Stochastic Environmental Research and Risk Assessment (2022) 36:2835–2845
https://doi.org/10.1007/s00477-021-02164-0

ORIGINAL PAPER

Modeling air quality level with a flexible categorical autoregression

Mengya Liu[1] · Qi Li[2] · Fukang Zhu[3]

Accepted: 21 December 2021 / Published online: 5 January 2022

Abstract
To study urban air quality, this paper proposes a novel categorical time series model, which is based on a linear combination of bounded Poisson distribution and discrete distribution to describe the dynamic and systemic features of air quality, respectively. Daily air quality level data of three major cities in China, including Beijing, Shanghai and Guangzhou, are analyzed. It is concluded that the air quality in Beijing is the worst among the three cities but is gradually improving, and its dynamics is also the most pronounced. Theoretically, the design of our model increases the flexibility of the probabilistic structure while ensuring a dynamic feedback mechanism without high computational stress. We estimate the parameters through an adaptive Bayesian Markov chain Monte Carlo sampling scheme and show the satisfactory finite sample performance of the model through simulation studies.

Keywords Air quality · Autoregression · Bayesian inference · Categorical time series

1 Introduction

Air quality has become a common concern, both for health-sensitive individuals and for the academics interested in it. The reason for concern is that air pollution is a major cause of death and disease, further posing a threat to economic development and inclusive prosperity. Exposure to air pollution is the fourth leading fatal health risk worldwide behind metabolic risks, dietary risks, and tobacco smoke, while in low- and middle-income countries, it is the third behind metabolic risks and dietary risks (World Bank and IHME 2016). In 2016, the global health cost of mortality and morbidity caused by exposure to ambient PM2.5 air pollution was $5.7 trillion, equivalent to 4.8% of global gross domestic product. By region, the cost in China and India is equivalent to 7.5–8% of GDP (World Bank 2020). Andrée (2020) showed that PM2.5 was a very important predictor of confirmed COVID-19 cases and associated hospital admissions. Because air pollution leads to the loss of productive labor, it is also an economic burden. In addition, air pollution disproportionately affects the poorest populations, which hinders the achievement of shared and inclusive prosperity.

Over the past two decades, China has been actively working to reduce average urban ambient air pollutant concentrations. However, challenges remain in managing air pollution according to either the World Health Organization guidelines or China's own grade I limit value. Therefore, it is necessary and valuable to conduct research on the air quality of major cities in China. In this paper, we focus on the air quality in recent years in three of China's most developed first-tier cities, including Beijing, Shanghai and Guangzhou. The air quality level is quantized into six categories in China: (1) excellent; (2) good; (3) slightly polluted; (4) moderately polluted; (5) heavily polluted; (6) severely polluted. Daily data on air quality levels of each city over time naturally form a categorical time series. This paper is about the analysis of such categorical time series $X_1, \ldots, X_n$ with the ordered categorical range $\{m_1, \ldots, m_M\}$, where $m_1 < \cdots < m_M$. Such data can also be viewed as an ordinal time series, see Weiß (2020) for details, which expresses the dissimilarity of ordinal categories through a distance metric.

✉ Fukang Zhu
zfk8010@163.com

[1] School of Mathematics and Statistics, Central China Normal University, Wuhan, China

[2] College of Mathematics, Changchun Normal University, Changchun, China

[3] School of Mathematics, Jilin University, 2699 Qianjin Street, Changchun 130012, China

This paper proposes an observation-driven model to study the data of air quality level, in which the observations are supposed to follow a novel distribution based on a linear combination of a bounded Poisson distribution and a discrete distribution. On the one hand, the dynamic structure relies on the intensity parameter of the bounded Poisson distribution, which is conditional on past information with the form of the autoregression. On the other hand, the discrete distribution characterizes systemic features in the observations that do not vary with time. This design increases the flexibility of the probabilistic structure of the model while ensuring the presence of a dynamic feedback mechanism. It is meaningful to distinguish between systemic and dynamic changes in air quality research. Specifically, in China, suspended dust, coal combustion, industrial dust, vehicle emissions, biomass burning and secondary particulate matter contribute to urban pollution sources (World Bank 2012), where suspended dust, coal combustion, biomass burning and secondary particulate matter are seasonal factors, but there exists a considerable non-seasonal part of them due to almost fixed climate and topography and the high maturity of the industrial structure in each city. Therefore, this part is treated as systemic, while industrial dust, vehicle emissions and the non-systemic part of seasonal factors are considered dynamic. According to data released by the Ministry of Public Security of China, 27.53 million new motor vehicles were registered in the first three quarters of 2021, an increase of 4.363 million units or 18.83% year-on-year, which supports the rationality of our classification in one way.

As for our proposed model, it is suitable for studying air quality data. The first advantage of our proposed model is its simplicity and practicality, mainly thanks to the fact that it does not require the conversion of observations into vector form and has fewer parameters to be estimated. There exist some categorical time series models that treat observations as a $(M-1)$-dimensional vector, and the conditional distribution given its past is multivariate naturally. Many other categorical time series models such as Markov chain models, generalized choice models, and spectral envelope models have also been studied in the literature. For earlier works, see, for example, Stoffer et al. (1993), Fokianos and Kedem (2003) and the references therein. For more recent ones, Kauppi and Saikkonen (2008), Moysiadis and Fokianos (2014), Fokianos and Moysiadis (2017) and Fokianos and Truquet (2019) conducted relevant studies. The theory of modeling categorical time series in vector form is gradually being refined, but there usually exist a large number of parameters to be estimated resulting in high computational costs. Therefore, the application of this type of model is difficult to implement. In contrast, our proposed model guarantees a time-varying feedback mechanism without high computational cost.

The second contribution of the proposed model is that it breaks, to some degree, the restriction of pre-defined distribution for models of bounded count time series or categorical time series. A large number of studies have been devoted to modeling bounded count time series with binomial distributions. For example, Weiß (2009), Cui and Lund (2010), Weiß and Pollett (2012) and Weiß and Kim (2013) modeled count data time series with a finite range based on the binomial thinning operator introduced by Steutel and van Harn (1979). And the binomial AR(1) model defined in Al-Osh and Alzaid (1991) based on the hypergeometric thinning operator is also available. Moreover, the integer-valued GARCH models with binomial marginals is an implementable approach, see Weiß and Pollett (2014) and Chen et al. (2020). Liu et al. (2022) proposed the zero-one-inflated bounded Poisson autoregressive model focusing on the normalcy-dominant phenomenon in the data of air quality levels, and achieved the ranking of air quality of major cities in China. However, the true probability structure of real data is complex and hard to determine, so the reliance on distribution type in the modeling process makes the risk of model mis-specification unavoidable, which in turn may result in the invalid estimation. In this paper, the introduction of the discrete distribution, which describes systemic features in the observations that do not vary with time, and weakens the restrictions on distribution to a certain extent. Compared with a fixed binomial marginal distribution or a bounded Poisson distribution, the linear combination of a bounded Poisson distribution and a discrete distribution makes the probability structure more adaptive to the data in practice.

For the estimation and inference, we develop Bayesian inference procedures via Markov chain Monte Carlo (MCMC) methods for the proposed model. Related Bayesian works can be found in Chen et al. (2016), Xu et al. (2020) and Gorgi (2020). The daily air quality level data for three major cities in China, including Beijing, Shanghai and Guangzhou, are analyzed. For this data set, we have two main concerns. The first is the overall difference in air quality in the three cities from 2016 to 2020. The second is the year-to-year change in air quality for each city. Accordingly, we draw conclusions separately.

The organization of this paper is as follows. Section 2 introduces the so called Novel Category model. Section 3 investigates Bayesian inference procedures. Simulations are provided in Sect. 4. Applications to the air quality level data are provided in Sect. 5.

2 Categorical time series models combining dynamic and systemic information

We start by considering a novel distribution with the following probability mass function, named the Novel Category (NC) distribution:

$$P(Y=k)=\delta\frac{\lambda^k/k!}{\Delta(\lambda)}+(1-\delta)\alpha_k,\quad k=0,\ldots,K,$$

where δ is the proportion parameter satisfying $\delta\in[0,1)$, $\lambda>0$ is the intensity parameter, $\Delta(\lambda)=\sum_{i=0}^{K}\frac{\lambda^i}{i!}$, the integer $K\geq 1$ is a given upper bound, $\alpha_i\geq 0$ for $i=0,\ldots,K$ satisfying $\sum_{i=0}^{K}\alpha_i=1$. For some $i^*\in\{0,\ldots,K\}$, α_{i^*} is set to be zero to ensure the identifiability of the model, i.e., there are at most K entries of $\boldsymbol{\alpha}=(\alpha_0,\ldots,\alpha_K)$ are non-zero. The above distribution is denoted as $\mathrm{NC}(\lambda,\boldsymbol{\alpha},\delta,K)$. It can be observed that the NC distribution is a linear combination of the bounded Poisson distribution $BP(\lambda,K)$ and the discrete distribution satisfying $P(Z=k)=\alpha_k$, where the probability mass function of $BP(\lambda,K)$ is $P(X=k)=\frac{\lambda^k/k!}{\Delta(\lambda)}$.

The NC distribution is suitable to fit categorical data. One reliable reason is its finite states $\{0,1,\ldots,K\}$, and the other is that the existence of parameter δ and α_is breaks restrictions of the existing probability structure, and no longer limited to Poisson form or even any other fixed form. The flexible probability structure improves the credibility of fitting sundry categorical data in real life. Meanwhile, the existence of λ facilitates the introduction of dynamic information, which will be explained in detail later.

Consider a categorical time series $\{Y_t\}_{t=1}^n$ that is conditionally NC distributed with time-varying λ_t as follows:

$$Y_t|\mathcal{F}_{t-1}\sim\ \mathrm{NC}\ (\lambda_t,\boldsymbol{\alpha},\delta,K),\tag{2.1}$$

where $\mathcal{F}_{t-1}$ is the σ-field generated by $\{Y_{t-1},Y_{t-2},\cdots\}$. By the above specification, the conditional mean of Y_t is:

$$E[Y_t|\mathcal{F}_{t-1}]=\delta\lambda_t\left(1-\frac{\lambda_t^K/K!}{\Delta(\lambda_t)}\right)+(1-\delta)\sum_{i=0}^{K}i\alpha_i.$$

It can be clearly observed that $E[Y_t|\mathcal{F}_{t-1}]$ is composed of two parts, including the time-varying term

$$\lambda_t\left(1-\frac{\lambda_t^K/K!}{\Delta(\lambda_t)}\right)\tag{2.2}$$

and the constant $\sum_{i=0}^{K}i\alpha_i$, which represent dynamic and systemic information, respectively. Further, we assume autoregressive structure for $\{\lambda_t\}_{t=1}^n$:

$$\lambda_t=\omega+\psi\lambda_{t-1}+\phi Y_{t-1},\tag{2.3}$$

where $\omega>0,\psi\geq 0$ and $\phi\geq 0$. For ease of discussion, only the first-order autoregressive structure for $\{\lambda_t\}$ is investigated. However, the generalization to higher-order autoregression is possible using similar stylized arguments. Note that the structure of (2.3) has been used by Chen et al. (2018) to examine the causal relationship between ambient fine particles and human influenza in Taiwan.

Equations (2.2) and (2.3) imply that the current state Y_t in (2.1) is comprehensively determined by two parts, including the time-varying part affected by past observations and the inherent part that does not change with time. The closer the proportion parameter δ is to zero, the more stable the probability structure, and the lighter the proportion of the part that changes over time. Conversely, The δ reflects the flexibility of the probability structure over time, with closer to 1 indicating higher flexibility. Next, we give an explicit definition.

Definition 2.1 A categorical time series $\{Y_t\}_{t=1}^n$ is said to follow the flexible categorical autoregressive (FCAR) model, if $\{Y_t\}_{t=1}^n$ satisfies (2.1) and (2.3).

The FCAR model introduces an autoregressive feedback mechanism in the linear combination of bounded Poisson and discrete distributions, which lays the foundation for realizing the analysis of the dynamics and systemic features of air quality. The subsequent section is concerned with the estimation of parameters in the FCAR model.

3 Bayesian inference

Before proceeding formally with parameter estimation, it is necessary to specify the dimensionality of the parameters to be estimated. Because of the restriction $\sum_{i=0}^{K}\alpha_i=1$ and the identifiability condition $\alpha_{i^*}\equiv 0$ for some $i^*\in\{0,\ldots,K\}$, only $K-1$ parameters of $\boldsymbol{\alpha}$ need to be estimated. For $i^*=0,1,\ldots,K-1$, we have $\alpha_K=1-\sum_{i=0}^{K-1}\alpha_i$; for $i^*=K$, we have $\alpha_{K-1}=1-\sum_{i=0}^{K-2}\alpha_i$. Define

$$\boldsymbol{\alpha}_s=\begin{cases}(\alpha_0,\alpha_1,\ldots,\alpha_{i^*-1},\alpha_{i^*+1},\ldots,\alpha_{K-1}), & \text{if } i^*=0,1,\ldots,K-2,\\ (\alpha_0,\alpha_1\ldots,\alpha_{K-3},\alpha_{K-2}), & \text{if } i^*=K-1,K.\end{cases}$$

Then, denote the time series of interest and the vector of $(K+3)$ unknown parameters by $\mathbf{Y}=(Y_1,\ldots,Y_n)$ and $\boldsymbol{\theta}=(\omega,\psi,\phi,\boldsymbol{\alpha}_s,\delta)^\top$, respectively.

Then, the log-likelihood function for the FCAR model with $\alpha_{i^*}\equiv 0$ is

$$L(\mathbf{Y}|\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\mathbf{Y}|\boldsymbol{\theta})$$
$$= \begin{cases} \sum_{t=1}^{n} \log\left(\delta \frac{\lambda_t^{Y_t}/Y_t!}{\Delta(\lambda_t)} + (1-\delta)\left(\alpha_{Y_t}\mathbb{I}_{\{Y_t \neq K\}} + \left(1 - \sum_{i=0}^{K-1}\alpha_i\right)\mathbb{I}_{\{Y_t=K\}}\right)\right), & \text{if } i^* \neq K, \\ \sum_{t=1}^{n} \log\left(\delta \frac{\lambda_t^{Y_t}/Y_t!}{\Delta(\lambda_t)} + (1-\delta)\left(\alpha_{Y_t}\mathbb{I}_{\{Y_t \neq (K-1)\}} + \left(1 - \sum_{i=0}^{K-2}\alpha_i\right)\mathbb{I}_{\{Y_t=K-1\}}\right)\right), & \text{if } i^* = K. \end{cases}$$

For simplicity of exposition, we rearrange $\boldsymbol{\theta}$ into three parts such that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\theta}_3)^\top$, where $\boldsymbol{\theta}_1 = (\omega, \psi, \phi)^\top$, $\boldsymbol{\theta}_2 = \boldsymbol{\alpha}_s^\top$ and $\boldsymbol{\theta}_3 = \delta$. For each $m = 1, 2$ and 3, the conditional posterior for $\boldsymbol{\theta}_m$ is proportional to the log-likelihood function multiplied by the prior density of $\boldsymbol{\theta}_m$,

$$P(\boldsymbol{\theta}_m|\mathbf{Y}, \boldsymbol{\theta}_{\breve{m}}) \propto L(\mathbf{Y}|\boldsymbol{\theta})P(\boldsymbol{\theta}_m), \tag{3.1}$$

where $\boldsymbol{\theta}_{\breve{m}}$ is the vector of all unknown parameters except $\boldsymbol{\theta}_m$, and $P(\boldsymbol{\theta}_m)$ is the prior density.

The choices of priors are not unique, but usually non-informative ones are appropriate, see Chen et al. (2016) and Xu et al. (2020). Specially, for $m = 1, 2$ and 3, we use indicator functions $\mathbb{I}_{\{\boldsymbol{\theta}_m \in \Omega_m\}}$ as uniform priors for $\boldsymbol{\theta}_m$, where

Ω_1 : $\omega, \phi, \psi > 0$ and $\phi + \psi < 1$;
Ω_2 : (the sum of each element of $\boldsymbol{\alpha}_s$) ≥ 1 and (each element of $\boldsymbol{\alpha}_s$) ≤ 1;
Ω_3 : $0 \leq \delta < 1$.

These generate flat priors on the parameters under required constraints. The non-standard posterior distributions (3.1) prompt us to adopt MCMC methods to fulfill computational inference, where the MC samples for groups of parameters are sampled successively from their conditional posterior distributions. To draw samples from the conditional posterior distributions with faster convergence and better mixing, we apply the random-walk Metropolis-Hastings in the first M iterations and the independent-kernel Metropolis-Hastings in the subsequent $N - M$ iterations. We complete the parameter estimation of the FCAR model along the lines described above and refer to Chen et al. (2016) for details.

4 Simulations

To examine the effectiveness of the proposed MCMC methods, we investigate the finite sample performance by Monte Carlo simulations in this section. The following three data generating processes (DGPs) of various sample sizes ($T = 300, 500, 1000$) are considered:

- DGP 1: Y_t follows the FCAR model with $K = 5$, $\alpha_1 = 0$ and
$$(\omega, \psi, \phi, \alpha_0, \alpha_2, \alpha_3, \alpha_4, \delta) = (0.3, 0.35, 0.2, 0.1, 0.55, 0.15, 0.1, 0.6);$$
- DGP 1: Y_t follows the FCAR model with $K = 5$, $\alpha_2 = 0$ and
$$(\omega, \psi, \phi, \alpha_0, \alpha_1, \alpha_3, \alpha_4, \delta) = (0.3, 0.35, 0.2, 0.1, 0.55, 0.15, 0.1, 0.6);$$
- DGP 3: Y_t follows the FCAR model with $K = 5$, $\alpha_1 = 0$ and
$$(\omega, \psi, \phi, \alpha_0, \alpha_2, \alpha_3, \alpha_4, \delta) = (0.3, 0.35, 0.2, 0.1, 0.55, 0.15, 0.1, 0.7).$$

We simulate 500 replications from each of the three DGPs. The sample of iterations in the random-walk Metropolis-Hastings is selected as $M = 10{,}000$ and the total sample of iterations is $N = 30{,}000$. Only $N - M$ iterations of the independent-kernel Metropolis-Hastings in every sample period is used for inference. The simulation results of DGPs 1, 2 and 3 are reported in Tables 1, 2 and 3, respectively. The true value, the average posterior mean, median, standard deviation (Std.), the posterior 2.5 and 97.5 percentiles are reported in each column from left to right in tables, and the last two items constitute a 95% credible interval (CI).

For all three DGPs and three sample sizes, the biases of the posterior means and the corresponding true values are reasonably small, as are the biases of the posterior median and the corresponding true values. This implies that both the posterior mean and posterior median estimators are applicable in the FCAR models. The standard deviations of the intercept parameter ω and δ are acceptably larger than that of other parameters. And as expected, all standard deviations decrease as the sample size increases. Moreover, all true values are covered by the 95% CI, and both the posterior 2.5 and 97.5 percentiles are closer to the true values with increase of the sample size. All the above results indicate that the Bayesian method is effectively applicable to the estimation of unknown parameters in the FCAR model.

5 Empirical analysis

In this section, we study the daily air quality level data for three major cities in China, including Beijing, Shanghai and Guangzhou. The air quality level is quantized by Chinese government into six categories: '0' stands for 'excellent'; '1' stands for 'good'; '2' stands for 'slightly polluted'; '3' stands for 'moderately polluted'; '4' stands for 'heavily polluted' and '5' stands for 'severely polluted'.

Table 1 Summary of estimation results for DGP 1

	True value	Mean	Median	Std.	2.5%	97.5%
$n = 300$						
ω	0.30	0.5837	0.5641	0.3044	0.0815	1.1805
ψ	0.35	0.4046	0.4053	0.2091	0.0395	0.7834
ϕ	0.20	0.1571	0.1513	0.0710	0.0348	0.3136
α_0	0.10	0.2164	0.2164	0.1069	0.0228	0.4274
α_2	0.55	0.5197	0.5207	0.0779	0.3645	0.6719
α_3	0.15	0.0751	0.0686	0.0491	0.0044	0.1833
α_4	0.10	0.0682	0.0653	0.0375	0.0070	0.1495
δ	0.60	0.6568	0.6567	0.0673	0.5262	0.7876
$n = 500$						
ω	0.30	0.4192	0.3889	0.2151	0.0872	0.8927
ψ	0.35	0.2992	0.2960	0.1530	0.0319	0.5992
ϕ	0.20	0.2458	0.2430	0.0680	0.1217	0.3865
α_0	0.10	0.1483	0.1414	0.0866	0.0105	0.3293
α_2	0.55	0.5600	0.5605	0.0590	0.4438	0.6723
α_3	0.15	0.1189	0.1208	0.0439	0.0291	0.2019
α_4	0.10	0.0550	0.0540	0.0229	0.0128	0.1026
δ	0.60	0.6176	0.6160	0.0486	0.5260	0.7155
$n = 1000$						
ω	0.30	0.3478	0.3402	0.1349	0.1103	0.6225
ψ	0.35	0.3458	0.3432	0.1368	0.0911	0.6164
ϕ	0.20	0.2213	0.2203	0.0444	0.1374	0.3118
α_0	0.10	0.1645	0.1626	0.0791	0.0232	0.3247
α_2	0.55	0.5550	0.5542	0.0508	0.4560	0.6552
α_3	0.15	0.0872	0.0876	0.0301	0.0261	0.1469
α_4	0.10	0.1004	0.1002	0.0207	0.0605	0.1409
δ	0.60	0.6208	0.6208	0.0340	0.5542	0.6849

Table 2 Summary of estimation results for DGP 2

	True value	Mean	Median	Std.	2.5%	97.5%
$n = 300$						
ω	0.40	0.4334	0.4076	0.2188	0.0906	0.9105
ψ	0.35	0.3768	0.3791	0.1868	0.0374	0.7162
ϕ	0.20	0.2355	0.2200	0.1029	0.0730	0.4751
α_0	0.10	0.1850	0.1918	0.0931	0.0147	0.3470
α_1	0.55	0.5326	0.5317	0.0592	0.4188	0.6530
α_3	0.15	0.0966	0.0927	0.0441	0.0213	0.1933
α_4	0.10	0.0805	0.0764	0.0344	0.0250	0.1583
δ	0.60	0.5444	0.5451	0.0960	0.3652	0.7220
$n = 500$						
ω	0.40	0.3550	0.3409	0.1351	0.1299	0.6518
ψ	0.35	0.3202	0.3215	0.1422	0.0497	0.5946
ϕ	0.20	0.2636	0.2563	0.0746	0.1377	0.4314
α_0	0.10	0.1344	0.1303	0.0805	0.0075	0.2929
α_1	0.55	0.5507	0.5499	0.0539	0.4470	0.6575
α_3	0.15	0.1304	0.1281	0.0393	0.0600	0.2131
α_4	0.10	0.0650	0.0626	0.0235	0.0259	0.1180
δ	0.60	0.6037	0.6083	0.0707	0.4576	0.7295
$n = 1000$						
ω	0.40	0.3414	0.3341	0.1046	0.1628	0.5652
ψ	0.35	0.3409	0.3405	0.1162	0.1123	0.5628
ϕ	0.20	0.2585	0.2542	0.0530	0.1664	0.3753
α_0	0.10	0.0959	0.0896	0.0619	0.0048	0.2244
α_1	0.55	0.5862	0.5856	0.0392	0.5113	0.6628
α_3	0.15	0.1351	0.1344	0.0294	0.0797	0.1944
α_4	0.10	0.0780	0.0768	0.0192	0.0437	0.1185
δ	0.60	0.6103	0.6143	0.0495	0.5054	0.6962

Naturally, the air quality level of each city forms a categorical time series.

Each sample set we considered covers 5 years of data from January 1, 2016 to December 31, 2020, with a total of 1827 observations. We have two concerns about this data set. The first is the overall difference in the air quality of three cities from 2016 to 2020. The second is how the air quality of each city changes year by year. The distinction between systemic and dynamic features is focused on during the analysis. Systemic features we considered is the systemic part of seasonal factors including suspended dust, coal combustion, biomass burning and secondary particulate matter, while industrial dust, vehicle emissions and the non-systemic part of seasonal factors are considered to be dynamic.

In advance, we report the plots of the ordinal Cohen's $\kappa(h)$ of three cities in Fig. 1, which is a measure of serial dependence in categorical time series defined in Weiß (2020). The slow decay with increasing time lag h implies that the data are consistent with the autoregressive structure (2.3).

It must be emphasized that when facing with real data, we cannot determine in advance which i^* satisfying $\alpha_{i^*} = 0$, otherwise the rationality of the model will be weakened. Therefore, in the empirical analysis, for each group of data, we will set $\alpha_{i^*} = 0$ in turn for $i^* = 0, \ldots, 5$ and then generate 6 candidate models, from which the one with the largest likelihood function will be selected as the final model suitable for that group of data.

To study the first concern, we use the FCAR model to fit 1827 observations of each city, respectively, and summarize results in Table 4, including the posterior mean, standard deviation, the posterior 2.5 and 97.5 percentiles. Based on these results, we elaborated the following three conclusions:

(1) The order of size of the proportion parameter δ in the three models is that Beijing (0.7182) > Guangzhou (0.5468) > Shanghai (0.5088). This implies that the

Table 3 Summary of estimation results for DGP 3

	True value	Mean	Median	Std.	2.5%	97.5%
$n = 300$						
ω	0.40	0.5664	0.5551	0.2521	0.1334	1.0495
ψ	0.35	0.3590	0.3553	0.2019	0.0253	0.7351
ϕ	0.20	0.1504	0.1454	0.0655	0.0370	0.2945
α_0	0.10	0.2310	0.2291	0.1249	0.0201	0.4760
α_2	0.55	0.5111	0.5122	0.0991	0.3123	0.6999
α_3	0.15	0.0617	0.0521	0.0462	0.0030	0.1719
α_4	0.10	0.0623	0.0574	0.0372	0.0066	0.1495
δ	0.60	0.7330	0.7332	0.0626	0.6109	0.8545
$n = 500$						
ω	0.40	0.2965	0.2822	0.1253	0.0932	0.5813
ψ	0.35	0.3491	0.3526	0.1300	0.0878	0.5928
ϕ	0.20	0.2755	0.2741	0.0542	0.1731	0.3851
α_0	0.10	0.0987	0.0802	0.0782	0.0036	0.2910
α_2	0.55	0.5659	0.5689	0.0693	0.4237	0.6953
α_3	0.15	0.1305	0.1310	0.0491	0.0319	0.2264
α_4	0.10	0.1246	0.1219	0.0371	0.0595	0.2067
δ	0.70	0.7427	0.7411	0.0453	0.6566	0.8342
$n = 1000$						
ω	0.40	0.3359	0.3244	0.1144	0.1451	0.5902
ψ	0.35	0.3568	0.3614	0.1142	0.1201	0.5681
ϕ	0.20	0.2412	0.2399	0.0411	0.1640	0.3250
α_0	0.10	0.1358	0.1261	0.0867	0.0072	0.3214
α_2	0.55	0.5651	0.5672	0.0594	0.4455	0.6740
α_3	0.15	0.1129	0.1140	0.0386	0.0345	0.1857
α_4	0.10	0.0940	0.0929	0.0234	0.0519	0.1431
δ	0.70	0.7242	0.7238	0.0338	0.6582	0.7911

probability structure of Beijing's daily air quality level has the most changes over time and is relatively unstable. Among the three cities, the probability structure of Shanghai is relatively stable. Beijing's air quality is most heavily influenced by these dynamics, including industrial dust, vehicle emissions and the non-systemic part of seasonal factors.

(2) From the comparison of the values of the three sets of parameters α_is ($i = 1, 2, 3, 4, 5$), it can be seen that the overall air quality of Shanghai is better than that of Beijing, but second to that of Guangzhou. The reasonable principle is the more probability is concentrated on α_i with smaller i, the better air quality in the city. It is well documented that soil dust and road dust contribute the most to PM10 in the urban atmosphere. In northern cities, the contribution of soil dust and road dust to PM10 concentration is higher than that in southern cities. In general, cities in the North China Plain, including Beijing, are polluted by dust emissions from industry, roads and buildings because diffusion conditions are not as good as those in the Northeast and Northwest (World Bank 2012)

(3) For the time-varying part, $\psi + \phi > 0.8$ occurs in all three cities, which is consistent with the expectation that the air quality of the day should be influenced by the previous day. The larger intercept term (0.3145) indicates that Beijing is at a disadvantage among these three cities.

(4) We compare FCAR model with two other models, including the zero-one-inflated bounded Poisson autoregressive (ZOIBPAR) model in Liu et al.

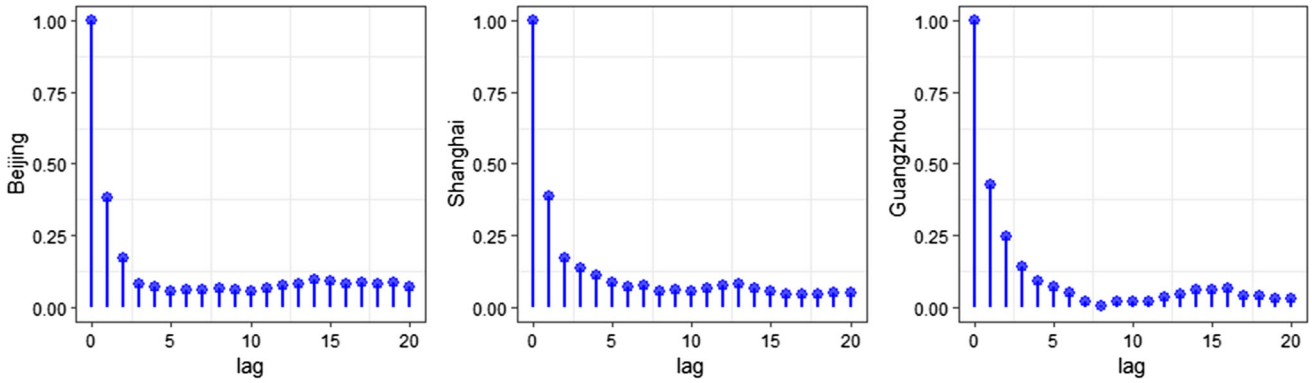

Fig. 1 The ordinal Cohen's $\kappa(h)$ of three cities with lags $h = 0, 1, \ldots, 20$

Table 4 Summary of estimation results for the FCAR Model

	Beijing					Shanghai					Guangzhou			
	$\alpha_5 = 0$					$\alpha_4 = 0$					$\alpha_4 = 0$			
	Mean	Std.	2.5%	97.5%		Mean	Std.	2.5%	97.5%		Mean	Std.	2.5%	97.5%
ω	0.3145	0.0493	0.2210	0.4148	ω	0.0184	0.0137	0.0010	0.0525	ω	0.0116	0.0091	0.0011	0.0341
ψ	0.0217	0.0187	0.0007	0.0723	ψ	0.0443	0.0316	0.0023	0.1229	ψ	0.0648	0.0253	0.0274	0.1124
ϕ	0.8283	0.0470	0.7298	0.9226	ϕ	0.8170	0.0385	0.7327	0.8728	ϕ	0.8030	0.0378	0.7398	0.8559
α_0	0.0246	0.0196	0.0008	0.0729	α_0	0.0065	0.0054	0.0002	0.0208	α_0	0.0064	0.0048	0.0007	0.0194
α_1	0.7453	0.0375	0.6705	0.8201	α_1	0.8503	0.0177	0.8171	0.8824	α_1	0.9033	0.0166	0.8690	0.9413
α_2	0.2146	0.0337	0.1440	0.2764	α_2	0.1358	0.0165	0.1072	0.1680	α_2	0.0865	0.0157	0.0500	0.1216
α_3	0.0111	0.0099	0.0005	0.0366	α_3	0.0055	0.0034	0.0011	0.0134	α_3	0.0027	0.0017	0.0002	0.0061
δ	0.7182	0.0303	0.6570	0.7776	δ	0.5088	0.0173	0.4746	0.5430	δ	0.5468	0.0135	0.5284	0.5752

Table 5 AICs and BICs for models considered

	Beijing		Shanghai		Guangzhou	
	AIC	BIC	AIC	BIC	AIC	BIC
FCAR	4645.645	4669.984	3640.476	3684.559	3423.746	3467.829
ZOIBPAR	4648.157	4675.709	3682.964	3710.516	3455.721	3483.273
BPAR	4748.518	4761.049	4067.796	4080.327	3872.581	3885.113

(2022) and a special FCAR model with $\alpha_i = 0$ for $\forall$ i named the bounded Poisson autoregressive (BPAR) model. In Table 5, it can be seen that FCAR model performs best in terms of Akaike information criterion (AIC) and Bayesian information criterion (BIC), which implies that the inclusion of systematic factors makes sense.

To check the adequacy of the specified model, we calculate estimated standardized Pearson residuals $e_t = \frac{Y_t - E[Y_t|\mathcal{F}_{t-1}]}{\sqrt{\mathrm{Var}[Y_t|\mathcal{F}_{t-1}]}}$ and report the ACF plots of the residuals in Fig. 2. Moreover, the Ljung-Box tests are also applied to check whether or not the residuals appear to be white noise, and the corresponding p-values are shown in Fig. 2. The results in Fig. 2 demonstrate that the fitted FCAR models are adequate.

Next, we fit the data of each city for each whole year (2016–2020) by the FCAR model to characterize the annual changes in air quality in the last 5 years. To elaborate more clearly, we show the results in Figs. 3, 4, 5 and 6, and obtain the following conclusions:

(1) Figure 3 shows δs of the three cities in the past 5 years. It can be seen that the δs in Beijing show a decreasing trend, which implies that the influence of systemic factors on air quality in Beijing is deepening. The δs in Guangzhou increases after 2017, implying that the air quality in Guangzhou is becoming more vulnerable to dynamic factors. The change of δs in Shanghai is relatively mild.

(2) It is obvious that the value of $\alpha_3 + \alpha_4 + \alpha_5$ is decreasing year by year in Fig. 4, which shows that the air quality of Beijing is showing signs of improvement. Based on $\alpha_3 + \alpha_4 + \alpha_5$ in Figs. 5 and 6, we can find that the air quality in Shanghai and Guangzhou has always been better than that in Beijing. From 2016 to 2019, $\alpha_0 + \alpha_1$ of Shanghai is rising, while α_2 is declining. This indicates further optimization of Shanghai's air quality, but with a slight rebound in 2020.

(3) Table 6 reports the posterior means and 95% CIs of $\hat{\omega}, \hat{\phi}$ and $\hat{\psi}$, respectively. The year-on-year changes of $\hat{\omega}$ once again imply that although the air quality of Beijing is the worst among the three cities, its air quality is gradually improving. The slight changes of $\hat{\omega}, \hat{\phi}$ and $\hat{\psi}$ over the 5 years indicate that the internal structure of the dynamic factors affecting air quality in Guangzhou is stable. This reflects the efficiency of Guangzhou's pollution control policy, which is well adaptive to changes in dynamic factors such as vehicle emissions and industrial upgrading.

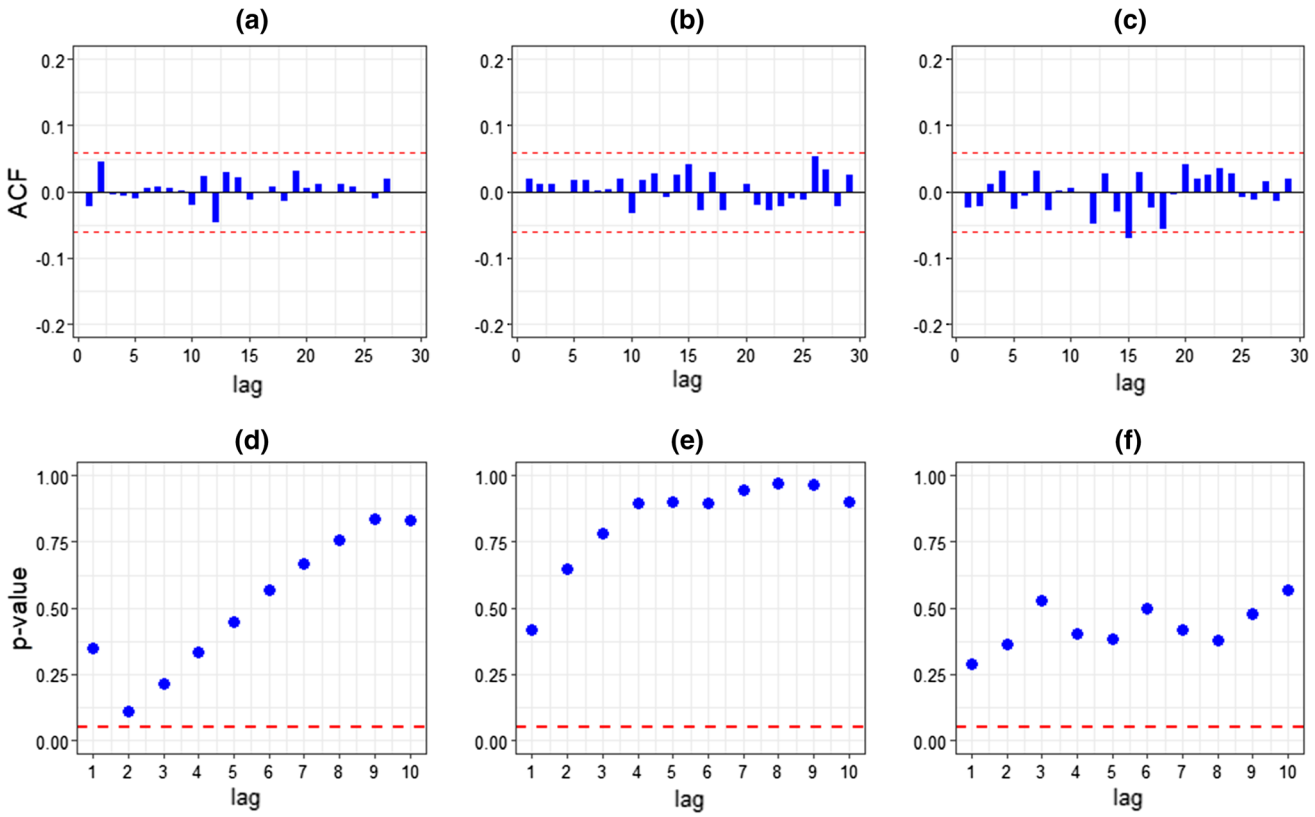

Fig. 2 The ACF plots of residuals (upper row) and the *p*-values of the Ljung-Box tests for residuals (lower row). **a**, **d** Beijing; **b**, **e** Shanghai; **c**, **f** Guangzhou

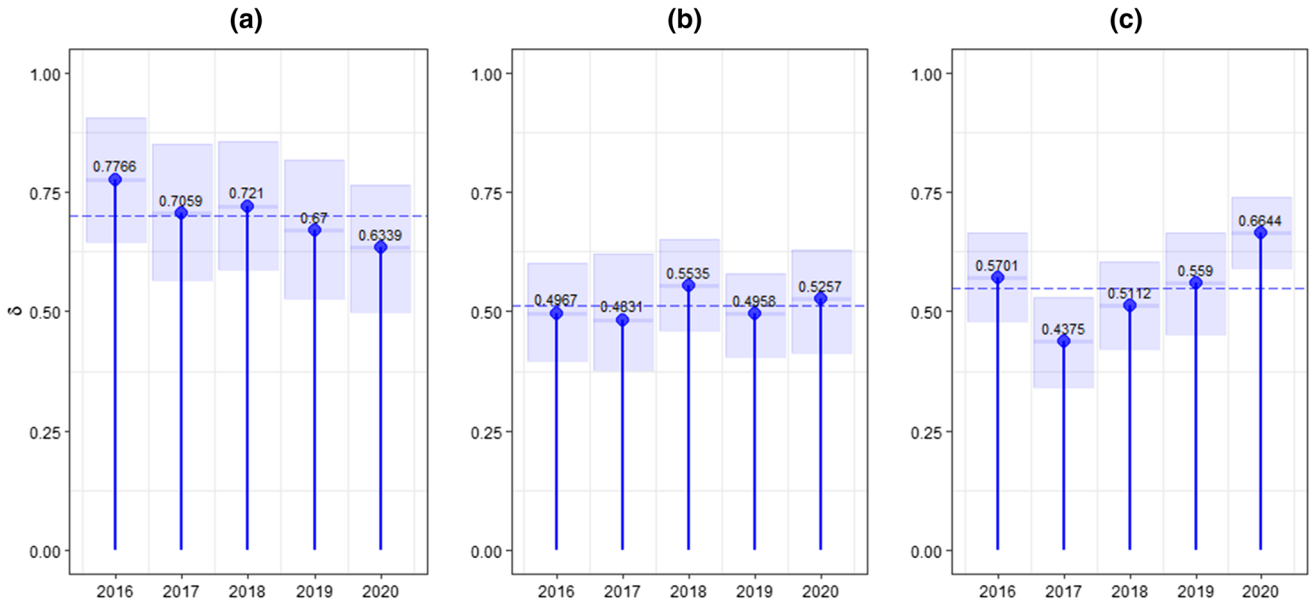

Fig. 3 $\hat{\delta}$s of three cities from 2016 to 2018. **a** Beijing; **b** Shanghai; **c** Guangzhou

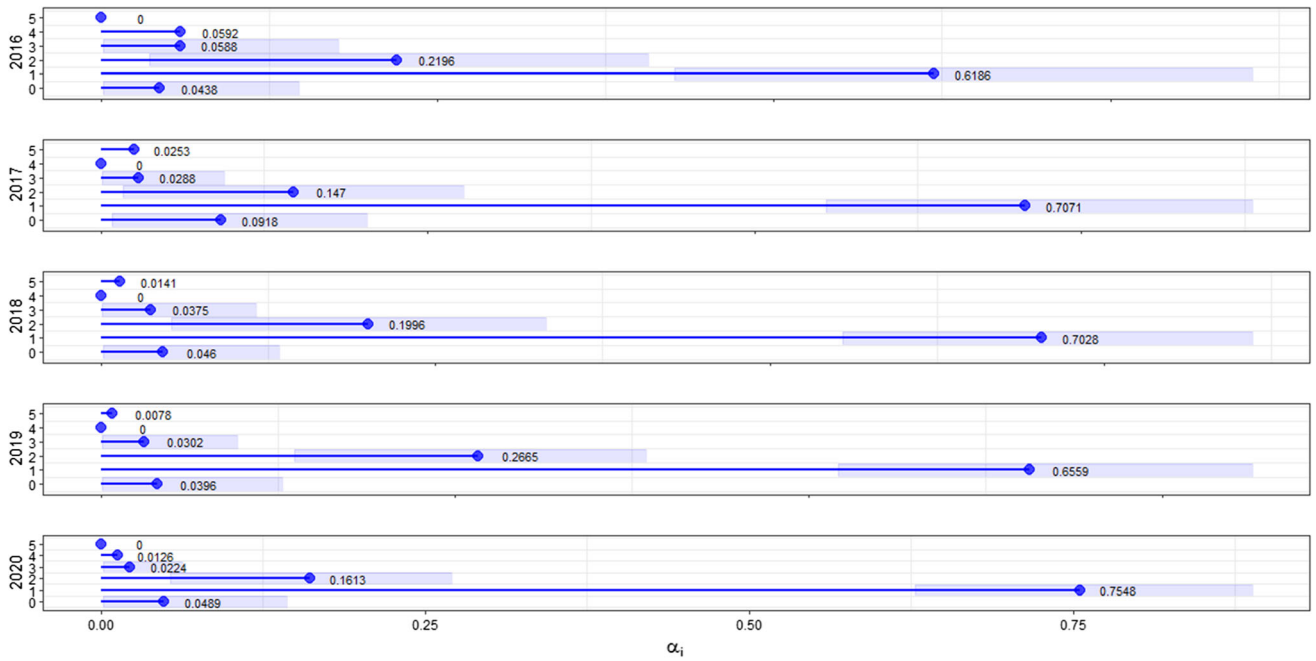

Fig. 4 $\hat{\alpha_i}$ of Beijing from 2016 to 2020, for $i = 0, \ldots, 5$

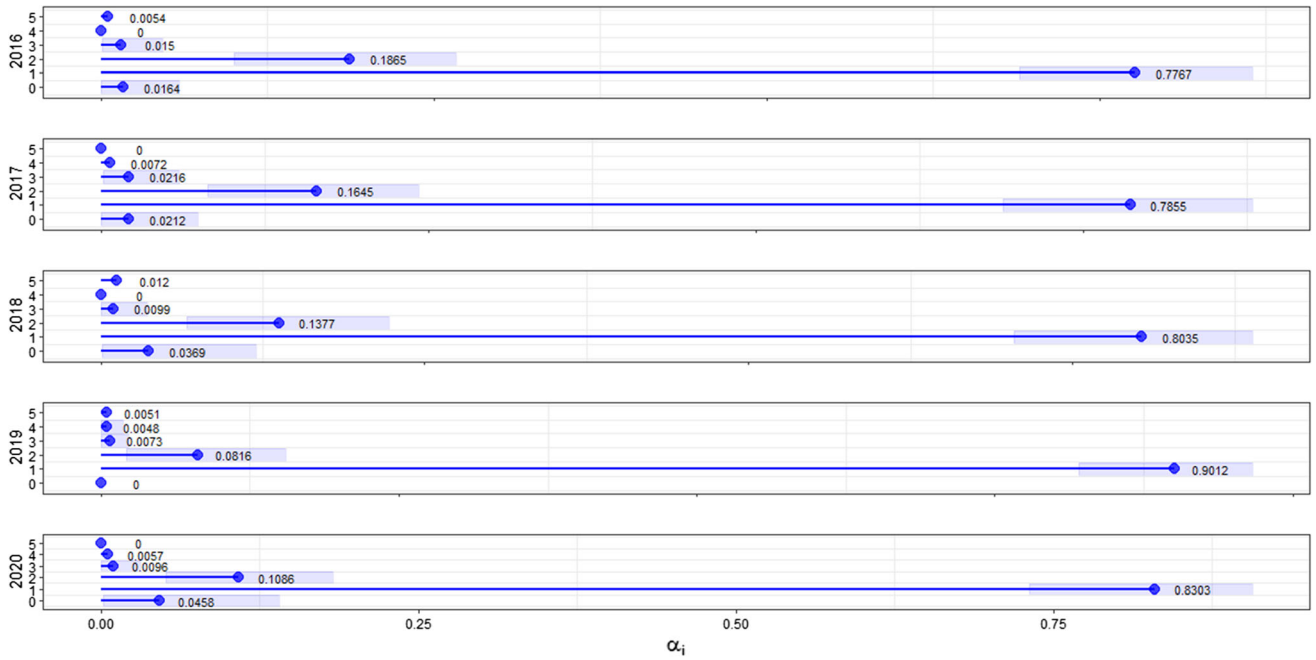

Fig. 5 $\hat{\alpha_i}$ of Shanghai from 2016 to 2020, for $i = 0, \ldots, 5$

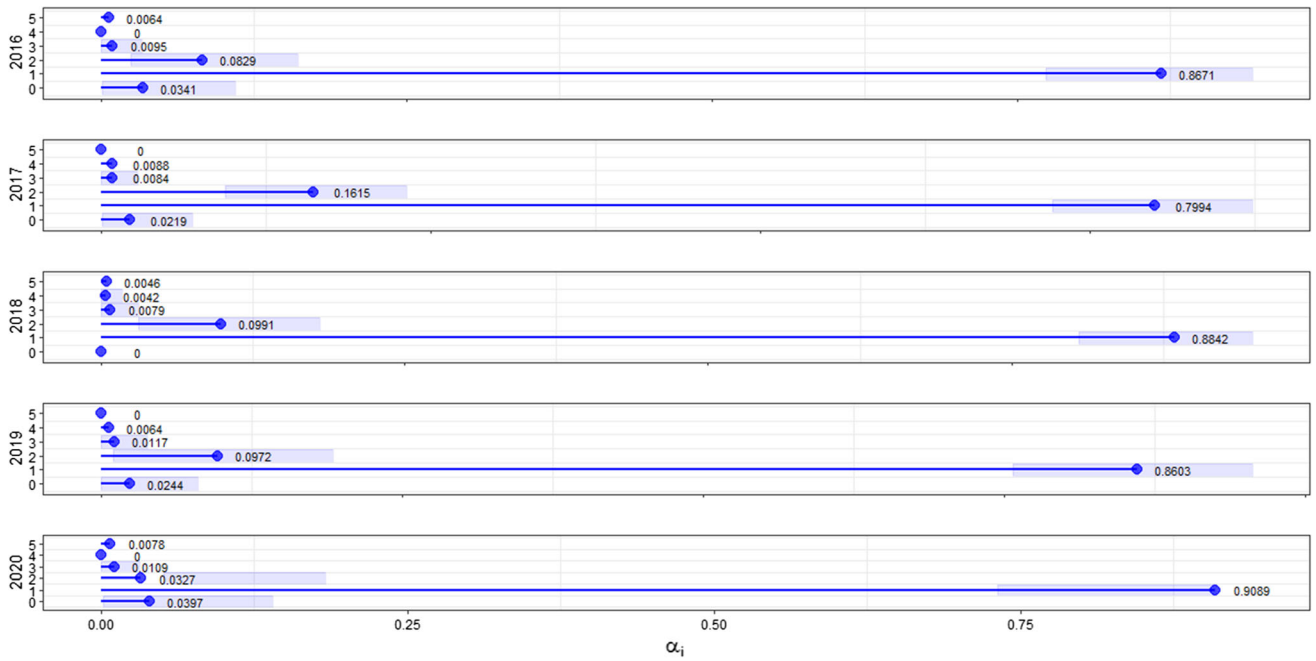

Fig. 6 $\hat{\alpha_i}$ of Guangzhou from 2016 to 2020, for $i = 0, \ldots, 5$

Table 6 Summary of estimation results for the CCAR Model

		2016	2017	2018	2019	2020
Beijing	ω	0.4491	0.3980	0.2887	0.2905	0.1504
	CI(ω)	(0.2474, 0.6793)	(0.2094, 0.6139)	(0.1226, 0.5013)	(0.1008, 0.4889)	(0.0123, 0.3382)
	ψ	0.0240	0.0634	0.0781	0.0462	0.1053
	CI(ψ)	(0.0006, 0.0818)	(0.0018, 0.1929)	(0.0029, 0.2298)	(0.0014, 0.1528)	(0.0056, 0.2741)
	ϕ	0.8529	0.7745	0.8047	0.7035	0.7078
	CI(ϕ)	(0.6668, 0.9774)	(0.5959, 0.9414)	(0.6150, 0.9453)	(0.5489, 0.8790)	(0.5306, 0.8903)
Shanghai	ω	0.0644	0.0845	0.0639	0.0501	0.0551
	CI(ω)	(0.0031, 0.1939)	(0.0029, 0.2485)	(0.0019, 0.1933)	(0.0020, 0.1737)	(0.0022, 0.1898)
	ψ	0.0582	0.1240	0.1298	0.0924	0.0587
	CI(ψ)	(0.0018, 0.1957)	(0.0043, 0.3156)	(0.0074, 0.3242)	(0.0039, 0.2756)	(0.0022, 0.1875)
	ϕ	0.8495	0.7577	0.6695	0.7411	0.6759
	CI(ϕ)	(0.6853, 0.9711)	(0.5690, 0.9460)	(0.4862, 0.8522)	(0.5697, 0.9068)	(0.5040, 0.8557)
Guangzhou	ω	0.0602	0.0629	0.0617	0.0630	0.0242
	CI(ω)	(0.0020, 0.2042)	(0.0013, 0.2001)	(0.0052, 0.2086)	(0.0024, 0.1892)	(0.0011, 0.0715)
	ψ	0.0618	0.0983	0.0742	0.1396	0.0774
	CI(ψ)	(0.0019, 0.1877)	(0.0041, 0.2664)	(0.0027, 0.2367)	(0.0137, 0.3212)	(0.0056, 0.1839)
	ϕ	0.7516	0.7496	0.6959	0.7470	0.7676
	CI(ϕ)	(0.5890, 0.9133)	(0.5647, 0.9298)	(0.5406, 0.8851)	(0.5618, 0.9100)	(0.6175, 0.9117)

Acknowledgements The authors greatly appreciate three anonymous referees for very valuable comments and suggestions that result in a substantial improvement of this paper. Liu's work is supported by China Postdoctoral Science Foundation (No. 2021M701366). Li's work is supported by Natural Science Foundation of Jilin Province (No. 20210101160JC), and Natural Science Foundation of Changchun Normal University. Zhu's work is supported by National Natural Science Foundation of China (Nos. 11871027, 11731015), and Natural Science Foundation of Jilin Province (No. 20210101143JC).

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Human and animal rights This research does not involve Human Participants and/or Animals.

References

Al-Osh MA, Alzaid AA (1991) Binomial autoregressive moving average models. Stoch Models 7:261–282

Andrée PJ (2020) Incidence of COVID-19 and connections with air pollution exposure: evidence from the Netherlands. Policy Research Working Paper 9221. World Bank, Washington, DC

Chen CWS, So MKP, Li JC, Sriboonchitta S (2016) Autoregressive conditional negative binomial model applied to over-dispersed time series of counts. Stat Methodol 31:73–90

Chen CWS, Hsien YH, Su HC, Wu JJ (2018) Causality test of ambient fine particles and human influenza in Taiwan: age group-specific disparity and geographic heterogeneity. Environ Int 111:354–361

Chen H, Li Q, Zhu F (2020) Two classes of dynamic binomial integer-valued ARCH models. Braz J Probab Stat 34:685–711

Cui Y, Lund R (2010) Inference in binomial AR(1) models. Stat Prob Lett 80:1985–1990

Fokianos K, Kedem B (2003) Regression theory for categorical time series. Stat Sci 18:357–376

Fokianos K, Moysiadis T (2017) Binary time series driven by a latent process. Econ Stat 2:117–130

Fokianos K, Truquet L (2019) On categorical time series models with covariates. Stoch Proc Appl 129:3446–3462

Francq C, Zakoïan J-M (2009) Testing the nullity of GARCH coefficients: correction of the standard tests and relative efficiency comparisons. J Am Stat Assoc 117:1265–1284

Gorgi P (2020) Beta-negative binomial auto-regressions for modelling integer-valued time series with extreme observations. J R Stat Soc Series B 82:1325–1347

Kauppi H, Saikkonen P (2008) Predicting US recessions with dynamic binary response models. Rev Econ Stat 90:777–791

Liu M, Zhu F, Zhu K (2022) Modeling normalcy-dominant ordinal time series: An application to air quality level. J Time Ser Anal. forthcoming, https://doi.org/10.1111/jtsa.12625

Moysiadis T, Fokianos K (2014) On binary and categorical time series models with feedback. J Multivar Anal 131:209–228

Steutel FW, van Harn K (1979) Discrete analogues of self-decomposability and stability. Ann Prob 7:893–899

Stoffer DS, Tyler DE, McDougall AJ (1993) Spectral analysis for categorical time series: scaling and the spectral envelope. Biometrika 80:611–622

Weiß CH (2009) A new class of autoregressive models for time series of binomial counts. Commun Stat Theor Method 38:447–460

Weiß CH (2020) Distance-based analysis of ordinal data and ordinal time series. J Am Stat Assoc 115:1189–1200

Weiß CH, Kim HY (2013) Binomial AR(1) processes: moments, cumulants, and estimation. Statistics 47:494–510

Weiß CH, Pollett PK (2012) Chain binomial models and binomial autoregressive processes. Biometrics 68:815–824

Weiß CH, Pollett PK (2014) Binomial autoregressive processes with density-dependent thinning. J Time Ser Anal 35:115–132

World Bank (2012) Integrated air pollution management in China: developing particulate matter control. Washington, DC. https://openknowledge.worldbank.org/handle/10986/11913 License: CC BY 3.0 IGO

World Bank (2020) The Global Health Cost of Ambient PM2.5 Air Pollution. World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/35721 License: CC BY 3.0 IGO

World Bank and Institute for Health Metrics and Evaluation (2016) The cost of air pollution: strengthening the economic case for action. World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/25013 License: CC BY 3.0 IGO

Xu X, Chen Y, Chen CWS, Lin X (2020) Adaptive log-linear zero-inflated generalized Poisson autoregressive model with applications to crime counts. Ann Appl Stat 14:1493–1515