature but also differences in the behaviour of the population, for example, the propensity to spend more or less time outdoors.

We manually marked the number of retail locations per capita as significant as it only formally misses the 5% significance level with a p-value of 0.0506. The negative coefficient (-0.15) confirms that a larger number of shops per capita in a given region leads to a lower density of customers in those shops, and therefore increases the opportunities for social distancing and reduces interaction. The share of employees in retail (G) is not significant, however, we expect it to be a proxy for a similar effect.

The last parameter that is significant and highly important in terms of its effect is the lambda of the SEM model. This suggests that there are one or more unobserved spatially autocorrelated factors, that neither we nor other researchers have considered. The effect of other variables is not significant. From the literature review, we have not seen high R^2 values and therefore full models even in studies concerning much smaller spatial units than Russian regions. We expect that a more complete model would be able to capture more individual effects at the municipal level.

Table 2. Comparison of the OLS, SEM and LAG estimates

	OLS	SEM	LAG
(Intercept)	0.00	-0.02	-0.02
	(0.08)	(0.10)	(0.07)
Migr_IntraReg_3Y_mean_per_capita_x10000	-0.10	-0.03	-0.06
	(0.09)	(0.08)	(0.08)
Doctors_per_capita_x1000	0.12	0.06	0.11
	(0.11)	(0.08)	(0.08)
Jul_Humid_Mean_10yr	-0.18	-0.22 *	-0.15
	(0.10)	(0.10)	(0.08)
Post_Working_Age_Population_Share	0.34 *	0.31 **	0.27 *
	(0.15)	(0.12)	(0.11)
Retail_N_per_capita_x1000	-0.16	-0.15*	-0.15
	(0.11)	(0.08)	(0.08)
Workers_in_G_EconAct_Share	0.13	0.14	0.13
	(0.13)	(0.10)	(0.11)
Workers_in_C_EconAct_Share	0.28	0.29 *	0.24 *
	(0.15)	(0.12)	(0.12)
Ecommerce_Users_Share	-0.15	-0.11	-0.15
	(0.10)	(0.08)	(0.08)
Population_Below_Living_Wage_Share	-0.11	-0.16	-0.12
	(0.11)	(0.09)	(0.10)
Lambda *λ*		0.35 **	
		(0.13)	
Rho p			0.24 *
			(0.12)
N	85	85	85
Pseudo-R2	0.59	0.63	0.62
Adjusted pseudo-R2	0.54	0.58	0.57
AICc	113.57	111.42	111.80

All continuous predictors are mean-centered and scaled by 1 standard deviation. Standard errors are heteroskedasticity robust.
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

As we can see from Fig. 7 below, the shares of employment in many economic activities are highly correlated with each other, as well as with the excess mortality. So hypothetically a model for excess mortality could be composed completely based on employment rates. We explored this option and can conclude that using just one of the economic activity types, either G (trade) or C (manufacturing), is the best option. Fitting the same set of spatial models using A, D, G and R economic activities results in similar R^2 and model error and even lower AICc values, however, it unnecessarily limits the model to just economic activities, which cannot be the only explanation for the excess mortality.

The employment in C (manufacturing) does not necessarily capture the whole employment and interaction structure, but it is just enough to explain the excess mortality without relying on other economic activities. Employment in G (retail) is just as important but was pushed out from the final model by the employment in C (mining) variable. The large workforce in retail leads to more opportunities and more necessity for close physical interaction between co-workers and with the customers. As a result, even though it was not found to be significant in the model, the overall correlation of the share in G with the excess mortality suggests that it does have an effect, which is not captured at the regional level.

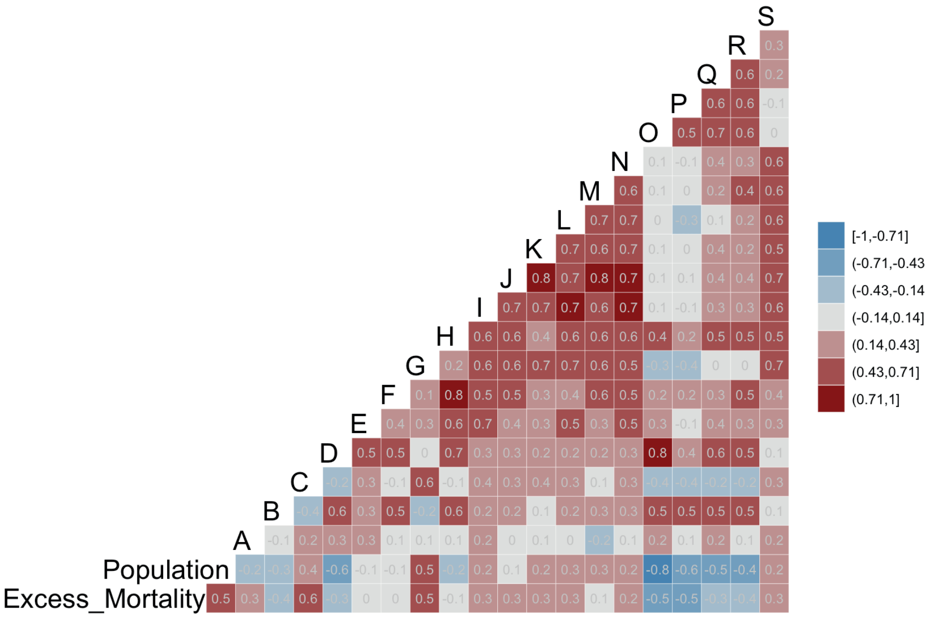

Fig. 7. Correlation matrix for the shares of employment in different economic activities

CONCLUSIONS

We identified the most important factors that caused the excess mortality between April 2020 and February 2021. The share of the elderly population was confirmed by our model as the obvious reason for excess deaths, followed closely by the share of employees in manufacturing (C economic activity according to European Skills, Competences, and Occupations (ESCO) v1 classificaiton). On the other hand, higher humidity, and a higher number of retail locations per capita reduce the excess mortality with a comparable impact.

Our final model is not complete and mostly focuses on a few factors, however, it is reliable in terms of the selection of these factors, which were identified based on the significance of their effect, as well as accounting for spatial autocorrelation. Still, there is room for improvement of the model.

Queen-type contiguity neighbourhood matrix is too simplified, so the spatial extensions of the baseline OLS model can potentially be improved by using a different type of spatial weights. For example, spatial weights based on the air-and rail-passenger flows for the year 2020 could be a better fit, as they would probably explain the pandemic transmission paths and intensity following the Hägerstrand's (1973) model of the spatial diffusion of innovations. Currently, we focused heavily on human interaction and possibly failed to take into account other factors, while compensating for spatial autocorrelation was not enough.

Due to data limitations, the current best model was built for the regional level. Because of the modifiable area unit problem, which is manifested in the excessive averaging of mortality and other variables over irregularly sized and populated regions, it might not be possible to improve the obtained results. We expect the same model to provide better results at a municipal level when the mortality data becomes available. ■

REFERENCES

Agnoletti M., Manganelli S. and Piras F. (2020). Covid-19 and rural landscape: The case of Italy. Landscape and Urban Planning, 204, 103955, DOI: 10.1016/j.landurbplan.2020.103955.

Amdaoud M., Arcuri G. and Levratto N. (2021). Are regions equal in adversity? A spatial analysis of spread and dynamics of COVID-19 in Europe. The European Journal of Health Economics, 22(4), 629-642, DOI: 10.1007/s10198-021-01280-6.

Andersen L.M., Harden S.R., Sugg M.M., Runkle J.D. and Lundquist T.E. (2021). Analyzing the spatial determinants of local Covid-19 transmission in the United States. Science of The Total Environment, 754, 142396, DOI: 10.1016/j.scitotenv.2020.142396.

Anselin L., Ibnu S. and Youngihn K. (2006). GeoDa: An Introduction to Spatial Data Analysis. Geographical Analysis 38(1), 5-22, DOI: 10.1111/j.0016-7363.2005.00671.x.

Ascani A., Faggian A. and Montresor S. (2021). The geography of COVID-19 and the structure of local economies: The case of Italy. Journal of Regional Science, 61(2), 407-441, DOI: 10.1111/jors.12510.

Bański J., Mazur M. and Kamińska W. (2021). Socioeconomic Conditioning of the Development of the COVID-19 Pandemic and Its Global Spatial Differentiation. International Journal of Environmental Research and Public Health, 18(9), 4802, DOI: 10.3390/ijerph18094802.

Chakraborti S., Maiti A., Pramanik S., Sannigrahi S., Pilla F., Banerjee A. and Das D.N. (2021). Evaluating the plausible application of advanced machine learnings in exploring determinant factors of present pandemic: A case for continent specific COVID-19 analysis. Science of The Total Environment, 765, 142723, DOI: 10.1016/j.scitotenv.2020.142723.

Chen Y., Chen M., Huang B., Wu C. and Shi W. (2021). Modeling the Spatiotemporal Association Between COVID-19 Transmission and Population Mobility Using Geographically and Temporally Weighted Regression. GeoHealth, 5, e2021GH000402, DOI: 10.1029/2021gh000402.

Desmet K. and Wacziarg R. (2021). JUE Insight: Understanding spatial variation in COVID-19 across the United States. Journal of Urban Economics, 103332, DOI: 10.1016/j.jue.2021.103332.

Ehlert A. (2021). The socio-economic determinants of COVID-19: A spatial analysis of German county level data. Socio-Economic Planning Sciences, 101083, DOI: 10.1016/j.seps.2021.101083.

Franch-Pardo I., Napoletano B.M., Rosete-Verges F. and Billa L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. Science of The Total Environment, 739, 140033, DOI: 10.1016/j.scitotenv.2020.140033.

Hägerstrand T. (1973). Innovation diffusion as a spatial process. University of Chicago press.

Hass F.S. and Jokar Arsanjani J. (2021). The Geography of the Covid-19 Pandemic: A Data-Driven Approach to Exploring Geographical Driving Forces. International Journal of Environmental Research and Public Health, 18(6), 2803, DOI: 10.3390/ijerph18062803.

Henning A., McLaughlin C., Armen S. and Allen S. (2021). Socio-spatial influences on the prevalence of COVID-19 in central Pennsylvania. Spatial and Spatio-Temporal Epidemiology, 37, 100411, DOI: 10.1016/j.sste.2021.100411.

Kelejian H.H. and Prucha I.R. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. The Journal of Real Estate Finance and Economics, 17(1), 99-121, DOI: 10/bxvhm4.

Kolosov V.A., Tikunov V.S. and Eremchenko E.N. (2021). Areas Of Socio-Geographical Study Of The Covid-19 Pandemic In Russia And The World. GEOGRAPHY, ENVIRONMENT, SUSTAINABILITY, 14(4), 109-116, DOI: 10.24057/2071-9388-2021-091.

Konstantinoudis G., Padellini T., Bennett J., Davies B., Ezzati M. and Blangiardo M. (2021). Long-term exposure to air-pollution and COVID-19 mortality in England: A hierarchical spatial analysis. Environment International, 146, 106316, DOI: 10.1016/j.envint.2020.106316.

Kotov E. (2022). e-kotov/ru-covid19-regional-excess-mortality article data and code. URL: https://github.com/e-kotov/ru-covid19-regional-excess-mortality. Zenodo, DOI: 10.5281/zenodo.6515455.

LeSage J. and Pace R.K. (2009). Introduction to Spatial Econometrics. CRC Press.

Luo Y., Yan J. and McClure S. (2021). Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. Environmental Science and Pollution Research, 28(6), 6587-6599, DOI: 10.1007/s11356-020-10962-2.

Maiti A., Zhang Q., Sannigrahi S., Pramanik S., Chakraborti S., Cerda A. and Pilla F. (2021). Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the contiguous United States. Sustainable Cities and Society, 68, 102784, DOI: 10.1016/j.scs.2021.102784.

Mogi R., Kato G. and Annaka S. (2020). Socioeconomic inequality and COVID-19 prevalence across municipalities in Catalonia, Spain [Preprint], DOI: 10.31235/osf.io/5jgzy.

Mollalo A., Vahedi B. and Rivera K.M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Science of The Total Environment, 728, 138884, DOI: 10.1016/j.scitotenv.2020.138884.

OpenStreetMap contributors (2017). OpenStreetMap. [online] Available at: https://www.openstreetmap.org [Accessed 01 Jul. 2021]

Oto-Peralías D. (2020). Regional correlations of COVID-19 in Spain [Preprint], DOI: 10.31219/osf.io/tjdgw.

Perone G. (2021). The determinants of COVID-19 case fatality rate (CFR) in the Italian regions and provinces: An analysis of environmental, demographic, and healthcare factors. Science of The Total Environment, 755, 142523, DOI: 10.1016/j.scitotenv.2020.142523.

Qi H., Xiao S., Shi R., Ward M.P., Chen Y., Tu W., Su Q., Wang W., Wang X. and Zhang Z. (2020). COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. Science of The Total Environment, 728, 138778, DOI: 10.1016/j.scitotenv.2020.138778.

Rahman M.H., Zafri N.M., Ashik F. and Waliullah M. (2020). Gis-Based Spatial Modeling to Identify Factors Affecting COVID-19 Incidence Rates in Bangladesh [SSRN Scholarly Paper], DOI: 10.2139/ssrn.3674984.

Raymundo C.E., Oliveira M.C., Eleuterio T. de A., André S.R., da Silva M.G., Queiroz E.R. da S. and Medronho R. de A. (2021). Spatial analysis of COVID-19 incidence and the sociodemographic context in Brazil. PLOS ONE, 16(3), e0247794, DOI: 10.1371/journal.pone.0247794.

Rodríguez-Pose A. and Burlina C. (2021). Institutions and the uneven geography of the first wave of the COVID-19 pandemic. Journal of Regional Science, 61(4), 728-752, DOI: 10.1111/jors.12541.

Sannigrahi S., Pilla F., Basu B. and Sarkar Basu A. (2020). The overall mortality caused by COVID-19 in the European region is highly associated with demographic composition: A spatial regression-based approach. https://ui.adsabs.harvard.edu/abs/2020arXiv200504029S.

Scarpone C., Brinkmann S.T., Große T., Sonnenwald D., Fuchs M. and Walker B.B. (2020). A multimethod approach for county-scale geospatial analysis of emerging infectious diseases: a cross-sectional case study of COVID-19 incidence in Germany. International Journal of Health Geographics, 19(1), 32, DOI: 10.1186/s12942-020-00225-1.

Sun F., Matthews S.A., Yang T.-C. and Hu M.-H. (2020). A spatial analysis of the COVID-19 period prevalence in U.S. counties through June 28, 2020: where geography matters? Annals of Epidemiology, 52, 54-59.e1, DOI: 10.1016/j.annepidem.2020.07.014.

Wang Q., Dong W., Yang K., Ren Z., Huang D., Zhang P. and Wang J. (2021). Temporal and spatial analysis of COVID-19 transmission in China and its influencing factors. International Journal of Infectious Diseases, 105, 675-685, DOI: 10.1016/j.ijid.2021.03.014.

Yarmol-Matusiak E.A., Cipriano L.E. and Stranges S. (2021). A comparison of COVID-19 epidemiological indicators in Sweden, Norway, Denmark, and Finland. Scandinavian Journal of Public Health, 49(1), 69-78, DOI: 10.1177/1403494820980264.

Zemtsov S.P. and Baburin V.L. (2020). Risks of morbidity and mortality during the COVID-19 pandemic in Russian regions. Population and Economics, 4(2), 158-181, DOI: 10.3897/popecon.4.e54055.

Appendix A. Variables used in the study

Group	Variable Name	Description	Source*
Dependent variable	Excess_mortality_apr_feb_per_capita	Excess per capita mortality over a period from April 2020 to February 2021 compared to the mean over previous 5 years	1
Demographic	Population	Mean population during a calendar year	1
Demographic	Urban_Population_Share	Share of urban population	1
Demographic	Migr_IntraReg_3Y_mean_per_capita_x10000	3-year mean intraregional migrants per 10 000 inhabitants	1
Demographic	Migr_Inflow_InterReg_3Y_mean_per_capita_x10000	3-year mean interregional arriving migrants per 10 000 inhabitants	1
Demographic	Migr_Inflow_International_3Y_mean_per_capita_x10000	3-year mean international arriving migrants per 10 000 inhabitants	1
Demographic	Migr_Outflow_InterReg_3Y_mean_per_capita_x10000	3-year mean interregional departing migrants per 10 000 inhabitants	1
Demographic	Migr_Outflow_International_3Y_mean_per_capita_x10000	3-year mean international departing migrants per 10 000 inhabitants	1
Demographic	Employees_in_Working_Age_Population_Share	Share of employed people in total working-age population	1
Demographic	Working_Age_Population_Share	Share of working-age population in total population	1
Demographic	Under_Working_Age_Population_Share	Share of under working-age population in total population	1
Demographic	Post_Working_Age_Population_Share	Share of post-working-age population in total population	1
Environment	Jan_Temp_Mean_10yr	10-year mean of temparature in January (2010-2020)	2
Environment	Jul_Temp_Mean_10yr	10-year mean of temparature in July (2010-2020)	2
Environment	Jan_Humid_Mean_10yr	10-year mean of humidity in January (2010-2020)	2
Environment	Jul_Humid_Mean_10yr	10-year mean of humidity in January (2010-2020)	2
Human Interaction	Road_Density	Total length of federal and regional level roads / area of the region	3
Human Interaction	Rail_Road_Density	Total length of standard width rail roads / area of the region	3
Human Interaction	Airport_Density	Total number of airports / area of the region	3, 4
Human Interaction	Buses_per_capita	Number of buses per person	1
Healthcare	Doctors_per_capita_x1000	Number of doctors per 1 000 inhabitants	1
Healthcare	Change_over_5yrs_Doctors_per_capita_x1000	Ratio of the number of doctors per 1 000 inhabitants, 2019 to 2015	1
Human Interaction	Modern_Retail_Area_per_capita_x1000	Total area of large retail (600+ square meters) per 1000 inhabitants	1
Human Interaction	Floor_Area_per_capita	Residential floor area per person	1
Human Interaction	Retail_N_per_capita_x1000	Total number of all retail stores per 1000 inhabitants	1
Human Interaction	Retail_Area_per_capita_x1000	Total area of all retail retail per 1000 inhabitants	1
Human Interaction	Cars_per_capita_x1000	Number of private passanger cars per 1 000 inhabitants	1
Human Interaction	Ecommerce_Users_Share	Share of people using ecommerce for shopping	1

Human Interaction	Digital_Gov_Serv_Users_Share	Share of people using digital government services	1
Human Interaction	Workers_in_A_EconAct_Share	Share of working-age population working in A - Agriculture	1
Human Interaction	Workers_in_B_EconAct_Share	Share of working-age population working in B - Mining	1
Human Interaction	Workers_in_C_EconAct_Share	Share of working-age population working in C - Manufacturing	1
Human Interaction	Workers_in_D_EconAct_Share	Share of working-age population working in D - Electricity, Gas, etc.	1
Human Interaction	Workers_in_E_EconAct_Share	Share of working-age population working in E - Water Supply	1
Human Interaction	Workers_in_F_EconAct_Share	Share of working-age population working in F - Construction	1
Human Interaction	Workers_in_G_EconAct_Share	Share of working-age population working in G - Wholesale and Retail Trade	1
Human Interaction	Workers_in_H_EconAct_Share	Share of working-age population working in H - Transportation and Storage	1
Human Interaction	Workers_in_I_EconAct_Share	Share of working-age population working in I - Accomodation and Food	1
Human Interaction	Workers_in_J_EconAct_Share	Share of working-age population working in J - IT and Communication	1
Human Interaction	Workers_in_K_EconAct_Share	Share of working-age population working in K - Finance and Insurance	1
Human Interaction	Workers_in_L_EconAct_Share	Share of working-age population working in L - Real Estate	1
Human Interaction	Workers_in_M_EconAct_Share	Share of working-age population working in M - Professional, Scientific, Technical	1
Human Interaction	Workers_in_N_EconAct_Share	Share of working-age population working in N - Administrative and Support	1
Human Interaction	Workers_in_O_EconAct_Share	Share of working-age population working in O - Public Administration and Defence	1
Human Interaction	Workers_in_P_EconAct_Share	Share of working-age population working in P - Education	1
Human Interaction	Workers_in_Q_EconAct_Share	Share of working-age population working in Q - Healthcare and Social Work	1
Human Interaction	Workers_in_R_EconAct_Share	Share of working-age population working in R - Arts, Entertainment and Recreation	1
Human Interaction	Workers_in_S_EconAct_Share	Share of working-age population working in S - Other Services	1
Socioeconomic	Salary_Region_to_Country_Ratio	Ratio of mean regional salary to mean national salary	1
Socioeconomic	Population_Below_Living_ Wage_Share	Share of population with income below the living wage	1
Socioeconomic	Mean_Real_Wage	Mean Real Wage	1
Socioeconomic	Income_per_capita	Income per person	1
Socioeconomic	SME_in_GRDP_Share	Share of Small and Medium Enterprise output in Gross regional domestic product	1
Socioeconomic	GRDP_in_GDP_Share	Share of Gross regional domestic product in Gross Domestic Product	1
Socioeconomic	Patents_per_capita_x10000	Number of patents per 10 000 inhabitants	1

*1 – Federal State Statistics Service – Rosstat, 2 – All-Russia Research Institute of Hydrometeorological Information - World Data Center (RIHMI-WDC), Roshydromet, 3 – OpenStreetMap, 4 – Aircraft owners and pilots association of Russia