



OPEN ACCESS

EDITED BY
Marwan Osman,
Cornell University, United States

REVIEWED BY
Rohan Kumar Raman,
ICAR-Research Complex for Eastern
Region, India
Imad Al Kassaa,
Fonterra, New Zealand
Ananda Kosuvaripalli,
JSS Banashankari Arts, Commerce and
Shanthi Kumar Gubbi Science College,
India

*CORRESPONDENCE
Asha Kamath
✉ asha.kamath@manipal.edu

SPECIALTY SECTION
This article was submitted to
Infectious Diseases: Epidemiology and
Prevention,
a section of the journal
Frontiers in Public Health

RECEIVED 15 June 2022
ACCEPTED 20 December 2022
PUBLISHED 10 January 2023

CITATION
Lobo J, Kamath A and Kalwaje
Eshwara V (2023) Degenerate Beta
autoregressive model for proportion
time-series with zeros or ones: An
application to antimicrobial resistance
rate using R shiny app.
Front. Public Health 10:969777.
doi: 10.3389/fpubh.2022.969777

COPYRIGHT
© 2023 Lobo, Kamath and Kalwaje
Eshwara. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Degenerate Beta autoregressive model for proportion time-series with zeros or ones: An application to antimicrobial resistance rate using R shiny app

Jevitha Lobo¹, Asha Kamath^{1*} and Vandana Kalwaje Eshwara²

¹Department of Data Sciences, Prasanna School of Public Health, Manipal Academy of Higher Education (MAHE), Manipal, Karnataka, India, ²Department of Microbiology, Kasturba Medical College of Manipal, Manipal Academy of Higher Education (MAHE), Manipal, Karnataka, India

Background: Antimicrobial resistance has emerged as one of the foremost public health troubles of the 21st century. This has ended in a public health disaster of the global situation, which threatens the exercise of present-day remedy. There is an urgent requirement for a cost-effective strategy to reduce antimicrobial resistance. Infectious disease control researchers most often analyze and predict antimicrobial resistance rate data that includes zeros or ones. Commonly used time-series analysis such as autoregressive moving average model is inappropriate for such data and may arrive at biased results.

Objective: This study aims to propose a time-series model for continuous rates or proportions when the interval of series includes zeros or ones and compares the model with existing models.

Data: The *Escherichia coli*, isolated from blood cultures showing variable susceptibility results to different antimicrobial agents, has been obtained from a clinical microbiology laboratory of a tertiary care hospital, Udupi district, Karnataka, during the years between 2011 and 2019.

Methodology: We proposed a Degenerate Beta Autoregressive model which is a mixture of continuous and discrete distributions with probability mass at zero or one. The proposed model includes autoregressive terms along with explanatory variables. The estimation is done using maximum likelihood with a non-linear optimization algorithm. An R shiny app has been provided for the same.

Results: The proposed Degenerate Beta Autoregressive model performed well compared to the existing autoregressive moving average models. The forecasted antimicrobial resistance rate has been obtained for the next 6 months.

Conclusion: The findings of this article could be beneficial to the infectious disease researchers to use an appropriate time-series model to forecast the resistance rate for the future and to have better or advance public health policies to control the rise in resistance rate.

KEYWORDS

Beta distribution, time-series model, mixture distribution, rates, proportions, inflated distribution, AMR, resistance

1. Introduction

Antimicrobial resistance (AMR) is a serious problem in many developing countries. The World Health Organization (WHO) has categorized AMR as a serious health problem affecting the patients of various countries. In 2015, the WHO surveyed in its six regions called country situation analysis to examine the current practices and to determine the gaps which increase antimicrobial resistance. Eleven and eight countries from Asian and African continents, respectively, from low- to middle-income participated in the study. The analysis showed that AMR is a major issue in both the continents, and in the South-East Asia, nosocomial infections are of particular concern. The main cause of resistance is the inappropriate use of antimicrobial medicines and poor healthcare facilities (1).

In the past, to study antibiotic usage and resistance time-series, models such as Box-Jenkins ARIMA (autoregressive-integrated moving average) (2) and transfer function models have been used. Some studies used time-series to forecast trend and seasonality of the data with the exponential smoothing technique.

However, there is a limitation of these models in terms of accuracy of prediction or violation of assumptions in the applicability. Hence, this study proposes to develop and test stochastic models to predict antibiotic resistance rate, which helps us in understanding the pattern of resistance to plan strategies for the rational use of antibiotics.

Oftentimes, the data like antimicrobial resistance include data in the interval $[0, 1)$ or $(0, 1]$, when the bacteria is highly susceptible or resistant to the antibiotic. To model data with rates/proportions regression, models are proposed by Ferrari and Cribari-Neto (3), Mitnik and Baek (4), Pumi et al. (5), and Artur and Bazán (6) and the time-series models are proposed by Rocha and Cribari-Neto (7) and Bayer et al. (8). But these models are not appropriate for the proportion data with zeros or ones. To model such kind of data, we look for a mixture of distributions. Ospina and Ferrari (9) and Cribari-Neto and Santos (10) introduced inflated Beta distributions and inflated Kumaraswamy distributions which is a mixture of discrete and continuous distributions. Ospina and Ferrari (11) and Bayer et al. (12) introduced the inflated Beta regression model and inflated Kumaraswamy regression model. Currently, there is scope to use a time-series model for proportion data in the interval $[0, 1)$ or $(0, 1]$.

This paper proposes to model the time-series data in the interval $[0, 1)$ or $(0, 1]$ using a mixture of Degenerate and Beta distributions through a frequentist approach. This study is an extension of β ARMA model proposed by Rocha and Cribari-Neto (7), where instead of Beta distribution, inflated Beta distribution is incorporated. The proposed model is compared with the ARIMA model.

The developed model is illustrated with an application based on antimicrobial resistance (AMR) data. Rates of *Escherichia coli* (*E. coli*) isolated from blood cultures showing variable susceptibility results to different antimicrobial agents are considered for the modeling in this study. *E. coli* is a gram-negative bacteria, most regularly isolated in patients with blood stream infection (BSI), and in severe instances, it may cause loss of life. The rates of BSI have accelerated steadily in recent years (13). However, knowledge of future resistance rate using forecasting may help in recommending new interventions or policy recommendations in hospital settings.

2. Materials and methods

2.1. Data description

The variable susceptibility of *Escherichia coli* (*E. coli*), isolated from blood cultures showing variable susceptibility results to different antimicrobial agents, has been obtained from a clinical microbiology laboratory of a tertiary care hospital, Udupi district, Karnataka, during the years between 2011 and 2019. Institutional ethical clearance was obtained from Kasturba Medical College and Kasturba Hospital Institutional Ethics Committee (IEC no. 832/2019). The laboratory generally receives more than 10,000 blood culture tests annually from patients who have been suspected of bacteremia/sepsis. The blood cultures yielded positive results in 20–30% of cases and *E. coli* is the most common among several other bacteria causing bacteremia in our patient population. We retrieved the data of antimicrobial susceptibility test (AST) results of this bacteria from the electronic records of the laboratory. Antimicrobial susceptibility tests were performed using the Kirby Bauer disc diffusion method until 2014 and by the Vitek-2 automated method after 2015, both of which are accepted and standardized test methods according to the Clinical Laboratory Standards Institute (Ref:CLSI supplement M100. Wayne, PA: Clinical and Laboratory Standards Institute; 2012). In the former method, the bacteria was considered resistant to the antimicrobial agent amoxicillin-clavulanic acid, based on the ESBL phenotypic test, although the individual agent appeared susceptible. However, in the latter method, results were reported unchanged as detected in the automated system. Laboratory had tested *E. coli* against different antimicrobial agents and the results were interpreted as either susceptible or resistant. The data include information on the monthly number of *E. coli* isolates obtained and the total number of isolates resistant to the different antibiotics which includes a total of 108 time points. The resistance proportion is calculated as the number of *E. coli* isolates that were resistant to a particular antimicrobial agent/total number of *E. coli* tested during that time interval. This study considered the data on the antibiotic amoxicillin-clavulanic acid (AMC) and Cefoperazone-sulbactam (CSL) since it met the requirements

of the model (i.e., time dependency and in the interval [0,1) or (0,1]).

2.2. Degenerate Beta autoregressive (DeβAR) model

2.2.1. Beta distribution

The parameterized probability density function of random variable X of Beta distribution is

$$f(x; \mu, \zeta) = \frac{\Gamma(\zeta)}{\Gamma(\mu\zeta)\Gamma((1-\mu)\zeta)} x^{\mu\zeta-1}(1-x)^{(1-\mu)\zeta-1}, 0 < x < 1 \tag{1}$$

Here, $0 < \mu < 1$ and $\zeta > 0$. The mean and variance of the random variable X is, $E(X = x) = \mu$ and $V(X = x) = \frac{V(\mu)}{1+\zeta}$; where $V(\mu) = \mu(1-\mu)$. Hence, here μ and ζ act as distribution mean and precision parameters. Here, when value of μ is fixed, as the value of ζ increases, the variance of x decreases.

2.2.2. Degenerate distribution

A Degenerate distribution is a one-point distribution where a random variable X has a single possible value.

The probability mass function of random variable X can be written as follows:

$$P(X = x) = \begin{cases} 1 & \text{if } x = c; \\ 0 & \text{elsewhere} \end{cases} \tag{2}$$

That is, a random variable, X, is degenerate if, for some constant, c, $P(X = c) = 1$.

This distribution has a single parameter, c, and it ranges from $-\infty$ to ∞ .

2.2.3. Inflated Beta distribution

Ospina and Ferrari (9) introduced inflated Beta distributions which is a mixture of Degenerate and Beta distributions. The probability density function is

$$bi_c(x; \omega, \mu, \zeta) = \begin{cases} \omega & \text{if } x = c; \\ (1-\omega)f(x; \mu, \zeta) & \text{if } x \in (0,1) \end{cases} \tag{3}$$

where, $0 < \omega < 1$ is the mixture parameter, $f(x; \mu, \zeta)$ is the p.d.f of Beta distribution and $c=0$ or 1 known value which follows Degenerate distribution. The cumulative distribution function (c.d.f) of mixture distribution is given by

$$BI_c(x; \omega, \mu, \zeta) = \omega \mathcal{I}_{[c,1]}(x) + (1-\omega)F(x; \mu, \zeta)$$

Where, $\mathcal{I}_A(x)$ is an indicator function that equals 1 if $x \in A$ and 0 if $x \notin A$. Here, $F(., \mu, \zeta)$ is the c.d.f. of the beta distribution.

The r^{th} moment and variance of inflated beta distribution is

$$E(x^r) = \omega c + (1-\omega)\mu_r$$

$$Var(x) = (1-\omega)\frac{V(\mu)}{\zeta+1} + \omega(1-\omega)(c-\mu)^2$$

where, $\mu_r = (\mu\zeta)_r/\zeta_r$.

2.2.4. Beta autoregressive moving average model

Rocha and Cribari-Neto (7) introduced β ARMA model to fit continuous time-series data in the interval (0, 1), which follows Beta distribution.

The proposed β ARMA(p, q) model is

$$g(\mu_t) = \beta_0 + s'_t\beta + \sum_{i=1}^p \phi_i \{g(x_{t-i}) - s'_{t-i}\beta\} + \sum_{j=1}^q \alpha_j r_{t-j} \tag{4}$$

where, $g(.)$ is the link function, β_0 is the intercept term, s_t 's are the regressor variables, and $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$ are set of parameters of regressors. The ϕ 's and the α 's are the autoregressive (AR) and moving average (MA) parameters, p and q are the AR and MA orders, and r_t is an error term, respectively.

In case of a real-life scenario when time-series data includes zeros or ones, it is challenging to use this model by assuming Beta distribution. To model such kind of continuous time-series data, we replaced Beta distribution with inflated Beta distribution proposed by Ospina and Ferrari (9). The density function of inflated Beta distribution with time can be written as follows:

Let X_t $t=1,2,\dots,n$ be the response variable of proportion data which includes zeros or ones. We assume that the proportion series is conditionally distributed as InBE (μ_t, ζ, ω) (where, InBE stands for "Inflated Beta") with probability density function defined as:

$$f_{X_t}(x_t|\mathcal{F}_{t-1}) = \omega \mathcal{I}_{(x_t=c)} + (1-\omega)\frac{\Gamma(\zeta)}{\Gamma(\mu_t\zeta)\Gamma((1-\mu_t)\zeta)} x_t^{\mu_t\zeta-1}(1-x_t)^{(1-\mu_t)\zeta-1} \tag{5}$$

or equivalently

$$f_{X_t}(x_t|\mathcal{F}_{t-1}) = \begin{cases} \omega & \text{if } x = c; \\ (1-\omega)\frac{\Gamma(\zeta)}{\Gamma(\mu_t\zeta)\Gamma((1-\mu_t)\zeta)} x_t^{\mu_t\zeta-1}(1-x_t)^{(1-\mu_t)\zeta-1} & \text{if } x \in (0,1) \end{cases}$$

which is a mixture of Beta and Degenerate distributions. Here, \mathcal{F}_{t-1} is the previous information set of response series. When the mixture parameter $\omega = 0$, inflated Beta distribution reduces

to Beta distribution. Based on Equations (3) and (4), the mean and variance of distribution can be written as follows:

For [0,1],

$$E(X_t|\mathcal{F}_{t-1}) = (1 - \omega)\mu_t$$

$$V(X_t|\mathcal{F}_{t-1}) = (1 - \omega)\frac{\mu_t(1 - \mu_t)}{\zeta + 1} + \{\omega(1 - \omega)\mu_t^2\}$$

For (0,1],

$$E(X_t|\mathcal{F}_{t-1}) = \omega + (1 - \omega)\mu_t$$

$$V(X_t|\mathcal{F}_{t-1}) = (1 - \omega)\frac{\mu_t(1 - \mu_t)}{\zeta + 1} + \{\omega(1 - \omega)(1 - \mu_t)^2\}$$

2.2.5. Proposed model

The proposed Degenerate Beta Autoregressive (DeβAR) model for the parameters of mixture distribution is

$$\eta_t = g(\mu_t) = s'_t\beta + \sum_{i=1}^p \phi_i\{g(x_{t-i}) - s'_{t-i}\beta\} \quad (6)$$

where, η_t is the linear or non-linear predictor of the model, where $g(\cdot)$ is the link function (we used logit), s_t 's are the regressor variables, and $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$ are the unknown parameters of regressor variables. $\phi_i = (\phi_{i1}, \phi_{i2}, \dots, \phi_{ip})'$ are the autoregressive parameters with order p . Let $\theta = (\beta', \zeta, \omega, \phi_i')$ be vector of unknown parameters with length $k+p+2$.

2.2.6. Parameter estimation

The parameters of the model are estimated by maximizing log-likelihood function.

Here, let $X_t, t=1,2,\dots,n$ be a random variable and \mathcal{F}_{t-1} be the set of past information. Then, the likelihood function of the parameters θ conditioning on the past p observations can be written as follows:

$$L(\theta; x) = \prod_{t=p+1}^n f_{X_t}(x_t|\mathcal{F}_{t-1})$$

The likelihood function for the parameters of Degenerate Beta AR model is given by

$$L(\theta; x) = \prod_{t=p+1}^n \{\omega\mathcal{I}_{x_t=c} + \mathcal{I}_{x_t\in(0,1)}(1 - \omega)\frac{\Gamma(\zeta)}{\Gamma(\mu_t\zeta)\Gamma((1 - \mu_t)\zeta)}x_t^{\mu_t\zeta-1}(1 - x_t)^{(1-\mu_t)\zeta-1}\}$$

Then, the log-likelihood of model is

$$\begin{aligned} \log(L(\theta; x)) &= l(\theta) \\ &= \sum_{t=p+1}^n \mathcal{I}_{x_t=c}\log(\omega) + \sum_{t=p+1}^n \mathcal{I}_{x_t\in(0,1)}\log\{(1 - \omega)\frac{\Gamma(\zeta)}{\Gamma(\mu_t\zeta)\Gamma((1 - \mu_t)\zeta)}x_t^{\mu_t\zeta-1}(1 - x_t)^{(1-\mu_t)\zeta-1}\} \end{aligned}$$

Here, take $\mathcal{I}_{x_t=c} = x_{ct}$, then $l(\theta)$ equals to

$$\begin{aligned} \sum_{t=p+1}^n x_{ct}\log(\omega) + \sum_{t=p+1}^n (1 - x_{ct})\log\{(1 - \omega)\frac{\Gamma(\zeta)}{\Gamma(\mu_t\zeta)\Gamma((1 - \mu_t)\zeta)}x_t^{\mu_t\zeta-1}(1 - x_t)^{(1-\mu_t)\zeta-1}\} \end{aligned}$$

Then, the score function is given by

$$U(\theta) = \frac{\partial}{\partial\theta}l(\theta) = 0$$

i.e., for $l=1,2,\dots,k$

$$\frac{\partial l(\theta)}{\partial\beta_l} = \sum_{t=p+1}^n \frac{\partial l(\theta)}{\partial\mu_t} \frac{\partial\mu_t}{\partial\eta_t} \frac{\partial\eta_t}{\partial\beta_l}$$

Note that, $\frac{\partial\mu_t}{\partial\eta_t} = \mu_t(1 - \mu_t)$ and $\frac{\partial\eta_t}{\partial\beta_l} = s_{tl} - \sum_{i=1}^p \phi_i s_{(t-i)l}$. Then,

$$\begin{aligned} \frac{\partial l(\theta)}{\partial\beta_l} &= \sum_{t=p+1}^n (1 - x_{ct})\zeta [\log(\frac{x_t}{1 - x_t}) - \{\psi(\mu_t\zeta) - \psi((1 - \mu_t)\zeta)\}]\mu_t(1 - \mu_t)(s_{tl} - \sum_{i=1}^p \phi_i s_{(t-i)l}) \end{aligned}$$

Here, let $x_t^* = \log(\frac{x_t}{1 - x_t})$ if $x_t \in (0,1)$ else $x_t^* = 0$ and $\psi(\mu_t\zeta) - \psi((1 - \mu_t)\zeta) = \mu_t^* - \mu_t^*$, where $\psi(\cdot)$ is a digamma function. Then,

$$\frac{\partial l(\theta)}{\partial\beta_l} = \sum_{t=p+1}^n (1 - x_{ct})\zeta(x_t^* - \mu_t^*)\mu_t(1 - \mu_t)(s_{tl} - \sum_{i=1}^p \phi_i s_{(t-i)l})$$

Similarly,

$$\begin{aligned} \frac{\partial l(\theta)}{\partial\zeta} &= \sum_{t=p+1}^n \zeta(1 - x_{ct})\{\mu_t(x_t^* - \mu_t^*) + \log(1 - x_t) - \psi((1 - \mu_t)\zeta) + \psi(\zeta)\} \end{aligned}$$

$$\frac{\partial l(\theta)}{\partial\omega} = \sum_{t=p+1}^n \frac{(x_{ct} - \omega)}{\omega(1 - \omega)}$$

TABLE 1 Simulation results.

Model	Sample size (n) Estimators	100		250		500	
		Bias	\sqrt{MSE}	Bias	\sqrt{MSE}	Bias	\sqrt{MSE}
D β AR(1)	β_0	0.001	0.057	0.0006	0.0364	0.0001	0.026
	ϕ_1	-0.002	0.072	-0.001	0.045	-0.0007	0.031
	ζ	-3.222	12.603	-1.308	6.883	-0.637	4.679
	ω	0.003	0.206	-0.0004	0.032	-0.00007	0.022
D β AR(2)	β_0	0.005	0.110	0.0004	0.048	0.0008	0.034
	ϕ_1	-0.002	0.125	-0.0006	0.053	-3.77E-05	0.037
	ϕ_2	-0.008	0.134	-0.002	0.055	-0.001	0.039
	ζ	-10.306	26.772	-1.758	7.159	-0.881	4.712
	ω	0.0003	0.073	-0.0003	0.032	-0.0001	0.022

For $i=1,2,\dots,p$

$$\frac{\partial l(\theta)}{\partial \phi_i} = \sum_{t=p+1}^n (1 - x_{ct}) \zeta_t (x_t^* - \mu_t^*) \mu_t (1 - \mu_t) (g(x_{t-i}) - s'_{t-i} \beta)$$

The maximum likelihood estimator of θ is obtained by equating $U(\theta) = 0$. Since there exists no closed form solution for these equations, a non-linear optimization algorithm like Newton's method or a Quasi-Newton algorithm such as limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS-B) has been used (14, 15).

Practically, we can use the gamlss function in R software in the package GAMLSS (generalized additive models for location scale and shape) (16) to get the initial values for estimating the parameters.

In this study, we have used the Quasi-Newton method (17) under which we performed the L-BFGS-B algorithm to obtain the optimum solution for the parameters.

Large sample inference: If the model specified by Equation (5) follows the regularity condition of maximum likelihood estimation (MLE) then, MLEs of θ and $J(\theta)$ (Fisher information matrix) are consistent. Assuming that $I(\theta) = \lim_{n \rightarrow \infty} \{n^{-1} J(\theta)\}$ exists and is non-singular, we have $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $N(0, I(\theta)^{-1})$.

2.3. Simulation study

In this simulation study, we featured finite-sample performance of the MLE. Due to time consumption, the simulated time-series data generated only for De β AR with lag 1 and lag 2.

$$\eta_t = \text{logit}(\mu_t) = \beta_0 + \phi_1 g(x_{t-1})$$

$$\eta_t = \text{logit}(\mu_t) = \beta_0 + \phi_1 g(x_{t-1}) + \phi_2 g(x_{t-2})$$

Here, ζ and ω are constant for all observations. We took $\beta_0 = 1.2$, $\phi_1 = -0.8$, $\phi_2 = -0.2$, $\zeta = 50$, and $\omega = 0.5$ as true parameters.

The Monte Carlo simulation with 15,000 replications was carried out each with sample size $n = 100, 250$, and 500 . Of the 15,000 replicants, 5,000 were the burn out. Parameter estimates were obtained as the convergent outcome of the remaining 10,000. The parameters are estimated by maximizing the log-likelihood function using the L-BFGS-B algorithm. The bias and root mean square error of the estimates are reported.

Table 1 represents the bias and root mean square error of the parameters. Here, we can observe that, as the sample size increases, the algorithm converged for all the samples. The mean of β_0, ϕ_1, ϕ_2 , and ω are close to the true values or the initial values. Also, the root mean square errors of all the estimators are decreased as the size of the sample increases, as anticipated.

2.4. Forecast evaluation criteria

Selecting a proper model among several competing candidates is a hassle in a lot of time-series analysis. In many cases, to look for out-of-sample forecast accuracy, the MAPE ("Mean absolute percentage error") is used for model selection or comparison. However, when data are close to zero, other forecast measurement criteria can be used, such as MAE ("mean absolute error"), MSE ("mean square error"), RMSE ("root mean square error"), and ex post forecast error (18). In the case of time-series analysis, the ACF plot is one of the model selection criteria, and the AIC ("Akaike information criterion") and the BIC ("Bayesian information criterion") will be used for an in-sample accuracy check.

TABLE 2 Descriptive statistics for the data on rate of *E. coli* resistant to AMC and CSL.

Antibiotic	Sample size	Range	Mean	Std.Deviation	Skewnes	Min	Max	Median
AMC	108	0.593	0.703	0.165	-0.0024	0.407	1	0.702
CSL	108	0.4	0.15	0.10	0.40	0	0.4	0.14

2.5. R shiny application

R shiny is a website application available freely to build under R-studio. It is an interactive website application that requires no web development skills¹. R shiny web app can be accessed at DeBAR.app².

3. Results

3.1. Exploratory data analysis

To understand the characteristics of the response series, the descriptive statistics for the data are represented in Table 2. Figure 1 represents the time-series plot of the bacteria *E. coli* isolated from blood cultures resistant to the antimicrobial agents AMC and CSL, where the sudden change between the years 2015 and 2016 is due to the improvisation in the testing method in April 2015, which involved a shift from manual testing method (i.e., Kirby Bauer disc diffusion method) to Vitek-2 automated method from the year 2015. This change point has been considered as a covariate in the model and represented as an indicator variable I_t . Figure 2 displays the histogram of the same. The proposed model is applicable to model stationary time-series data. The ACF plot will be used to identify the stationarity in the series. If non-stationarity exists, then significant deterministic elements can be added as a regressor variable in the model. As the current study data include change points in the series, stationarity cannot be identified through the ACF plot.

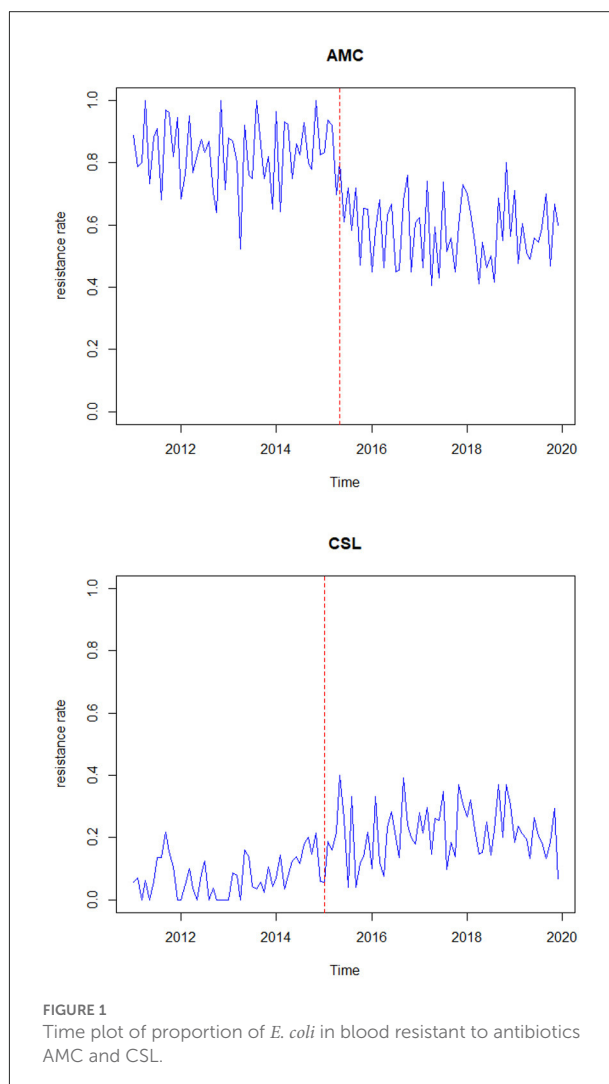
3.2. Time-series analysis

The data have been modeled using Degenerate Beta Autoregressive ($De\beta AR(p)$) model proposed in Section 2 and the analysis has been carried out using the software R. Quasi-Newton algorithm (L-BFGS-B) has been used to maximize the partial log-likelihood function.

First, the following $De\beta AR(p)$ models, where $p=1, 2, 3, \dots, 12$ fitted to the AMC data,

1 <https://shiny.rstudio.com/>

2 <https://jevithalobo.shinyapps.io/column/>



$$Model1: \eta_t = \text{logit}(\mu_t) = \beta_0 + \alpha I_t + \phi_1 \{g(x_{t-1})\}$$

$$Model2: \eta_t = \text{logit}(\mu_t) = \beta_0 + \alpha I_t + \sum_{i=1}^2 \phi_i \{g(x_{t-i})\}$$

$$Model3: \eta_t = \text{logit}(\mu_t) = \beta_0 + \alpha I_t + \sum_{i=1}^3 \phi_i \{g(x_{t-i})\}$$

$$Model12: \eta_t = \text{logit}(\mu_t) = \beta_0 + \alpha I_t + \sum_{i=1}^{12} \phi_i \{g(x_{t-i})\}$$

where, x_{t-i} $i=1,2,\dots,12$ is the lagged response series and I_t is the change point (Indicator) variable with coefficient α .

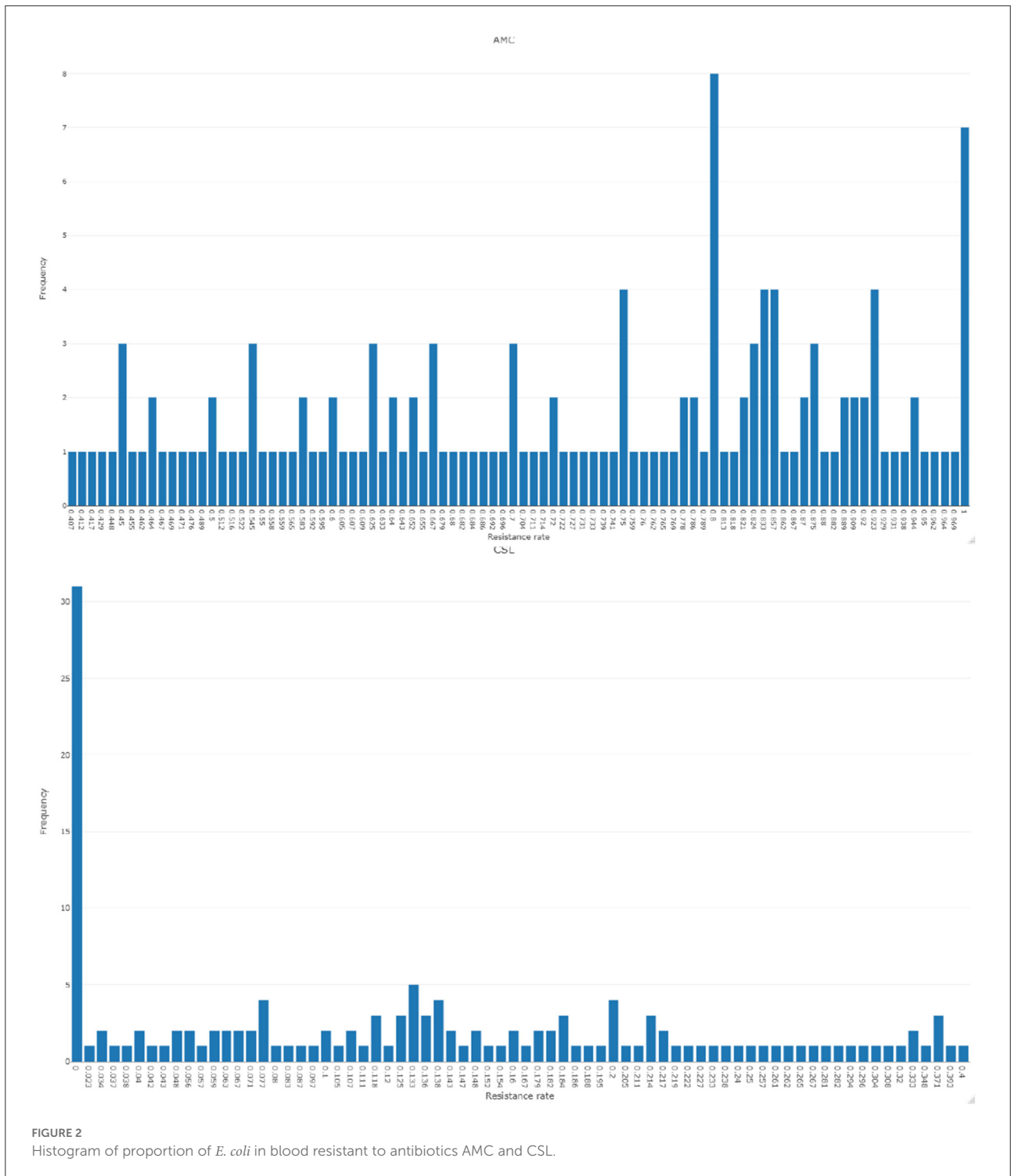


FIGURE 2 Histogram of proportion of *E. coli* in blood resistant to antibiotics AMC and CSL.

Among the 12 models, the best model is selected using the AIC and the BIC for which these values are minimum. The selected model is then compared with the existing ARIMA model.

The AIC and the BIC select Model 1 as the best among the 12 models as its values are minimum compared to the others (Table 4). From the significance test (Table 3), it can be seen that for the data AMC, lag-1 is significant at 5% level

TABLE 3 Fitted De β AR model for rate of *E. coli* resistant to AMC and CSL.

Antibiotic	Parameter	Estimate	Std. Error	z stat	p-value
	β_0	1.689	0.156	10.808	$0.00e^{+00}$
	ϕ_1	-0.109	0.080	-1.367	$1.74e^{-02}$
AMC	$\log(\rho)$	2.798	0.554	5.043	$2.15e^{-06}$
	$\text{logit}(\Omega)$	-3.188	0.099	-32.043	$0.00e^{+00}$
	α	-1.338	0.150	-8.889	$3.59e^{-14}$
	β_0	-2.185	0.166	-13.112	$0.00e^{+00}$
	ϕ_1	-0.0215	0.068	-0.313	$7.547e^{-02}$
CSL	$\log(\rho)$	3.099	0.020	148.21	$0.00e^{+00}$
	$\text{logit}(\Omega)$	-2.102	0.319	-6.580	$2.46e^{-09}$
	α	0.877	0.128	6.834	$7.55e^{-10}$

of significance (l.o.s) and for data CSL, lag-1 is significant at 10% level of significance (l.o.s) along with the indicator variable and parameters of the model, respectively. Next, the model is compared with the ARIMA model. Autocorrelation function (ACF) plots for residuals, forecast accuracy criteria (MAE, MSE, and MAPE), and information criteria (AIC and BIC) were used to select the best model for the given data.

Models considered were,

Model 1: De β AR(p), with lag at 1

$$\eta_t = \text{logit}(\mu_t) = \beta_0 + \alpha I_t + \phi_1 g(x_{t-1})$$

Model 2: ARIMA(0, 0, 1) model using Arcsine transformation

$$y_t = \beta_0 + \alpha I_t - \theta_1 \epsilon_{t-1}$$

In the analysis, as a final step, data have been forecasted by keeping out the last six observations (hold-out data) from the time-series, and Model 1 and Model 2 have fitted to the selected (test data) series and forecasted for the next 6 months. The out-of-sample forecast accuracy are calculated using the hold-out and forecasted data from the models and comparisons between models have been carried out. The hold-out series of AMC is 0.545, 0.592, 0.7, 0.469, 0.667, and 0.6. The forecasted series from Model 1 is 0.597, 0.594, 0.6, 0.595, 0.595, and 0.6. The forecasted series from Model 2 is 0.602, 0.579, 0.579, 0.579, 0.579, and 0.579. Similarly, the hold-out series of CSL is 0.205, 0.184, 0.133, 0.184, 0.294, and 0.067. The forecasted series from Model 1 is 0.192, 0.194, 0.193, 0.193, 0.194, and 0.195. The forecasted series from Model 2 is 0.223, 0.215, 0.213, 0.214, 0.214, and 0.213.

The study found that for both the data, the forecast accuracy for Model 1 is better compared to the Model 2 (Table 5) and residual plot of Model 1 follows white noise assumptions (i.e., the residual series should have mean 0 and no autocorrelation within the series), whereas for AMC data (Figure 3), Model 2 is

serially correlated at lag 4. Thus, we select Model 1 as the best fit model for both the data.

Thus, the estimated De β AR(1) model for AMC is

$$\hat{\mu}_t = \frac{\exp(\beta_0 + \alpha I_t + \phi_1 g(x_{t-1}))}{1 + \exp(\beta_0 + \alpha I_t + \phi_1 g(x_{t-1}))}$$

$$E(x) = \hat{\omega} + (1 - \hat{\omega})\hat{\mu}_t$$

where, $(\hat{\beta}_0, \hat{\alpha}, \hat{\phi}_1, \hat{\omega}) = (1.68, -1.34, -0.11, 0.03)$. The forecasted series for next 6 months is 0.604, 0.587, 0.613, 0.587, 0.604, and 0.582.

Similarly, the estimated De β AR(1) model for CSL is

$$\hat{\mu}_t = \frac{\exp(\beta_0 + \alpha I_t + \phi_1 g(x_{t-1}))}{1 + \exp(\beta_0 + \alpha I_t + \phi_1 g(x_{t-1}))}$$

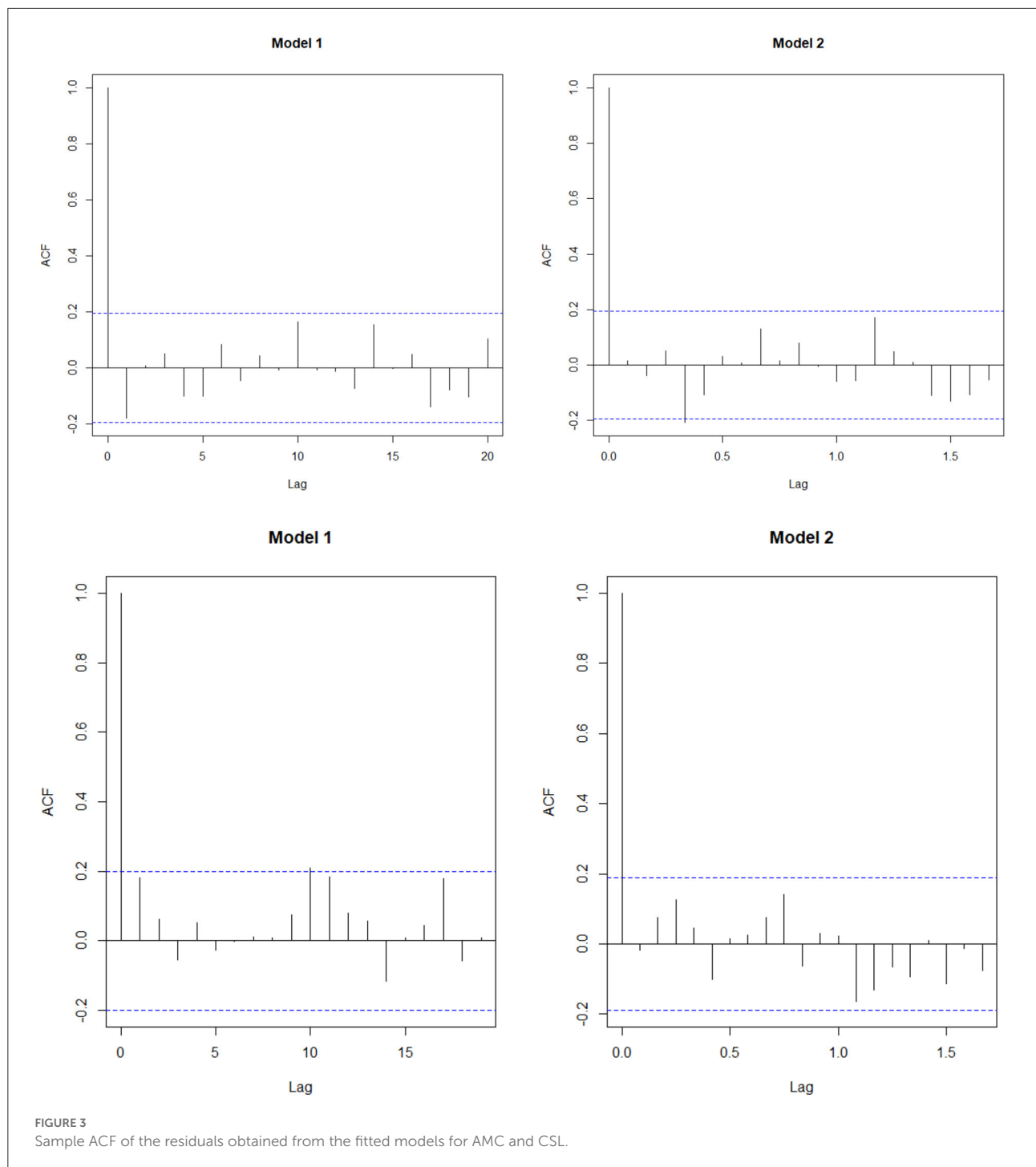
$$E(x) = (1 - \hat{\omega})\hat{\mu}_t$$

where, $(\hat{\beta}_0, \hat{\alpha}, \hat{\phi}_1, \hat{\omega}) = (-2.185, 0.877, -0.021, 0.108)$. The forecasted series for next 6 months is 0.19, 0.20, 0.19, 0.194, 0.192, and 0.189.

The R code of the analysis has been provided in the supporting file.

4. Discussion

AMR is difficult to control and expensive to deal with, and the outcome is that it can lead to death or severe disability. Many researchers have used different statistical techniques to look at the relationship between antimicrobial use and resistance. To list a few of them, a study by Athanasiou and Kopsini (19) in 2018 systematically reviewed the statistical methods used to analyze the AMR rate time-series data. Many of the studies



used the ARIMA model (20) and the transfer function model and few studies used the multiple linear regression model. Infectious disease control researchers most often analyze and predict antimicrobial resistance rate data, which includes zeros or ones [resistance rate with zeros can be seen in the figure of the article by Lopez-Lozano et al. (21)]. The commonly used ARIMA model is inappropriate for non-Gaussian (not

normally distributed) time-series data and may arrive at biased results. Rocha and Cribari-Neto (7) and Bayer et al. (8) introduced the Beta autoregressive moving average model and the Kumaraswamy autoregressive moving average model to fit rate/proportion time-series data in the interval (0, 1). To analyze proportion data with zeros or ones, Ospina and Ferrari (11) proposed an inflated beta regression model, and as an

TABLE 4 AIC and BIC for $D\beta$ AR(p) models for AMC and CSL data.

Antibiotic Model	AMC		CSL	
	AIC	BIC	AIC	BIC
Model 1	-126.19	-112.78	-153.69	-140.28
Model 2	-121.92	-105.83	-149.96	-133.86
Model 3	-117.64	-98.86	-152.58	-133.81
Model 4	-122.33	-100.87	-147.01	-125.55
Model 5	-119.55	-95.42	-151.37	-127.23
Model 6	-115.17	-88.35	-146.66	-119.84
Model 7	-110.99	-81.48	-142.14	-112.63
Model 8	-108.40	-76.22	-140.54	-108.35
Model 9	-105.93	-71.07	-138.07	-103.21
Model 10	-110.76	-73.22	-134.39	-96.84
Model 11	-106.62	-66.39	-133.56	-93.32
Model 12	-103.46	-60.55	-136.87	-93.96

TABLE 5 Model selection criterion.

Antibiotic	Model	Log-likelihood	MAE	MSE	MAPE	AIC	BIC
AMC	Model 1	68.09	0.089	0.011	13.93	-126.19	-112.78
	Model 2	57.81	0.109	0.018	18.40	-107.62	-97.12
CSL	Model 1	75.64	0.053	0.005	—	-153.69	-140.28
	Model 2	63.39	0.064	0.006	—	-118.79	-108.29

alternative, Bayer et al. (12) proposed an inflated Kumaraswamy regression model. But none of the studies addressed the autocorrelation in the response series in the interval $[0, 1)$ or $(0, 1]$. So, to fill this gap, in this article, we proposed the Degenerate Beta Autoregressive ($D\beta$ AR) model.

For this intention, we collected time-series data on *E. coli* isolates resistant to the different antimicrobial agents, and for the current study, only amoxicillin-clavulanic acid (AMC) and Cefoperazone-sulbactam (CSL) were considered. The primary focus of this study is to forecast the AMR rate for the time-series data in the interval $(0, 1)$ or $[0, 1)$. The proposed $D\beta$ AR(p) models (where $p=1,2,\dots,12$) fitted to the data and best order for "p" was selected for which the values of the AIC and the BIC were minimum. The best model among these was then compared with the existing ARIMA model and the best among both was decided based on the forecasting evaluation criteria (MSE, MAE, and MAPE).

From Table 4, we can see that for the data AMC, the values of the AIC and the BIC are minimum for Model 1 (i.e., AIC = -126.19 and BIC = -112.78) compared to other models. The study found that the $D\beta$ AR (1) model performed well compared to the remaining 11 models. Next, the selected $D\beta$ AR(1) model is compared with the existing ARIMA (0, 0, 1)

model selected from the autogeneration. From Table 5, we can see the values of MAE, MSE, and MAPE are minimum for the proposed Model 1 compared to the existing Model 2 along with the AIC and BIC. Hence, the study selected the $D\beta$ AR(1) model as the best among all. Similar procedure followed for the data of CSL.

By using the proposed $D\beta$ AR(1) model, the AMR rate was forecasted for next 6 months. The study results indicate the forecasted resistance rate of *E. coli* to the antimicrobial AMC ranges between 58 and 62% for the next 6 months, implying constant variations in the resistance rate.

The $D\beta$ AR(p) model introduced in this article would be beneficial to healthcare providers to implement early public health measures to control and prevent the rise in resistance rate.

5. Conclusion and future direction

Antimicrobial resistance is an emerging issue of public health. However, taking appropriate precautions or interventions in advance may help in reducing the resistance rate. Forecasting using an appropriate time-series model may

help in predicting the expected resistance rate for the future which is further useful for policy-making.

This study proposed the Degenerate Beta Autoregressive model (De β AR) to model antimicrobial resistance rate data, which is an extension of the β ARMA model proposed by Rocha and Cribabri-Neto (2009). The proposed model can be used to fit continuous time-series data in the interval $[0, 1)$ and $(0, 1]$, for example, rates or proportions. The model is applicable when the series is stationary in nature. The parameters of the model are estimated by maximizing the likelihood function and closed-form solutions for the score function are obtained by using the non-linear optimization algorithm (L-BFGS-B). The application of the model is presented using AMR data.

The outcome from a time-series model helps the healthcare policymakers to implement an appropriate intervention in advance to reduce the risk of rise in resistance rate.

In future, the study results can be improvised by considering antimicrobial consumption as a regressor variable and the proposed model can be improvised by incorporating moving average terms and seasonal components.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Ethical Clearance was obtained from Kasturba Medical College and Kasturba Hospital Institutional Ethics Committee (IEC No. 832/2019). Written informed

consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

JL, AK, and VK framed the objective of the study. VK collected the data. JL conducted the analysis, drafted initial manuscript, and developed R shiny app. AK and VK revised the manuscript. All authors have read and approved the final manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.969777/full#supplementary-material>

References

1. WHO. *Worldwide Country Situation Analysis: Response to Antimicrobial Resistance*. Geneva: WHO (2015). Available online at: <https://ghc.fiu.edu/wpcontent/uploads/sites/38/2018/04/Supplement-Material.pdf>
2. Zeng S, Xu Z, Wang X, Liu W, Qian L, Chen X, et al. Time series analysis of antibacterial usage and bacterial resistance in China: observations from a tertiary hospital from 2014 to 2018. *Infect Drug Resist.* (2019) 12:2683. doi: 10.2147/IDR.S220183
3. Ferrari S, Cribabri-Neto F. Beta regression for modelling rates and proportions. *J Appl Stat.* (2004) 31:799–815. doi: 10.1080/0266476042000214501
4. Mitnik PA, Baek S. The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Stat Pap.* (2013) 54:177–92. doi: 10.1007/s00362-011-0417-y
5. Pumi G, Rauber C, Bayer FM, Kumaraswamy regression model with Aranda-Ordaz link function. *Test.* (2020) 33:1–21. doi: 10.1007/s11749-020-00700-8
6. Lemonte AJ, Bazán JL. New class of Johnson distributions and its associated regression model for rates and proportions. *Biometr J.* (2016) 58:727–46. doi: 10.1002/bimj.201500030
7. Rocha AV, Cribabri-Neto F. Beta autoregressive moving average models. *Test.* (2009) 18:529. doi: 10.1007/s11749-008-0112-z
8. Bayer FM, Bayer DM, Pumi G. Kumaraswamy autoregressive moving average models for double bounded environmental data. *J Hydrol.* (2017) 555:385–96. doi: 10.1016/j.jhydrol.2017.10.006
9. Ospina R, Ferrari SL. Inflated beta distributions. *Stat Pap.* (2010) 51:111. doi: 10.1007/s00362-008-0125-4
10. Cribabri-Neto F, Santos J. Inflated Kumaraswamy distributions. *Anais da Academia Brasileira de Ciências.* (2019) 91:955. doi: 10.1590/0001-3765201920180955
11. Ospina R, Ferrari SL. A general class of zero-or-one inflated beta regression models. *Comput Stat Data Anal.* (2012) 56:1609–23. doi: 10.1016/j.csda.2011.10.005
12. Bayer FM, Cribabri-Neto F, Santos J. Inflated Kumaraswamy regressions with application to water supply and sanitation in Brazil. *Stat Neerlandica.* (2021) 75: 453–81. doi: 10.1111/stan.12242

13. Daga AP, Koga VL, de Matos CM, Perugini MR, Pelisson M, Kobayashi RK, et al. Escherichia coli bloodstream infections in patients at a university hospital: virulence factors and clinical characteristics. *Front Cell Infect Microbiol.* (2019) 9:191. doi: 10.3389/fcimb.2019.00191
14. Nocedal J, Wright SJ. *Numerical Optimization*. New York, NY: Springer (1999).
15. Xiao Y, Wei Z, Wang Z. A limited memory BFGS-type method for large-scale unconstrained optimization. *Comput Math Appl.* (2008) 56:1001–9. doi: 10.1016/j.camwa.2008.01.028
16. Stasinopoulos DM, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Softw.* (2008) 23:1–46. doi: 10.18637/jss.v023.i07
17. Schoenberg R. *Optimization With the Quasi-Newton Method*. Maple Valley, WA: Aptech Systems, Inc. (2001).
18. Doszyń M. Forecasting randomly distributed zero-inflated time series. *Folia Oeconomica Stetinensia.* (2017) 17:7–19. doi: 10.1515/fofi-2017-0001
19. Athanasiou CI, Kopsini A. Systematic review of the use of time series data in the study of antimicrobial consumption and Pseudomonas aeruginosa resistance. *J Glob Antimicrob Resist.* (2018) 15:69–73. doi: 10.1016/j.jgar.2018.06.001
20. Choi B. *ARMA Model Identification*. Berlin: Springer Science & Business Media (2012).
21. Lopez-Lozano JM, Monnet DL, Yagüe A, Burgos A, Gonzalo N, Campillos P, et al. Modelling and forecasting antimicrobial resistance and its dynamic relationship to antimicrobial use: a time series analysis. *Int J Antimicrob Agents.* (2000) 14:21–31. doi: 10.1016/S0924-8579(99)00135-1