pp. 1–17 (2016)

Fitting the SEIR model of seasonal influenza outbreak to the incidence data for Russian cities

V. N. Leonenko* and S. V. Ivanov*

Abstract — In this paper we present a computational algorithm aimed at fitting a SEIR populational model to the influenza outbreaks incidence in Russian cities. The input data are derived from the long-term records on the incidence of acute respiratory diseases in Moscow, St. Petersburg, and Novosibirsk. It is shown that the classical SEIR model could provide a satisfactory fit for the majority of employed influenza outbreak incidence data sets ($R^2 > 0.91$ for the 64 curves out of 67). Nevertheless, the model fitting algorithm in its current implementation has a number of drawbacks, which are discussed in the paper along with the ways of overcoming them.

Keywords: Mathematical modelling, epidemiology, influenza, model fitting.

MSC 2010: 37N25, 65C20, 92C60

Seasonal acute respiratory infections (ARI) are among the oldest and the most common human infectious diseases. The most dangerous of these infections is influenza, which causes repetitive outbreaks both in temperate and tropical regions resulting in high worker/school absenteeism and productivity losses. The outbreaks of influenza result in 3 to 5 million cases of severe illness annually worldwide, and the mortality is from 250 to 500 thousand individuals [23]. Unlike influenza, most of the ARIs affect the human population throughout the year and do not cause distinct epidemic outbreaks. At the same time, the symptoms of severe cases of ARI are very similar to those of influenza, and the laboratory analysis is required to tell one disease from another. That is why in most of the healthcare surveillance systems, including the Russian one, the statistics on ARI and influenza incidence is calculated cumulatively as ‘ARI incidence’.

The problem of foremost importance connected with the research on influenza is to predict the epidemic outbreak dynamics, which could facilitate the fight with the disease and its possible negative consequences (such as increases of heart attacks and strokes [5]). Although the seasonality of influenza outbreaks is widely known, its mechanism still does not have satisfactory explanation. A lot of factors are named that may influence the starting moment of the outbreak and its dynamics over time, but the extent of influence of each factor on the outbreak parameters is arguable. In the contemporary works on the topic, most authors incorporate into their models

*National Research University of Information Technologies, Mechanics, and Optics, St. Petersburg 197101, Russia

This work was financially supported by Ministry of Education and Science of the Russian Federation, Agreement No. 14.584.21.0015 (11.11.2015), project ID: RFMEFI58415X0015

the dependence between the temperature (humidity) and the force of infection [8, 21, 24]. During the last years the social reasons that could increase or decrease the infection spread were also taken into consideration [16].

In one of the pioneering papers on influenza propagation, written by Baroyan and Rvachev, the authors assumed and demonstrated with the help of simulations [3, 20], that the speed and direction of the propagation of influenza epidemics are connected mostly with the peculiarities of the contact patterns of individuals and the level of their immunity, whereas the weather factor does not play the leading role and thus could be left beyond scope. The Baroyan–Rvachev model, despite its simplicity, appeared capable of making plausible predictions of the influenza spread in the USSR in 1971–1980. Nevertheless, since 1980's the predictive abilities of the model were seriously undermined due to the growing herd immunity in urban populations and the consequent changes in the influenza A virus propagation patterns [10]. This fact, combined with the socio-economical instability in the USSR, followed by its collapse and the corresponding problems in the Soviet (later Russian) healthcare, stopped the employment of the predictive modelling approach.

As a part of our research, concerning the modelling of seasonal acute respiratory infections dynamics, we aim at finding out whether the dynamics of influenza outbreaks in the Russian cities could be satisfactorily described by a simple general populational model without considering the weather and behavioural effects and how this model could be possibly modified to better fit the epidemic data. This is considered as a first step towards the prediction of country-wide influenza epidemics in the same fashion that it was made by Baroyan and Rvachev.

1. The model

In order to describe the dynamics of influenza epidemic process, we have chosen a simple populational model based on classical Kermack–McKendrick formulation [1]. Since the flu has an incubation period and the individuals recovered from the illness acquire the immunity from the particular virus strain [23], the population of an urban area under consideration is represented by the 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 ratios of total population N:

S — ratio of susceptible individuals;

E — ratio of exposed individuals;

I — ratio of infectious individuals;

R — ratio of removed individuals.

Following [3, 20], we state that a certain ratio of population of every city under consideration is not vulnerable to flu — that includes the people with immunity gained from the previous infections and those who are not immune by themselves but are protected by the herd immunity. The ratio of the population which is vulnerable to flu infection is denoted by $\alpha \in (0;1)$.

The dynamics of the group sizes over time is specified by a system of ordinary differential equations:

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta SI\\ \frac{\mathrm{d}E}{\mathrm{d}t} &= \beta SI - \gamma E\\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \gamma E - \delta I\\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \delta I\end{aligned} \tag{1.1}$$

$$\begin{gathered}S(t_0) = S_0 \geqslant 0, \qquad E(t_0) = E_0 \geqslant 0, \qquad I(t_0) = I_0 \geqslant 0\\ S_0 + E_0 + I_0 = \alpha\\ R(t_0) = 1 - \alpha.\end{gathered} \tag{1.2}$$

The term βSI corresponds to the process of infection of susceptible individuals. The term γE corresponds to the process of acquiring of infectivity by the exposed individuals. The term δI describes the recovery process of the infectives. We consider the intensity coefficients β, γ, and δ non-negative. Since the duration of the epidemic process is relatively short, we assume the influence of birth and migration processes on the disease dynamics negligible and do not include these processes into the model.

Further without loss of generality we assume $t_0 = 0$.

2. Outbreak incidence data

The original data-set provided by the Research Institute of Influenza [19] contains weekly ARI incidence (including flu) in three Russian cities from 1986 to 2014. Before we start the model fitting, we have to refine the incidence data by restoring the missed values and fixing under-reporting. Also we need to extract the flu incidence from the cumulative ARI incidence data. The corresponding algorithms are described in detail in [14], here we introduce briefly the sequence of operations.

- Under-reporting correction. Since during the holidays the infected people avoid visiting healthcare facilities, the corresponding weekly incidence is lower than the actual number of newly infected. This under-reporting bias, along with the missing data, could be corrected by means of cubic interpolation [3].

- Bringing the incidence data to daily format. The daily incidence is estimated by the cubic interpolation of the weekly incidence, assuming that $n_{\mathrm{inf}}^{\mathrm{Thu}} =$

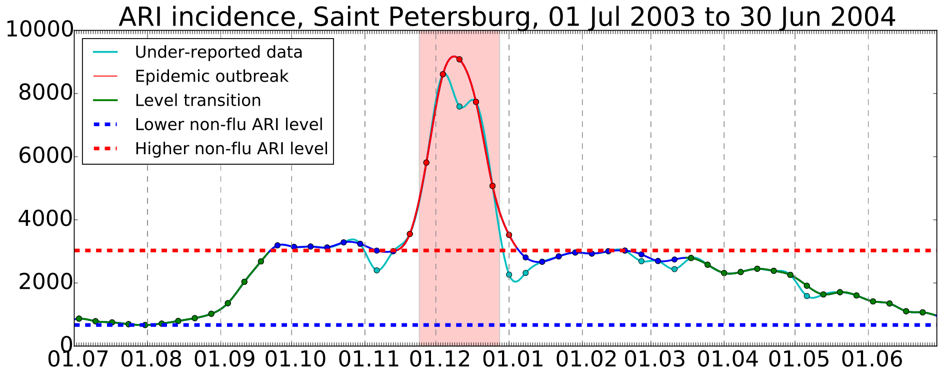

Figure 1. Typical ARI incidence curve.

$n^{W}_{\text{inf}}/7$, where $n^{\text{W}}_{\text{inf}}$ is weekly incidence taken from the database and $n^{\text{Thu}}_{\text{inf}}$ is the daily incidence for Thursday of the corresponding week.

- Extracting the 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 the higher non-flu ARI incidence level a_2, which corresponds to the average number of newly infected in non-epidemic period (Fig. 1, red horizontal dashed line). The ARI epidemic curves, which are recognized as flu outbreaks (Fig. 1, red solid line), should have their peaks well above the higher ARI level. Also they should comply with the time period during which the ARI prevalence exceeds the non-epidemic ARI threshold assessed by the Flu Research Institute (Fig. 1, red rectangle). The beginning and ending of the extracted curve are chosen to match the level a_2.

3. The fitting algorithm

3.1. Description of fitting parameters

The list of parameters involved in the fitting procedure (see Table 1) includes five epidemiological parameters, α, β, γ, δ, and I_0, from model (1.1)–(1.2), and two auxiliary parameters, Δ and k_{inc}, corresponding to horizontal and vertical positioning of the modelled incidence curve relatively to the epidemic data points. The necessity of the latter arises from the fact that the incidence curve in some cases could have a baseline below the level a_2 (see Fig. 2).

The fitting procedure is based on several simplifying assumptions:

- We assume that the epidemic starts with the appearance of a small number of infected individuals in the population $I(0) = \text{const}$, whereas $E(0) = 0$ (thus $S(0) = \alpha - I(0)$). Since the epidemic outbreak could start literally from one individual (that is confirmed by modelling experiments [3]), for the sake of model fitting the value I_0 is assumed to be (somewhat arbitrarily) a small fixed number.

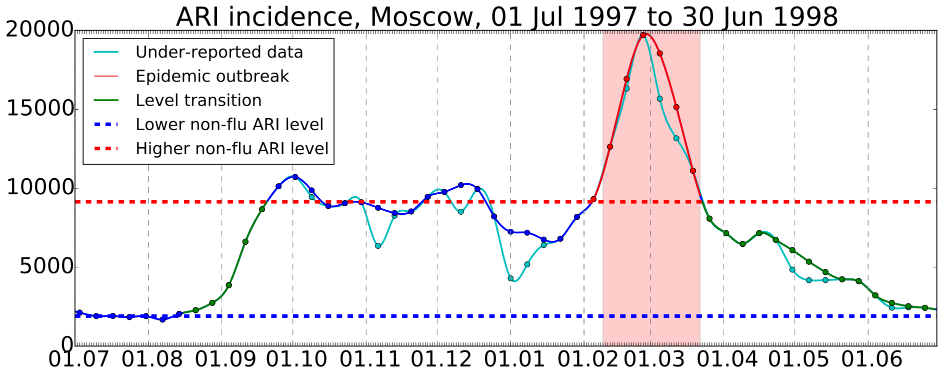

Figure 2. ARI incidence curve showing the discrepancy between the outbreak curve baseline and higher ARI level.

Table 1.
Parameters for model fitting.

Definition	Description	Value	Unit
	Epidemiological parameters		
α	Initial ratio of susceptible individuals in the population	Estimated	—*
β	Intensity of infection	Estimated	1/(person · day)
γ	Intensity of transition to infective form of the disease	Varied	1/day
δ	Intensity of recovery	Varied	1/day
I_0	Initial ratio of infected	0.0001	—*
	Curve positioning parameters		
k_{inc}	Relative vertical bias of the modelled incidence curve position	$[0.8; 1.0]$	—*
Δ	Absolute horizontal bias of the modelled incidence curve position	$5, \ldots, 54$	day

* dimensionless

- We suppose that in the period of the influenza outbreak the disease incidence due to acute respiratory infections of non-flu nature remains stable and corresponds to the higher ARI level registered before the epidemic outbreak (though in the general case it may not be true).

- Since we do not have any *a priori* information on the distribution of bias for the incidence data, we assume that the bias is independent in each point and normally distributed, which makes it possible to apply the least squares method to fit the model curve to data, following [3, 9].

It is worth noting that the values of γ and δ represent the general features of disease progression and could be considered independent of the epidemic season and the city under study (although the differences between these parameter values exist for different influenza A strains, for the purpose of our research this difference could be considered negligible). At the same time, the values of α and β are to be estimated independently for each outbreak case.

3.2. Algorithm structure

Let $Y^{(\mathrm{dat})}$ 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, which equals the observed duration of the outbreak.

The limited-memory BFGS optimization method is used to find the best fit [15]. For each value of Δ the algorithm varies the values of parameters $\alpha, \beta, \gamma, \delta, k_{\mathrm{inc}}$ to get the model output, which minimizes the distance between the modelled and real incidence points:

$$F(Y^{(\mathrm{mod})}, Y^{(\mathrm{dat})}) = \sum_{i=1}^{n} (y_i^{(\mathrm{mod})} - y_i^{(\mathrm{dat})})^2.$$

Here $y_i^{(\mathrm{dat})}$ and $y_i^{(\mathrm{mod})}$ represent the absolute flu incidence on the ith day taken from the input data-set and derived from the model, respectively.

Since the existence of several local minima is possible, the algorithm has to be started several times with different initial values of input variables. The best fit is chosen as a minimum of distances from all the algorithm runs.

The algorithm operations are performed in the following order. For each $\Delta \in 5, \ldots, 54$:

- For each fixed combination of values $\{\alpha, \beta, \gamma, \delta, k_{\mathrm{inc}}\}$ generated by BFGS optimization procedure:

  1. Find the numerical solution of model (1.1)–(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 modelling flu incidence in relative numbers: $y^{(\mathrm{mod,rel})}(t) = N_{E \to I}(t)$. Since from (1.1)–(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 obtain:

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

3. As we work with the 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 modelled 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:

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

4. We assume that the data incidence points from the data-set are shifted by Δ days from the model curve start. Thus, we are to compare the distance between the following data-sets:

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

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

$$y_i^{(\text{mod})} = y_i^{(\text{mod,rel})} \cdot N_L(m) \tag{3.1}$$

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

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

In the described manner for each value of Δ the BFGS algorithm finds the least distance F_Δ. We define $\Delta_{\min}$: $F(\Delta_{\min}, \dots) = \min F(\Delta, \dots)$, and the parameter set $\{\alpha, \beta, \gamma, \delta, k_{\text{inc}}\}$, corresponding to $\Delta_{\min}$. 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 (3.1).

4. Numerical experiments

The algorithm was implemented in a form of a set of scripts written in Python programming language (Python 3.x with `numpy` and `matplotlib` libraries was used). The higher ARI level was estimated with the help of `scipy.optimize.curve_fit` procedure and the limited-memory BFGS optimization method for curve fitting was performed via `scipy.optimize.minimize` routine.

Table 2.
Parameter point values and value ranges.

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Input data				
Cities num.	3			1
Seasons num.	28			1
Outbreak-dependent parameters				
α	$[10^{-2}; 1.0]$			
β	$[10^{-7}; 50.0]$			
Outbreak-independent parameters				
γ	$[10^{-7}; 50.0]$	$[0.2; 8.0]$	0.39	$[0.2; 8.0]$*
δ	$[10^{-7}; 50.0]$	$[0.08; 0.33]$	0.133	$[0.08; 0.33]$*
Goodness of fit				
Avg. R^2	0.977	0.966	0.946	0.983

* fixed values of vectors for each start of the fitting procedure

In order to test the algorithm we used the weekly ARI incidence data for three Russian cities (Moscow, St. Petersburg, and Novosibirsk) from July 1986 to June 2014. By means of epidemic curve allocation algorithm we have extracted the incidence data for the epidemic outbreaks, which gave us 67 epidemic outbreaks in total (there were no epidemics during some seasons). To characterize the goodness of fit we have chosen the coefficient of determination $R^2 \in (0,1]$ — the parameter widely used for fitting SIR models to data (for instance, see [22]). This coefficient shows the percentage of the response variable variation that is explained by a model. The closer to 1 the value of R^2 is, the better fit is supposed to be achieved. The aim of the numerical experiments was to estimate the variation of α and β corresponding to the fitted modelled incidence curve, making various assumptions on the values of γ and δ. The parameter point values and value ranges for each experiment are presented in Table 2. Value intervals are denoted by square brackets. Further details on experiment procedures are given below.

4.1. Experiment 1

The aim of the first experiment was to find the distribution of the values of outbreak-dependent parameters α and β in absence of strict limitations on the values of outbreak-independent parameters γ and δ and to check out whether the intervals for these values will be plausible from the epidemiological point (for the corresponding estimations see Appendix A). The obtained quality of fit is demonstrated in Fig. 4.

As seen from Fig. 3a, the values of γ and δ of the fitted model curves are spread along the horizontal axes, and the big values of γ as a rule correspond to the small values of δ. The corresponding interval values of α (not shown on the graph) and β almost coincide with their initial value intervals listed in Table 2 (from 0.011 to

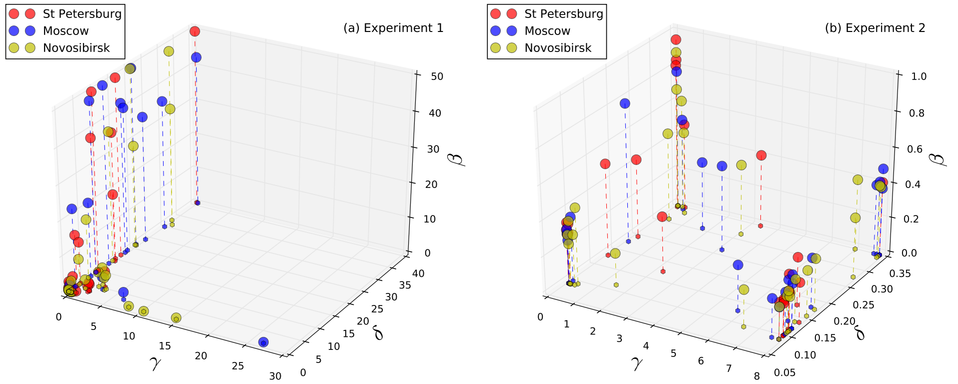

Figure 3. The joint distribution of γ, δ, and β: (a) Experiment 1, (b) Experiment 2. The big dots represent the values of parameter vectors (γ, δ, β), the small ones correspond to their projections onto the plane $\beta = 0$.

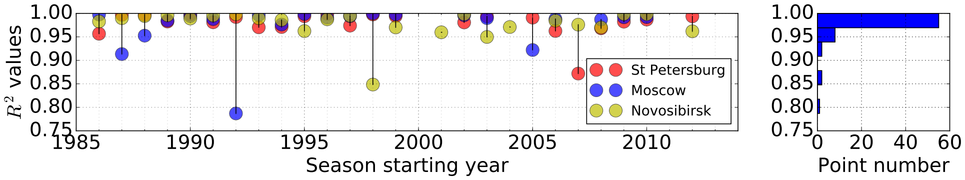

Figure 4. The distribution of R^2 values for Experiment 1.

0.972, or 97% of the initial interval width, for α; and from 0.107 to 50, i.e. 99,8% of the initial interval width, for β). Thus we can conclude that if we take arbitrarily broad value intervals for the parameters, the model curves with the best fit to data could have unrealistic parameter values.

4.2. Experiment 2

The aim of the second experiment is to find out whether the values of α and β will be realistic in case if we narrow the limits on γ and δ. The limitations on the latter parameters are put according to the estimations from the epidemiological data performed in Appendix A. As seen from Fig. 3b, the joint distribution of optimal γ and δ shows the strong influence of initial constraints on the optimization procedure: there are many points corresponding to minimum or maximum of the possible values. At the same time, due to these constraints the distribution of α and β (Fig. 5a) conforms to the estimated plausible intervals. Figure 6 demonstrates that the quality of fit in this experiment has declined slightly compared to Experiment 1.

4.3. Experiment 3

Reflecting upon the results obtained in the previous experiments, we have assumed that the number of free parameters in the model is too big, thus the fairly good fit could be achieved with various combinations of parameter values, including those

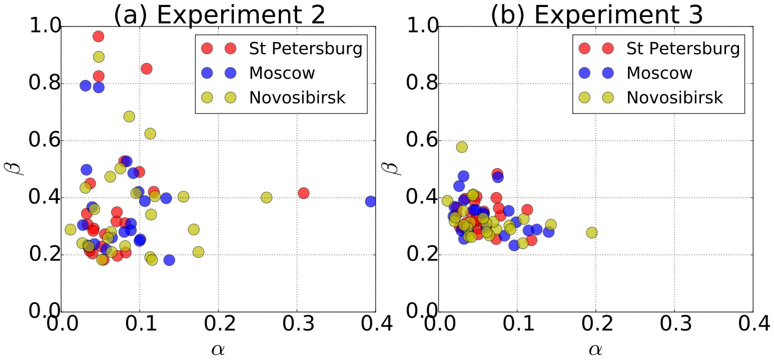

Figure 5. The distribution of α and β: (a) Experiment 2; (b) Experiment 3.

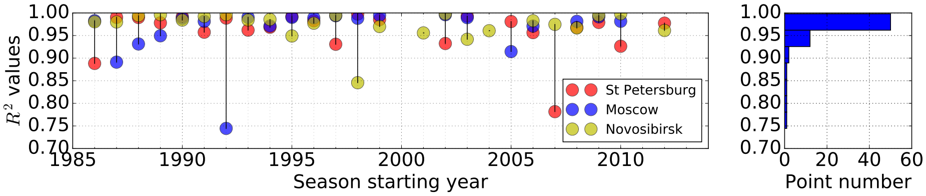

Figure 6. The distribution of R^2 values for Experiment 2.

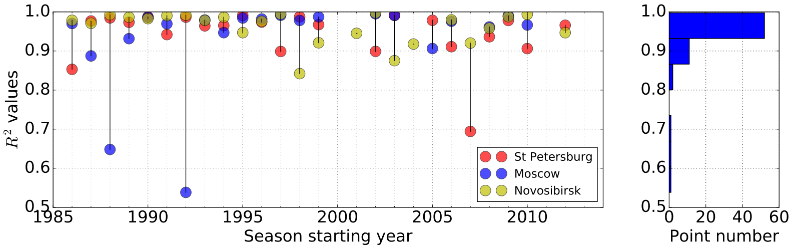

Figure 7. The distribution of R^2 values for Experiment 3.

that do not have epidemiological sense. To resolve that issue we have decided to reduce the number of varied parameters employed in the model fitting procedure. To achieve that goal we have fixed the values of γ and δ. The taken values $\gamma = \gamma^*$, $\delta = \delta^*$ correspond to their point estimations made on the basis of the epidemiological data (see Appendix A). The resulting distribution of the outbreak-dependent parameters is shown in Fig. 5b. The experiment shows that the version of our model with reduced state space can still give the realistic output and satisfactory quality of fit (see Fig. 7).

4.4. Experiment 4

Since the choice of values for γ and δ was made with the help of not very strict, though plausible, estimations, the question arises whether there could exist other

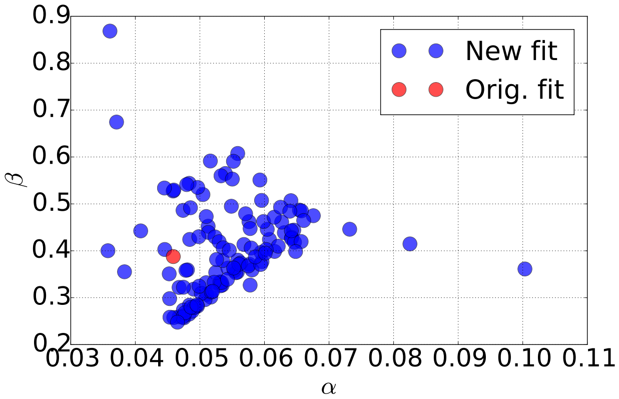

Figure 8. The joint distribution of the epidemiological parameters, Experiment 4. The red dot corresponds to the result of model fitting to the incidence curve with fixed $\gamma = \gamma^{*}$, $\delta = \delta^{*}$ from Experiment 3 (see Table 2).

pairs of realistic values for γ and δ which will give us the curve fit equal or better to the one we have obtained in the previous experiment. And if the answer is positive, what will be the variation of the corresponding outbreak-dependent values α and β for these new input parameter sets. To shed some light on that matter, we have conducted a new experiment, details of which are given below. Due to the computational limitations the experiment was performed for a sole case of epidemic outbreak: St. Petersburg, winter of 2003–2004 (see Fig. 1).

- We have assumed that the outbreak-independent parameters γ and δ were not defined by the fixed values, but by the random variables $\tilde{\gamma}$ and $\tilde{\delta}$. These variables had uniform distribution on the intervals that were used as constraints for γ and δ in Experiment 2 (see Table 2). The theoretical values of $\tilde{\gamma}$ and $\tilde{\delta}$, thus, were equal to 4.1 and 0.205, respectively, the relative standard deviations (i.e., standard deviations divided by means) were 0.549 and 0.352. A sample of 125 corresponding value vectors ($\tilde{\gamma}_i$, $\tilde{\delta}_i$) was generated via Monte Carlo simulation.

- For $i = 1, \dots, 125$ the model was fitted to incidence data of a given epidemic outbreak with fixed $\gamma = \tilde{\gamma}_i$, $\delta = \tilde{\delta}_i$ and varied α and β.

The means and relative standard deviations calculated for the input and output samples are presented in Table 3. The resulting joint distribution of α and β is shown in Fig. 8. As the table demonstrates, the relative standard deviation of both output values is less than the relative SD for the input. The picture shows that except several outliers the output value vectors are densely concentrated in the limited area. The form of the point cloud (which seems to be diagonally oriented) allows us to assume somewhat speculatively that there could be a slight dependence between the optimal fit values of β and α. It is worth noting that the similar fact (i.e., the dependence

Table 3.
Sample mean and relative SD of the output compared to input.

	Output		Input	
	α	β	γ	δ
Experiment 3				
Exact value	0.046	0.388	0.39	0.133
Experiment 4				
Mean	0.054	0.402	4.141	0.206
Relative SD	0.156	0.249	0.553	0.35

between the fraction of individuals without the immunity to a particular flu virus and the force of infection during the outbreak caused by this virus) was noted and extensively used in [3] for fitting the SIR model to flu incidence data in Soviet cites.

Figure 9 demonstrates the distribution of relative biases of output values from those calculated in the point $\gamma = \gamma^*$, $\delta = \delta^*$. It could be found that the areas of the same bias value tend to spread horizontally. So, despite the bigger variance, the variable γ has a lesser influence on the output values than δ. At the same time it is worth mentioning, that generally, as it was also shown in Fig. 8 and Table 2, the changing of neither γ nor δ does not affect sufficiently the optimal values of α and β.

The changes of R^2 registered during the Experiment 4 could be considered negligible: sample mean value is equal to 0.983 (which is slightly less compared to 0.993 obtained for the same incidence curve in the Experiment 3), relative SD is 2.6%.

Summing up, the results of this experiment bring us to preliminary conclusion that the form of the model incidence curve should be predominantly defined by the relation between α and β themselves rather than by γ or δ (provided that their values are taken from the plausible interval), but this hypothesis requires further justification.

5. Discussion

The distribution of R^2 for Experiments 1–3 shown in Figs. 4, 6, and 7 demonstrated that the quality of fit is satisfactory and declines slowly due to narrowing of the intervals for γ and δ. If we do not take into account three outliers, the low border for R^2 is 0.91 for Experiment 1, 0.89 for Experiment 2, and 0.84 for Experiment 3. The outliers correspond to the peculiar forms of the epidemic curves that cannot be reproduced by the ordinary SEIR model (see Fig. 10). These epidemic curves require further investigation, for they could demonstrate the examples of multistrain epidemics or the data biases due to the influence of external factors.

The consequent limitation of the model parameter space through Experiments 1–3 without considerable loss of fit quality demonstrated that the initial number of free parameters in the fitting algorithm seems to be too big. Even the model version with fixed values of γ and δ could be simplified even further. One can argue that one of

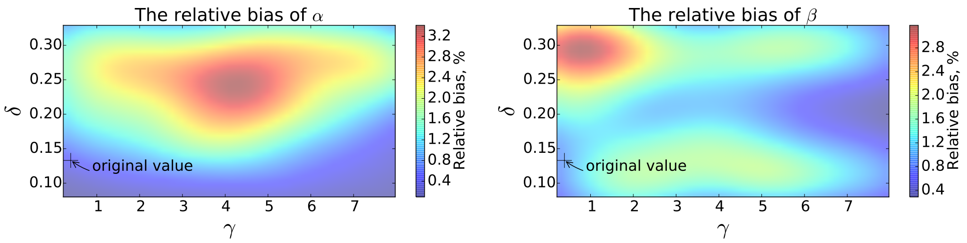

Figure 9. The relative biases of α and β from their values achieved in the point $\gamma = \gamma^*$, $\delta = \delta^*$ (marked by the cross).

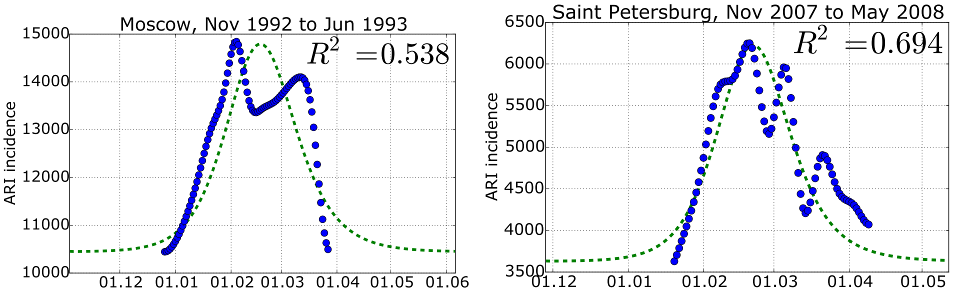

Figure 10. Examples of the epidemic curves corresponding to the cases of unsatisfactory model fitting (Experiment 3). Blue dots correspond to the incidence data, green dashed line represents the model curve.

the possible sources of the 'undesirable freedom' of the algorithm is the existence of the curve fitting parameters Δ and k_{inc} which theoretically could serve as additional correctors, making it possible to fit improper model curve to incidence data. It goes without saying that ideally the model fitting should rely on the precise information on the moment of epidemic start and the initial number of infected. However, the exact data cannot be obtained due to the absence of distinct diagnosis of influenza and other acute respiratory illnesses. Under these circumstances the use of fitting parameters Δ and k_{inc} allows the algorithm to estimate approximately the moment of epidemic start and the corresponding level of non-epidemic ARI incidence for the input incidence data with improper curve edges by relying on the expected regular form of the incidence curve. Small values of Δ and $k_{\text{inc}} \approx 1$ as a rule correspond to the 'clear' cases, when it is easy to distinguish the flu outbreak curve edges from the seasonal ARI level (see Fig. 1), whereas the incidence data with non-smooth edges (Fig. 2) is usually fitted by the curve with big Δ and small k_{inc}, indicating that the epidemic outbreak had started earlier than it was detected by the curve allocation algorithm. It is worth noting that if the task of the researcher is limited to the retrospective analysis of the epidemic curves (like it is in this paper), the moment of epidemic start probably may be estimated more accurately by employing the laboratory studies on dominant ARI virus strains (if they are at researcher's disposal, which was not our case). Unfortunately, the data of laboratory studies tend to become available with a considerable time lag comparing to reported ARI incidence, which makes them hardly applicable for the real-time outbreak dynamics prediction.

The drawback of this work which is to be mentioned is that we do not consider the bias in data gained as a result of conversion of the weekly incidence data to daily one. However, we hope to get rid of this issue after getting the real daily epidemic incidence data instead of generating synthetic ones. Despite the fact that the interpolated daily data is surely more ‘smooth’ than the real one, we presume that our set of algorithms will be still suitable to handle the real daily incidence data-set, as the daily epidemic curves demonstrated in [3], after filtering the fluctuations caused by the weekly cycle of individuals, resemble the synthetic data we work with.

An interesting problem which was not covered in this paper is to find the relation between the values of the outbreak-dependent parameters and the time and place of the outbreak (i.e. the city and the epidemic season). As seen from Figs. 3 and 5, for our epidemic incidence data-set there is no articulated difference in parameter distribution for different cities. The preliminary analysis of the yearly changes in the outbreak-dependent parameters (not included in this paper) did not give us any stable patterns as well. We hope to elaborate on that topic and possibly get some insights on the matter with the help of more broad data-set containing 49 cities instead of the limited one used in the presented research.

6. Conclusions and future works

In this paper we have presented the algorithm aimed at fitting the SEIR model to the incidence data for Russian cities, namely, Moscow, Saint Petersburg and Novosibirsk. The numerical experiments have demonstrated that our algorithm could provide a satisfactory fit for the 64 curves out of 67 influenza outbreak incidence data-sets we used ($R^2 > 0.84\ldots 0.91$, depending on the initial parameter values). The remained epidemic curves, which could not be plausibly described by the model, should be either handled by incorporating the additional factors into the model or left beyond the scope — in the latter case we will have to admit that the descriptive force of the model is considerably limited. Nevertheless, it is worth mentioning that even for these outliers the time moment of the epidemic peak and the maximum incidence value are estimated fairly well by the fitted model curves (see Fig. 10). This fact gives us a hope that despite the possible biases the flu dynamics in Russian cities still shows a general agreement with the output of Kermack–McKendrick epidemic models. The authors plan to elaborate more on that topic, particularly implementing the following improvements.

Modify the model to make it find plausible output values during the fitting procedure ‘in a natural way’, without the limitations on input variance. That could be achieved using additional *a priori* data related to the outbreak features for the fitting procedure, applying several criteria for the goodness of fit in addition to R^2 (e.g., the correctness of the prediction of epidemic peak and the length of epidemics), or by reducing the number of model parameters. Another idea worth trying consists in using a part of the incidence curve to fit the model instead of the full set of incidence points returned by the epidemic curve extraction algorithm. For instance, in [3] it was mentioned that the use of the first 'half-wave' of the incidence curve instead of the full curve helped the authors to increase the quality of flu incidence forecast. Also if we omit the incidence points at the beginning and the end of the curve, that could also improve the fitting quality, because at the stages of low flu incidence the epidemic statistics is badly affected by the existence of the parallel process of non-flu ARI dynamics [9].

Compare different minimization criteria for the 'distance' between the data points and the model curve. So far we have used the least squares method, mainly for the sake of simplicity and better algorithm performance, despite the absence of *a priori* information on the distribution law of bias in the incidence data. However, in fact it will be more correct to employ more general methods that do not rely on the normal distribution of the bias.

Create the parallel modification of the algorithm. The loss of algorithm performance due to higher computational expenses on the more sophisticated distance measurements could be compensated by shifting from serial to parallel execution of the algorithm. For instance, the iterations of the 'for' cycle on variable Δ could be launched independently in several threads — this approach was already employed for the sake of epidemiological modelling by one of the authors and helped to obtain a considerable speed up [13].

Acknowledgements

The authors are thankful to Kirill Stolyarov, Research Institute of Influenza, for providing us with Russian ARI incidence data and consultations on the nature of ARI and ILI seasonal dynamics, and the two anonymous referees who helped to improve significantly the quality of the paper.

References

1. R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control, (Vol. 28)*. Oxford University Press, 1992.
2. M. Artzrouni, V. N. Leonenko, and T. A. Mara, A syringe-sharing model for the spread of HIV: application to Omsk, Western Siberia. *Math. Med. Biol.* (2015), doi: 10.1093/imammb/dqv036.
3. O. V. Baroyan, L. A. Genchikov, L. A. Rvachev, and V. A. Shashkov, An attempt at large-scale influenza epidemic modelling by means of a computer. *Bull. Int. Epid. Assoc.* **18** (1969), 22–31.
4. M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir and L. Finelli, Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC Infect. Dis.* **14** (2014), No. 1, 1.
5. Centers for Disease Control and Prevention. People with Heart Disease and Those Who Have

Had a Stroke Are at High Risk of Developing Complications from Influenza (the Flu). URL: http://www.cdc.gov/flu/heartdisease/

6. Centers for Disease Control and Prevention. National Early Season Flu Vaccination Coverage, United States, November 2015. URL: http://www.cdc.gov/flu/fluvaxview/nifs-estimates-nov2015.htm

7. Centers for Disease Control and Prevention. Flu Vaccine Nearly 60 Percent Effective. URL: http://www.cdc.gov/media/releases/2016/flu-vaccine-60-percent.html

8. D. He, J. Dushoff, R. Eftimie, and D. J. Earn, Patterns of spread of influenza A in Canada. *Proc. R. Soc. Lond. Biol.* **280** (2013) 1170. doi: 10.1098/rspb.2013.1174.

9. Yu. G. Ivannikov and A. T. Ismagulov, The epidemiology of influenza. Almaty, Kazakhstan, 1983 (in Russian).

10. Yu. G. Ivannikov and P. I. Ogarkov, An experience of mathematical computing forecasting of the influenza epidemics for big territory. *Zhurnal Infectologii* **4** (2012) No. 3, 101–106. (in Russian).

11. O. I. Kiselev, L. M. Tsybalova, and V. I. Pokrovskiy, *Influenza: Epidemiology, Diagnosis, Treatment, Prophylaxis*. Moscow, 2012 (in Russian).

12. S. Kumar, K. Piper, D. D. Galloway, J. L. Hadler, and J. J. Grefenstette, Is population structure sufficient to generate area-level inequalities in influenza rates? An examination using agent-based models. *BMC Public Health* **15** (2015), 947.

13. 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 Comput. Sci.* **51** (2015), 150–159.

14. V. N. Leonenko, S. V. Ivanov, and Yu. K. Novoselova, A Computational approach to investigate patterns of acute respiratory illness dynamics in the regions with distinct seasonal climate transitions. *Procedia Comput. Sci.* **80** (2016), 2402–2412.

15. D. Liu and J. Nocedal, On the limited memory BFGS method for large-scale optimization. *Math. Program.* **45** (1989), 503–528.

16. P. Manfredi and A. D'Onofrio, Modelling the interplay between human behaviour and the spread of infectious diseases. Springer-Verlag, New York, 2013.

17. G. I. Marchuk, *Mathematical Models in Immunology*. Moscow, Nauka, 1991 (in Russian).

18. I. G. Marinich, L. S. Karpova, and V. A. Kondratyev, Methodological recommendations for the operational analysis and forecasting of the epidemiological situation on influenza and ARI. Moscow, 2005 (in Russian.).

19. Research Institute of Influenza website. URL: http://influenza.spb.ru/en/.

20. L. A. Rvachev and I. M. Longini, A mathematical model for the global spread of influenza. *Math. Biosci.* **75** (1985) No. 1, 3–22.

21. J. Shaman and A. Karspeck, Forecasting seasonal outbreaks of influenza. *PNAS* **109** (2012) No. 50, 20425–20430.

22. S. P. Van Noort, R. Aguas, S. Ballesteros, and M. G. M. Gomes, The role of weather on the relation between influenza and influenza-like illness. *J. Theor. Biol.* **298** (2012), 131–137.

23. WHO. Influenza (seasonal). Fact sheet No. 211, March 2014. URL: http://www.who.int/mediacentre/factsheets/fs211/en/.

24. R. Yaari, G. Katriel, A. Huppert, J. B. Axelsen, and L. Stone, Modelling seasonal influenza: the role of weather and punctuated antigenic drift. *J. R. Soc. Interface* **10** (2013) No. 84, 20130298.

Appendix A. Estimation of the values for input parameters

Since there is still a lack of quantitative information on the herd immunity and the individual partial immunity gained as a result of previous epidemic outbreaks, the upper bound for α could be roughly obtained as $\alpha = 1 - \eta\mu$, where η is the ratio of vaccinated in the population and μ is the vaccine effectiveness. Unfortunately the authors failed to find the corresponding data for Russia. For the solely illustrative purposes we estimate the vaccine effectiveness and coverage based on the US data during the epidemic season of 2015–2016: $\eta = 0.399$ [6], $\mu = 0.6$ [7], thus $0 < \alpha \leqslant 0.77$.

The value intervals for β, γ, δ were derived from the epidemiological data on influenza in the same fashion as it was made by one of the authors in [2] for HIV propagation model. The details on that process follow.

- The variable γ, as an intensity of transition from E to I, is an inverse of the average time of staying in the group E, or of the average flu incubation period duration. Since various sources give an estimation for the flu incubation period of 0–2 days [12], several hours to 3 days [9, 11], 1–5 days [17], we can take period of 3 hours to 5 days as the plausible value interval. That gives us the value interval for γ from 0.2 to 8.0. The average flu incubation period according to our value interval will be approximately 2.56 days, which gives us the point estimate for γ equal to 0.39.
- Similarly, δ is an inverse of the average infectious period, which is said to be 4–5 days [11], 3–6 days [12], 5–7 days [9], or 8–12 days [17]. Assuming that the infectious period is from 3 to 12 days, we get the value interval [0.08; 0.33] for δ. The average flu infectious period will be 7.5 days, which gives us the point estimation for δ equal to 0.133.
- The coefficient of infection β is connected with the basic reproduction number [1] of the illness. For the model (1.1)–(1.2) the basic reproduction number has the form

$$R_0 = \frac{\beta}{\delta}.$$

  Relying on the assessments of the basic reproduction number for seasonal ([1.19;1.37]) and pandemic ([1.47;2.27]) influenza from [4], and using the value interval for δ estimated above, we obtain the estimate for the values of β from 0.095 to 0.75.