PREDICTION OF INFLUENZA PEAKS IN RUSSIAN CITIES: COMPARING THE ACCURACY OF TWO SEIR MODELS

Vasiliy N. Leonenko*

ITMO University, 49 Kronverksky Pr
197101, St. Petersburg, Russia

Sergey V. Ivanov

ITMO University, 49 Kronverksky Pr
197101, St. Petersburg, Russia

Abstract. This paper is dedicated to the application of two types of SEIR models to the influenza outbreak peak prediction in Russian cities. The first one is a continuous SEIR model described by a system of ordinary differential equations. The second one is a discrete model formulated as a set of difference equations, which was used in the Baroyan-Rvachev modeling framework for the influenza outbreak prediction in the Soviet Union. The outbreak peak day and height predictions were performed by calibrating both models to varied-size samples of long-term data on ARI incidence in Moscow, Saint Petersburg, and Novosibirsk. The accuracy of the modeling predictions on incomplete data was compared with a number of other peak forecasting methods tested on the same dataset. The drawbacks of the described prediction approach and possible ways to overcome them are discussed.

1. **Introduction.** Acute respiratory infections (ARIs) are among the oldest and the most widely spread human infectious diseases. The most notorious of them, influenza, causes repetitive epidemic outbreaks with ARI incidence dramatically exceeding the average seasonal level. Outbreaks of influenza result in 3 to 5 million cases of severe illness annually worldwide, and the mortality rate is from 250 to 500 thousand individuals per year [29]. Influenza also causes an increase of heart attacks and strokes [4], as well as other disease complications. Even during an epidemic outbreak, only 15 to 20% of the total ARI cases are attributed to influenza [22], and diagnosis of influenza or another acute respiratory infection with similar symptoms is possible only through laboratory testing [3]. Due to those issues, the common clinical diagnosis 'influenza-like illness' (ILI) is often used, which includes all severe ARI cases fitting to a certain description. The criteria of ILI vary slightly in different national healthcare systems. According to the WHO, ILI is an acute respiratory infection with measured fever of ≥ 38 C and cough with onset within the past 10 days [30]. Although is is not completely accurate from the epidemiological point, for the sake of simplicity we consider 'flu', 'influenza', and 'ILI' synonyms.

The earliest attempts of mathematical modeling of influenza-like illness outbreaks took place in the late 1960s and this area of mathematical epidemiology is still popular today. Despite the efforts of various scientific groups to clarify the mechanism

of flu propagation dynamics, many unresolved questions remains. Today researchers try to enhance the descriptive abilities of their models by taking into account different external factors, such as weather [11], [24], [31], variation of virus strains [26], individual contact patterns, and others. A thorough review on the corresponding research papers and the flu-related factors considered can be found in [25].

One of the particular applications of the calibrated flu dynamics models is outbreak prediction. For that purpose, in addition to the variations of classical Kermack-McKendrick SEIR models [10], [32], other various approaches are used including agent-based modeling [13], metapopulational modeling [7], and social media data analysis [8], [12]. A detailed review on influenza forecasting can be found in [5]. One of the issues that the researchers face is the lack of reliable long-term flu incidence data provided by well-established influenza surveillance systems. As a result, the majority of studies are performed on data from Western Europe ([9], [27]) and Northern America ([11], [24], [28]). However, there are several exceptions, such as [2], where flu incidence in Chile is regarded. The use of the data from a limited set of geographical areas sometimes brings researchers to disputable assumptions. For instance, when flu incidence data from tropical regions became available in last decade, it demonstrated that some well-established hypotheses on the nature of ILI dynamics emerged from the temperate regions data have limited applicability [25]. Therefore, it is fruitful to expand the number of ILI incidence data sources.

Russia is one of the countries that can potentially contribute to the field. Influenza is considered a reportable disease by the Russian healthcare. In Saint-Petersburg (earlier Leningrad) ARI incidence has been collected since 1935, which gives one of the longest flu surveillance periods known [15]. In 1957, the all-USSR surveillance center was established. Its aim was to collect weekly and daily reports on ARI incidence from the local healthcare units throughout the country, with the number of covered cities being constantly increased over the years. The efficiency of the Soviet system of ARI cases registration led to the possibility of predicting the flu outbreaks in Soviet cities with a fairly high accuracy [6]. The employed model, created by Baroyan and Rvachev, was utilized by the specialists of Research Institute of Influenza [14] specifically for the sake of within-USSR flu propagation modeling [1]. Later it was applied to worldwide propagation of the pandemic flu [23]. The Baroyan-Rvachev model was a combination of the discrete Kermack-McKendrick SEIR model ("the local model") and a linear model of inter-city migration flows ("the transport model"). Although the structure itself was not novel, it matched real disease dynamics and made it possible to achieve accurate forecasts of the starting moments and peaks of influenza outbreaks in Soviet cities. For instance, from all the cases of epidemic outbreaks in 1970's, the day of the outbreak start was predicted without errors in 56.1% of cases and with a bias less than a week in 92.2% of cases, the same numbers for the day of the outbreak peak were 53.0% and 87.4% correspondingly [15].

However, from early 1980's the Soviet modeling framework for flu forecasting showed the signs of growing incoherence with the epidemic outbreak patterns observed in Soviet cities. According to [15], its malfunction was caused by the growing levels of herd immunity to flu due to increasing speed of its circulation around the globe. The core idea of Baroyan-Rvachev approach was that the fraction of non-immune individuals was the same in all Soviet cities and depended only on currently circulating virus strain. Since herd immunity levels grow with a rate dependent from different factors, including the structure of contact networks within an urban area,

the aforementioned assumption is less applicable now than in Soviet times. Professor Ivannikov, who was in charge of utilizing the Soviet forecasting system, claimed that a new model for the influenza peak prediction in Russia should be built to provide accurate forecasts [16]. Unlike the Baroyan-Rvachev model, it should rely heavily on the analysis of the local urban incidence data, thus making it possible to overcome the issue of unequal herd immunity levels within the country. Nevertheless, this idea has never been tested on actual data.

In this paper, the authors aim at assessing the accuracy of the peak predictions with the help of two different SEIR models calibrated with local incidence data, and discussing the ways of improving it.

2. **Research idea.** In [17] we formulated the continuous SEIR ("Susceptible-Exposed-Infected-Recovered") model for flu outbreak dynamics and calibrated it to the long-term Russian ARI incidence data. It was shown that the classical SEIR model, without any modifications accounting for the influence of external factors, may provide a satisfactory fit for the majority of influenza outbreak incidence datasets (the value of the coefficient of determination was $R^2 > 0.91$ for 64 epidemic curves out of 67). The current paper represents the next step of the research aimed at assessing the predictive force of the simple epidemic models for the case of Russian ARI incidence data. Our goals were:

- to assess the possibility of accurate prediction of the epidemic outbreak peaks in any of the Russian cities relying on the incomplete incidence data for the current outbreak in this city only — apart from Baroyan-Rvachev approach, where the ARI incidence in all Soviet cities along with migration flow data was used to obtain predictions;
- to find out whether it is possible to obtain the desired accuracy without incorporating external factors into the model;
- to assess the number of incidence points for model calibration and, consequently, the average time before the actual peak, required to obtain the accurate prediction.

The particular steps we had to take to reach the aforementioned goals consisted of the following:

- to perform the retrospective forecast of influenza dynamics in three Russian cities (Moscow, Saint Petersburg and Novosibirsk) with the help of the SEIR model calibrated on incomplete data using long-term incidence data from Research Institute of Influenza;
- to assess the accuracy of the epidemic peak parameters prediction, particularly the day of the peak and its prospected height, and its dependence on the number of incidence points used for the model calibration.

Assuming that the accuracy of prediction may largely depend on the model serving as a core of the fitting algorithm, we have decided to employ the local submodel of Baroyan-Rvachev modeling framework [1], [15] in addition to our continuous SEIR model, and to compare their predictive abilities.

3. **Models description.**

3.1. **The continuous SEIR model.** For the sake of describing the dynamics of influenza epidemic process, we have utilized a simple populational model represented by a system of ordinary differential equations. Since the flu has an incubation

period, and recovered individuals acquire immunity from the particular virus strain [29], the population of an urban area under consideration is represented by set of four groups of individuals: susceptible (vulnerable to flu infection), exposed (asymptomatic and non-infectious), infectious (symptomatic, spreading the flu) and removed (immune to the flu). The sizes of groups are measured in fractions of total population N: let S be the fraction of susceptible individuals, E — the fraction of exposed individuals, I — the fraction of infectious individuals, and R — the fraction of removed individuals. Following [1], [23], we state that a certain fraction of the population in every city under consideration is not vulnerable to flu. This subgroup includes people with immunity gained from previous infections and those who are not immune on their own but are protected by herd immunity. The fraction of population which is vulnerable to the currently circulating flu virus strain is denoted by $\alpha \in [0; 1]$, and the remaining population is considered immune to the infection. The dynamics of the groups' sizes over time are set by the following equations:

$$\frac{dS}{dt} = -\beta SI, \tag{1}$$

$$\frac{dE}{dt} = \beta SI - \gamma E,$$

$$\frac{dI}{dt} = \gamma E - \delta I,$$

$$\frac{dR}{dt} = \delta I,$$

$$S(t_0) = S_0 \geq 0, E(t_0) = E_0 \geq 0, I(t_0) = I_0 \geq 0,$$

$$S_0 + E_0 + I_0 = \alpha, R(t_0) = 1 - \alpha. \tag{2}$$

Since the duration of the epidemic process is relatively short, we consider the influence of birth and migration processes on the disease dynamics negligible and do not include these processes into the model. The description of parameters used for fitting the model is given in Table 1. Further in the text we consider $t_0 = 0$ without loss of generality.

3.2. **The Baroyan-Rvachev model.** The local submodel used in the Baroyan-Rvachev prediction framework was represented by the system of difference equations, with the time step equal to one day. Following the notations introduced in [15], let x_t be the fraction of susceptibles in the population, y_t be the number of newly infected individuals at the moment t and $\overline{y_t}$ – the cumulative number of *infectious* persons (i.e. those who can transmit flu) by the time t. Then, the equation system may be written in the following manner:

$$\overline{y_t} = \sum_{\tau=0}^{T} y_{t-\tau} g_\tau, \tag{3}$$

$$y_{t+1} = \frac{\beta}{\rho} x_t \overline{y_t},$$

$$x_{t+1} = x_t - y_{t+1},$$

$$x_0 = \alpha\rho. \tag{4}$$

The piecewise constant function g_τ gives a fraction of infectious individuals in the group of individuals infected τ days before the current moment t. The function reflects the change of individual infectiousness over time from the moment of

acquiring influenza. It is assumed that there exists some moment $\bar{t}$: $\forall t \geq \bar{t}$ $g_\tau = 0$, which reflects the moment of recovery.

4. **Outbreak incidence data.** The original dataset provided by the Research Institute of Influenza [14] contains weekly cumulative incidence for all the ARI types (including flu) in three Russian cities from 1986 to 2014. Before the model fitting, we have to refine the incidence data by restoring the missed values and fixing the under-reporting. We also need to extract flu incidence from the cumulative ARI incidence data. Corresponding algorithms are described in detail in [18], here we introduce briefly the sequence of operations.

- Under-reporting correction. Since infected people avoid visiting healthcare facilities during holidays, the corresponding weekly prevalence is lower than the actual number of newly infected. This under-reporting bias can be corrected by means of cubic interpolation [1] using the incidence registered in the adjacent weeks. The sporadic gaps in incidence data are filled in the same fashion.
- Bringing the incidence data to daily format. The daily incidence is found with the help of cubic interpolation of weekly incidence. We assume that $n_{inf}^{Thu} = n_{inf}^{W}/7$, where n_{inf}^{W} is the weekly incidence taken from the database and n_{inf}^{Thu} is the daily incidence for Thursday of the corresponding week.
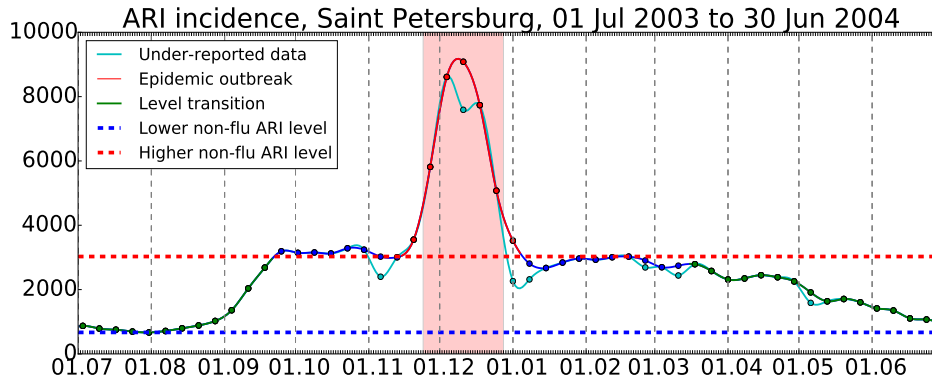

FIGURE 1. An example of the epidemic curve extraction from the interpolated ARI incidence.

- Extracting data on influenza outbreak from the cumulative seasonal ARI data with the help of a separate epidemic curve allocation algorithm. At first, the algorithm finds higher non-flu ARI incidence level a_2, which corresponds to the average number of newly infected in non-epidemic period (figure 1, red horizontal dashed line). ARI epidemic curves, which are detected as flu outbreaks (figure 1, red solid line), should have their peaks well above the higher ARI level. They should also comply with the time period during which the ARI prevalence exceeds the non-epidemic ARI threshold assessed in the Flu Research Institute (figure 1, red rectangle). The beginning and ending of the extracted curve is chosen to match the level a_2. The first incidence point of the curve is considered to be the first day of the epidemic outbreak.

5. **Fitting models to data.**

5.1. **General procedure.** Let $Z^{(dat)}$ be the set of incidence data points loaded from the input file and corresponding to one particular outbreak. Assume that the number of points is t_1, which equals the observed duration of the outbreak. The basic idea of the model fitting procedure is the same for the both models. The algorithms vary the values of model parameters to achieve the model output, which minimizes the distance between the modeled and real incidence points:

$$F(Z^{(mod)}, Z^{(dat)}) = \sum_{i=0}^{t_1} (z_i^{(mod)} - z_i^{(dat)})^2, \tag{5}$$

Here $z_i^{(dat)}$ and $z_i^{(mod)}$ are the absolute incidence numbers for the i-th day taken from the input dataset and derived from the model correspondingly. The limited-memory BFGS optimization method is used to find the best fit [21]. Since the existence of several local minima is possible, the algorithm has to be launched several times with different initial values of input variables. The best fit is chosen as a minimum among the distances achieved from all the algorithm runs. To characterize the goodness of fit we utilize the coefficient of determination $R^2 \in (0, 1]$. This coefficient shows the fraction of the response variable variation that is explained by a model [27].

Before optimizing the model parameters, we need to match accurately the model timeline ($t = 0, 1, \dots$) to the timeline of the epidemic outbreak incidence dataset. In this paper we use two approaches for that, namely:

- **Aligning the timelines by outbreak starting day.** We assume that the moment $t = 0$ of the model coincides with the first incidence point of the curve.
- **Aligning the timelines by peak day.** We assume that the peak moment of the modeled epidemic curve coincides with the epidemic peak day from the dataset.

The first one is a part of a fitting algorithm for the continuous SEIR model, the second one is used for the Baroyan-Rvachev model calibration.

The issue that affects timelines alignment is inaccuracy of procedure input. The outbreak starting day detection depends on the curve extraction algorithms employed (fig. 2) and cannot be established accurately due to absence of distinct diagnosis of influenza and other acute respiratory illnesses. The peak moment is known only in the case we fit the model to data on past epidemic outbreaks, apart from performing predictions for the ongoing outbreak, but even then, biases in incidence registration (like the aforementioned under-reporting during holidays) can lead to incorrect determination of the peak moment.

In this paper we compensate the uncertainty in the input (outbreak starting day and peak day obtained from the dataset) by introducing curve positioning parameters. These parameters are the part of fitting algorithms, not the models themselves. Thus, their function is to not change the shape of the model curves, but rather to adjust the position of the modeled curve relatively to the epidemic incidence data. The main issue of using curve fitting parameters is that they give additional degrees of freedom to the fitting algorithm. Thus, they make it possible to fit various model curves to incidence data and expand the range of possible model parameter values.

The details on the curve positioning parameters used in each of two fitting algorithms are given in the subsequent section.
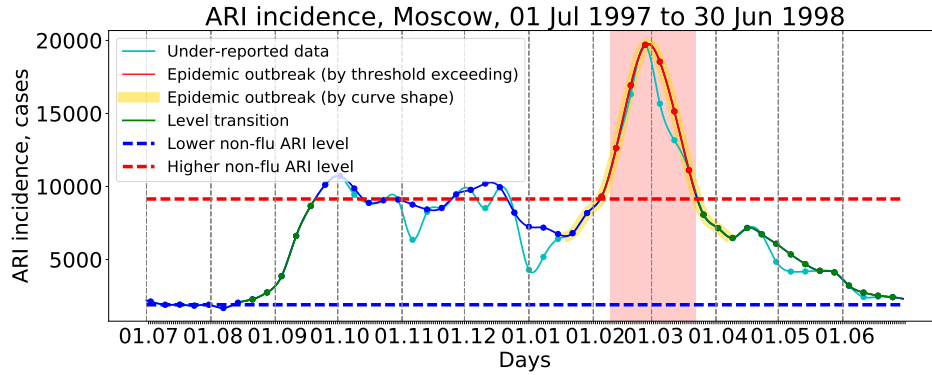
FIGURE 2. ARI incidence curve showing the discrepancy between different curve allocation algorithms

TABLE 1. Parameters of the fitting algorithm for the continuous SEIR model

Definition	Description	Value	Unit
	Epidemiological parameters		
α	Initial ratio of susceptible individuals in the population	Estimated	-
β	Intensity of infection	Estimated	$1/(\text{person} \cdot \text{day})$
γ	Intensity of transition to infective form of the disease	0.39	$1/\text{day}$
δ	Intensity of recovery	0.133	$1/\text{day}$
I_0	Initial ratio of the infected	0.0001	-
	Curve positioning parameters		
k_{inc}	Relative vertical bias of the modeled incidence curve position	Estimated	-
Δ_s	Absolute horizontal bias of the modeled incidence curve epidemic start position compared to the data	Estimated	day

5.2. **The continuous SEIR model fitting.** The list of parameters involved in the fitting procedure (table 1), apart from five model parameters (α, β, γ, δ and I_0), includes two curve positioning parameters, Δ and k_{inc}. Δ has sense of the difference between the day of the outbreak start detected by the curve allocation algorithm and the outbreak start demonstrated by the model. k_{inc} is the relative difference between the baseline ARI level found from the data and the corresponding level obtained as a result of model fitting. The necessity of variation of the two parameters (both for horizontal and vertical model curve positioning) arises from possible controversies in the assessment of the epidemic outbreak period depending on the chosen curve detection algorithm. For example, apart from figure 1, where the base of visually distinguished epidemic curve coincides with the higher non-flu ARI level, in figure 2 the curve seemingly starts well below the level a_2. That is why if the outbreak start obtained from real data is the left end of visually distinguished curve (yellow solid line) rather than the left end of automatically detected curve (red solid line), the model curve has to be shifted both in horizontal and vertical

direction to match the epidemic incidence data. Accordingly, small values of Δ and $k_{inc} \approx 1$ correspond to the 'clear' cases, when it is easy to distinguish the flu outbreak curve edges from the seasonal ARI level. On the contrary, the incidence data with non-smooth edges (fig. 2) is fitted by the curve with big Δ and small k_{inc}, indicating that the epidemic outbreak started earlier than it was detected by the curve allocation algorithm.

Relying on conclusions made from earlier numerical experiments with the model fitting [17], we have decided to fix the values of γ and δ, thus reducing the model state space and possibly making the model less prone to overfitting.

The fitting algorithm for the continuous SEIR model was introduced for the first time in [17]. The algorithm operations are performed as follows. For each $\Delta \in 5, .., 54$:

- For each fixed combination of values $\{\alpha, \beta, \gamma, \delta, k_{inc}\}$ generated by BFGS optimization procedure:
  1. Find the numerical solution of the model (1) – (2) with the initial conditions $S(0) = \alpha - I_0, E(0) = 0, I(0) = I_0, R(0) = 1 - \alpha$.
  2. Calculate the modeling flu incidence in relative numbers: $y^{(mod,rel)}(t) = N_{E \to I}(t)$. Since from (1) – (2)

  $$E(t) = E(t-1) + N_{S \to E}(t) - N_{E \to I}(t)$$

  and

  $$N_{S \to E}(t) = S(t-1) - S(t),$$

  we achieve:

  $$y^{(mod,rel)}(t) = -\Delta S(t) - \Delta E(t), t = 1, 2, \ldots$$
  $$\Delta S(t) = S(t) - S(t-1),$$
  $$\Delta E(t) = E(t) - E(t-1),$$

  3. As we are working with disease incidence attributed only to influenza outbreaks, excluding the non-epidemic cases of ARI infections, we need to subtract the non-epidemic incidence from the overall ARI incidence data. For that purpose we need to derive the baseline level for the modeled outbreak start y_{base} from the value for higher ARI incidence level a_2, considering the relative bias k_{inc}, and to subtract it from the data incidence points:

  $$y_{base} := k_{inc} \cdot a_2,$$
  $$y_i^{(dat)} := y_i^{(dat)} - y_{base}, i \in \overline{0, T-1}$$

  4. Consider that the data incidence points from the dataset are shifted by Δ days from the model curve start. Thus we are to compare the distance between the following datasets:

  $$Y^{(dat)} = \{y_0^{(dat)}, y_1^{(dat)}, \ldots, y_{T-1}^{(dat)}\},$$
  $$Y^{(mod)} = \{y^{(mod)}(\Delta), y^{(mod)}(\Delta + 1), \ldots, y^{(mod)}(\Delta + T - 1)\}.$$

  5. Convert the relative model incidence values to absolute values:

  $$y_i^{(mod)} = y_i^{(mod,rel)} \cdot N_L(m), \tag{6}$$

  where $N_L(m)$ is the total population of the city L in the year m equal to the starting year of the epidemic season under consideration.

6. Calculate the value of the fit function $F(Y^{(mod)}, Y^{(dat)})$, $F = F(\Delta)$.

In the described manner the BFGS algorithm finds the least distance F_Δ for every value of Δ. We define Δ_{min}: $F(\Delta_{min}, \dots) = \min F(\Delta, \dots)$, and the parameter set $\{\alpha, \beta, \gamma, \delta, k_{inc}\}$, corresponding to Δ_{min}. These values are the final result of our optimization procedure.

After the optimization algorithm has established the best fitting model parameter values, the model can be used to estimate the dynamics of population groups $S(t)$, $E(t)$, $I(t)$ and $R(t)$ over time. The group quantities are converted to absolute format in the same way as it is done with influenza incidence in (6).

5.3. **The Baroyan-Rvachev model fitting.** The parameter description of the fitting algorithm which corresponds to Baroyan-Rvachev model is given in Table 2. In addition to the parameters taken from the model (3)–(4), a curve positioning parameter Δ_p is introduced. In the ideal case, the best fit of the model curve to data should give the model curve peak occurring at the same day as the real peak seen from the incidence data. In that case, Δ_p is fixed and equals zero, so there is no need to vary it. However, sometimes the best fit is achieved when the modeled and the real peak moments differ by several days due to discrepancies between the real outbreak process and its theoretical model, that is why we made this parameter variable. Also the algorithm ability to vary Δ_p will be used in Section 6 for the sake of peak prediction.

TABLE 2. Parameters of the fitting algorithm for Baroyan-Rvachev model

Definition	Description	Value	Unit
	Model parameters		
α	Initial ratio of susceptible individuals in the population	Estimated	-
β	Intensity of infection	Estimated	-
I_0	Initial ratio of infected in the population	Estimated	-
T	Duration of infection	Fixed	day
g_τ	A fraction of infectious individuals among those who were infected τ days before the current moment	Fixed	-
ρ	Population size	Fixed	persons
	Curve positioning parameters		
Δ_p	Absolute horizontal bias of the modeled incidence curve peak position compared to the data	Estimated	day

The important advantage of the Baroyan-Rvachev model fitting algorithm compared to the continuous SEIR model fitting is that it has fewer parameters to be varied. Moreover, it has been proven [1] that without the loss of fit quality we can vary the sole auxiliary value $s = \alpha\beta$ instead of variation α and β separately. This fact results from the biologically plausible assumption that the virulence of the current circulating influenza strain and the immunity level to this strain in the population are interconnected [15]. Another idea of the algorithm is that $s = s(k)$, where k is a semi-empirical parameter which approximates the initial fraction of infectious individuals in the first stage of the outbreak [1]. Finally we come to idea

that it is k that should be varied, and s is consequently calculated as a function of k (see the formula below).

Another benefit of the algorithm is that it relies on the peak day alignment rather than starting day alignment, so it is not affected by incorrect outbreak starting day detection and we do not need to add a vertical positioning parameter.

The description of the algorithm follows.

- For each fixed combination of values $\{k, I_0\}$ generated by BFGS optimization procedure, and for every Δ_p:
  1. Derive the value of s from the current value of k using the following formula:
  $$s = \frac{k^{T+1}}{\sum_{\tau=0}^{T} k^{T-\tau} g_\tau} \tag{7}$$
  2. Set the preliminary model parameter values, α' and β', to make them conform to the equation $\alpha' \beta' = s$, for instance:
  $$\alpha' = 1, \beta' = s$$
  3. Find the preliminary numerical solution of the model (3) – (4) with the parameter values $\alpha = \alpha'$, $\beta = \beta'$ and the initial conditions $y_0 = I(0), y_{-1} = \cdots = y_{-T} = 0$.
  4. Derive the preliminary number of newly infected each day from the model output: $z_i^{(mod)'} := \overline{y_i}, \ i \in \overline{0, t_1 - 1}$.
  5. Derive the baseline level for the modeled outbreak start z_{base} from the value for higher ARI incidence level a_2 and subtract it from the data incidence points:
  $$z_i^{(dat)} := z_i^{(dat)} - a_2, i \in \overline{0, t_1 - 1}$$
  6. According to the algorithm, we need to match in time the model peak with the incidence data peak. For that purpose we find the shift δ_{adj}:
  $$\delta_{adj} = t_{peak}^{(mod)} - (t_{peak}^{(dat)} + \Delta_p), \tag{8}$$
  where $t_{peak}^{(dat)}$ and $t_{peak}^{(mod)}$ are peak days for the data incidence and the modeled incidence correspondingly, Δ_p is the difference between the modeled peak and the data peak after the shift. After performing a shift, we are to compare the distance between the following datasets:
  $$Z^{(dat)} = \{z_0^{(dat)}, z_1^{(dat)}, \ldots, z_{N-1}^{(dat)}\},$$
  $$Z^{(mod)'} = \{z^{(mod)'}(\delta_{adj}), z^{(mod)'}(\delta_{adj} + 1), \ldots, z^{(mod)'}(\delta_{adj} + N - 1)\},$$
  where N is the total number of incidence points.
  7. Assigning optimal values to α, β. It was mathematically justified in [1] that for $\alpha, \beta : \alpha\beta = s$
  $$\min_{\alpha, \beta} F(Z^{(mod)}(\alpha, \beta, \Delta_p), Z^{(dat)}) = F(Z^{(mod)}(\tilde{\alpha}, \tilde{\beta}, \Delta_p), Z^{(dat)}),$$
  $$\max Z^{(mod)}(\tilde{\alpha}, \tilde{\beta}, \Delta_p) = \max Z^{(dat)} + \Delta_p;$$
  where $\tilde{\alpha} = \alpha' a$, $\tilde{\beta} = \frac{\beta'}{a}$, a is the correction coefficient calculated according to the formula:
  $$a = \frac{\sum_{i=0}^{t_1} z_i^{(dat)}}{\sum_{i=0}^{t_1} \left(z_i^{(mod)'}\right)^2}.$$

To avoid launching the simulation for the second time, now with the values $\alpha = \tilde{\alpha}, \beta = \tilde{\beta}$, one may obtain the new model incidence values $z_i^{(mod)}$ by multiplying the corresponding preliminary values by a [15], that is:

$$z_i^{(mod)} = a z_i^{(mod)'}, z_i^{(mod)} \in Z^{(mod)}(\tilde{\alpha}, \tilde{\beta}, \Delta_p), z_i^{(mod)'} \in Z^{(mod)'}.$$

In that manner we find the optimal parameter values and the corresponding model curve $Z^{(mod)}$.

It is worth mentioning that α has the sense of fraction, $\alpha \in [0; 1]$. In the case if $\tilde{\alpha} > 1$, we artificially set it to 1.

8. Calculate the value of the fit function $F(Z^{(mod)}, Z^{(dat)})$ according to the formula (5).

The BFGS algorithm finds the least distance F_{Δ_p} in the described manner for every value of Δ_p. We define $\Delta_p^{(min)}$: $F(\Delta_p^{(min)}, \dots) = \min F(\Delta_p, \dots)$, and the parameter set $\{\alpha, \beta, I_0\}$, corresponding to $\Delta_p^{(min)}$. These values are the final result of our optimization procedure.

The described fitting algorithm for Baroyan-Rvachev model originates from [15], with several modifications that were made to unify it with the same procedure for the SEIR model. Particularly:

- The curve positioning parameter Δ_p was introduced (the similar parameter was mentioned in [15], but it was not explicitly included into the fitting procedure).
- The iteration over the values of variable k with a fixed step was replaced by BFGS optimization algorithm.
- The value of I_0 was changed from fixed to varied. This change was made due to the ambiguity of flu epidemic outbreak start detection in seasonal incidence data [17]. This allows the algorithm to fit the model to the early outbreak stages more accurately.

The modifications described enhanced both the accuracy and the performance of the algorithm and made it more suitable for our task of peak prediction.

6. **Numerical experiments.** Both algorithms are implemented as scripts collection written in Python programming language (Python 3.x with `numpy` and `matplotlib` libraries was used). The higher ARI level was assessed with the help of `scipy.optimize.curve_fit` procedure. The limited-memory BFGS optimization method for curve fitting was performed via `scipy.optimize.minimize` routine. The parameter value ranges used for the model fitting are given in Table 3. The range for k used in the fitting procedure for the Baroyan-Rvachev model was empirically found by the authors of the original model [1], the range for $\theta_{peak}^{(dat)}$ was assessed based on the actual moments of epidemic peaks taken from the Russian ARI incidence data, the ranges for I_0 and Δ_p were set somewhat arbitrary. The values of $g(\tau)$ were set according to the statistics available from the Soviet healthcare units [1]: $g(0) = g(1) = 0$, $g(3) = 0.9$, $g(3) = 0.9$, $g(4) = 0.55$, $g(5) = 0.3$, $g(6) = 0.15$, $g(7) = 0.05$, $g(8) = g(9) = \cdots = 0$. According to [1], another set of values for $g(\tau)$ was also tested, and the fitting algorithm showed low sensitivity to the change of these values on the Soviet ARI incidence data of 1970–1980s. This fact allowed us to use the same form of $g(\tau)$ in our experiments.

We have conducted the numerical experiments on weekly ARI incidence data for three Russian cities (Moscow, Saint Petersburg, and Novosibirsk) from July 1986
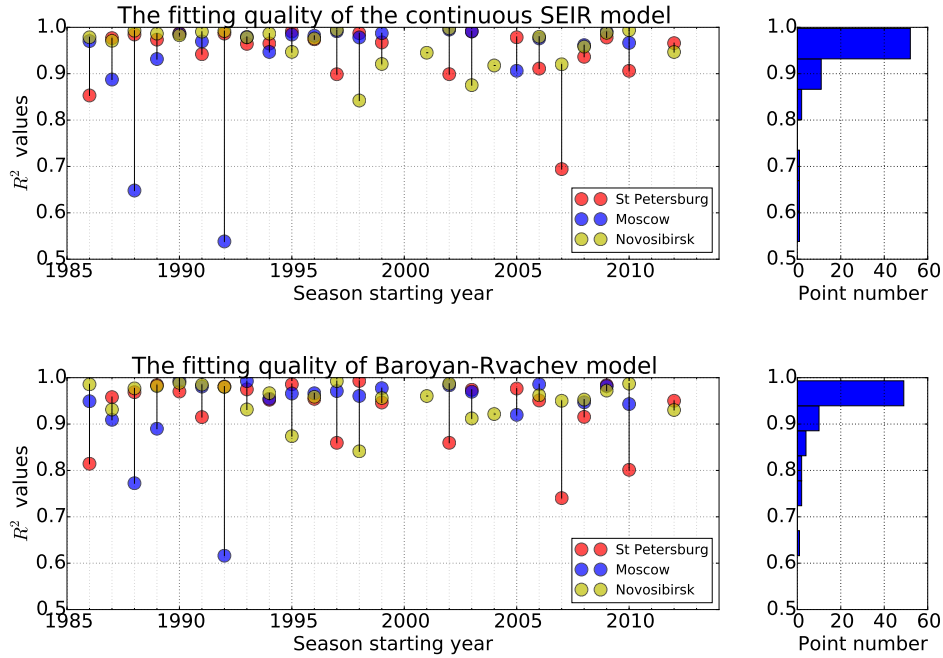
FIGURE 3. The fitting quality of the algorithms in the case of complete outbreak data.

to June 2014. By means of epidemic curve allocation algorithm, we extracted the incidence data for the epidemic outbreaks, which gave us 67 epidemic outbreaks in total (there were no epidemics during some seasons).

6.1. **Retrospective fitting to complete data.** To test the Baroyan-Rvachev fitting algorithm, we applied it to the complete outbreak data and compared the resulting accuracy of fit with the one achieved by the continuous SEIR model. The values of R^2 reflecting the quality of model fitting to the incidence data are demonstrated in Figure 3. On average, both models tend to demonstrate satisfactory fit quality. The exceptional cases correspond to a number of epidemic curves with peculiar forms than cannot be fitted well by one-peak models [17].

6.2. **Peak prediction on incomplete data.** After comparing the two models and the corresponding fitting algorithms on complete incidence data, we compared the predictive force of the models for the outbreak peak forecasting. For this purpose we reduced the incidence datasets for each epidemic season, reproducing the case of incomplete incidence data (Figure 4). The sample sizes were varied starting from 5 incidence points (that corresponds to the attempt of peak prediction at the fifth day of the outbreak, provided that the actual incidence data is provided by healthcare units by the end of each day).

At that stage of the experiment an important issue of the Baroyan-Rvachev fitting algorithm described in section 5.3 was revealed. Unlike the algorithm from section 5.2, it relies on the explicit knowledge of the day of outbreak peak. Obviously, during the exploitation of Baroyan-Rvachev modeling framework, the local submodel was calibrated on the 'half-wave' of flu incidence (i.e. on the data from the outbreak start

Table 3. Varied model parameters

Definition	Description	Value	Variation type
	Continuous SEIR model		
α	Initial ratio of susceptible individuals in the population	$[10^{-2}; 1.0]$	BFGS optimization
β	Intensity of infection	$[10^{-7}; 50.0]$	BFGS optimization
k_{inc}	Relative vertical bias of the modeled incidence curve position	$[0.8; 1.0]$	BFGS optimization
Δ_s	Absolute horizontal bias of the modeled incidence curve epidemic start position compared to the data	$5, \ldots, 54$	Iteration
	Baroyan-Rvachev model		
k	The service parameter defining the product of α and β	$[1.02; 1.6]$	BFGS optimization
I_0	Initial ratio of infected in the population	$[10^{-1}; 50.0]$	BFGS optimization
Δ_p*	Absolute horizontal bias of the modeled incidence curve peak position compared to the data	$-3, \ldots, 3$	Iteration
$\theta_{peak}^{(dat)}**$	Prospected incidence curve peak day	$17, \ldots, 83$	Iteration

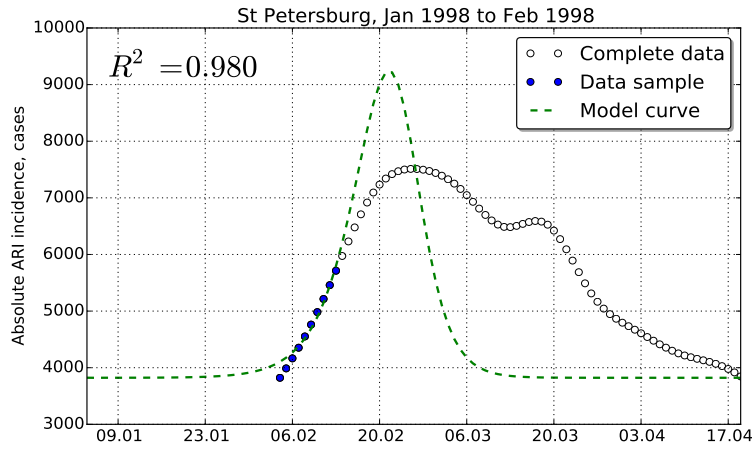* For complete incidence data ** For incomplete incidence data

Figure 4. An example of epidemic peak prediction by Baroyan-Rvachev model

till its peak), thus the value of the outbreak peak day was always available. When the fitting was made to complete data in the previous experiment, we knew this value too. In the current experiment, on the contrary, the day of the outbreak peak is meant to be the output parameter of the algorithm, along with the peak height. Hence, we had to modify the initial algorithm to make it suitable for prediction purposes. We modified the formula (8) in the following way:

$$\delta_{adj} = t_{peak}^{(mod)} - \theta_{peak}^{(dat)}.$$

The varied parameter $\theta_{peak}^{(dat)}$ reflects our guess of when the future incidence peak is likely to occur. Thus, instead of iterating through the set of possible values of Δ_p, as in fitting to complete data, we have to iterate through the set of different $\theta_{peak}^{(dat)}$ with a larger number of values. This fact influences significantly the algorithm performance, making it 5 –10 times slower than the one for the continuous SEIR model. On the contrary, in the previous experiment dedicated to model fitting to complete data, the Baroyan-Rvachev algorithm worked sufficiently faster than the latter.

Let n be the number of days taken from the beginning of epidemic outbreak, while $t_{peak}^{(dat)}$ is the day of the peak in the incidence data. For each $n \in \overline{5, t_{peak} - 1}$, the models are fitted to the incidence data set $Z^{(dat)} = \{z_i^{(dat)} | i \in \overline{1, n}\}$, the model curve peak is assessed and the following values are calculated:

- the prediction bias of the peak day dt,
- the ratio between the modeled and real outbreak peak heights dh.

To assess the accuracy of peak prediction results, we have used the 1970's Soviet flu outbreak prediction framework criteria [15] already applied by the authors in [19]:

- 'Square'. The prediction is thought to be accurate if $dt \in -8..8$ and $dh \in (0.5; 2.0)$.
- 'Vertical stripe'. The accurate prediction should have $dt \in -7..7$.
- 'Horizontal stripe'. The accurate prediction should have $dh \in (0.7; 1.5)$.

For every fixed outbreak, we have calculated the sample size l (or, which is the same, the number of days since the outbreak start), which allows the algorithm to comply with the chosen accuracy criteria for all consequent sample sizes starting from l, i.e. $l + 1, l + 2, \ldots$ The particular fitting algorithm is considered to be the more accurate of the two if it tends to demonstrate lesser values of l. The example of accuracy checking according to the vertical stripe criterion for the peak predictions of the outbreaks in Moscow is shown in Figure 5.

The obtained values of l were accumulated to derive the overall statistics, particularly, what chance is for the algorithm studied to achieve the prediction accuracy compliant to the fixed accuracy criterion for the fixed sample size (i.e. fixed number of days from the outbreak start). The overall results for the three cities are demonstrated in Figure 6. The detailed results for each of the cities may be found in Appendix (section 9).

As one can see, the accordance of the predictions to the 'horizontal stripe' criterion, i.e. the quality of peak height prediction, may be named satisfactory for the Baroyan-Rvachev model (98% of compliance for the predictions achieved one day before the peak) and unsatisfactory for the continuous SEIR model (51% of compliance). We believe that the Baroyan-Rvachev model, being more 'rigid', tends to reproduce better the overall trend of incidence data, whereas the continuous SEIR model is more prone to the reaction on the outliers in the data. Thus, the modeled incidence curves tend to change their slope in a greater extent in the latter case, resulting in bigger biases of the peak predictions.

The reason for the fact that the prediction compliance to accuracy criteria may still be low near the peak is due to the peculiar shape of some outbreak incidence curves which cannot be properly fitted by the one-peaked incidence model (see Figure 3 and [17] for more details).
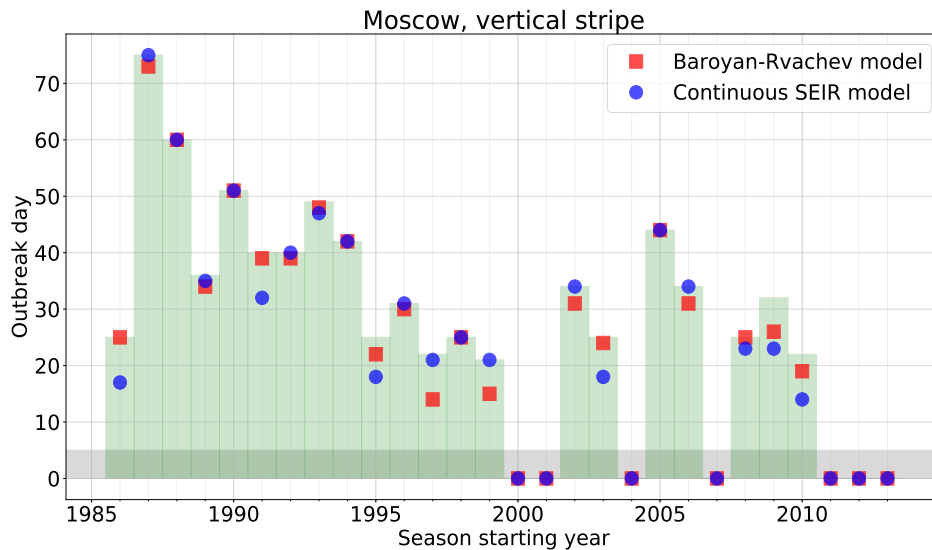
FIGURE 5. An example of estimating forecast accuracies. The height of the green bars corresponds to the duration of the outbreak before reaching the peak, the markers indicate the day when the accuracy criterion was reached by the particular model. The absent bars correspond to the years without an epidemic outbreak.

The accuracy percentage for the 'vertical stripe' and, consequently, 'square', is generally unsatisfactory for the both models. The SEIR model gives slightly better forecasts on the initial stages of the outbreak, whereas the Baroyan-Rvachev model fits better to big samples (when we almost have the epidemic "half-wave").

7. **Discussion.** The numerical experiments have shown that the prediction methods demonstrated in the paper may be applied to assess the height of the peak, but are incapable to predict the peak time. (An interesting fact is that the same issue, although caused by another reasons, also holds true for the last Baroyan and Rvachev forecasts performed in early 1980s [15] and for our attempts to modify their algorithm to be used without the transport data [19]). Apparently, we cannot expect great accuracy from the prediction obtained in such a straightforward manner, especially if the number of incidence points used to calibrate the model is not large. In this case, an incidence point sample can be fitted by various model curves with almost equal goodness of fit R^2, which results in a big variety of peak predictions. Obviously, the fact that a curve has the best value of R^2 among the curves with similar fit accuracy does not imply that it gives the best peak prediction. To enhance the predictive ability of the model, we need to add more constraints on the model curves, possibly based on some *a priori* knowledge on the past epidemic outbreaks in the particular urban area. This may include preliminary assessment of α and β by studying the dynamics of herd immunity over time and the peculiarities of contact patterns within the area correspondingly.

Among the technical limitations that we face while applying the algorithms, the most important was connected with the fitting algorithm performance. Because of the long duration of algorithm execution on the amount of incidence data employed,
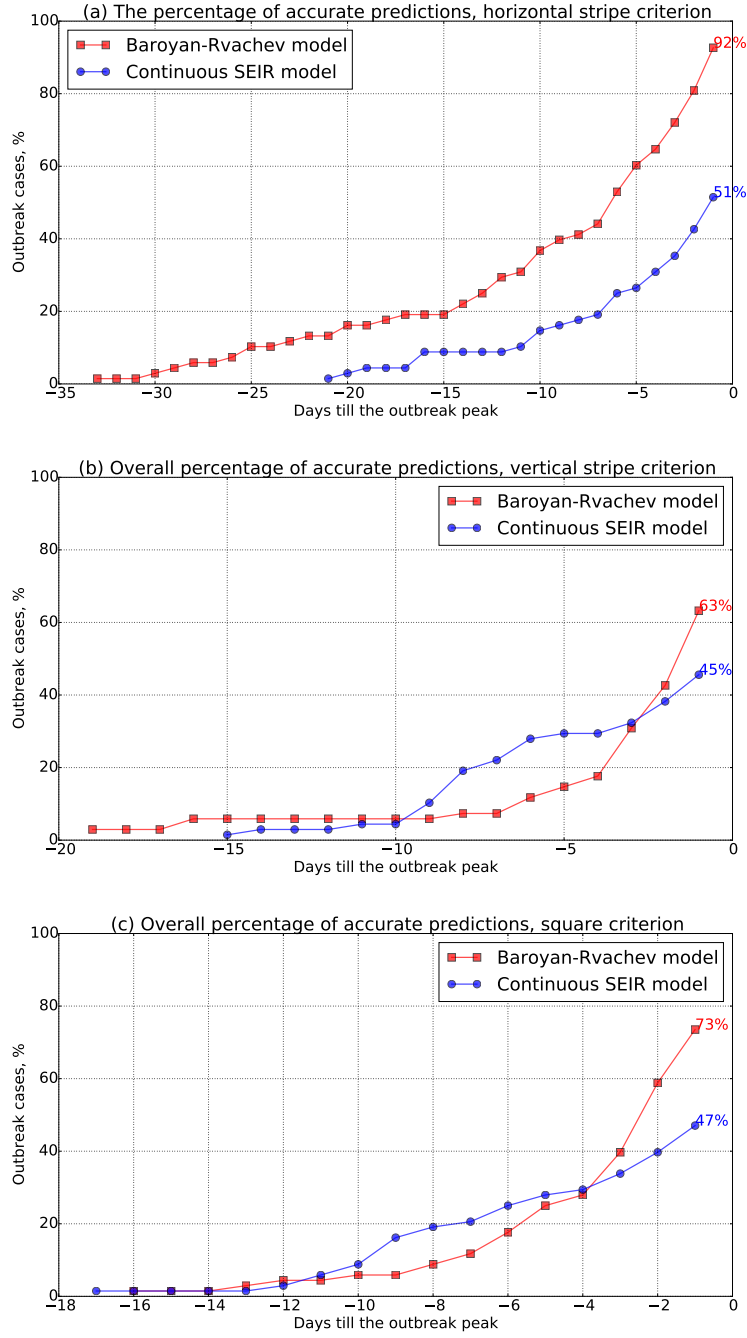
FIGURE 6. The dependence of the percentage of accurate predictions from the number of days before the peak, in overall for the three cities

we had to limit the number of runs with different initial values of input. In some cases, that may lead to an unsatisfactory fit due to the fact that the optimization function may have several local minima. We hope to increase the algorithm speed and reach a higher fit accuracy by employing the parallel techniques, such as thread distribution over the computer cores, in the same way as we made it in earlier works for a number of epidemic model algorithms [20].

The drawback of this work, that was already mentioned in [17], is that we did not consider the bias in data gained as a result of conversion of the weekly to daily incidence data. Despite the fact that the "synthetic" daily data is surely more "smooth" than the original one, we presume that our set of algorithms will be suitable to handle the real daily incidence dataset. Our assumption is supported by the fact that after filtering the fluctuations caused by the weekly cycle of individuals the daily epidemic curves resemble the synthetic data we work with (see [1]).

9. **Appendix. The comparison of prediction methods accuracy for the particular cities.** Figures 7–9 demonstrate the prediction accuracy according to three accuracy criteria for Saint Petersburg, Moscow, and Novosibirsk. For comparison purposes, we have added prospected prediction quality for the same cities obtained by calibrating the models to the earlier occurred outbreaks of the same epidemic season (see [19] for more details).

Figure 7 shows that the prediction accuracy of the peak height ('horizontal stripe' criterion) is better for the method employed in [19] than for the forecasting method described in this article. At the same time, the former method may be employed in a limited number of cases. Particularly, the employment of the method requires the existence of the city with the climax of the epidemic outbreak reached, otherwise we cannot calibrate the model to make a prediction. Thus, peak forecasting based on the incomplete data is more versatile, although less accurate. Note that the prediction accuracy of the method from [19] depends to large extent on the city which was used to calibrate the model. For instance, the reader can see that the predictions for Moscow obtained by calibrating the model to Novosibirsk data is significantly worse than the prediction obtained by using the incidence data from Saint Petersburg.

The prediction of peak days is unsatisfactory for the both methods, with the accuracy percentage of incomplete data forecasting becoming higher than the one of the method from [19] on average in the "day -3" (i.e. approximately three days before the peak).

Comparing the forecasting methods, we have come to the following question: may it be that the majority of the predictions is accurate due to the fact that the peak data (day and height) has very limited variance over the years? For instance, may we match the 'vertical stripe' accuracy criterion (a peak height prediction is 0.7 to 1.5 of the real peak height) by taking the average height of the previous epidemic peaks in that city? To answer this question, we have utilized the two statistical approaches:

- "Prediction by last peak data". We "predict" that the peak in the current season is likely to have exactly the same height and will happen after the same number of days from the epidemic outbreak start, as in the previous one.

- "Prediction by average peak data". We "predict" that the peak height and day are to coincide with the average height and average day calculated from the peak data over all the previous years.

The accuracy of the predictions obtained in the described way was compared with the modeling prediction obtained by the method from [19] (for each city we took the best accuracy from the two predictions based on the models calibrated on two cities; for instance, for Saint Petersburg we took the accuracy obtained on Moscow data, etc.) and by the method described in this article (we took the accuracy obtained by Baroyan-Rvachev model calibrated on the dataset correspondent to "day -1", i.e. the day before the actual peak). The comparison results are shown in Figure 10.

As one can see, in the case of the peak height prediction ("horizontal stripe" criterion) the accuracy of the modeling methods is significantly higher than of the primitive statistical approaches mentioned above. In case of peak day prediction, the modeling method from [19] demonstrates the accuracy, which is equal or worse than the one of the statistical approaches. This result supports our assumption that we cannot use the described modeling methods to assess the peak days. The fact that the Baroyan-Rvachev prediction accuracy is rather high for both "vertical stripe" and "square" criteria, does not prove this assumption wrong, because, as we can see on Figures 8–9, the accuracy of the method falls fast when we take fewer incidence points for the model calibration.
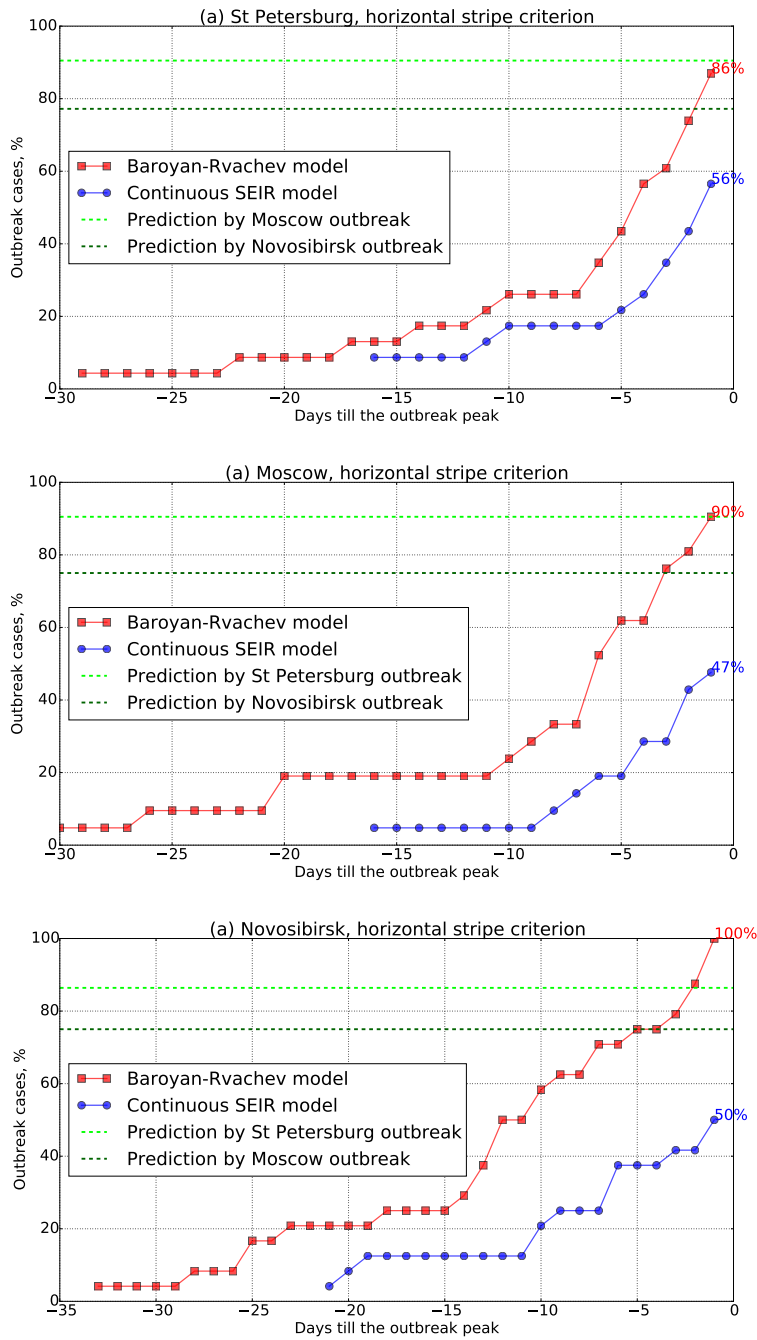
FIGURE 7. The dependence of the percentage of accurate predictions from the number of days before the peak, horizontal stripe
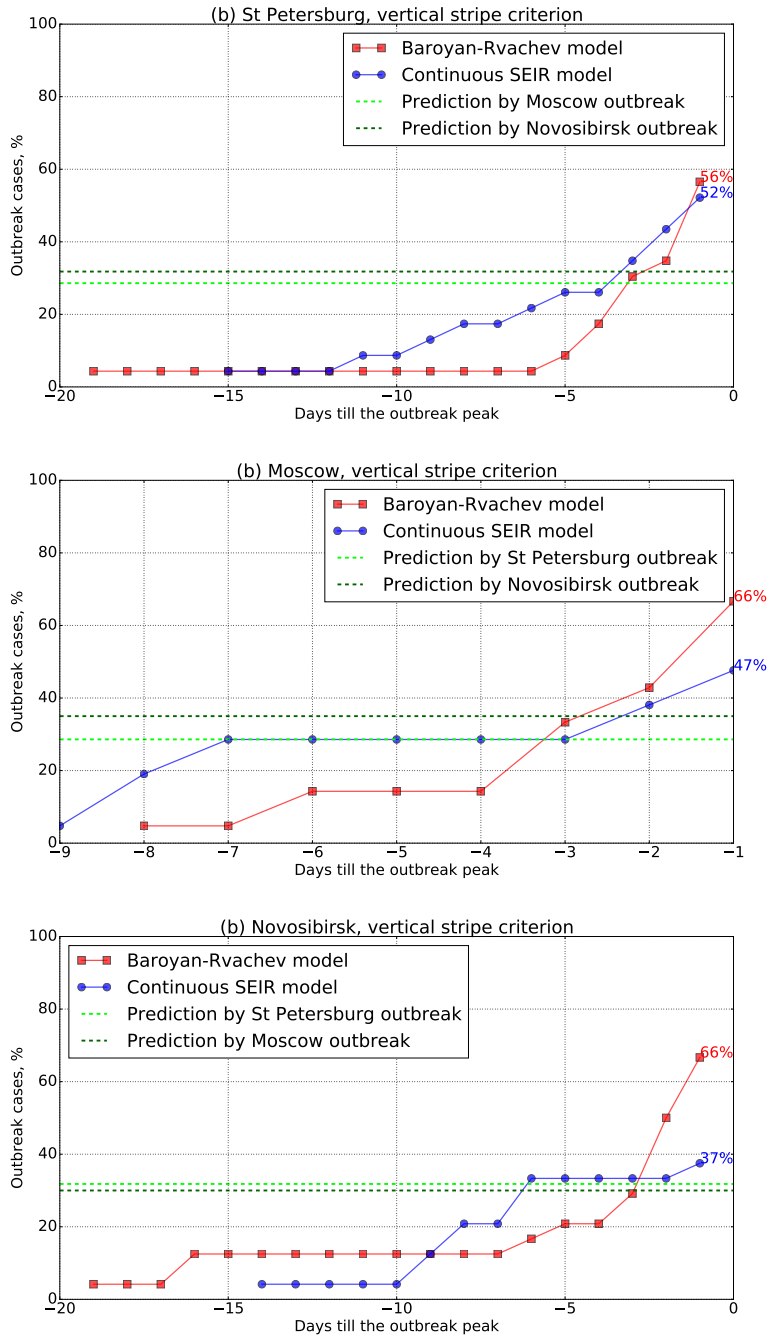
Figure 8. The dependence of the percentage of accurate predictions from the number of days before the peak, vertical stripe
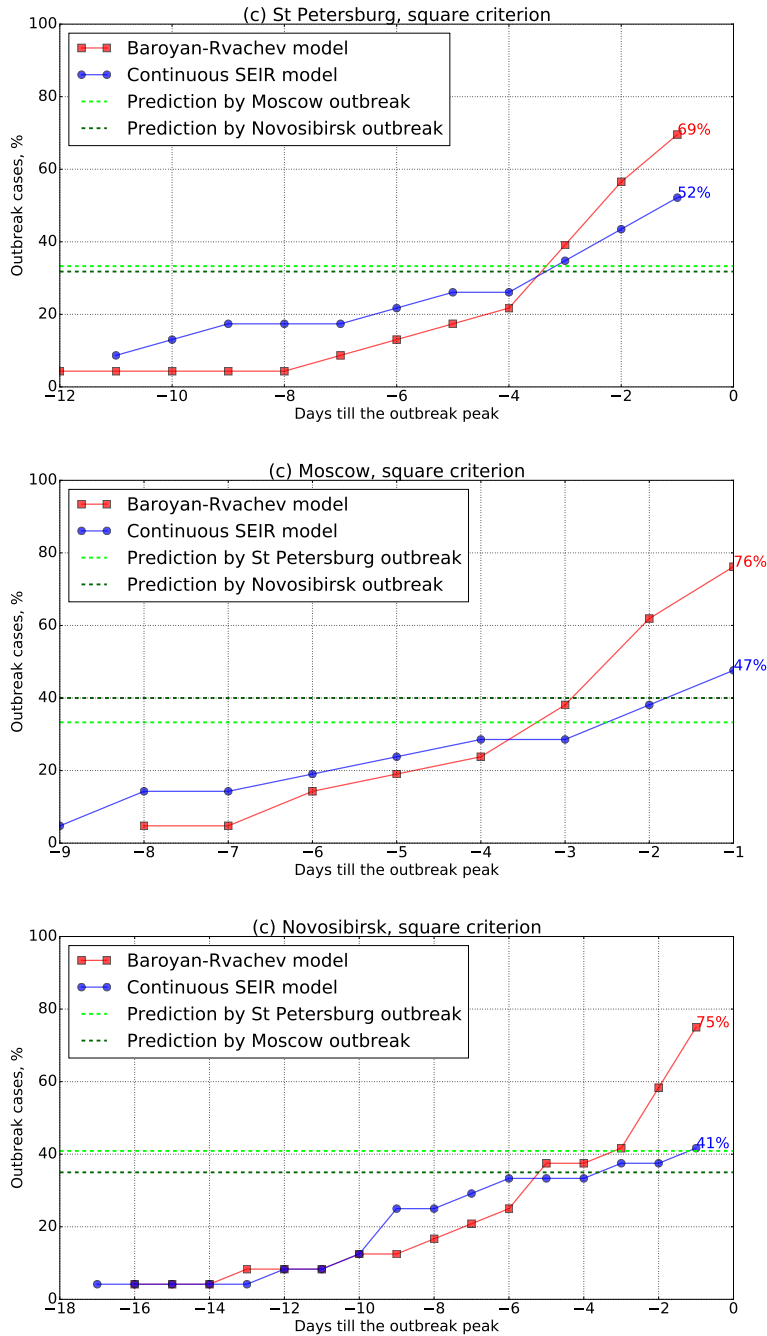
FIGURE 9. The dependence of the percentage of accurate predictions from the number of days before the peak, square
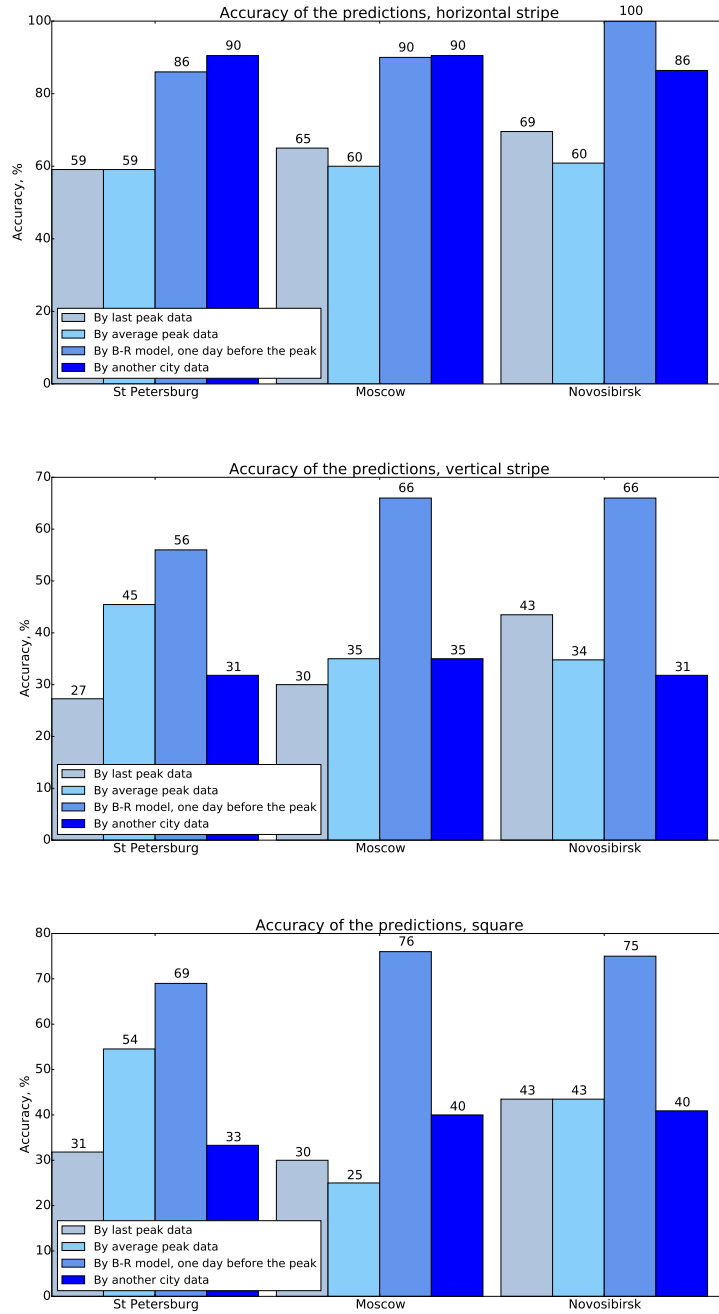
FIGURE 10. The comparison of the statistical and modeling fore-casting methods

REFERENCES

[1] O. Baroyan, U. Basilevsky, V. Ermakov, K. Frank, L. Rvachev and V. Shashkov, Computer modelling of influenza epidemics for large-scale systems of cities and territories, in *Proc. WHO Symposium on Quantitative Epidemiology, Moscow*, 1970.

[2] R. Burger, G. Chowell, P. Mulet and L. Villada, Modelling the spatial-temporal progression of the 2009 A/H1N1 influenza pandemic in Chile, *Mathematical Biosciences and Engineering*, **13** (2016), 43–65.

[3] CDC, Influenza signs and symptoms and the role of laboratory diagnostics, [online], http://www.cdc.gov/flu/professionals/diagnosis/labrolesprocedures.htm.

[4] CDC, People with heart disease and those who have had a stroke are at high risk of developing complications from influenza (the flu), [online], http://www.cdc.gov/flu/heartdisease/.

[5] J.-P. Chretien, D. George, J. Shaman, R. A. Chitale and F. E. McKenzie, Influenza forecasting in human populations: A scoping review, *PloS one*, **9** (2014), e94130.

[6] A. D. Cliff, P. Haggett and J. K. Ord, *Spatial Aspects of Influenza Epidemics*, Routledge, 1986.

[7] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron and A. Vespignani, Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions, *PLoS Med*, **4** (2007), e13.

[8] S. Cook, C. Conrad, A. L. Fowlkes and M. H. Mohebbi, Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic, *PLoS one*, **6** (2011), e23610.

[9] N. Goeyvaerts, L. Willem, K. Van Kerckhove, Y. Vandendijck, G. Hanquet, P. Beutels and N. Hens, Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence, *Epidemics*, **13** (2015), p1.

[10] I. Hall, R. Gani, H. Hughes and S. Leach, Real-time epidemic forecasting for pandemic influenza, *Epidemiology and Infection*, **135** (2007), 372–385.

[11] D. He, J. Dushoff, R. Eftimie and D. J. Earn, Patterns of spread of influenza A in Canada, *Proceedings of the Royal Society of London B: Biological Sciences*, **280** (2013), 20131174.

[12] K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande and S. Y. Del Valle, Forecasting the 2013–2014 influenza season using wikipedia, *PLoS Comput Biol*, **11** (2015), e1004239.

[13] A. Hyder, D. L. Buckeridge and B. Leung, Predictive validation of an influenza spread model, *PloS one*, **8** (2013), e65459.

[14] F. Institute, Research Institute of Influenza website, [online], http://influenza.spb.ru/en/.

[15] Y. G. Ivannikov and A. T. Ismagulov, *Epidemiologiya Grippa (The Epidemiology of Influenza)*, Almaty, Kazakhstan, 1983, In Russian.

[16] Y. Ivannikov and P. Ogarkov, An experience of mathematical computing forecasting of the influenza epidemics for big territory, *Journal of Infectology*, **4** (2012), 101–106, In Russian.

[17] V. N. Leonenko and S. V. Ivanov, Fitting the SEIR model of seasonal influenza outbreak to the incidence data for Russian cities, *Russian Journal of Numerical Analysis and Mathematical Modelling*, **31** (2016), 267–279.

[18] V. N. Leonenko, S. V. Ivanov and Y. K. Novoselova., A computational approach to investigate patterns of acute respiratory illness dynamics in the regions with distinct seasonal climate transitions, *Procedia Computer Science*, **80** (2016), 2402–2412.

[19] V. N. Leonenko, Y. K. Novoselova and K. M. Ong, Influenza outbreaks forecasting in Russian cities: Is Baroyan-Rvachev approach still applicable?, *Procedia Computer Science*, **101** (2016), 282–291.

[20] V. N. Leonenko, N. V. Pertsev and M. Artzrouni, Using high performance algorithms for the hybrid simulation of disease dynamics on CPU and GPU, *Procedia Computer Science*, **51** (2015), 150–159.

[21] D. C. Liu and J. Nocedal, On the limited memory bfgs method for large scale optimization, *Mathematical programming*, **45** (1989), 503–528.

[22] A. Romanyukha, T. Sannikova and I. Drynov, The origin of acute respiratory epidemics, *Herald of the Russian Academy of Sciences*, **81** (2011), 31–34.

[23] L. A. Rvachev and I. M. Longini, A mathematical model for the global spread of influenza, *Mathematical Biosciences*, **75** (1985), 1–22.

[24] J. Shaman, V. E. Pitzer, C. Viboud, B. T. Grenfell and M. Lipsitch, Absolute humidity and the seasonal onset of influenza in the continental United States, *PLoS Biol*, **8** (2010), e1000316.

[25] J. Tamerius, M. I. Nelson, S. Z. Zhou, C. Viboud, M. A. Miller and W. J. Alonso, Global influenza seasonality: Reconciling patterns across temperate and tropical regions, *Environmental Health Perspectives*, **119** (2011), 439–445.

[26] J. Truscott, C. Fraser, S. Cauchemez, A. Meeyai, W. Hinsley, C. A. Donnelly, A. Ghani and N. Ferguson, Essential epidemiological mechanisms underpinning the transmission dynamics of seasonal influenza, *Journal of The Royal Society Interface*, **9** (2011), 304–312.

[27] S. P. van Noort, R. Águas, S. Ballesteros and M. G. M. Gomes, The role of weather on the relation between influenza and influenza-like illness, *Journal of Theoretical Biology*, **298** (2012), 131–137.

[28] C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller and B. T. Grenfell, Synchrony, waves, and spatial hierarchies in the spread of influenza, *Science*, **312** (2006), 447–451.

[29] WHO, Influenza (seasonal). Fact sheet No. 211, March 2014., [online], `http://www.who.int/mediacentre/factsheets/fs211/en/`.

[30] WHO, Surveillance case definitions for ILI and SARI, [online], `http://www.who.int/influenza/surveillance_monitoring/ili_sari_surveillance_case_definition/en/`.

[31] R. Yaari, G. Katriel, A. Huppert, J. Axelsen and L. Stone, Modelling seasonal influenza: The role of weather and punctuated antigenic drift, *Journal of The Royal Society Interface*, **10** (2013), 20130298.

[32] W. Yang, B. J. Cowling, E. H. Lau and J. Shaman, Forecasting influenza epidemics in hong kong, *PLoS Comput Biol*, **11** (2015), e1004383.

E-mail address: `vnleonenko@yandex.ru`
E-mail address: `sergei.v.ivanov@gmail.com`