Analysis and prediction of confirmed COVID-19 cases in China with uncertain time series

Tingqing Ye¹ · Xiangfeng Yang²

Accepted: 7 September 2020 / Published online: 16 September 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

This paper presents an uncertain time series model to analyse and predict the evolution of confirmed COVID-19 cases in China, excluding imported cases. Compared with the results of the classical time series model, the uncertain time series model could better describe the COVID-19 epidemic by using an uncertain hypothesis test to filter out outliers. This improvement is reflected in the two observations. One is that the estimated variance of the disturbance term in the uncertain time series model is more appropriate and acceptable than that in the classical time series model, and the other is that the disturbance term of the classical time series model cannot be regarded as a random variable but as an uncertain variable.

Keywords Uncertainty theory · Uncertain time series · Uncertain hypothesis test · COVID-19

1 Introduction

COVID-19 has become a pandemic and a public health emergency. As of March 23, 2020, a total of 67,801 confirmed cases and 3160 deaths have been reported in mainland China. Respiratory droplets and human-to-human contact are the major routes of transmission of COVID-19. Based on initial confirmed COVID-19 cases, how can we model its evolution?

There exist two mathematical methods: probabilistic statistics and uncertain statistics. The difference between probabilistic statistics and uncertain statistics is that the

✉ Xiangfeng Yang
yangxf@uibe.edu.cn

Tingqing Ye
yetq18@mails.tsinghua.edu.cn

¹ Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

² School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China

former is in the framework of the probability theory, while the latter is in the structure of the uncertainty theory. Probability theory is a “multiplication” mathematical system, and uncertainty theory is a “minimum” mathematical system. The origin of uncertainty theory could be traced to the pioneering work of Liu (2007), who established uncertainty theory based on normality, duality, and subadditivity axioms. In addition, to address the product uncertain measure, Liu (2009) introduced the fourth axiom—product axiom of uncertainty theory. Since then, uncertainty theory has become an ideal mathematical system and has been applied in many fields.

This paper will adopt an uncertain statistics methodology to interpret and analyse confirmed COVID-19 cases in China. Uncertain regression analysis as a field of uncertain statistics was first presented by Yao and Liu (2018). They applied the least squares method to estimating the parameters of the uncertain regression model. Later, Lio and Liu (2018) investigated the residual and confidence intervals in an uncertain regression model. As an extension of an uncertain regression model, a multivariate uncertain regression model was studied by Song and Fu (2018), Ye and Liu (2020a), and Zhang et al. (2020). In addition, some researchers analysed the other uncertain regression models, such as the uncertain Verhulst–Pearl model by Liu (2019b), the uncertain Gompertz regression model by Hu and Gao (2020), the uncertain revised regression model by Fang and Hong (2020), and the uncertain Chapman–Richards growth model by Liu and Jia (2020). Other scholars have suggested the various estimation approaches, such as the least absolute deviations estimation by Liu and Yang (2020b), the Tukeys biweight estimation by Chen (2020), and the uncertain maximum likelihood estimation by Lio and Liu (2020). To test the fitness of estimation in an uncertain regression model, Ye and Liu (2020b) introduced an uncertain hypothesis test. It is worth mentioning that Liu (2020) investigated an uncertain growth model and applied it to depicting the cumulative number of COVID-19 infections in China.

Uncertain time series analysis, another field of uncertain statistics, was first proposed by Yang and Liu (2019), who considered time series data as uncertain variables and implemented an uncertain autoregressive model. They used a least squares method to estimate the unknown coefficients of the uncertain autoregressive model. Subsequently, Yang et al. (2020) proposed a least absolute deviations method, and Zhao et al. (2020) derived the analytic solution of uncertain autoregressive model based on the principle of least squares. Yang and Ni (2020) introduced another uncertain time series model, the uncertain moving average model, and converted the 1-order uncertain moving average model into an uncertain autoregressive model. Liu and Yang (2020a) offered a cross-validation process to determine the order of the uncertain autoregressive model. Tang (2020) extended the uncertain autoregressive model in the sense of vector situation and developed an uncertain vector autoregressive model.

In this study, we aimed to implement an uncertain time series model to analyse and predict the evolution of confirmed COVID-19 cases. The rest of this paper is organized as follows. Section 2 employs a classical time series analysis to predict the cumulative number of confirmed COVID-19 cases in China, but the result is not ideal. Section 3 introduces uncertain time series analysis, including least squares estimations, residual analysis, the uncertain hypothesis test, the forecast value, and the confidence interval. Section 4 applies uncertain time series analysis to confirmed COVID-19 cases to obtain the uncertain time series model without outliers. Section 5 compares the results

Table 1 The cumulative number of confirmed COVID-19 cases excluding imported cases in China from February 13 to March 23, 2020

63,851	72,436	76,288	78,064	79,824	80,389	80,668	80,725	80,738	80,739
66,492	74,185	76,936	78,497	80,026	80,516	80,685	80,729	80,739	80,740
68,500	74,576	77,150	78,824	80,151	80,591	80,699	80,733	80,739	80,740
70,548	75,465	77,658	79,251	80,270	80,632	80,708	80,737	80,739	80,744

produced by classical time series analysis and uncertain time series analysis. Finally, Sect. 6 provides the conclusions of this paper.

2 Classical time series analysis for COVID-19 cases

In this section, classical time series analysis is applied to modeling the cumulative number of confirmed COVID-19 cases by local transmission in China. We collected the data on the cumulative number of confirmed COVID-19 cases in China, excluding imported cases from February 13 to March 23, 2020 (see Table 1) via the website of the National Health Commission of China National Health (2020). Although the Chinese government has been publishing data since January 20, the confirmed cases before February 13 are not real-time data due to the limitation of laboratory testing capacity for COVID-19.

To model the data in Table 1, we denote the cumulative number of confirmed COVID-19 cases in China from February 13 to March 23, 2020 by $X_1, X_2, \dots, X_{40}$ (Fig. 1). In other words, for each t ($t = 1, 2, \dots, 40$), X_t represents the cumulative number of confirmed COVID-19 cases in China on day t after February 12, 2020. For example, when $t = 1$, X_1 represents the cumulative number of confirmed COVID-19 cases in China on February 13, 2020, and when $t = 40$, X_{40} represents the cumulative number of confirmed COVID-19 cases in China on March 23, 2020.

According to the classical method, we first perform a stationarity test. Unfortunately, the data $X_1, X_2, \dots, X_{40}$ do not pass the stationarity test. To obtain stationary data, we should derive the 1-order difference of X_t . Thus we write

$$Y_t = X_{t+1} - X_t, \quad t = 1, 2, \dots, 39.$$

See Fig. 2.

However, the data $Y_1, Y_2, \dots, Y_{39}$ still do not pass the stationarity test. Thus, we denote the 2-order difference of X_t by

$$Z_t = Y_{t+1} - Y_t = (X_{t+2} - X_{t+1}) - (X_{t+1} - X_t) = X_{t+2} + X_t - 2X_{t+1} \quad (1)$$

for $t = 1, 2, \dots, 38$ (Fig. 3).

Fortunately, the data $Z_1, Z_2, \dots, Z_{38}$ now pass the stationarity test and the white noise test. Let us consider an ARMA(p, q) model. Based on BIC criteria, we have

Fig. 1 Cumulative confirmed COVID-19 cases in China

Fig. 2 1-order difference data

$(p, q) = (4, 0)$. By using MATLAB ¹, the ARMA(4,0) model is

$$Z_t = -26.252 - 0.2357Z_{t-1} - 0.0050Z_{t-2} + 0.1229Z_{t-3} + 0.6556Z_{t-4} + \varepsilon_t \quad (2)$$

$$\varepsilon_t \sim IID(4.2443, 201.41^2).$$

¹ MATLAB R2020a, 9.8.0.1323502, maci64, Econometrics Toolbox, “arma” and “estimate” functions.

Fig. 3 2-order difference data

The residuals of the ARMA(4,0) model (2) are shown in Fig. 4. Then, according to the ARMA(4,0) model (2), the forecast value of X_{41} (the cumulative number of confirmed COVID-19 cases in China on March 24, 2020) is

$$\hat{X}_{41} = 80721.$$

In the ARMA(4,0) model (2), the estimated variance of the disturbance term is 201.41^2 . It is too large to be accepted. In addition, a residual plot in classical time series analysis is usually required to be a null plot which has no separated points and describes the random variable with constant mean and variance. However, it seems that the residual plot for the ARMA(4,0) model (2) shown in Fig. 4 does not look like a null plot. In this case, if we still use classical time series analysis, then the distribution function is not close enough to the frequency we get. Thus the disturbance term cannot be regarded as a random variable, and the classical time series analysis is not appropriate for predicting the future the cumulative number of confirmed COVID-19 cases in China. That is why we are going to try to use uncertain time series analysis in the next sections.

3 Uncertain time series analysis

In this section, we will introduce the uncertain time series analysis, including least squares estimations, residual analysis, uncertain hypothesis test, forecast value, and confidence interval. Since we choose ARMA(4,0) model in the classical time series analysis, we only introduce the uncertain autoregressive model for the convenience of calculation and comparison.

Fig. 4 Residuals of ARMA(4,0) model (2)

3.1 Uncertain time series model

Yang and Liu (2019) proposed the uncertain autoregressive model,

$$Z_t = a_0 + \sum_{i=1}^k a_i Z_{t-i} + \varepsilon_t \quad (3)$$

where $a_0, a_1, \dots, a_k$ are unknown parameters, k is the order of the autoregressive model, and $\{\varepsilon_1, \varepsilon_2, \dots\}$ is a sequence of the disturbance terms which are iid uncertain variables with a common uncertainty distribution $\mathcal{N}(e, \sigma)$, in which e and σ are unknown parameters. The uncertainty distribution of $\mathcal{N}(e, \sigma)$ is

$$\Phi(x) = \left(1 + \exp\left(\frac{\pi(e-x)}{\sqrt{3}\sigma}\right) \right)^{-1},$$

and the inverse uncertainty distribution of $\mathcal{N}(e, \sigma)$ is

$$\Phi^{-1}(\alpha) = e + \frac{\sigma\sqrt{3}}{\pi} \ln \frac{\alpha}{1-\alpha}.$$

In order to be consistent with the classical time series model in writing, we write the uncertain autoregressive model (3) as an uncertain ARMA($k, 0$) model.

3.2 Least squares estimation and residual analysis

When $Z_1, Z_2, \dots, Z_n$ are observed, Yang and Liu (2019) defined the least squares estimations of $a_0, a_1, \dots, a_k$ in the uncertain ARMA($k, 0$) model (3) is the solution of the minimization problem

$$\min_{a_0, a_1, \dots, a_k} \sum_{t=k+1}^n \left(Z_t - a_0 - \sum_{i=1}^k a_i Z_{t-i} \right)^2. \quad (4)$$

Denote the optimal solutions of Eq. (4) by $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_k$. Then the fitted ARMA($k, 0$) model is

$$Z_t = \hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{t-i}.$$

For each index t ($t = k + 1, k + 2, \dots, n$), the t th residual is

$$\hat{\varepsilon}_t = Z_t - \hat{a}_0 - \sum_{i=1}^k \hat{a}_i Z_{t-i}. \quad (5)$$

Then we use the average of the expected values of residuals, i.e.,

$$\hat{e} = \frac{1}{n-k} \sum_{t=k+1}^n \hat{\varepsilon}_t \quad (6)$$

to estimate the expected value e of the disturbance term, and

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{t=k+1}^n (\hat{\varepsilon}_t - \hat{e})^2 \quad (7)$$

to estimate the variance σ^2 of the disturbance term.

3.3 Uncertain hypothesis test

The uncertain hypothesis test, first proposed by Ye and Liu (2020b), is a mathematical tool that uses uncertainty theory to rationally judge whether some hypotheses are correct or not according to the observed data. Next, the uncertain hypothesis test will be employed to test the appropriateness of estimations of the unknown parameters in the uncertain ARMA($k, 0$) model.

In Sect. 3.2, we obtained the estimations $\hat{e}$ and $\hat{\sigma}$. Naturally, we are interested in testing whether the estimations $\hat{e}$ and $\hat{\sigma}$ are appropriate. Note that $\hat{e}$ and $\hat{\sigma}$ are estimations of the expected value e and the standard variance σ of the disturbance

term, respectively. Therefore, we want to test the appropriateness of the estimations $\hat{e}$ and $\hat{\sigma}$ by testing whether $e = \hat{e}$ and whether $\sigma = \hat{\sigma}$. Let us consider the following two-sided hypotheses

$$H_0 : e = \hat{e} \text{ and } \sigma = \hat{\sigma} \text{ versus } H_1 : e \neq \hat{e} \text{ or } \sigma \neq \hat{\sigma}.$$

For a given significance level α (e.g., 0.01), Ye and Liu (2020b) suggested that the test is

$$W = \left\{ (\hat{e}_{k+1}, \hat{e}_{k+2}, \dots, \hat{e}_n) \in \mathbb{R}^{n-k} \left| \begin{array}{l} \hat{e} + \frac{\hat{\sigma}\sqrt{3}}{\pi} \ln \frac{\alpha}{1-\alpha} \leq \hat{e}_t \leq \hat{e} + \frac{\hat{\sigma}\sqrt{3}}{\pi} \ln \frac{1-\alpha}{\alpha}, \quad t = k+1, k+2, \dots, n \end{array} \right. \right\}^c. \quad (8)$$

If $(\hat{e}_{k+1}, \hat{e}_{k+2}, \dots, \hat{e}_n) \in W$, then we reject H_0 , and the estimations $\hat{e}$ and $\hat{\sigma}$ do not pass the test. Otherwise, we accept H_0 , and the estimations $\hat{e}$ and $\hat{\sigma}$ pass the test.

3.4 Data modification

Assume the estimations $\hat{e}$ and $\hat{\sigma}$ do not pass the test, i.e., $(\hat{e}_{k+1}, \hat{e}_{k+2}, \dots, \hat{e}_n) \in W$. For each index t with $k+1 \leq t \leq n$, the data Z_t is called an outlier if

$$\hat{e}_t \notin \left[\hat{e} + \frac{\sqrt{3}\hat{\sigma}}{\pi} \ln \frac{\alpha}{1-\alpha}, \hat{e} + \frac{\sqrt{3}\hat{\sigma}}{\pi} \ln \frac{1-\alpha}{\alpha} \right],$$

otherwise, Z_t is called a normal point. Since the outliers cause estimations $\hat{e}$ and $\hat{\sigma}$ not to pass the test, we want to modify the outliers and then re-estimate the parameters with the modified data. The detailed process of data modification is as follows.

For each index t ($t = k+1, k+2, \dots, n$), if Z_t is not an outlier, then it remains unchanged. Otherwise the data modification can be broken down into four cases.

Case I: If there is a normal point following Z_t and Z_t is not on the line $l(s)$ formed by the normal point nearest to Z_t forward and the normal point nearest to Z_t backward in position, then replace the value of Z_t with the value of $l(t)$. For example, assume Z_{k+1} and Z_{k+2} are two outliers while Z_k and Z_{k+3} are not, and assume neither Z_{k+1} nor Z_{k+2} is not on the line $l(s)$ formed by Z_k and Z_{k+3} (Fig. 5). Then we replace the values of Z_{k+1} and Z_{k+2} with

$$l(k+1) = Z_k + 1 \cdot \frac{Z_{k+3} - Z_k}{3}, \quad l(k+2) = Z_k + 2 \cdot \frac{Z_{k+3} - Z_k}{3},$$

respectively.

Fig. 5 Schematic diagram of case I

Case II: If there is a normal point following Z_t and Z_t is on the line $l(s)$ formed by the normal point nearest to Z_t forward and the normal point nearest to Z_t backward in position, then replace the value of Z_t with fitted value

$$\hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{t-i}.$$

For example, assume Z_{k+1} and Z_{k+2} are two outliers while Z_k and Z_{k+3} are not, and assume Z_{k+1} and Z_{k+2} are both on the line $l(s)$ formed by Z_k and Z_{k+3} (Fig. 6). Then we replace the values of Z_{k+1} and Z_{k+2} with fitted values

$$\hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{k+1-i}, \quad \hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{k+2-i},$$

respectively.

Case III: If there is no normal point following Z_t and Z_t is not on the line $l(s)$ formed by the normal point nearest to Z_t forward and the normal point second nearest to Z_t forward in position, then replace the value of Z_t with the value of $l(t)$. For example, assume Z_{n-1} and Z_n are two outliers while Z_{n-3} and Z_{n-2} are not, and assume neither Z_{n-1} nor Z_n is on the line $l(s)$ formed by Z_{n-3} and Z_{n-2} (Fig. 7). Then we replace the values of Z_{n-1} and Z_n with

$$l(n-1) = Z_{n-2} + 1 \cdot (Z_{n-2} - Z_{n-3}), \quad l(n) = Z_{n-2} + 2 \cdot (Z_{n-2} - Z_{n-3}),$$

respectively.

Fig. 6 Schematic diagram of case II

Fig. 7 Schematic diagram of case III

Case IV: If there is no normal point following Z_t and Z_t is on the line $l(s)$ formed by the normal point nearest to Z_t forward and the normal point second nearest to Z_t forward in position, then replace the value of Z_t with fitted value

$$\hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{t-i}.$$

Fig. 8 Schematic diagram of case IV

For example, assume Z_{n-1} and Z_n are two outliers while Z_{n-3} and Z_{n-2} are not, and assume Z_{n-1} and Z_n are both on the line $l(s)$ formed by Z_{n-3} and Z_{n-2} . See Fig. 8. Then we replace the values of Z_{n-1} and Z_n with fitted values

$$\hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{n-1-i}, \quad \hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{n-i},$$

respectively.

After a round of data modification, a new set of data $Z_1, Z_2, \dots, Z_n$ is obtained. Afterwards, we use the new data to recalculate $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_k, \hat{\varepsilon}_{k+1}, \hat{\varepsilon}_{k+2}, \dots, \hat{\varepsilon}_n, \hat{\varepsilon}$, and $\hat{\sigma}$ with the method proposed in Sect. 3.2 and test the estimations $\hat{\varepsilon}$ and $\hat{\sigma}$ with the method in Sect. 3.3. If the estimations do not pass the test, then we continue to modify the data and repeat the process above until the estimations pass the test. Then, the final uncertain ARMA($k, 0$) model is

$$Z_t = \hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{t-i} + \varepsilon_t, \quad \varepsilon_t \stackrel{iid}{\sim} \mathcal{N}(\hat{\varepsilon}, \hat{\sigma}). \quad (9)$$

3.5 Forecast value and confidence interval

According to the final uncertain ARMA($k, 0$) model (9), the forecast value of Z_{n+1} is

$$\hat{Z}_{n+1} = \hat{a}_0 + \sum_{i=1}^k \hat{a}_i Z_{n+1-i} + \hat{e}. \quad (10)$$

Taking β as a confidence level (e.g., 95%), we find the minimum c such that

$$\Phi_{n+1}(\hat{Z}_{n+1} + c) - \Phi_{n+1}(\hat{Z}_{n+1} - c) \geq \beta \quad (11)$$

where

$$\Phi_{n+1}(x) = \left(1 + \exp \left(\frac{\pi(\hat{Z}_{n+1} + \hat{e} - x)}{\sqrt{3}\hat{\sigma}} \right) \right)^{-1}.$$

Then the β -confidence interval of Z_{n+1} is

$$[\hat{Z}_{n+1} - c, \hat{Z}_{n+1} + c].$$

3.6 Algorithm flowchart

This section condenses the method proposed in Sects. 3.1–3.5 into an algorithm to address the outliers found in the uncertain hypothesis test and obtain the estimations that can pass the test.

Step 1. (Parameter estimation) Use the data $Z_1, Z_2, \dots, Z_n$ to compute least squares estimations of $a_0, a_1, \dots, a_k$ in the uncertain ARMA($k, 0$) model with Eq. (4), denoted by $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_k$.

Step 2. (Residual analysis) Calculate $\hat{e}_{k+1}, \hat{e}_{k+2}, \dots, \hat{e}_n, \hat{e}$, and $\hat{\sigma}^2$ with Eqs. (5)–(7).

Step 3. (Uncertain hypothesis test) Take a significance level α (e.g., 0.01), and construct the test W with Eq. (8).

Step 4. (Data modification) If $(\hat{e}_{k+1}, \hat{e}_{k+2}, \dots, \hat{e}_n) \notin W$, then go to Step 5. Otherwise, start the data modification. For each t ($t = k + 1, k + 2, \dots, n$), reset the data Z_t by the method proposed in Sect. 3.4 if Z_t is an outlier, otherwise it remains unchanged. Then, return to Step 1.

Step 5. (Forecast value and confidence interval) Calculate the forecast value $\hat{Z}_{n+1}$ of Z_{n+1} with Eq. (10). Take β (e.g., 95%) as a confidence level, and find the minimum c satisfying Eq. (11). Then, the β -confidence interval of Z_{n+1} is $[\hat{Z}_{n+1} - c, \hat{Z}_{n+1} + c]$.

4 Uncertain time series analysis for COVID-19 cases

In this section, we will apply the uncertain time series analysis to modeling the cumulative number of confirmed COVID-19 cases in China from February 13 to March 23, 2020 with the uncertain ARMA($k, 0$) model.

Table 2 Average testing error (ATE) with cross validation

k	ATE(k)	k	ATE(k)	k	ATE(k)	k	ATE(k)
1	90,619	6	5124	11	11,904	16	15,742
2	106,409	7	7432	12	8489	17	5161
3	134,265	8	9125	13	61,587	18	21,467
4	11,099	9	17,087	14	124,562	19	6933
5	4996	10	24,393	15	59,443	20	12,844

Table 3 Residuals in 1st iteration

t	$\hat{\varepsilon}_t$	t	$\hat{\varepsilon}_t$	t	$\hat{\varepsilon}_t$	t	$\hat{\varepsilon}_t$	t	$\hat{\varepsilon}_t$	t	$\hat{\varepsilon}_t$
6	5.7	12	112	18	-101.5	24	-13.1	30	14.1	36	15.6
7	-4.9	13	49.4	19	-46	25	20	31	9.3	37	14
8	-36.7	14	14.4	20	131.6	26	19.9	32	16.8	38	18.2
9	-61.3	15	195.9	21	14.8	27	24.2	33	13.8		
10	50	16	-286.6	22	-31.3	28	10.1	34	14		
11	-46.6	17	-153.1	23	-10.8	29	12.8	35	15.2		

4.1 Determining order by cross validation

In order to select the order k in uncertain ARMA($k, 0$) model, we employ the method of ν fold rolling origin cross validation proposed by Liu and Yang (2020a) based on the data $Z_1, Z_2, \dots, Z_{38}$. Consider the maximum order 20 and $\nu = 26$. Then corresponding average testing errors for different orders are shown in Table 2.

From Table 2, we can see the average testing error is minimal when $k = 5$. Thus we choose uncertain ARMA(5, 0) model.

4.2 1st iteration

According to the data $Z_1, Z_2, \dots, Z_{38}$ in Sect. 2, we can obtain the least squares estimations

$$\begin{aligned}\hat{a}_0 &= -14.345, \hat{a}_1 = -0.3820, \hat{a}_2 = -0.1431, \\ \hat{a}_3 &= 0.0168, \hat{a}_4 = 0.3194, \hat{a}_5 = 0.0666\end{aligned}$$

with Eq. (4). Thus, the fitted ARMA(5, 0) model is

$$Z_t = -14.345 - 0.3820Z_{t-1} - 0.1431Z_{t-2} + 0.0168Z_{t-3} + 0.3194Z_{t-4} + 0.0666Z_{t-5}.$$

In addition, we can compute the t th residuals

$$\hat{\varepsilon}_t = Z_t + 14.345 + 0.3820Z_{t-1} + 0.1431Z_{t-2} - 0.0168Z_{t-3} - 0.3194Z_{t-4} - 0.0666Z_{t-5},$$

for $t = 6, 7, \dots, 38$, which are shown in Table 3.

Fig. 9 Residuals in 1st iteration

The estimated expected value $\hat{e}$ and estimated variance $\hat{\sigma}^2$ of the disturbance term ε_t are

$$\hat{e} = \frac{1}{33} \sum_{t=6}^{38} \hat{\varepsilon}_t = 0.0000, \quad \hat{\sigma}^2 = \frac{1}{33} \sum_{t=6}^{38} (\hat{\varepsilon}_t - \hat{e})^2 = 78.593^2,$$

respectively. In short, the uncertain ARMA(5, 0) model is

$$\begin{aligned} Z_t = & -14.345 - 0.3820Z_{t-1} - 0.1431Z_{t-2} + 0.0168Z_{t-3} + 0.3194Z_{t-4} \\ & + 0.0666Z_{t-5} + \varepsilon_t, \quad \varepsilon_t \stackrel{iid}{\sim} \mathcal{N}(0.0000, 78.593). \end{aligned} \quad (12)$$

Taking the level of significance $\alpha = 0.01$, we obtain

$$\Phi^{-1}(\alpha) = -199.11, \quad \Phi^{-1}(1 - \alpha) = 199.11$$

where $\Phi^{-1}(\alpha)$ is the inverse uncertainty distribution of $\mathcal{N}(0.0000, 78.593)$. It follows from Eq. (8) that the test is

$$W = \left\{ (\hat{\varepsilon}_6, \hat{\varepsilon}_7, \dots, \hat{\varepsilon}_{38}) \in \mathbb{R}^{33} \mid -199.11 \leq \hat{\varepsilon}_t \leq 199.11, \quad t = 6, 7, \dots, 38 \right\}^c.$$

Figure 9 shows the residuals in Table 3 and the two bounds (-199.11 and 199.11).

From Fig. 9, we can see that only $\hat{\varepsilon}_{16} = -286.6$ does not fall into the interval $[-199.11, 199.11]$. Thus, Z_{16} is an outlier, and $(\hat{\varepsilon}_6, \hat{\varepsilon}_7, \dots, \hat{\varepsilon}_{38}) \in W$ which implies that the estimations $\hat{e}$ and $\hat{\sigma}$ in the model (12) do not pass the test. Then, we reset Z_{16}

Fig. 10 Reset data in 1st iteration

with the method for case I in Sect. 3.4, i.e., replace Z_{16} by

$$Z_{15} + \frac{Z_{17} - Z_{15}}{2} = 146 + \frac{-77 - 146}{2} = 34.5.$$

The rest of the data remain unchanged. Figure 10 shows the reset data.

4.3 2nd iteration

Based on the reset data in Fig. 10, we can compute the estimations

$$\begin{aligned} \hat{a}_0 &= 2.7052, \hat{a}_1 = -0.3532, \hat{a}_2 = -0.1155, \hat{a}_3 = 0.0298, \\ \hat{a}_4 &= 0.3490, \hat{a}_5 = 0.0484, \hat{e} = 0.0000, \hat{\sigma} = 48.208, \end{aligned}$$

and all residuals $\hat{e}_6, \hat{e}_7, \dots, \hat{e}_{38}$. It follows from Eq. (8) that the test is

$$W = \left\{ (\hat{e}_6, \hat{e}_7, \dots, \hat{e}_{38}) \in \mathbb{R}^{33} \mid -122.13 \leq \hat{e}_t \leq 122.13, t = 6, 7, \dots, 38 \right\}^c.$$

Figure 11 shows the residuals and the two bounds (-122.13 and 122.13).

From Fig. 11, we can see that only $\hat{e}_{15} = 186.94$ does not fall into the interval $[-122.13, 122.13]$. Thus, Z_{15} is an outlier, and $(\hat{e}_6, \hat{e}_7, \dots, \hat{e}_{38}) \in W$ which implies that the estimations $\hat{e}$ and $\hat{\sigma}$ do not pass the test. Then we reset the outlier with the method for case I in Sect. 3.4, i.e., replace Z_{15} by

$$Z_{14} + \frac{Z_{17} - Z_{15}}{2} = 100 + \frac{34.5 - 100}{2} = 67.25.$$

Fig. 11 Residuals in 2nd iteration

Fig. 12 Reset data in 2nd iteration

The rest of the data remain unchanged. Figure 12 shows the reset data.

4.4 15th iteration

After repeating the process 15 times as described in Sects. 4.2 and 4.3, we can compute the estimations

Fig. 13 Residuals in 15th iteration

$$\begin{aligned}\hat{a}_0 &= -0.1432, \hat{a}_1 = -0.3791, \hat{a}_2 = -0.1416, \hat{a}_3 = 0.0423, \\ \hat{a}_4 &= 0.3802, \hat{a}_5 = 0.0712, \hat{\varepsilon} = 0.0000, \hat{\sigma} = 10.588,\end{aligned}$$

and all residuals $\hat{\varepsilon}_6, \hat{\varepsilon}_7, \dots, \hat{\varepsilon}_{38}$. It follows from Eq. (8) that the test is

$$W = \left\{ (\hat{\varepsilon}_6, \hat{\varepsilon}_7, \dots, \hat{\varepsilon}_{38}) \in \mathbb{R}^{33} \mid -26.823 \leq \hat{\varepsilon}_t \leq 26.823, t = 6, 7, \dots, 38 \right\}^c.$$

Figure 13 shows the residuals and the two bounds (-26.823 and 26.823).

From Fig. 13, we can see that all residuals fall into the interval $[-26.823, 26.823]$, that is, $(\hat{\varepsilon}_6, \hat{\varepsilon}_7, \dots, \hat{\varepsilon}_{38}) \notin W$ implying that the estimations $\hat{\varepsilon}$ and $\hat{\sigma}$ pass the test. Therefore, the final uncertain ARMA(5, 0) model is

$$\begin{aligned}Z_t &= -0.1432 - 0.3791Z_{t-1} - 0.1416Z_{t-2} + 0.0423Z_{t-3} + 0.3802Z_{t-4} \\ &\quad + 0.0712Z_{t-5} + \varepsilon_t, \quad \varepsilon_t \stackrel{iid}{\sim} \mathcal{N}(0.0000, 10.588).\end{aligned}\quad (13)$$

Compared with the estimated variance 201.41^2 in the model (2) produced by the classical time series analysis, the estimated variance 10.588^2 in the model (13) provided by the uncertain time series analysis is much smaller. In summary, the data have been modified for a total of 15 times, and a total of 8 days of data have been modified. Table 4 shows the results of the data modification.

It follows from the model (13) and the method in Sect. 3.2 that the forecast value of Z_{39} is

$$\hat{Z}_{39} = 1.6145,$$

Table 4 Results of data modification

t	Original data	Modified data	t	Original data	Modified data
11	− 102	− 31	17	− 77	− 42
14	100	117	18	− 6	44
15	146	− 21	21	− 52	2
16	− 371	− 63	22	− 34	30

and the 95% confidence interval of Z_{39} is

$$[-19.771, 23.000].$$

From the difference Eq. (1), the forecast value of X_{41} is

$$\hat{X}_{41} = 80,750,$$

and the 95% confidence interval of X_{41} is

$$[80,719, 80,781].$$

5 Comparison and analysis

This section will compare the results produced by the classical time series analysis and the uncertain time series analysis. In Sect. 2, we get the classical ARMA(4,0) model

$$Z_t = -26.252 - 0.2357Z_{t-1} - 0.0050Z_{t-2} + 0.1229Z_{t-3} + 0.6556Z_{t-4} + \varepsilon_t$$

$$\varepsilon_t \sim IID(4.2443, 201.41^2),$$

and the forecast value of X_{41} (the cumulative number of confirmed COVID-19 cases in China on March 24, 2020), i.e.,

$$\hat{X}_{41} = 80721.$$

In Sect. 3.6, the final uncertain ARMA(5, 0) model is

$$Z_t = -0.1432 - 0.3791Z_{t-1} - 0.1416Z_{t-2} + 0.0423Z_{t-3} + 0.3802Z_{t-4}$$

$$+ 0.0712Z_{t-5} + \varepsilon_t, \quad \varepsilon_t \stackrel{iid}{\sim} \mathcal{N}(0.0000, 10.588).$$

and the forecast value of X_{41} is

$$\hat{X}_{41} = 80,750.$$

All the results are shown in Table 5.

Table 5 Results produced by classical time series analysis and uncertain time series analysis

	Variance ($\hat{\sigma}^2$)	Forecast value (March 24)
Using classical model	201.41 ²	80,721
Using uncertain model	10.588 ²	80,750

On the one hand, from Table 5, we can see that the estimated variance obtained by the classical time series analysis, 201.41², is too large to be accepted while the estimated variance obtained by the uncertain time series analysis, 10.588², makes more sense. In terms of variance, uncertain time series analysis is more appropriate than classical time series analysis for predicting the cumulative number of confirmed COVID-19 cases, which is due to the uncertain hypothesis test, which can automatically filter out outliers. On the other hand, in Sect. 2, we have illustrated that the residual plot for the classical ARMA(4,0) model does not look like a null plot. In this case, if we use classical time series analysis, then the distribution function is not close enough to the frequency we get. Thus the disturbance term of classical time series model cannot be regarded as a random variable. Fortunately, Liu (2019a) suggested that you should use probability theory if your distribution function is close enough to the frequency, otherwise you have to use uncertainty theory. Therefore the disturbance term should be regarded as an uncertain variable instead of a random variable, and the uncertain time series analysis should be used to predict future cases of COVID-19 rather than the classical time series analysis. Furthermore, the 95% confidence interval can also be obtained with the uncertain time series analysis.

6 Conclusions

The COVID-19 swept the world suddenly like a rainstorm. Increasing numbers of people suffer from his disease. It is important to study the trend of COVID-19 development. This paper used the cumulative number of confirmed COVID-19 cases in China, excluding imported cases from February 13 to March 23, 2020 to predict cumulative confirmed cases in the future by using the classical time series analysis. However, although the data and the model passed the white noise test and the stationarity test, the results were not ideal. This is mainly reflected in two observations. One is that the estimated variance of the disturbance term was too large to be accepted. The other is that the disturbance term could not be regarded as a random variable because the residual plot for the classical time series model did not look like a null plot. To address these two problems, we employed uncertain time series analysis, including least estimations, residual analysis, uncertain hypothesis test, forecast value, and confidence interval. The uncertain hypothesis test and the ability to automatically identify outliers reduced the estimated variance of the disturbance term to an acceptable value. Thus, the uncertain time series analysis was more appropriate than the classical time series analysis for predicting the cumulative number of confirmed COVID-19 cases in China. In the future, uncertain time series analysis can also be applied to modeling

other statistics about COVID-19 in China or the world, such as cumulative COVID-19 deaths and cumulative cured COVID-19 cases.

Acknowledgements This work was supported by the National Natural Science Foundation of China Grant No. 61873329, and Program for Young Excellent Talents in UIBE Grant No. 18YQ06.

References

- Chen, D. (2020). Tukeys biweight estimation for uncertain regression model with imprecise observations. *Soft Computing*. <https://doi.org/10.1007/s00500-020-04973-x>.
- Fang, L., & Hong, Y. (2020). Uncertain revised regression analysis with responses of logarithmic, square root and reciprocal transformations. *Soft Computing*, 24(4), 2655–2670.
- Hu, Z., & Gao, J. (2020). Uncertain Gompertz regression model with imprecise observations. *Soft Computing*, 24(4), 2543–2549.
- Lio, W., & Liu, B. (2018). Residual and confidence interval for uncertain regression model with imprecise observations. *Journal Intelligent & Fuzzy Systems*, 35(2), 2573–2583.
- Lio, W., & Liu, B. (2020). Uncertain maximum likelihood estimation with application to uncertain regression analysis. *Soft Computing*, 24, 9351–9360.
- Liu, Z. (2020). Uncertain growth model for the cumulative number of COVID-19 infections in China. *Fuzzy Optimization and Decision Making*. <https://doi.org/10.1007/s10700-020-09340-x>.
- Liu, Z., & Jia, L. (2020). Cross-validation for the uncertain Chapman–Richards growth model with imprecise observations. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*.
- Liu, B. (2007). *Uncertainty theory* (2nd ed.). Berlin: Springer.
- Liu, B. (2009). Some research problems in uncertainty theory. *Journal of Uncertain System*, 3(1), 3–10.
- Liu, B. (2019a). Uncertain urn problems and Ellsberg experiment. *Soft Computing*, 23, 6579–6584.
- Liu, S. (2019b). Leave- p -out cross validation test for uncertain Verhulst–Pearl model with imprecise observations. *IEEE Access*, 7, 131705–131709.
- Liu, Z., & Yang, X. (2020a). Cross validation for uncertain autoregressive model. *Communications in Statistics-Simulation and Computation*. <https://doi.org/10.1080/03610918.2020.1747077>.
- Liu, Z., & Yang, Y. (2020b). Least absolute deviations estimation for uncertain regression with imprecise observations. *Fuzzy Optimization and Decision Making*, 19(1), 33–52.
- National Health Commission of China. (2020). Coronavirus disease (COVID-19) situation reports. Retrieved April 24, 2020 from http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml.
- Song, Y., & Fu, Z. (2018). Uncertain multivariable regression model. *Soft Computing*, 22(17), 5861–5866.
- Tang, H. (2020). Uncertain vector autoregressive model with imprecise observations. *Soft Computing*. <https://doi.org/10.1007/s00500-020-04991-9>.
- Yang, X., & Liu, B. (2019). Uncertain time series analysis with imprecise observations. *Fuzzy Optimization and Decision Making*, 18(3), 263–278.
- Yang, X., & Ni, Y. (2020). Least-squares estimation for uncertain moving average model. *Communications in Statistics Theory and Methods*. <https://doi.org/10.1080/3610926.2020.1713373>.
- Yang, X., Park, G. K., & Hu, Y. (2020). Least absolute deviations estimation for uncertain autoregressive model. *Soft Computing*. <https://doi.org/10.1007/s00500-020-05079-0>.
- Yao, K., & Liu, B. (2018). Uncertain regression analysis: An approach for imprecise observations. *Soft Computing*, 22(17), 5579–5582.
- Ye, T., & Liu, B. (2020b). Uncertain hypothesis test with application to uncertain regression analysis, Technical report.
- Ye, T., & Liu, Y. (2020a). Multivariate uncertain regression model with imprecise observations. *Journal of Ambient Intelligence and Humanized Computing*. <https://doi.org/10.1007/s12652-020-01763-z>.
- Zhang, C., Liu, Z., & Liu, J. (2020). Least absolute deviations for uncertain multivariate regression model. *International Journal of General Systems*. <https://doi.org/10.1080/03081079.2020.1748615>.
- Zhao, X., Peng, J., Liu, J., & Zhou, X. (2020). Analytic solution of uncertain autoregressive model based on principle of least squares. *Soft Computing*, 24(4), 2721–2726.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.