Influenza outbreaks forecasting in Russian cities: is Baroyan-Rvachev approach still applicable?

Vasiliy N. Leonenko¹, Yulia K. Novoselova², and Karen M. Ong³

¹ ITMO University, Saint Petersburg, Russian Federation
vnleonenko@yandex.ru

² ITMO University, Saint Petersburg, Russian Federation
juliaeternity@gmail.com

³ New York University School of Medicine, New York, the United States of America
karen.ong@med.nyu.edu

Abstract

In the current work we assess the accuracy of the outbreak peak prediction expected in a fixed Russian city with the help of the populational SEIR models fitted to the past local outbreaks happened earlier in the country during the same season. This approach was successfully used in USSR by Baroyan and Rvachev to predict flu outbreaks throughout 1970’s. However, in early 1980’s it became increasingly inaccurate and due to that the approach was not applied since. Our aim is to find out, using the long-term data on acute respiratory infections incidence for Moscow, Saint Petersburg and Novosibirsk as an input, to what extent this method of outbreak peak prediction is applicable nowadays, and discuss the methods of increasing its accuracy. We found that SEIR population modeling is still valid for prediction of peak heights, but not peak days.

Keywords: data analysis, mathematical epidemiology, acute respiratory infection, seasonal influenza, Python

1 Introduction and motivation

Acute respiratory infections (or, shortly, ARI s) are known as one of the oldest and the most popular human infectious diseases. The most notorious of them, influenza, has an ability to cause repetitive epidemic outbreaks with ARI incidence dramatically exceeding the average seasonal level. 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 [14]. Influenza also causes increase of heart attacks and strokes [4], as well as other disease complications. It’s important to mention that even during an epidemic outbreak only 15 to 20% of the total ARI cases are attributed to influenza [11], and the certain diagnosis whether a person is infected by influenza or by another acute respiratory infection with similar symptoms is possible only
through laboratory testing [3]. Due to that issues the common clinical diagnosis ‘influenza-like illness’ (ILI) is often used, which includes all severe ARI cases fitting a certain description (the term often used in Russian healthcare is ‘clinical influenza’). The criteria of ILI vary slightly in different national healthcare systems. According to WHO, ILI is an acute respiratory infection with measured fever of ≥ 38 C and cough with onset within the last 10 days [15].

In the late 1960’s, multiple mathematical models of influenza-like illness outbreaks were created. One of the most prominent studies of that time, accomplished by Baroyan and Rvachev, was connected with the flu propagation within the cities of Soviet Union [2] (later applied to worldwide propagation of the pandemic flu [12]). The Baroyan-Rvachev model was a combination of the Kermack-McKendrick SEIR model and a linear model of inter-city migration flows. Although the structure itself was not novel, it matched the true incidence data and achieved accurate forecasts of the outbreaks starts and peaks for one hundred Soviet cities included into the modeling system. For instance, the day of the outbreak start was predicted without mistakes 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 [7]. The accuracy of forecasts was achieved due to the efficient Soviet system of ARI cases registration [5] which gathered and collected reports on ARI incidence from the local healthcare units. The data for ILI outbreak model calibration was extracted from the obtained overall ARI and ILI incidence using statistical methods [7].

However, since early 1980’s the Soviet modeling complex for flu forecasting, created in Research Institute of Influenza [6] and based on Baroyan-Rvachev model, showed the signs of growing incoherence with the epidemic outbreak patterns observed in Soviet cities [8]. The reason for that, according to one of the versions, is in the growing levels of collective immunity to flu due to increasing speed of its circulation around the globe. The dynamics of growth of collective immunity could be dependent from different factors, including the structure of contact networks within an urban area, that’s why the original assumption that fraction of non-immune individuals is the same for all Soviet cities and depends only on currently circulating virus strain, seems to be less applicable [8]. According to [7], to maintain the original prediction accuracy, the values of parameters α and β (the fraction of non-immune individuals in the population and the force of infection correspondingly) was to be assessed separately for every city according to available a priori information on past outbreaks in this area. Dramatic events of the late 1980’s – early 1990’s, including the deaths of Baroyan and Rvachev and the collapse of the Soviet Union, adversely affected ARI surveillance and led to the cessation of use of the ARI forecasting system. The plans for testing and validating the model predictions came to a halt. For the past thirty years, the incidence data has not been fitted to the Baryoan-Rvachev model, so its predictive accuracy for modern data is unknown.

In this paper the authors are adapting to the modern data some of the ideas used in Baroyan-Rvachev forecast modeling complex for the sake of predicting the day and height of the flu outbreak peaks in Russian cities. Our aim was to assess the prediction accuracy of the named outbreak parameters and to find out to what extent the assumptions of Baroyan and Rvachev could be still applied for the sake of modeling the influenza propagation within Russia.

2 Research statement

The original idea of the Baroyan-Rvachev modeling procedure for a fixed flu epidemic season was based on the idea that influenza is spread between the Soviet cities via transport network,

\[1\] Although is is not completely accurate from the epidemiological point, for the sake of simplicity further in this study we consider ‘flu’, ‘influenza’ and ‘ILI’ synonyms.
i.e. infected individuals from the cities affected by the epidemics travel to the cities where there is no epidemics yet and cause outbreaks there. The forecasting algorithm had the following structure:

- Using the incidence data from local healthcare units, detect the city $A$ which was affected by flu epidemics first.
- Wait until there would be enough incidence data points to fit the model and find the values of model parameters $\alpha$, $\beta$.
- Using the data on transport flows, find the sequence of cities $B, C, \ldots$, where the epidemic outbreak will occur next, and the corresponding days of the outbreak start in these cities.
- Model the epidemics in these cities starting from the days detected in the previous stage using known values of $\alpha$ and $\beta$.
- Predict the day and height of the epidemic peaks in the cities using the modeling results.

Ideally the whole described algorithm should have been reconstructed and launched on new data. However, we are restricted by a number of limitations connected with data availability. The two main issues are described in detail further.

**Absence of migration flow data.** As it was mentioned earlier, the starting days of epidemic outbreaks are predicted in the original model using the intensities of migrations from the regions with ongoing epidemics. In Soviet times the flow data was assessed using the statistics on the number of sold airplane, train and bus tickets delivered by the Ministry of transport [2]. Now this data is unavailable in open sources. The official request of Research Institute of Influenza to Russian Ministry of transport was not fruitful either. Due to this circumstances we have decided to analyze the incidence data retrospectively, using the initial days of epidemic start assessed with the help of the epidemic outbreak curves extraction algorithm (see detailed explanation in the next section).

**Lower incidence data detalization.** In Soviet times the daily ARI incidence data was collected during the epidemic period for all the urban areas covered by the surveillance system. Also for the capitals of Soviet republics, along with Leningrad and Vladivostok [7], the daily data was available throughout the year, whereas the non-epidemic ARI dynamics for other urban areas was characterized by weekly incidence data. In modern times the openly available incidence data generated by the Research Institute for Influenza is given with the time step equal to one week for all the Russian cities covered by surveillance, both for non-epidemic and epidemic periods. As far as employing the weekly incidence data leaves a very small chance of performing detailed and accurate forecast, and we still hope to obtain the daily data, which apparently exists but has restricted access, we have decided to use interpolation procedures to create ‘synthetic’ daily data from the weekly ones. Needless to say that the mentioned procedures affect badly the quality of data, especially the definition of the outbreak start.

### 3 Algorithm structure

#### 3.1 Outbreak incidence data

The original dataset provided by the Research Institute of Influenza [6] contains weekly ARI incidence (including flu) in three Russian cities from 1986 to 2014. Before we are able to start the model fitting, we have to refine the incidence data by restoring the missed values and
fixing the under-reporting. Also we need to extract the flu incidence from the cumulative ARI incidence data. The corresponding algorithms are described in detail in [10], here we introduce briefly the sequence of operations.

- Under-reporting correction. As far as during the holidays the infected people avoid visiting healthcare facilities, the corresponding weekly prevalence 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 [2].

- Bringing the incidence data to daily format. The daily incidence is assessed by the cubic interpolation of the weekly incidence, assuming that $n_{Thu}^{Thu} = n_{inf}^W / 7$, where $n_{inf}^W$ is weekly incidence taken from the database and $n_{inf}^{Thu}$ is the daily incidence for Thursday of the corresponding week.

![ARI incidence curve.](image)

- 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 (figure 1, red horizontal dashed line). The ARI epidemic curves, which are detected as flu outbreaks (figure 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 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.

3.2 The model

For the sake of describing the dynamics of influenza epidemic process, we have chosen a simple populational model based on classical Kermack-McKendrick formulation [1]. 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 is defined by the 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 $R$ — fraction of removed individuals.

Following [2], [12], we state that the certain percentage 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 collective immunity. The fraction of the population which is vulnerable to flu infection is denoted by $\alpha \in (0;1)$.

The dynamics of the groups’ quantities over time is set by a system of ordinary differential equations:

\[
\begin{align*}
\frac{dS}{dt} &= -\beta SI, \\
\frac{dE}{dt} &= \beta SI - \gamma E, \\
\frac{dI}{dt} &= \gamma E - \delta I, \\
\frac{dR}{dt} &= \delta I, \\
S(t_0) &= S_0 \geq 0, E(t_0) = E_0 \geq 0, I(t_0) = I_0 \geq 0, \\
S_0 + E_0 + I_0 &= \alpha, \\
R(t_0) &= 1 - \alpha.
\end{align*}
\]

The term $\beta SI$ corresponds to the process of infection of susceptible individuals. The term $\gamma E$ corresponds to the process of acquiring of infectivity by the exposed individuals. The term $\delta I$ reflects the recovery process of the infectives. We consider the intensity coefficients $\beta$, $\gamma$ and $\delta$ non-negative. As far as 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.

### 3.3 The fitting algorithm

The list of model parameters (table 3.3) includes five epidemiological parameters, $\alpha$, $\beta$, $\gamma$, $\delta$ and $I_0$, from the model 1–2, and two auxiliary parameters, $\Delta$ and $k_{inc}$, which are in charge of horizontal and vertical positioning of the modeled incidence curve relatively to the epidemic data points. It could be assumed that the outbreak always starts from small fixed number of infected persons which let us assign some arbitrary fixed value to $I_0$. The values of $\gamma$ and $\delta$ reflect the properties of the individual disease advance, thus they could be considered outbreak-independent and thus, like in the case of $I_0$, are not involved in the fitting procedure. Their values are derived using external sources [9].

The limited-memory BFGS optimization method is used to find the best fit. For the each value of $\Delta$ the algorithm varies the values of parameters $\alpha$, $\beta$, $k_{inc}$ to achieve the model output, which minimizes the euclidean distance between the modeled and real incidence points. 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 algorithm was implemented in a form of a collection of scripts 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 and the limited-memory BFGS optimization method for curve fitting was performed via scipy.optimize.minimize routine.
### Table 1: Parameters for model fitting

| Definition            | Description                                                                 | Value  | Unit   |
|-----------------------|-----------------------------------------------------------------------------|--------|--------|
| **Epidemiological parameters** |                                                                             |        |        |
| $\alpha$              | Fraction of susceptible individuals in the population                       | Estimated | -*     |
| $\beta$               | Intensity of infection                                                      | Estimated | 1/person · day |
| $\gamma$              | Intensity of transition to infective form of the disease                    | 4.1    | 1/day  |
| $\delta$              | Intensity of recovery                                                       | 0.205  | 1/day  |
| $I_0$                 | Initial fraction of infected                                                | 0.0001 | -*     |
| **Curve positioning parameters** |                                                                              |        |        |
| $k_{inc}$             | A relative vertical bias of the modeled incidence curve position          | [0.8; 1.0] | -*     |
| $\Delta$              | An absolute horizontal bias of the modeled incidence curve position        | 5, ..., 54 | day    |

* dimensionless

### 4 Numerical experiments

Our aim was to assess the abilities of the model to predict the epidemic peaks using the weekly ARI incidence data for three Russian cities (Moscow, Saint-Petersburg and Novosibirsk) from July 1986 to June 2014. The mentioned cities have the biggest population sizes in Russia. Being also the important industrial and economical centers of the country, they are highly interconnected and thus could directly transmit infection among themselves. The experiment was performed in the following manner:

![Peak prediction for Moscow, 1993-1994](image)

Figure 2: An example of peak prediction of the flu outbreak in Moscow (blue dots indicate the interpolated ARI incidence curve) using the models fitted for the data of St Petersburg (red line) and Novosibirsk (yellow line).

- 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).

- For every extracted epidemic curve (which corresponds to the outbreak in one of the three cities in a fixed epidemic season) we have performed the model fitting and assessed the values of $\alpha$, $\beta$, $k_{inc}$ and $\Delta$.

- Consider that we have city where we want to predict the day and height of the outbreak peak – for example, suppose it is Moscow. Let us assume that in the season under consideration Saint Petersburg and/or Novosibirsk have already experienced the outbreak and we have found two pairs of parameter values obtained by model incidence curve calibration: $(\alpha_{spb}, \beta_{spb})$ and $(\alpha_{nsk}, \beta_{nsk})$. At the same time, the epidemic outbreak in Moscow has just started and we know only several incidence points, including the starting one (which gives us the moment of epidemic start in Moscow). In this case we cannot find directly the parameter values for Moscow, i.e. $(\alpha_{msk}, \beta_{msk})$. But according to Baroyan and Rvachev (see section 2), $\alpha_{msk} \approx \alpha_{spb} \approx \alpha_{nsk}$ and $\beta_{msk} \approx \beta_{spb} \approx \beta_{nsk}$. Hence we can use the pair of parameter values from Saint Petersburg and/or Novosibirsk (depending on where the outbreak has started earlier and hence the incidence data is more complete) to build the model curves and obtain the approximate forecasts in absence of incidence data from Moscow (see fig. 3 — the curve named ‘Moscow’ shows the retrospective model fit for incidence data of Moscow, i.e. the curve built with the exact values of $(\alpha_{msk}, \beta_{msk})$).

- Repeat the previous procedure for every epidemic curve and measure two values: the prediction bias of the peak day $dt$ and the ratio between the modeled and the real outbreak peak heights $dh$.

To assess the accuracy of prediction results, we have used the criteria employed in 1970’s for the Soviet flu outbreak prediction complex [7], three in total, keeping their original designations:

- ‘Square’. The prediction is thought to be accurate if $dt \in -8..8$ and $dh \in (0.5;2.0)$.

- ‘Vertical stripe’. Accurate prediction should have $dt \in -7..7$.

- ‘Horizontal stripe’ Accurate prediction should have $dh \in (0.7;1.5)$.

For comparison purposes we have included the prediction accuracies received in 1970’s with the help of prediction complex: particular point assessments for the three cities regarded in this paper (gray dots in the figure 3 and the corresponding statistics in table 2) and the cumulative accuracy statistics for all the Soviet cities included in the system (tables 2 and 3)

| Table 2: Prediction accuracy according to ‘square’ criterion |
|---------------------------------------------------------------|
|                                | Saint Petersburg | Moscow | Novosibirsk |
|--------------------------------|------------------|--------|-------------|
| Saint Petersburg data          | 33.3%            |        | 40.9%       |
| Moscow data                    | 33.3%            | -      | 35%         |
| Novosibirsk data               | 31.8%            | 40.0%  | -           |
| Soviet model                   | 100%             | 100%   | 75%         |
| Soviet model, all cities       | 80%              |        |             |

The following conclusions can be made from the obtained experiment results.
Figure 3: Biases of the peak prediction for Saint Petersburg, Moscow and Novosibirsk compared to those of Soviet-period predictions. Dashed rectangle corresponds to the ‘square’ accuracy criterion, the ‘stripe’ criteria are shown by transparent fill.

Table 3: Prediction accuracy according to ‘stripe’ criteria

|                  | Saint Petersburg | Moscow | Novosibirsk | All cities, Soviet model |
|------------------|------------------|--------|-------------|--------------------------|
| Msk Nsk SPb Nsk SPb Msk SPb Msk | 'Vertical stripe' 28.6% 31.8% 28.6% 35.0% 31.8% 30.0% 87.4% | 'Horizontal stripe' 90.5% 77.2% 90.5% 75.0% 86.4% 75.0% 69.0% |
• As it can be seen, most of the point assessments for all three cities obtained by the presented algorithm do not comply neither with 'square' criteria nor with 'vertical stripe' criteria because of the big variation in $dt$ (the biases from the real peak day are up to several months), whereas all but one assessment made by Baroyan-Rvachev model in 1970’s are accurate enough to suit both. This result was somewhat predictable and, for authors’ opinion, predominantly caused by high level of uncertainty during the extraction of the edges of the epidemic curves by curve extraction algorithm and consequent mistakes in assessment of the first day of epidemics. The positive consequence of that effect is that by means of the value interval of epidemic peak biases we could judge the curve extraction algorithm accuracy and compare different versions of that algorithms. It is worth mentioning that relying on dates of epidemic outbreak start delivered by the specialists from Research Institute of Influenza rather than on our artificially found curve edges could have improved the quality of peak time prediction.

• The compliance with the criteria ‘horizontal stripe’ is surprisingly good: the percentage of predictions that suit the criteria is even higher than that of the Soviet-time predictions in overall (see table 3), although still lower than of the Soviet predictions for the three cities under study.

• When Novosibirsk data is used to predict the forthcoming outbreak peak in Moscow or Saint Petersburg, the results are less accurate than in the case when Moscow peak is predicted by Saint Petersburg data and vice versa. This observation conforms to the revealed fact that the degree of resemblance of the forms of epidemic curves in two cities is inversely proportional to their distance [13]. It is important to say that this is not the only possible explanation and the matter should be investigated further.

5 Conclusions

We have applied our algorithm for incidence analysis, epidemic curve extraction and the model fitting to assess the quality of peak predictions on the data of Moscow, Novosibirsk and Saint Petersburg. We have found, that due to the absence of migration flow model and the inaccuracies in selection of the first outbreak days the accuracy of peak day prediction is unsatisfactory. On the other hand, the accuracy of peak height has the level comparable to those achieved by Baroyan-Rvachev model in 1970’s (i.e. when it was still reliable). Our future plans is to improve the prediction accuracy by implementing following measures:

• Using incidence data instead of weekly ones
• Employing transport model
• In case if we fail to get the migration flows, improve the curve extraction algorithm for the sake of better assessing the outbreak starting days
• Using a priori data on past epidemics for the sake of correcting the peak forecasting

We also hope to elaborate on that matter further by using the incidence data from the extended database (49 cities in total).

The issue that was not regarded in this paper, which affected badly the adequacy of Baroyan-Rvachev modeling complex, is that at times the outbreaks simply did not occur in the cities where, according to the model, they should have happened [8]. This effect still has no certain
explanation and its presence undermines the whole idea of inter-city epidemic modeling, so we have a big interest for elaborating on that matter, presumably with the help of acquired migration data and calibrated transport models.

6 Acknowledgements

This paper is financially supported by the grant of Ministry of Education and Science of the Russian Federation Goszadanie No 2014/190 (3969).

References

[1] Roy M Anderson, Robert M May, and B Anderson. Infectious diseases of humans: dynamics and control, volume 28. Wiley Online Library, 1992.

[2] OV Baroyan, UV Basilevsky, VV Ermakov, KD Frank, LA Rvachev, and VA Shashkov. Computer modelling of influenza epidemics for large-scale systems of cities and territories. In Proc. WHO Symposium on Quantitative Epidemiology, Moscow, 1970.

[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] Andrew David Cliff, Peter Haggett, and John Keith Ord. Spatial aspects of influenza epidemics. Routledge, 1986.

[6] Flu Institute. Research Institute of Influenza website. [online]. http://influenza.spb.ru/en/.

[7] Yu. G. Ivannikov and A. T. Ismagulov. Epidemiologiya grippa (The epidemiology of influenza). Almaty, Kazakhstan, 1983. In Russian.

[8] Yu.G. Ivannikov and P.I. Ogarkov. An experience of mathematical computing forecasting of the influenza epidemics for big territory. Journal of Infectology, 4(3):101–106, 2012. In Russian.

[9] 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, 2016. In press.

[10] 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 Computer Science, 80:2402–2412, 2016.

[11] AA Romanyukha, TE Sannikova, and ID Drynov. The origin of acute respiratory epidemics. Herald of the Russian Academy of Sciences, 81(1):31–34, 2011.

[12] Leonid A Rvachev and Ira M Longini. A mathematical model for the global spread of influenza. Mathematical biosciences, 75(1):3–22, 1985.

[13] Cécile Viboud, Ottar N Bjørnstad, David L Smith, Lone Simonsen, Mark A Miller, and Bryan T Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. Science, 312(5772):447–451, 2006.

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

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