Statistics based predictions of coronavirus 2019-nCoV spreading in mainland China

Igor Nesteruk

Institute of Hydromechanics, National Academy of Sciences of Ukraine, Zheliabova 8/4, UA-03680 Kyiv, Ukraine

National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Prosp.Peremohy 37, UA-03056, Kyiv, Ukraine.
inesteruk@yahoo.com

ABSTRACT

Background. The epidemic outbreak cased by coronavirus 2019-nCoV is of great interest to researches because of the high rate of spread of the infection and the significant number of fatalities. A detailed scientific analysis of the phenomenon is yet to come, but the public is already interested in the questions of the duration of the epidemic, the expected number of patients and deaths. For long time predictions, the complicated mathematical models are necessary which need many efforts for unknown parameters identification and calculations. In this article, some preliminary estimates will be presented.

Objective. Since the reliable long time data are available only for mainland China, we will try to predict the epidemic characteristics only in this area. We will estimate some of the epidemic characteristics and present the most reliable dependences for victim numbers, infected and removed persons versus time.

Methods. In this study we use the known SIR model for the dynamics of an epidemic, the known exact solution of the linear equations and statistical approach developed before for investigation of the children disease, which occurred in Chernivtsi (Ukraine) in 1988-1989.

Results. The optimal values of the SIR model parameters were identified with the use of statistical approach. The numbers of infected, susceptible and removed persons versus time were predicted.

Conclusions. Simple mathematical model was used to predict the characteristics of the epidemic caused by coronavirus 2019-nCoV in mainland China. The further research should focus on
updating the predictions with the use of fresh data and using more complicated mathematical models.

**Key-words.** coronavirus epidemic in China, coronavirus 2019-nCoV, mathematical modeling of infection diseases, SIR-model, parameter identification, statistical methods.

**INTRODUCTION**

Here, we consider the development of epidemic outbreak cased by coronavirus 2019-nCoV (see e.g., [1-3]). Since the reliable long time data are available only for mainland China, we will try to predict the number of victims \( V \) of this virus only in this area. The first estimations of \( V \) exponential growth, typical for the initial stages of every epidemic (see e.g., [4]) have been done in [3]. For long time predictions, more complicated mathematical models are necessary. For example, a susceptible-exposed-infectious-recovered (SEIR) model was used in [2]. Nevertheless, the complicated models need more efforts for unknown parameters identification. This procedure may be especially difficult, if reliable data are limited.

In this study, we use the known SIR model for the dynamics of an epidemic [4-8] To the parameter identification, we will use the exact solution of the SIR set of linear equations and statistical approach developed in [4]. These methods were applied for investigation of the children disease, which occurred in Chernivtsi (Ukraine) in 1988-1989. We will estimate some of the epidemic characteristics and present the most reliable dependences for victim numbers, infected and removed persons versus time.

**MATERIALS AND METHODS**

**Data**

We shall analyze the daily data for the number of confirmed cases in mainland China, which origins from China National Health Commission [1]. We show in the table Table 1 the corresponding time moments \( t_j \) from 0 to 24 and the number of victims \( V_j \) (confirmed cumulative cases of coronavirus 2019-nCoV infection), which were used for calculations. Table 1 shows that the precise time of the epidemic beginning \( t_0 \) is unknown. Therefore, the optimization procedures have to determine the optimal value of this parameter as well as for other parameters of SIR model.

**Table 1.** The information from official table of from China National Health Commission [1]. The corresponding time moments \( t_j \) and the number of victims \( V_j \) (confirmed cumulative cases of coronavirus 2019-nCoV infection), which were used for calculations.
| $t_j$ | $V_j$ | Date |
|------|------|------|
|      |      | **January 2020** |
| 0    | 45   | 16   |
| 1    | 62   | 17   |
| 2    | 121  | 18   |
| 3    | 198  | 19   |
| 4    | 291  | 20   |
| 5    | 440  | 21   |
| 6    | 571  | 22   |
| 7    | 830  | 23   |
| 8    | 1287 | 24   |
| 9    | 1975 | 25   |
| 10   | 2744 | 26   |
| 11   | 4515 | 27   |
| 12   | 5974 | 28   |
| 13   | 7711 | 29   |
| 14   | 9692 | 30   |
| 15   | 11791| 31   |
|      |      | **February 2020** |
| 16   | 14380| 1     |
| 17   | 17205| 2     |
| 18   | 20440| 3     |
| 19   | 24324| 4     |
| 20   | 28018| 5     |
| 21   | 31161| 6     |
| 22   | 34568| 7     |
| 23   | 37198| 8     |
**Exact solution of SIR-equations**

The SIR-model for an infectious disease can be written as follows, \[6,7\]:

\[
\begin{align*}
\dot{S} &= -\alpha SI \\
\dot{I} &= \alpha SI - \rho I \\
\dot{R} &= \rho I
\end{align*}
\]

(1) \hspace{2cm} (2) \hspace{2cm} (3)

The number of susceptible persons is \(S\), infected - \(I\), removed - \(R\); the infection and immunization rates are \(\alpha\) and \(\rho\) respectively. Since \(\dot{S} + \dot{I} + \dot{R} = 0\) (see, Eqs. (1-3)), the sum \(N = S + I + R\) must be constant for all moments of time and can be treated as the amount on susceptible persons before the outbreak of an epidemic, since \(I = R = 0\) at \(t < t_0\). It must be noted that the constant \(N\) is not the volume of population \(N_{total}\), but only the initial number if people sensitive and not protected to some specific disease. In particular, the ratio \(N / N_{total}\) may be rather small. For example, the number of people on the board of Diamond Princess is 3711, and the number of confirmed cases is 70 (February 10, 2020). It means that the percentage of susceptible can be estimated by 1.89%.

To determine the initial conditions for the set of equations (1-3), let us suppose that

\[I(t_0) = 1, \quad R(t_0) = 0, \quad S(t_0) = N - 1\]

(4)

It follows from (1) and (2) that

\[
\frac{dI}{dS} = \frac{\nu}{S} - 1, \quad \nu = \frac{\rho}{\alpha}
\]

(5)

Integration of (5) with the initial conditions (4) yields:

\[I = \nu \ln S - S + N - \nu \ln(N - 1)\]

(6)

Function \(I\) has a maximum at \(S = \nu\) and tends to zero at infinity, see \[6, 7\]. In comparison, the number of susceptible persons at infinity \(S_\infty > 0\), and can be calculated with the use of (6) from a non-linear equation

\[S_\infty = (N - 1)e^{-\frac{S_\infty - N}{\nu}}\]

(7)

In \[4\] the equations (1-3) were solved by introducing the function \(V(t) = I(t) + R(t)\), corresponding to the number of victims. The integration of corresponding equation:

\[\dot{V} = \alpha SI = \alpha(N - V)\left[\nu \ln(N - V) + V - \nu \ln(N - 1)\right]\]

(8)

yields:

\[t = \frac{F_1(V, N, \nu) + \alpha t_0}{\alpha}\]

(9)
Thus, for every set of parameters $N$, $\nu$, $\alpha$, $t_0$ and a fixed value of $V$ the integral (10) can be calculated and the corresponding moment of time can be determined from (9). Then $I$ can be calculated from (6) by putting $S=N-V$ and function $R$ from $R=V-I$.

**Statistical approach for parameter identification. Linear regression**

As in paper [4], we shall use the fact that the random function $F_i(V, N, \nu)$ has a linear distribution with (see (9)). Then we can apply the linear regression (see [9]) for every pair of parameters $N$ and $\nu$ and the corresponding values of $t_0$ and $\alpha$. The optimal (the most reliable) values of $N$ and $\nu$ correspond to the maximum value of the correlation coefficient $r$ (see [4]).

**RESULTS**

The optimal values of parameters were calculated:

$N=90611; \quad \nu=65546.5; \quad \alpha=1.477985357571669e-05; \quad t_0=-7.720998173432072.$

The corresponding correlation coefficient is very high $r=0.997966487046645$. The solution of (7) yields the value $S_\infty=45579$. The corresponding number of infected $I$, susceptible $S$ and removed $R$ persons versus time (starting from January 16, 2020) were calculated and shown in Fig. 1. The blue line represent the number of victims $V=I+R$ and is in good agreement with confirmed number of victims $V=I+R$ (blue markers, see [1]).
**Fig. 1.** Numbers of infected $I$ (green line), susceptible $S$ (red line) and removed $R$ (black line) persons versus time in days (starting from January 16, 2020). The blue line represents the number of victims $V=I+R$. Blue markers show the confirmed number of victims, reported by China National Health Commission [1].

**DISCUSSION**

The obtained value of the correlation coefficient is very close to unit 1. This fact can make the results rather reliable. Nevertheless, this the value of $r$ is very close to the optimal one $r=0.99796487046645$ for the values of parameters located in the vicinity of the optimal point $N=90611; \nu=65546.5$, since the maximum of regression coefficient at this point is not sharp. This fact can question the procedure of the parameter identification. The calculations must be refreshed after obtaining new data.

Another weak feature of the method is connected with the fact that the estimation of susceptible persons, who are still present in the population $S_{\infty} = N - V_{\infty} = 45579$, is very large. It means that these persons can couch the infection. Such situation needs additional analysis, in particular, with the use of more complicated models (see, e.g., [10]).

**CONCLUSIONS**

Simple mathematical model was used to predict the characteristics of the epidemic caused by coronavirus 2019-nCoV in mainland China. The further research should focus on updating the predictions with the use of fresh data and using more complicated mathematical models.

**Acknowledgements**

I would like to express my sincere thanks to professors Dirk Langemann (Technische Universitaet Braunschweig) and Juergen Prestin (Universitaet zu Luebeck) for their support in developing the used optimization approach. I would like to thank also professors Alberto Redaelli, Giuseppe Passoni and Gianfranco Fiore (Politecnico di Milano), S. Pereverzyev (RICAM, Linz, Austria) for involving me in very interesting biomedical investigations in frames of EU-financed Horizon-2020 projects EUMLS (Grant agreement PIRSES-GA-2011-295164-EUMLS) and AMMODIT (Grant Number MSCA-RISE 645672).
References

1. https://en.wikipedia.org/wiki/Timeline_of_the_2019%E2%80%9320_Wuhan_coronavirus_outbreak. Retrieved February 10.

2. Joseph T Wu, Kathy Leung, Gabriel M Leung. Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-nCoV Outbreak Originating in Wuhan, China: A Modelling Study. Lancet 2020 Jan 31. DOI: 10.1016/S0140-6736(20)30260-9

3. Shi Zhao, Qianyin Lin, Jinjun Ran, Salihu S Musa, Guangpu Yang, Weiming Wang, Yijun Lou, Daozhou Gao, and others. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. International Journal of Infectious Diseases. Published online: January 30, 2020. DOI:https://doi.org/10.1016/j.ijid.2020.01.050

4. I. Nesteruk, Statistics based models for the dynamics of Chernivtsi children disease, AMMODIT Conference, Kyiv, Ukraine, January, 2017, Naukovi Visti NTUU KPI, 2017, No. 5, pp. 26-34, DOI: https://doi.org/10.20535/1810-0546.2017.5.108577

5. W.D. Kermack and A.G. McKendrick, “A Contribution to the Mathematical Theory of Epidemics”, *J. Royal Statistical Society*, Ser. A, v. 115, pp. 700-721, 1927.

6. J.D. Murray, *Mathematical Biology I/II*, Springer, New York, 2002.

7. N.T.J Bailey, *The Mathematical Theory of Epidemics*. Griffin Book Co., 1957.

8. D. Langemann, I. Nesteruk and J. Prestin, “Comparison of mathematical models for the dynamics of the Chernivtsi children disease”, *Mathematics in computers and simulation*, v. 123, pp. 68-79, 2016.

9. N.R. Draper and H. Smith, *Applied Regression Analysis (3rd ed.)*, John Wiley, 1998.

10. P. Waltman, *Deterministic Threshold Models in the Theory of Epidemics, Lecture Notes in Biomathematics*, v. 1, Springer, 1974.