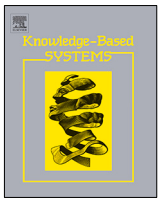
Knowledge-Based Systems 237 (2022) 107798

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Fixed effects panel interval-valued data models and applications

Ai-bing Ji *, Jin-jin Zhang, Xing He, Yu-hang Zhang

College of Mathematics and Information Science, Hebei University, Baoding 071002, China

ARTICLE INFO

Article history:
Received 24 March 2021
Received in revised form 17 November 2021
Accepted 20 November 2021
Available online 5 December 2021

Keywords:
Interval data analysis
Interval-valued regression
Panel data model
Forecasting

ABSTRACT

Interval-valued data is a complex data type which can be got by summarizing large datasets, linear regression models for interval-valued data have been widely studied. Panel data models combining cross-section and time series real-valued data have become increasingly popular in economic research and data mining. It is very important to construct the regression models for panel data with uncertainty and range variability. This paper introduces panel data regression model for interval-valued data and constructs three kinds of panel interval-valued data regression models: the centre model of fixed effects panel interval-valued data regression, the min–max model of fixed effects panel interval-valued data regression and its special model, the centre and range model of fixed effects panel interval-valued data regression. Then combining the parameters estimation of interval-valued regression and analysis of covariance for panel data, this paper presents the parameters estimations for three kinds of panel interval-valued data regression models. Finally, our proposed panel interval-valued data regression models are applied in forecasting of Air Quality Index, the experimental evaluation of actual data sets shows the advantages and the performance of our proposed panel interval-valued data models.

1. Introduction

Interval-valued data are applied in many situations [1,2] in which data represent variability (minimum and maximum of daily temperature), uncertainty (confidence intervals), etc. The research of interval analysis assumes that observations and estimations in practice are usually uncertain or incomplete and can be represented as intervals. Interval-valued data have also been used in symbolic data analysis (SDA) [3]. One source of interval-valued data is the aggregation of huge data in a reduced number of groups [4], which makes it possible to consider the variability in the data. So how to analyse interval-valued data is meaningful.

Several approaches were used to analyse interval-valued data. Some authors applied neural network models to manage interval-valued data [5,6]. In SDA, Billard and Diday [7] introduced dispersion measures and central tendency of interval-valued data. Interval-valued linear regression models were also built based on certain predefined criterion [8–14]. Billard and Diday [8] presented the first algorithm for fitting interval-valued linear regression, and this algorithm consisted of fitting a linear regression model to the midpoints of the interval values and the parameters were obtained by minimization of the mid-point error. Later, Billard and Diday [9] discussed the Min–Max Method, which defined two models, each for one response bound, with response lower bounds depending on the lower bounds of regressor variables and the response upper bounds depending on the upper bounds of regressor variables. Lima Neto et al. [10] introduced the Centre and Range Method (CRM), also proposed two linear models: one for the midpoints and the other for the ranges. Lima Neto et al. [11] later extended CRM to the Constrained Centre and Range Method (CCRM) which guaranteed the mathematical coherence of predicted values. Nonlinear regression models for interval-valued data were also established based on dynamic clustering algorithm [12].

Modelling and forecasting interval-valued time series (ITS) received considerable attention in econometrics and statistics [15–17]. Song et al. [18] discussed feature selection approaches for dynamic interval-valued ordered data.

For more complicated economical or administrative activities, which are represented by panel interval-valued data, i.e. cross-section and time series interval-valued data, how to model it is still an interesting problem, although there are some regression models and time series models for interval-valued data.

Statistical models combining cross-section and time series real-valued data become increasingly popular in economic research, a panel data set offers a certain number of advantages over traditional pure cross-section or pure time series data sets. A detailed account of the benefits of panel data can be found in the book by Hsiao [19], and panel data models allow us to construct and test more complicated behavioural models than pure cross-section or time series data. So far, the data in panel data models are all real-valued [20–24], it is very necessary to build a panel data model for interval-valued data.

* Corresponding author.
E-mail address: jabpjh@163.com (A.-b. Ji).

https://doi.org/10.1016/j.knosys.2021.107798
0950-7051/

By introducing fixed effects panel data regression models for interval-valued data, this paper presents three kinds of panel interval-valued data regression models and the estimation of their parameters, and this is the first attempt to discuss panel interval-valued data models. Both interval-valued regression models and panel data regression models are extended to panel interval-valued data regression models. Compared with panel data regression model, panel interval-valued data regression model can deal with uncertain data or variable data; compared with interval-valued regression model, panel interval-valued data regression model can reduce the degree of collinearity among explanatory variables, and improve the effectiveness of model estimation. Panel interval-valued data regression model can construct and test more complex behaviours. Panel interval-valued data regression models are used to analyse various actual problems, such as Air Quality Index (AQI) and stocks forecasting.

The main contributions of this paper include:

- Introducing interval-valued data into the panel data models solves the problem of uncertainty and range variability.
- Three novel fixed effects panel interval-valued data regression models are constructed respectively:
  (1) The centre model of fixed effects panel interval-valued data regression (P-CM);
  (2) The min–max model of fixed effects panel interval-valued data regression (P-Min–Max) and its special model (S-P-Min–Max);
  (3) The centre and range model of fixed effects panel interval-valued data regression (P-CRM).
- Each model reduces the degree of collinearity among explanatory variables and can predict the interval of the response variables, respectively. The applications in the forecasting of AQI show the advantages and the performance of our proposed panel interval-valued data models.

The rest of this paper are organized as follows: some related preliminaries are provided in Section 2; the proposed three kinds of panel interval-valued data regression models are presented in Section 3; the actual interval-valued data sets are analysed and compared in Section 4 to illustrate the performance of the proposed models.

2. Preliminary

In this section, some related preliminaries are provided.

2.1. Interval and their arithmetic

The statistical treatment of interval-valued data is considered them as elements belonging to the space $K_c(R) = \{[a, b] : a \le b, a, b \in R\}$. Each compact interval $A \in K_c(R)$ can be expressed by means of its (inf, sup)-representation, i.e. $A = [infA, supA]$, with $infA \le supA$. Alternatively, the notation $A = (midA, sprA)$ with $sprA \ge 0$, where $midA = \frac{supA+infA}{2}$ is the midpoint of the interval, and $sprA = \frac{\sup A - \inf A}{2}$ denotes the spread or radius of A. Statistical developments with interval-valued data are generally based on the (mid, spr)-parametrization, since the non-negativity condition for the spr component is usually easier to handle than the order condition for the inf and sup components of the (inf, sup)-characterization.

In order to manage intervals, a natural arithmetic is defined on $K_c(R)$ by means of the Minkowski addition $A + B = \{a + b : a \in A, b \in B\}$ and the product by scalars $\lambda A = \{\lambda a : a \in A\}$, for any $A, B \in K_c(R)$ and $\lambda \in R$. The space $(K_c(R), +, \cdot)$ is not linear but semi-linear due to the lack of symmetric elements with respect to the addition.

For two intervals $A = [a^-, a^+]$, $B = [b^-, b^+]$ in $K_c(R)$, the operations are as follows:

- $A + B = [a^- + b^-, a^+ + b^+] = (midA + midB, sprA + sprB)$,
- $A - B = A + (-B) = [a^- - b^+, a^+ - b^-] = (midA - midB, sprA + sprB)$,
- For $\lambda \in R$, $\lambda B = \begin{cases} [\lambda b^-, \lambda b^+], \lambda \ge 0 \\ [\lambda b^+, \lambda b^-], \lambda < 0 \end{cases}$ and in (mid, spr)-parametrization $\lambda B = (\lambda midB, |\lambda| sprB)$.

2.2. Linear regression model for interval-valued data

All existing linear regression approaches with interval-valued data used certain fixed reference points to model interval data [8–14], three methods are introduced and a novel special case for Min–Max method is presented.

Let $E = \{e_1, e_2, \ldots, e_n\}$ be a set of examples which are described by $p + 1$ interval-valued quantitative variables: dependent variables Y and independent variables $X_1, X_2, \ldots, X_p$. Each example $e_i \in E\ (i = 1, 2, \ldots, n)$ is represented as an interval quantitative feature vector $z_i = (x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, where $x_{ij} = [a_{ij}, b_{ij}] \in K_c(R)\ (j = 1, 2, \ldots, p)$, $y_i = [y_i^L, y_i^U] \in K_c(R)$.

2.2.1. Centre method

For the example $z_i = (x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, $(i = 1, 2, \ldots, N)$, where $x_{ij} = [a_{ij}, b_{ij}] \in K_c(R)$, $y_i = [y_i^L, y_i^U] \in K_c(R)$. Let $x_i^c = (x_{i1}^c, x_{i2}^c, \ldots, x_{ip}^c)$, where $x_{ij}^c = \frac{a_{ij}+b_{ij}}{2}$. $y_i^c = \frac{y_i^L+y_i^U}{2}$ $(i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$.

The linear regression relationship of $X_1, X_2, \ldots, X_p$ related to Y was constructed by using the midpoints of response and regressor intervals:

$$y_i^c = \beta_0^c + \beta_1^c x_{i1}^c + \beta_2^c x_{i2}^c + \cdots + \beta_p^c x_{ip}^c + \varepsilon_i^c$$

The values of $\beta_0^c, \beta_1^c, \beta_2^c, \ldots, \beta_p^c$ were estimated by ordinary least square (OLS) [8].

2.2.2. Min–Max method

The Min–Max method suggested estimating the lower and upper bounds of the intervals by using different vectors of parameters [5]. It was equivalent to supposing independence between the values of lower and upper bounds of the intervals.

$X_1, X_2, \ldots, X_p$ related to Y was studied according to the linear regression relationship:

$$y_i^L = \beta_0^L + \beta_1^L a_{i1} + \beta_2^L a_{i2} + \cdots + \beta_p^L a_{ip} + \varepsilon_i^L \quad (1)$$

$$y_i^U = \beta_0^U + \beta_1^U b_{i1} + \beta_2^U b_{i2} + \cdots + \beta_p^U b_{ip} + \varepsilon_i^U \quad (2)$$

The sum of the squares of deviations in the Min–Max model was denoted by

$$S_1 = \sum_{i=1}^{n} \left(\varepsilon_i^L\right)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^U\right)^2$$

Ordinary least square (OLS) was used to find the values of $\beta_0^L, \beta_1^L, \beta_2^L, \ldots, \beta_p^L$ and $\beta_0^U, \beta_1^U, \beta_2^U, \ldots, \beta_p^U$ by minimizing the expression S_1.

The estimations of the parameters $\beta_0^L, \beta_1^L, \beta_2^L, \ldots, \beta_p^L$ and $\beta_0^U, \beta_1^U, \beta_2^U, \ldots, \beta_p^U$ were equivalent to estimating $\beta_0^L, \beta_1^L, \beta_2^L, \ldots, \beta_p^L$ based on left points a_{ij} $(i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$ and $\beta_0^U, \beta_1^U, \beta_2^U, \ldots, \beta_p^U$ based on right points b_{ij} $(i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$ respectively.

$$A_2 = \begin{pmatrix} n, & 0, & \sum_i a_{i1}, & \sum_i a_{i2}, & \dots & \sum_i a_{ip} \\ 0, & n, & \sum_i b_{i1}, & \sum_i b_{i2}, & \dots & \sum_i b_{ip} \\ \sum_i a_{i1} & \sum_i b_{i1} & \sum_i [(a_{i1})^2 + (b_{i1})^2] & \sum_i [a_{i2}a_{i1} + b_{i2}b_{i1}] & \dots & \sum_i [a_{ip}a_{i1} + b_{ip}b_{i1}] \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ \sum_i a_{i(p-1)} & \sum_i b_{i(p-1)} & \sum_i [a_{i1}a_{i(p-1)} + b_{i1}b_{i(p-1)}] & \sum_i [a_{i2}a_{i(p-1)} + b_{i2}b_{i(p-1)}] & \dots & \sum_i [a_{ip}a_{i(p-1)} + b_{ip}b_{i(p-1)}] \\ \sum_i a_{ip} & \sum_i b_{ip} & \sum_i [a_{i1}a_{ip} + b_{i1}b_{ip}] & \sum_i [a_{i2}a_{ip} + b_{i2}b_{ip}] & \dots & \sum_i [a_{ip}^2 + b_{ip}^2] \end{pmatrix}$$

$$B_2 = \left(\sum_i y_i^L, \quad \sum_i y_i^U, \quad \sum_i (y_i^L a_{i1} + y_i^U b_{i1}) \quad \dots \quad \sum_i (y_i^L a_{ip} + y_i^U b_{ip})\right)^T$$

Box I.

Taking into consideration the system with the same mechanism, as a special case of Min–Max method, the coefficients of lower and upper bounds of the intervals are assumed to be the same in Eqs. (1), (2), but different in the intercept terms, that is

$$y_i^L = \beta_0^- + \beta_1 a_{i1} + \beta_2 a_{i2} + \cdots + \beta_p a_{ip} + \varepsilon_i^U \tag{3}$$

$$y_i^U = \beta_0^+ + \beta_1 b_{i1} + \beta_2 b_{i2} + \cdots + \beta_p b_{ip} + \varepsilon_i^U \tag{4}$$

$\beta_0^-, \beta_0^+, \beta_1, \beta_2, \dots, \beta_p$ are obtained by minimizing this expression S_2,

$$S_2 = \sum_{i=1}^{n} \left(\varepsilon_i^L\right)^2 + \sum_{i=1}^{n} \left(\varepsilon_i^U\right)^2$$

$$= \sum_{i=1}^{n} \left[\left(y_i^L - \beta_0^- - \sum_{j=1}^{p} \beta_j a_{ij} \right)^2 + \left(y_i^U - \beta_0^+ - \sum_{j=1}^{p} \beta_j a_{ij} \right)^2 \right]$$

$$\hat{\beta} = (\hat{\beta}_0^-, \hat{\beta}_o^+, \hat{\beta}_1, \dots, \hat{\beta}_p)^T = A_2^{-1} B_2$$

where A_2 is a matrix $(p+2)\times(p+2)$ and B_2 is a vector $(p+2)\times 1$, denoted as equations in Box I.

For a new given example x_{it} described by $x_{it} = \left(x_{it1}, x_{it2}, \dots, x_{itp}\right)$, where $x_{itk} = [a_{itk}, b_{itk}] (k = 1, 2, \dots, p)$, according to Eqs. (1), (2) or (3), (4), the response variable $\hat{y}_i = \left[\hat{y}_i^L, \hat{y}_i^U\right]$ is predicted as follows:

$$\hat{y}_i^L = \hat{\beta}_0^L + \hat{\beta}_1^L a_{i1} + \hat{\beta}_2^L a_{i2} + \cdots + \hat{\beta}_p^L a_{ip},$$

$$\hat{y}_i^U = \hat{\beta}_0^U + \hat{\beta}_1^U b_{i1} + \hat{\beta}_2^U b_{i2} + \cdots + \hat{\beta}_p^U b_{ip}$$

or

$$\hat{y}_i^L = \hat{\beta}_0^- + \hat{\beta}_1 a_{i1} + \hat{\beta}_2 a_{i2} + \cdots + \hat{\beta}_p a_{ip},$$

$$\hat{y}_i^U = \hat{\beta}_0^+ + \hat{\beta}_1 b_{i1} + \hat{\beta}_2 b_{i2} + \cdots + \hat{\beta}_p b_{ip}.$$

Sometimes Min–Max method and it is special case do not guarantee the mathematical coherence of the predicted interval bounds, then the response variable is predicted as follows:

$$\hat{y}_i = \left[\hat{y}_i^L, \hat{y}_i^U\right] = \begin{cases} \left[\dfrac{\hat{y}_i^L + \hat{y}_i^U}{2}, \dfrac{\hat{y}_i^L + \hat{y}_i^U}{2}\right] & \text{if } \hat{y}_i^L > \hat{y}_i^U; \\ [\hat{y}_i^L, \hat{y}_i^U] & \text{if } \hat{y}_i^L \le \hat{y}_i^U. \end{cases}$$

2.2.3. Centre and range method

Lima Neto et al. [10] presented the Centre and range method (CRM) to estimate the parameters vector using the information contained in the midpoints and ranges of the intervals.

For the examples $z_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}, y_i\right)$, $i = 1, 2, \dots, n$, where $x_{ij} = \left[a_{ij}, b_{ij}\right] \in K_c(R)$, $y_i = \left[y_i^L, y_i^U\right] \in K_c(R)$. Let $x_i^c = \left(x_{i1}^c, x_{i2}^c, \dots, x_{ip}^c\right)$ and $x_i^r = \left(x_{i1}^r, x_{i2}^r, \dots, x_{ip}^r\right)$, where $x_{ij}^c = \frac{a_{ij}+b_{ij}}{2}, x_{ij}^r = \frac{b_{ij}-a_{ij}}{2}$. $y_i^c = \frac{y_i^L+y_i^U}{2}, y_i^r = \frac{y_i^U-y_i^L}{2}$.

y_i^c, y_i^r were considered as dependent variables, $x_i^c = \left(x_{i1}^c, x_{i2}^c, \dots, x_{ip}^c\right)$ and $x_i^r = \left(x_{i1}^r, x_{i2}^r, \dots, x_{ip}^r\right) (i = 1, 2, \dots, n)$ as independent predictor variables. They were related in the following linear regression relationship:

$$y_i^c = \beta_0^c + \beta_1^c x_{i1}^c + \beta_2^c x_{i2}^c + \cdots + \beta_p^c x_{ip}^c + \varepsilon_i^c \tag{5}$$

$$y_i^r = \beta_0^r + \beta_1^r x_{i1}^r + \beta_2^r x_{i2}^r + \cdots + \beta_p^r x_{ip}^r + \varepsilon_i^r \tag{6}$$

In the CRM method, the sum of squares of deviations was given by

$$S_3 = \sum_{i=1}^{n} \left(\left(\varepsilon_i^c\right)^2 + \left(\varepsilon_i^r\right)^2 \right)$$

$$= \sum_{i=1}^{n} (y_i^c - \beta_0^c - \beta_1^c x_{i1}^c - \cdots - \beta_p^c x_{ip}^c)^2 + \sum_{i=1}^{n} (y_i^r - \beta_0^r - \beta_1^r x_{i1}^r - \cdots - \beta_p^r x_{ip}^r)^2.$$

$\beta_0^c, \beta_1^c, \beta_2^c, \dots, \beta_p^c$ and $\beta_0^r, \beta_1^r, \beta_2^r, \dots, \beta_p^r$ were estimated by minimizing the expression S_3.

In CRM, maybe $\hat{y}_i^r$ is negative. To avoid this situation, Lima Neto et al. [11] extended CRM to include positive constraints to the coefficients of interval ranges, which guaranteed the predicted upper bounds were greater than or equal to their respective lower bounds. But the obtained estimators can be biased, implying a poor adjustment of the model to the real linear relationship of data. For a given new example X, the assignment method of the predicted value $\hat{y}_i = \left[\hat{y}_i^L, \hat{y}_i^U\right]$ is adjusted as follows:

$$\hat{y}_i = \left[\hat{y}_i^L, \hat{y}_i^U\right] = \begin{cases} \left[\hat{y}_i^c, \hat{y}_i^c\right] & \text{if } \hat{y}_i^r \le 0; \\ \left[\hat{y}_i^c - \hat{y}_i^r, \hat{y}_i^c + \hat{y}_i^r\right] & \text{if } \hat{y}_i^r > 0. \end{cases}$$

2.3. Linear regression model for panel data

The term "panel data" refers to the pooling of observations on a cross-section of firms, countries, households, etc. over several periods. This can be achieved by surveying a number of individuals and following them over time.

Panel data set contains repeated observations for the same unit y_{it} and X_{it} for units, $i = 1, 2, \dots, N$ and periods $t = 1, 2, \dots, T$, also called longitudinal data.

Compared with time series or cross-section data sets, panel data model holds some advantages. (1) Panel data model supposes that individuals (firms, states or countries) are heterogeneous, and it can control individuals' heterogeneity. Time series and cross-section studies not controlling this heterogeneity run

the risk of obtaining biased results. (2) Panel data give more informative data, more variability, more degrees of freedom and less collinearity among the variables. Time-series studies are plagued with multi-collinearity. (3) Panel data models allow us to construct and test more complicated behavioural models than pure cross-section or time series data. (4) Panel data are more able to identify and measure effects that are simply not detectable in pure cross-section or pure time series data.

A panel data regression differs from a regular time series or cross-section regression in that it has a double subscript on its variables

$$y_{it} = \alpha + \beta^T X_{it} + u_{it}, i = 1, 2, \ldots, N; t = 1, 2, \ldots, T \tag{7}$$

with i denoting individuals (such as firms, countries, etc.) and t denoting time, the i subscript, therefore, denotes the cross-section dimension whereas t denotes the time series dimension. α is a scalar, β is $p \times 1$ vector and X_{it} are the itth observation on p explanatory variables.

Most of the panel data applications utilize a one-way error component model for the disturbances, with

$$u_{it} = \mu_i + \nu_t + \varepsilon_{it},$$

where μ_i denotes the unobservable individual-specific effect, ν_t denotes the unobservable time-specific effect and ε_{it} denotes the remainder disturbance. In the following, the focus is on the panel data model with individual-specific effects, that is

$$y_{it} = \alpha_i + \beta^T X_{it} + \varepsilon_{it}, i = 1, 2, \ldots, N; t = 1, 2, \ldots, T \tag{8}$$

where α_i is an unobserved effect for individual i and is time invariant. The parameters α_i, β were estimated by least-squares dummy-variable (LSDV) [19].

$$\begin{aligned}\hat{\beta} = &\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\bar{X}_i\right)\left(X_{it}-\bar{X}_i\right)^T\right]^{-1} \\ &\times\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\bar{X}_i\right)\left(y_{it}-\bar{y}_i\right)\right]\end{aligned} \tag{9}$$

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}\bar{X}_i, i = 1, 2, \ldots, N,$$

where $X_{it} = (x_{it1}, x_{it2}, \ldots, x_{itk})^T$, $\bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$, $\bar{y}_i = \frac{1}{T}\sum_{i=1}^{T} y_{it}$.

3. Fixed effects panel interval-valued data model

For panel interval-valued data set $S = \{(X_{it}, y_{it}) |\ i = 1, 2, \ldots, N;\ \ t = 1, 2, \ldots, T\}$, $y_{it} = \left[y_{it}^L, y_{it}^U\right]$ is assumed as the observed interval-valued dependent variables, $X_{it} = \left(x_{it1}, x_{it2}, \ldots, x_{itp}\right)^T$ as $p \times 1$ interval-valued independent vectors with $x_{itj} = \left[a_{itj}, b_{itj}\right]$, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, $j = 1, 2, \ldots, p$.

In the following, using certain fixed reference points, the linear regression models with fixed individual-specific effects are to be constructed for the panel interval-valued data set.

$$S = \{(x_{it}, y_{it}) |\ i = 1, 2, \ldots, N;\ \ t = 1, 2, \ldots, T\}.$$

In this section, three kinds of fixed-effects panel interval-valued data models are proposed. Using the midpoints of the interval to represent interval-valued data, the centre model of fixed effects panel interval-valued data regression (P-CM) is constructed based on the midpoints of the panel interval-valued data. The Min–Max model of fixed effects panel interval-valued data regression (P-Min–Max) improves the P-CM model by establishing two models to fit the lower- and upper-bound of panel interval-valued data. The centre and range model of fixed effects panel interval-valued data regression (P-CRM) also uses two independent models to fit the panel interval-valued data: one for the midpoints of interval and another for ranges of the interval. These three kinds of fixed effects panel interval-valued data regression models hold their own advantages and disadvantages.

3.1. The centre model of fixed effects panel interval-valued data regression

For the panel interval-valued data set $S = \{(X_{it}, y_{it}) |\ i = 1, 2, \ldots, N;\ \ t = 1, 2, \ldots, T\}$, let $x_{itk}^c = \frac{a_{itk}+b_{itk}}{2}$, $y_{it}^c = \frac{y_{it}^L+y_{it}^U}{2}$.

y_{it}^c is considered as dependent variable and $x_{itk}^c (k = 1, 2, \ldots, p)$ as independent predictor variables. Based on the centre value of the panel interval-valued data, the linear regression model with fixed individual-specific effects is as follows:

$$y_{it}^c = \alpha_i + \beta_1^c x_{it1}^c + \beta_2^c x_{it2}^c + \cdots + \beta_p^c x_{itp}^c + \varepsilon_{it}^c \tag{10}$$

where $\varepsilon_{it}^c \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, N$; $t = 1, 2, \ldots, T$.

Averaging Eq. (10) over time gives

$$\bar{y}_{i.}^c = \alpha_i + \beta_1^c \bar{x}_{i.1}^c + \beta_2^c \bar{x}_{i.2}^c + \cdots + \beta_p^c \bar{x}_{i.p}^c + \bar{\varepsilon}_{i.}^c \tag{11}$$

where $\bar{y}_{i.}^c = \frac{1}{T}\sum_{t=1}^{T} y_{it}^c$, $\bar{x}_{i.k}^c = \frac{1}{T}\sum_{t=1}^{T} x_{itk}^c (k = 1, 2, \ldots, p)$.

Subtracting (11) from (10) gives

$$y_{it}^c - \bar{y}_{i.}^c = \beta_1^c(x_{it1}^c - \bar{x}_{i.1}^c) + \beta_2^c(x_{it2}^c - \bar{x}_{i.2}^c) + \cdots + \beta_p^c(x_{itp}^c - \bar{x}_{i.p}^c) + (\varepsilon_{it}^c - \bar{\varepsilon}_{i.}^c)$$

Thus, in the centre model of fixed effects panel interval-valued data, the sum of squares of deviations is given by

$$\begin{aligned} S_4 &= \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\varepsilon_{it}^c - \bar{\varepsilon}_{i.}^c\right)^2 \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}^c - \bar{y}_{i.}^c - \sum_{j=1}^{p}\beta_j^c(x_{itj}^c - \bar{x}_{i.j}^c)\right)^2 \end{aligned}$$

The values of $\beta_1^c, \beta_2^c, \ldots, \beta_p^c$ are estimated by minimizing the expression S_4,

$$\hat{\beta}^T = (\hat{\beta}_1^c, \hat{\beta}_2^c, \ldots, \hat{\beta}_p^c)^T = A_4^{-1}B_4$$

$$\hat{\alpha}_i = \bar{y}_{i.}^c - (\beta_1^c \bar{x}_{i.1}^c + \beta_2^c \bar{x}_{i.2}^c + \cdots + \beta_p^c \bar{x}_{i.p}^c),$$

where, see equations in Box II.

For a new given example $X_{it} = \left(x_{it1}, x_{it2}, \ldots, x_{itp}\right)^T$, with $x_{itj} = \left[a_{itj}, b_{itj}\right]$, the response variables $\hat{y}_{it} = \left[\hat{y}_{it}^L, \hat{y}_{it}^U\right]$ are predicted as follows:

$$\hat{y}_{it}^L = \hat{\alpha}_i + \left(X_{it}^L\right)^T \hat{\beta}$$

$$\hat{y}_{it}^U = \hat{\alpha}_i + \left(X_{it}^U\right)^T \hat{\beta}$$

where $\left(X_{it}^L\right)^T = \left(a_{it1}, a_{it2}, \ldots, a_{itp}\right)$, $\left(X_{it}^U\right)^T = \left(b_{it1}, b_{it2}, \ldots, b_{itp}\right)$.

3.2. The Min–Max model of fixed effects panel interval-valued data regression

In Min–Max panel interval-valued data regression model, suppose a panel interval-valued data model with individual-specific effects is as follows:

$$y_{it}^L = \alpha_i^L + \beta_1^L a_{it1} + \beta_2^L a_{it2} + \cdots + \beta_p^L a_{itp} + \varepsilon_{it}^L \tag{12}$$

$$y_{it}^U = \alpha_i^U + \beta_1^U b_{it1} + \beta_2^U b_{it2} + \cdots + \beta_p^U b_{itp} + \varepsilon_{it}^U \tag{13}$$

$\varepsilon_{it}^L \sim N(0, \sigma^2)$, $\varepsilon_{it}^U \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, N$; $t = 1, 2, \ldots, T$.

P-Min–Max model consists of two different models. The regressor lower bounds are used to build a model for the response lower bounds. The same is done for the upper bounds.

$$A_4 = \begin{pmatrix} \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it1}^c - \bar{x}_{i.1}^c)(x_{it1}^c - \bar{x}_{i.1}^c)], & \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it2}^c - \bar{x}_{i.2}^c)(x_{it1}^c - \bar{x}_{i.1}^c)], & \dots, & \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{itp}^c - \bar{x}_{i.p}^c)(x_{it1}^c - \bar{x}_{i.1}^c)] \\ \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it1}^c - \bar{x}_{i.1}^c)(x_{it2}^c - \bar{x}_{i.2}^c)], & \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it2}^c - \bar{x}_{i.2}^c)(x_{it2}^c - \bar{x}_{i.2}^c)], & \dots, & \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{itp}^c - \bar{x}_{i.p}^c)(x_{it2}^c - \bar{x}_{i.2}^c)] \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it1}^c - \bar{x}_{i.1}^c)(x_{itp}^c - \bar{x}_{i.p}^c)], & \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it2}^c - \bar{x}_{i.2}^c)(x_{itp}^c - \bar{x}_{i.p}^c)], & \dots, & \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{itp}^c - \bar{x}_{i.p}^c)(x_{itp}^c - \bar{x}_{i.p}^c)] \end{pmatrix}$$

$$B_4 = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it}^c - \bar{y}_{i.}^c)(x_{it1}^c - \bar{x}_{i.1}^c)], \quad \dots, \quad \sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it}^c - \bar{y}_{i.}^c)(x_{itp}^c - \bar{x}_{i.p}^c)]\right)^T$$

Box II.

Based on the infimum of y_{it} and x_{itk} ($i = 1, 2, \dots, N$; $t = 1, 2, \dots, T$; $k = 1, 2, \dots, p$), the parameters α_0^L, $\beta^L = \left(\beta_1^L, \beta_2^L, \dots, \beta_p^L\right)^T$ are estimated by least-squares dummy-variable (LSDV) [19].

$$\hat{\beta}^L = \left(\hat{\beta}_1^L, \hat{\beta}_2^L, \dots, \hat{\beta}_p^L\right)^T = \left(A_4^L\right)^{-1} B_4^L \tag{14}$$

$$\hat{\alpha}_i^L = \bar{y}_i^L - \hat{\beta}^L \bar{A}_i, i = 1, 2, \dots, N,$$

where $A_4^L = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(A_{it} - \bar{A}_i\right)\left(A_{it} - \bar{A}_i\right)^T$, $B_4^L = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(A_{it} - \bar{A}_i\right)\left(y_{it}^L - \bar{y}_i^L\right)$, $A_{it} = \left(a_{it1}, a_{it2}, \dots, a_{itp}\right)^T$, $\bar{A}_i = \frac{1}{T}\sum_{t=1}^{T} A_{it}$, $\bar{y}_i^L = \frac{1}{T}\sum_{t=1}^{T} y_{it}^L$.

So does it to obtain the parameters α_0^U, $\beta^U = \left(\beta_1^U, \beta_2^U, \dots, \beta_p^U\right)^T$ based on the supremum of y_{it} and x_{itj} ($i = 1, 2, \dots, N$; $t = 1, 2, \dots, T$; $k = 1, 2, \dots, p$).

$$\hat{\beta}^U = \left(\hat{\beta}_1^U, \hat{\beta}_2^U, \dots, \hat{\beta}_p^U\right)^T = \left(A_4^U\right)^{-1} B_4^U \tag{15}$$

$$\hat{\alpha}_i^U = \bar{y}_i^U - \hat{\beta}^U \bar{B}_i, i = 1, 2, \dots, N,$$

where $A_4^U = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(B_{it} - \bar{B}_i\right)\left(B_{it} - \bar{B}_i\right)^T$, $B_4^U = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(B_{it} - \bar{B}_i\right)\left(y_{it}^U - \bar{y}_i^U\right)$, $B_{it} = \left(b_{it1}, b_{it2}, \dots, b_{itp}\right)^T$, $\bar{B}_i = \frac{1}{T}\sum_{t=1}^{T} B_{it}$, $\bar{y}_i^U = \frac{1}{T}\sum_{t=1}^{T} y_{it}^U$.

Considering some systems with the same mechanism, as a special case of Min–Max panel interval-valued data model **(S-P-Min–Max)**, the coefficients of lower and upper bounds of the intervals are assumed to be the same in Eqs. (12), (13), but different in the intercept terms, that is

$$y_{it}^L = \alpha_i^L + \beta_1 a_{it1} + \beta_2 a_{it2} + \cdots + \beta_p a_{itp} + \varepsilon_{it}^L \tag{16}$$

$$y_{it}^U = \alpha_i^U + \beta_1 b_{it1} + \beta_2 b_{it2} + \cdots + \beta_p b_{itp} + \varepsilon_{it}^U \tag{17}$$

$\varepsilon_{it}^L \sim N(0, \sigma^2)$, $\varepsilon_{it}^U \sim N(0, \sigma^2)$, $i = 1, 2, \dots, N$; $t = 1, 2, \dots, T$.

Averaging (16) and (17) over time gives

$$\bar{y}_{i.}^L = \alpha_i^L + \beta_1 \bar{a}_{i1} + \beta_2 \bar{a}_{i2} + \cdots + \beta_p \bar{a}_{ip} + \bar{\varepsilon}_{i.}^L \tag{18}$$

$$\bar{y}_{i.}^U = \alpha_i^U + \beta_1 \bar{b}_{i1} + \beta_2 \bar{b}_{i2} + \cdots + \beta_p \bar{b}_{ip} + \bar{\varepsilon}_{i.}^U \tag{19}$$

where $\bar{y}_{i.}^L = \frac{1}{T}\sum_{t=1}^{T} y_{it}^L$, $\bar{a}_{i.k} = \frac{1}{T}\sum_{t=1}^{T} a_{itk}$, $\bar{y}_{i.}^U = \frac{1}{T}\sum_{t=1}^{T} y_{it}^U$, $\bar{b}_{i.k} = \frac{1}{T}\sum_{t=1}^{T} b_{itk}$.

Subtracting (18) from (16) gives

$$y_{it}^L - \bar{y}_{i.}^L = \beta_1(a_{it1} - \bar{a}_{i.1}) + \cdots + \beta_p(a_{itp} - \bar{a}_{i.p}) + (\varepsilon_{it}^L - \bar{\varepsilon}_{i.}^L) \tag{20}$$

Subtracting (19) from (17) gives

$$y_{it}^U - \bar{y}_{i.}^U = \beta_1(b_{it1} - \bar{b}_{i.1}) + \cdots + \beta_p(b_{itp} - \bar{b}_{i.p}) + (\varepsilon_{it}^U - \bar{\varepsilon}_{i.}^U) \tag{21}$$

The values of $\beta_1, \beta_2, \dots, \beta_p$ can be estimated by minimizing this expression S_5,

$$S_5 = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\varepsilon_{it}^L - \bar{\varepsilon}_{i.}^L\right)^2 + \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\varepsilon_{it}^U - \bar{\varepsilon}_{i.}^U\right)^2 = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(y_{it}^L - \bar{y}_{i.}^L - \sum_{j}^{P}\beta_j(a_{itJ} - \bar{a}_{i.j})\right)^2 + \left(y_{it}^U - \bar{y}_{i.}^U - \sum_{j=1}^{p}\beta_j(b_{itj} - \bar{b}_{i.j})\right)^2\right]$$

$$\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T = A_5^{-1} B_5,$$

$$\hat{\alpha}_i^L = \bar{y}_{i.}^L - \hat{\beta}_1 \bar{a}_{i.1} - \hat{\beta}_2 \bar{a}_{i.2} - \cdots - \hat{\beta}_p \bar{a}_{i.p},$$

$$\hat{\alpha}_i^U = \bar{y}_{i.}^U - \hat{\beta}_1 \bar{b}_{i.1} - \hat{\beta}_2 \bar{b}_{i.2} - \cdots - \hat{\beta}_p \bar{b}_{i.p},$$

where

$$A_5 = \begin{pmatrix} A_{11} & A_{21} & \dots & A_{p1} \\ A_{12} & A_{22} & \dots & A_{p2} \\ \dots & \dots & \dots & \dots \\ A_{1p} & A_{2p} & \dots & A_{pp} \end{pmatrix},$$

$$B_5 = \left(\gamma_1, \gamma_2, \dots, \gamma_p\right)^T,$$

with

$$A_{nm} = \sum_{i=1}^{N}\sum_{t=1}^{T}[(a_{itm} - \bar{a}_{i.m})(a_{itn} - \bar{a}_{i.n}) + \left(b_{itm} - \bar{b}_{i.m}\right)\left(b_{itn} - \bar{b}_{i.n}\right)],$$

$n, m = 1, 2, \dots, p$,

$$\gamma_j = \sum_{i=1}^{N}\sum_{t=1}^{T}[\left(y_{it}^L - \bar{y}_{i.}^L\right)\left(a_{itj} - \bar{a}_{i.j}\right) + \left(y_{it}^U - \bar{y}_{i.}^U\right)\left(b_{itj} - \bar{b}_{i.j}\right)],$$

$j = 1, 2, \dots, p$.

For a new given example x_{it} described by $x_{it} = \left(x_{it1}, x_{it2}, \dots, x_{itp}\right)$, where $x_{itk} = [a_{itk}, b_{itk}]$, $(k = 1, 2, \dots, p)$, according to Eqs. (12), (13) or by (16), (17),

$$\hat{y}_{it}^L = \hat{\alpha}_i^L + \hat{\beta}_1^L a_{it1} + \cdots + \hat{\beta}_2^L a_{it2} + \hat{\beta}_p^L a_{itp},$$

$$\hat{y}_{it}^{U} = \hat{\alpha}_i^{U} + \hat{\beta}_1^{U} b_{it1} + \hat{\beta}_2^{U} b_{it2} + \cdots + \hat{\beta}_p^{U} b_{itp}$$

or

$$\hat{y}_{it}^{L} = \hat{\alpha}_i^{L} + \hat{\beta}_1 a_{it1} + \hat{\beta}_2 a_{it2} + \cdots + \hat{\beta}_p a_{itp},$$

$$\hat{y}_{it}^{U} = \hat{\alpha}_i^{U} + \hat{\beta}_1 b_{it1} + \hat{\beta}_2 b_{it2} + \cdots + \hat{\beta}_p b_{itp}.$$

Sometime P-Min–Max model and S-P-Min–Max model do not guarantee the mathematical coherence of the predicted interval bounds, then the response variables are predicted as follows:

$$\hat{y}_{it} = \left[\hat{y}_{it}^{L}, \hat{y}_{it}^{U}\right] = \begin{cases} \left[\dfrac{\hat{y}_{it}^{L} + \hat{y}_{it}^{U}}{2}, \dfrac{\hat{y}_{it}^{L} + \hat{y}_{it}^{U}}{2}\right] & \text{if } \hat{y}_{it}^{L} > \hat{y}_{it}^{U}; \\ [\hat{y}_{it}^{L}, \hat{y}_{it}^{U}] & \text{if } \hat{y}_{it}^{L} \le \hat{y}_{it}^{U}. \end{cases}$$

3.3. The centre and range model of fixed effects panel interval-valued data regression

For panel interval-valued data set $S = \{(X_{it}, y_{it}) \mid i = 1, 2, \ldots, N; \quad t = 1, 2, \ldots, T\}$, let $X_{it}^{c} = \left(x_{it1}^{c}, x_{it2}^{c}, \ldots, x_{itp}^{c}\right)$, $X_{it}^{r} = \left(x_{it1}^{r}, x_{it2}^{r}, \ldots, x_{itp}^{r}\right)$, where $x_{itj}^{c} = \frac{a_{itj}+b_{itj}}{2}$, $x_{itj}^{r} = \frac{b_{itj}-a_{itj}}{2}$. $y_{it}^{c} = \frac{y_{it}^{U}+y_{it}^{L}}{2}$, $y_{it}^{r} = \frac{y_{it}^{U}-y_{it}^{L}}{2}$.

$y_{it}^{c}(y_{it}^{r})$ is assumed as dependent variable and $X_{it}^{c}(X_{it}^{r})$ as independent predictor variables. They are related in the following linear regression relationship:

$$y_{it}^{c} = \alpha_i^{c} + \beta_1^{c} x_{it1}^{c} + \beta_2^{c} x_{it2}^{c} + \cdots + \beta_p^{c} x_{itp}^{c} + \varepsilon_{it}^{c} \tag{22}$$

$$y_{it}^{r} = \alpha_i^{r} + \beta_1^{r} x_{it1}^{r} + \beta_2^{r} x_{it2}^{r} + \cdots + \beta_p^{r} x_{itp}^{r} + \varepsilon_{it}^{r} \tag{23}$$

where $\varepsilon_{it}^{c} \sim N(0, \sigma^2)$, $\varepsilon_{it}^{r} \sim N(0, \sigma^2)$

P-CRM model consists of two independent models: one for the interval midpoints and another for the interval ranges.

Averaging over time in (22) gives

$$\bar{y}_{i.}^{c} = \alpha_i^{c} + \beta_1^{c} \bar{x}_{i.1}^{c} + \beta_2^{c} \bar{x}_{i.2}^{c} + \cdots + \beta_p^{c} \bar{x}_{i.p}^{c} + \bar{\varepsilon}_{i.}^{c} \tag{24}$$

Averaging over time in (23) gives

$$\bar{y}_{i.}^{r} = \alpha_i^{r} + \beta_1^{r} \bar{x}_{i.1}^{r} + \beta_2^{r} \bar{x}_{i.2}^{r} + \cdots + \beta_p^{r} \bar{x}_{i.p}^{r} + \bar{\varepsilon}_{i.}^{r} \tag{25}$$

where $\bar{y}_{i.}^{c} = \frac{1}{T}\sum_{t=1}^{T} y_{it}^{c}$, $\bar{x}_{i.k}^{c} = \frac{1}{T}\sum_{t=1}^{T} x_{itk}^{c}$, $\bar{y}_{i.}^{r} = \frac{1}{T}\sum_{t=1}^{T} y_{it}^{r}$, $\bar{x}_{i.k}^{r} = \frac{1}{T}\sum_{t=1}^{T} x_{itk}^{r}$.

Subtracting (24) from (22) gives

$$y_{it}^{c} - \bar{y}_{i.}^{c} = \beta_1^{c}(x_{it1}^{c} - \bar{x}_{i.1}^{c}) + \cdots + \beta_p^{c}(x_{itp}^{c} - \bar{x}_{i.p}^{c}) + (\varepsilon_{it}^{c} - \bar{\varepsilon}_{i.}^{c})$$

Subtracting (25) from (23) gives

$$y_{it}^{r} - \bar{y}_{i.}^{r} = \beta_1^{r}(x_{it1}^{r} - \bar{x}_{i.1}^{r}) + \cdots + \beta_p^{r}(x_{itp}^{r} - \bar{x}_{i.p}^{r}) + (\varepsilon_{it}^{r} - \bar{\varepsilon}_{i.}^{r})$$

Thus, in the P-CRM model of fixed effects panel interval-valued data, the sum of squares of deviations is given by

$$\begin{aligned} S_6 &= \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(\varepsilon_{it}^{c} - \bar{\varepsilon}_{i.}^{c}\right)^2 + \left(\varepsilon_{it}^{r} - \bar{\varepsilon}_{i.}^{r}\right)^2\right] \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(y_{it}^{c} - \bar{y}_{i.}^{c} - \sum_{j=1}^{P}\beta_j^{c}(x_{itj}^{c} - \bar{x}_{i.j}^{c})\right)^2 \right. \\ &\quad \left. + \left(y_{it}^{r} - \bar{y}_{i.}^{r} - \sum_{j=1}^{p}\beta_j^{r}(x_{itj}^{r} - \bar{x}_{i.j}^{r})\right)^2\right] \end{aligned}$$

The values of $\beta_1^{c}, \beta_2^{c}, \ldots, \beta_p^{c}; \beta_1^{r}, \beta_2^{r}, \ldots, \beta_p^{r}$ can be estimated by minimizing this expression S_6

$$\begin{aligned} &(\hat{\beta}_1^{c}, \hat{\beta}_2^{c}, \ldots, \hat{\beta}_p^{c}; \hat{\beta}_1^{r}, \hat{\beta}_2^{r}, \ldots, \hat{\beta}_p^{r})^{T} = A_6^{-1} B_6 \\ &= \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}^{-1} \begin{pmatrix} \gamma_1 & \cdots & \gamma_p & \eta_1 & \cdots & \eta_p \end{pmatrix}^{T} \end{aligned}$$

$$\hat{\alpha}_i^{c} = \bar{y}_{i.}^{c} - \hat{\beta}_1^{c}\bar{x}_{i.1}^{c} - \hat{\beta}_2^{c}\bar{x}_{i.2}^{c} - \cdots - \hat{\beta}_p^{c}\bar{x}_{i.p}^{c},$$

$$\hat{\alpha}_i^{r} = \bar{y}_{i.}^{r} - \hat{\beta}_1^{r}\bar{x}_{i.1}^{r} - \hat{\beta}_2^{r}\bar{x}_{i.2}^{r} - \cdots - \hat{\beta}_p^{r}\bar{x}_{i.p}^{r}.$$

where

$$A = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{p1} \\ A_{12} & A_{22} & \cdots & A_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1p} & A_{2p} & \cdots & A_{pp} \end{pmatrix},$$

$$B = \begin{pmatrix} B_{11} & B_{21} & \cdots & B_{p1} \\ B_{12} & B_{22} & \cdots & B_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ B_{1p} & B_{2p} & \cdots & B_{pp} \end{pmatrix},$$

and 0 means that the $p \times p$ zero matrix here,

$$A_{nm} = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(x_{itm}^{c} - \bar{x}_{i.m}^{c}\right)\left(x_{itn}^{c} - \bar{x}_{i.n}^{c}\right)\right],$$

$$B_{nm} = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(x_{itm}^{r} - \bar{x}_{i.m}^{r}\right)\left(x_{itn}^{r} - \bar{x}_{i.n}^{r}\right)\right], n, m = 1, 2, \ldots, p.$$

$$\gamma_j = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(y_{it}^{c} - \bar{y}_{i.}^{c}\right)\left(x_{itj}^{c} - \bar{x}_{i.j}^{c}\right)\right],$$

$$\eta_j = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\left(y_{it}^{r} - \bar{y}_{i.}^{r}\right)\left(x_{itj}^{r} - \bar{x}_{i.j}^{r}\right)\right], j = 1, 2, \ldots, p.$$

For a new given example x_{it} described by $x_{it} = \left(x_{it1}, x_{it2}, \ldots, x_{itp}\right)$, where $x_{itk} = [a_{itk}, b_{itk}], (k = 1, 2, \ldots, p)$, according to Eqs. (22), (23), $\hat{y}_{it}^{c}, \hat{y}_{it}^{r}$ are derived as follows.

$$\hat{y}_{it}^{c} = \hat{\alpha}_i^{c} + \hat{\beta}_1^{c} x_{it1}^{c} + \hat{\beta}_2^{c} x_{it2}^{c} + \cdots + \hat{\beta}_p^{c} x_{itp}^{c}$$

$$\hat{y}_{it}^{r} = \hat{\alpha}_i^{r} + \hat{\beta}_1^{r} x_{it1}^{r} + \hat{\beta}_2^{r} x_{it2}^{r} + \cdots + \hat{\beta}_p^{r} x_{itp}^{r}$$

Negative $\hat{y}_{it}^{r}$ may result in incoherence of the predicted interval bounds. In order to guarantee the mathematical coherence of the predicted interval bounds, the predicted values are adjusted as follows:

$$\hat{y}_{it} = \left[\hat{y}_{it}^{L}, \hat{y}_{it}^{U}\right] = \begin{cases} [\hat{y}_{it}^{c} - \hat{y}_{it}^{r}, \hat{y}_{it}^{c} + \hat{y}_{it}^{r}] & \text{if } \hat{y}_{it}^{r} \ge 0; \\ [\hat{y}_{it}^{c}, \hat{y}_{it}^{c}] & \text{if } \hat{y}_{it}^{r} < 0. \end{cases}$$

4. Applications in forecasting of air quality index (AQI)

Air pollution is caused by harmful suspended particles released into the atmosphere. The air pollution problem in most cities is very severe and has been the focus of the public and government [25].

Some scholars studied the air quality from different aspects. Luo [26] found out the primary pollutant of city was PM10, followed by SO2. Xu et al. [27] showed that O3 and NO2 exhibited a moderately negative correlation. Xu [28] used the principal component analysis to explore the correlation between various pollutants in the air quality monitoring index in Xi'an. Xu [29] used daily average data of air pollution indicators in Changsha and Haikou to obtain an optimal linear regression model.

The studies mentioned above are limited to a separate city and use daily average data of air pollution indicators. The used calculation process of air quality index is as follows: (1) compared with the grading concentration limit of each pollutant (GB3095-2012), the Individual Air Quality Index (IAQI) is calculated by the measured concentration values of CO, NO2, O3, PM10, PM2.5 and SO2, (2) AQI is the maximum value of the IAQI of various pollutants. This brings up a problem: In the current measurement,

$$A_4 = \begin{pmatrix} 118.720 & 1897.420 & -6.059 & 8910.749 & 6165.432 & 764.430 \\ 1897.420 & 69439.917 & 1584.667 & 214216.458 & 152054.792 & 15517.008 \\ -6.059 & 1584.667 & 37584.208 & 1644.758 & 20593.275 & 471.683 \\ 8910.749 & 214216.458 & 1644.758 & 1436243.017 & 765614.758 & 53382.075 \\ 6165.432 & 152054.792 & 20593.275 & 765614.758 & 593461.017 & 22183.542 \\ 764.430 & 15517.008 & 471.683 & 53382.075 & 22183.542 & 33606.175 \end{pmatrix}$$

$$B_4 = (\ 1565.643,\ \ 61067.783,\ \ 4088.633,\ 203740.45,\ \ 189149.316,\ \ 4962.3\ )^T$$

Box III.

Table 1
Individual-specific effects α_i of each city in P-CM model.

Cities	i	$\hat{\alpha}_i^c$
Shanghai	1	13.260
Chongqing	2	4.820
Beijing	3	13.430
Tianjin	4	13.941

Table 2
Individual effect values $\hat{\alpha}_i^L$, $\hat{\alpha}_i^U$ of each city in P-Min–Max model.

Cities	i	$\hat{\alpha}_i^L$	$\hat{\alpha}_i^U$
Shanghai	1	12.190	24.962
Chongqing	2	3.506	14.083
Beijing	3	13.395	18.086
Tianjin	4	11.040	31.134

only the PM10 or PM2.5 affects the final AQI. This paper tries to solve this problem with new methods.

The concentration of all kinds of pollutants changes with space and time. The panel interval-valued data can be used to describe this variation, this paper aims to construct panel interval-valued data models for AQI. Based on the AQI-related data in Shanghai, Chongqing, Beijing and Tianjin in China, this study selected AQI-related data from 4 representative cities for 40 consecutive days (2018.11.01–2018.12.10). Among them, the data of the first 30 days are used to train the models, and the remaining data are used to test the models.

For the panel interval-valued data set $S = \{(x_{it}, y_{it}) |\ i = 1, \ldots, 4;\quad t = 1, 2, \ldots, 40\}$, $y_{it} = \left[y_{it}^L, y_{it}^U\right]$ is considered to be the observed interval-valued dependent variable, which is AQI. y_{it}^L represents the minimum value of AQI in ith city on date t and y_{it}^U represents the maximum value of AQI in ith city on date t, $X_{it} = (x_{it1}, x_{it2}, \ldots, x_{it6})^T$ is an interval-valued independent vectors, which represent the values of CO, NO2, O3, PM10, PM2.5 and SO2 respectively, $x_{itj} = \left[a_{itj}, b_{itj}\right]$, $i = 1, \ldots, 4$, $t = 1, 2, \ldots, 40$, $j = 1, 2, \ldots, 6$, a_{itk} indicates the minimum value of kth pollutant in ith city on date t, and b_{itk} indicates the maximum value of kth pollutant in ith city on date t. Because the original data set is too large, it is not shown here, and can be found in the attached table. Only models results are displayed.

4.1. AQI forecasting of P-CM

The specific construction process of this model is shown in Section 3.1. Data processing can be lead to matrix: (see equations in Box III.)

The regression coefficients are computed as follows:

$$\hat{\beta} = (\hat{\beta}_1^c, \hat{\beta}_2^c, \hat{\beta}_3^c, \hat{\beta}_4^c, \hat{\beta}_5^c, \hat{\beta}_6^c)^T = A_4^{-1} B_4$$
$$= \left(\ 1.784,\ -0.069,\ -0.015,\ 0.261,\ 0.938,\ -0.158\ \right)^T$$

The individual-specific effects can be computed as follows:

$$\hat{\alpha}_i = \bar{y}_{i.}^c - (1.784\bar{x}_{i.1}^c - 0.069\bar{x}_{i.2}^c + 0.015\bar{x}_{i.3}^c + 0.261\bar{x}_{i.4}^c + 0.938\bar{x}_{i.5}^c - 0.158\bar{x}_{i.6}^c)$$

where $\bar{y}_{i.}^c = \frac{1}{30}\sum_{t=1}^{30} y_{it}^c$, $\bar{x}_{i.k}^c = \frac{1}{30}\sum_{t=1}^{30} x_{itk}^c$, $(k = 1, \ldots, 6; i = 1, \ldots, 4)$. The individual-specific effects (a_i) of each city in this model are shown in Table 1.

The AQI P-CM forecasting models can be constructed as follows: $\hat{y}_{it} = [\hat{y}_{it}^L, \hat{y}_{it}^U]$, where

$$\hat{y}_{it}^L = \hat{\alpha}_i^c + 1.784a_{it1} - 0.069a_{it2} + 0.015a_{it3} + 0.261a_{it4} + 0.938a_{it5} - 0.158a_{it6},$$

$$\hat{y}_{it}^U = \hat{\alpha}_i^c + 1.784b_{it1} - 0.069b_{it2} + 0.015b_{it3} + 0.261b_{it4} + 0.938b_{it5} - 0.158b_{it6}.$$

4.2. AQI forecasting of P-Min–Max model and S-P-Min–Max model

4.2.1. P-Min–Max model

The specific construction process of this model is shown in Section 3.2, by processing the lower bounds of the interval, the related matrix can be obtained as equations in Box IV.

The regression coefficients are computed as follows:

$$\hat{\beta}^L = \left(\hat{\beta}_1^L\ \ \hat{\beta}_2^L\ \ \hat{\beta}_3^L\ \ \hat{\beta}_4^L\ \ \hat{\beta}_5^L\ \ \hat{\beta}_6^L\right)^T = (A_4^L)^{-1} B_4^L$$
$$= \left(8.144,\ \ -0.065,\ \ -0.002,\ \ 0.186,\ \ 0.945,\ \ -0.195\right)^T$$

Similarly, by processing the upper bounds of the interval, the related matrix can be got as equations in Box V.

The regression coefficients are computed as follows:

$$B_4^U = (969.253, 40759.4, -6307.9, 102146.433, 103829.367, 2637.667)^T$$

The regression coefficients are computed as follows:

$$\hat{\beta}^U = \left(\hat{\beta}_1^U\ \ \hat{\beta}_2^U\ \ \hat{\beta}_3^U\ \ \hat{\beta}_4^U\ \ \hat{\beta}_5^U\ \ \hat{\beta}_6^U\right)^T = (A_4^U)^{-1} B_4^U$$
$$= \left(-15.719,\ 0.009,\ -0.166,\ 0.347,\ 1.020,\ -0.711\right)^T$$

and the individual effect values $\hat{\alpha}_i^L, \hat{\alpha}_i^U$ are computed as follows:

$$\hat{\alpha}_i^L = \bar{y}_i^L - \hat{\beta}^L \bar{A}_i$$

$$\hat{\alpha}_i^U = \bar{y}_i^U - \hat{\beta}^U \bar{B}_i, i = 1, 2, \ldots, N,$$

where $A_{it} = (a_{it1}, a_{it2}, \ldots, a_{it6})^T$, $B_{it} = \left(b_{it1}, b_{it2}, \ldots, b_{itp}\right)^T$, $\bar{A}_i = \frac{1}{30}\sum_{t=1}^{30} A_{it}$, $\bar{B}_i = \frac{1}{30}\sum_{t=1}^{30} B_{it}$, $\bar{y}_i^L = \frac{1}{30}\sum_{t=1}^{30} y_{it}^L$, $\bar{y}_i^U = \frac{1}{30}\sum_{t=1}^{30} y_{it}^U$.

The individual effect values $\hat{\alpha}_i^L, \hat{\alpha}_i^U$ of each city in this model are shown in Table 2.

$$A_4^L = \begin{pmatrix} 24.456 & 776.628 & 76.082 & 2493.962 & 2297.016 & 109.323 \\ 776.628 & 52571.433 & 20087.833 & 97769.333 & 86656.067 & 6122.533 \\ 76.082 & 20087.833 & 62053.4 & 16904.267 & 19593.067 & 4012.3 \\ 2493.962 & 97769.333 & 16904.267 & 502140.9 & 312562.267 & 11284.367 \\ 2297.016 & 86656.067 & 19593.067 & 312562.267 & 314787.733 & 5249.067 \\ 109.323 & 6122.533 & 4012.3 & 11284.367 & 5249.067 & 4314.2 \end{pmatrix}$$

$$B_4^L = \left(2771.609,\ 102108.367,\ 20135.4,\ 401845.767,\ 368975.067,\ 6722.667\right)^T$$

Box IV.

$$A_4^U = \begin{pmatrix} 10.369 & 410.147 & -56.275 & 782.054 & 850.545 & 28.581 \\ 410.147 & 24313.533 & -3843.433 & 31886 & 35642.033 & 1546.3 \\ -56.275 & -3843.433 & 7628.733 & -2615.6 & -4949.033 & -91.967 \\ 782.074 & 31886 & -2615.6 & 129391.467 & 69390.833 & 2830.167 \\ 850.545 & 35642.033 & -4949.033 & 69390.833 & 91720.5 & 2301.267 \\ 28.582 & 1546.3 & -91.967 & 2830.167 & 2301.267 & 1539.142 \end{pmatrix}$$

Box V.

$$A_5 = \begin{pmatrix} 34.825 & 1186.775 & 19.806 & 3276.016 & 3147.561 & 137.904 \\ 1186.775 & 76884.967 & 16244.4 & 129655.333 & 122298.1 & 7668.833 \\ 19.806 & 16244.4 & 69682.133 & 14288.667 & 14644.033 & 3920.333 \\ 3276.016 & 129655.333 & 14288.667 & 631532.367 & 381953.1 & 14114.533 \\ 3147.561 & 122298.1 & 14644.033 & 381953.1 & 406508.233 & 7550.333 \\ 137.904 & 7668.833 & 3920.333 & 14114.533 & 7550.333 & 4698.9 \end{pmatrix}$$

$$B_5 = (3740.863,\ 142867.767,\ 13827.5,\ 503992.2,\ 472804.433,\ 9360.333)^T$$

Box VI.

Then the AQI P-Min–Max forecasting models can be constructed as follows: $\hat{y}_{it} = [\hat{y}_{it}^L, \hat{y}_{it}^U]$, where

$$\hat{y}_{it}^L = \hat{\alpha}_i^L + 8.144a_{it1} - 0.065a_{it2} - 0.002a_{it3} + 0.186a_{it4} + 0.949a_{it5} - 0.195a_{it6},$$

$$\hat{y}_{it}^U = \hat{\alpha}_i^U - 15.719b_{it1} + 0.009b_{it2} - 0.166b_{it3} + 0.347b_{it4} + 1.020b_{it5} - 0.711b_{it6}.$$

4.2.2. S-P-Min–Max model

The specific construction process of this model is shown in Section 3.1. Data processing can lead to matrix as equations in Box VI.

The regression coefficients are computed as follows:

$$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)^T = A_5^{-1}B_5 = \left(3.313,\ -0.053,\ -0.024,\ 0.223,\ 0.948,\ -0.192\right)^T$$

and the individual effect values $\hat{\alpha}_i^L$, $\hat{\alpha}_i^U$ are computed as follows:

$$\hat{\alpha}_i^L = \bar{y}_{i.}^L - (3.313\bar{a}_{i.1} - 0.053\bar{a}_{i.2} - 0.024\bar{a}_{i.3} + 0.223\bar{a}_{i.4} + 0.948\bar{a}_{i.5} - 0.192\bar{a}_{i.6})$$

$$\hat{\alpha}_i^U = \bar{y}_{i.}^U - (3.313\bar{b}_{i.1} - 0.053\bar{b}_{i.2} - 0.024\bar{b}_{i.3} + 0.223\bar{b}_{i.4} + 0.948\bar{b}_{i.5} - 0.192\bar{b}_{i.6}).$$

The individual effect values $\hat{\alpha}_i^L$, $\hat{\alpha}_i^U$ of each city in S-P-Min–Max model are shown in Table 3.

Table 3
Individual effect values $\hat{\alpha}_i^L$, $\hat{\alpha}_i^U$ of each city in S-P-Min–Max model.

Cities	i	$\hat{\alpha}_i^L$	$\hat{\alpha}_i^U$
Shanghai	1	15.132	14.717
Chongqing	2	5.402	6.715
Beijing	3	14.893	16.393
Tianjin	4	15.654	15.966

Then the AQI S-P-Min–Max forecasting models can be constructed follows: $\hat{y}_{it} = [\hat{y}_{it}^L, \hat{y}_{it}^U]$, where

$$\hat{y}_{it}^L = \hat{\alpha}_i^L + 3.313a_{it1} - 0.053a_{it2} - 0.024a_{it3} + 0.223a_{it4} + 0.948a_{it5} - 0.192a_{it6},$$

$$\hat{y}_{it}^U = \hat{\alpha}_i^U + 3.313b_{it1} - 0.053b_{it2} - 0.024b_{it3} + 0.223b_{it4} + 0.948b_{it5} - 0.192b_{it6}.$$

Sometimes P-Min–Max model and its special model do not guarantee the mathematical coherence of the predicted interval bounds, then the response variables are predicted as follows:

$$\hat{y}_{it} = \left[\hat{y}_{it}^L, \hat{y}_{it}^U\right] = \begin{cases} \left[\dfrac{\hat{y}_{it}^L + \hat{y}_{it}^U}{2}, \dfrac{\hat{y}_{it}^L + \hat{y}_{it}^U}{2}\right] & \text{if } \hat{y}_{it}^L > \hat{y}_{it}^U; \\ [\hat{y}_{it}^L, \hat{y}_{it}^U] & \text{if } \hat{y}_{it}^L \le \hat{y}_{it}^U. \end{cases}$$

$$A_{61} = \begin{pmatrix} 14.125 & 508.548 & -44.549 & 1359.774 & 1311.901 & 58.976 \\ 508.548 & 30672.958 & 885.258 & 52853.3 & 52270.642 & 3543.375 \\ -44.549 & 885.258 & 18477.067 & 1727.842 & 3928.108 & 1145.025 \\ 1359.774 & 52853.3 & 1727.842 & 244995.542 & 151395.933 & 6415.783 \\ 1311.901 & 52270.642 & 3928.108 & 151395.933 & 161458.992 & 3895.075 \\ 58.976 & 3543.375 & 1145.025 & 6415.783 & 3895.075 & 1539.142 \end{pmatrix}$$

$$A_{62} = A_{63} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{64} = \begin{pmatrix} 3.287 & 84.840 & 54.452 & 278.234 & 261.880 & 9.977 \\ 84.840 & 7769.525 & 7236.942 & 11974.367 & 8878.408 & 291.042 \\ 54.452 & 7236.942 & 16364 & 5416.492 & 3393.908 & 815.142 \\ 278.234 & 11974.367 & 5416.492 & 70770.642 & 39580.617 & 641.483 \\ 261.880 & 8878.408 & 3393.908 & 39580.617 & 41795.125 & -119.908 \\ 9.977 & 291.042 & 815.142 & 641.483 & -119.908 & 810.308 \end{pmatrix}$$

$$B_6 = (1565.643 \quad 61067.783 \quad 4088.633 \quad 203740.45 \quad 189149.317 \quad 4962.3 \quad 304.789 \quad 10366.1 \quad 2825.117 \quad 48255.65 \quad 47252.9 \quad -282.133)^T$$

Box VII.

Table 4
Individual effect values $\hat{\alpha}_i^c$, $\hat{\alpha}_i^r$ of each city in P-CRM model.

Cities	i	$\hat{a}_i^c$	$\hat{a}_i^r$
Shanghai	1	1.563	13.260
Chongqing	2	0.075	4.820
Beijing	3	1.224	13.430
Tianjin	4	0.962	13.941

4.3. AQI forecasting of P-CRM

The specific construction process of this model is shown in Section 3.3. Data processing can lead to matrix: $A_6 = \begin{pmatrix} A_{61} & A_{62} \\ A_{63} & A_{64} \end{pmatrix}$, where (see equations in Box VII.)

The regression coefficients are computed as follows:

$$\hat{\beta} = (\hat{\beta}_1^c, \hat{\beta}_2^c, \hat{\beta}_3^c, \hat{\beta}_4^c, \hat{\beta}_5^c, \hat{\beta}_6^c, \hat{\beta}_1^r, \hat{\beta}_2^r, \hat{\beta}_3^r, \hat{\beta}_4^r, \hat{\beta}_5^r, \hat{\beta}_6^r,)^T = A_6^{-1} B_6$$
$$= (1.784, \quad -0.069, \quad 0.015, \quad 0.261, \quad 0.938, \quad -0.158, \quad 5.590, \quad 0.115, \quad -0.121, \quad 0.102, \quad 0.984, \quad -0.272)^T$$

The individual effect values $\hat{\alpha}_i^c, \hat{\alpha}_i^r$ of each city can be computed as follows:

$$\hat{\alpha}_i^c = \bar{y}_{i.}^c - (1.784\bar{x}_{i.1}^c - 0.069\bar{x}_{i.2}^c + 0.015\bar{x}_{i.3}^c + 0.261\bar{x}_{i.4}^c + 0.938\bar{x}_{i.5}^c - 0.158\bar{x}_{i.6}^c)$$

$$\hat{\alpha}_i^r = \bar{y}_{i.}^r - (5.590\bar{x}_{i.1}^r + 0.115\bar{x}_{i.2}^r - 0.121\bar{x}_{i.3}^r + 0.102\bar{x}_{i.4}^r + 0.984\bar{x}_{i.5}^r - 0.272\bar{x}_{i.6}^r).$$

The individual effect values $\hat{\alpha}_i^c, \hat{\alpha}_i^r$ of each city in P-CRM model are shown in Table 4.

Then the AQI P-CRM forecasting models can be constructed as follows: $\hat{y}_{it} = (\hat{y}_{it}^c, \hat{y}_{it}^r)$, where

$$\hat{y}_{it}^c = \hat{\alpha}_i^c - 1.784x_{it1}^c - 0.069x_{it2}^c + 0.015x_{it3}^c + 0.261x_{it4}^c + 0.938x_{it5}^c - 0.158x_{it6}^c,$$

$$\hat{y}_{it}^r = \hat{\alpha}_i^r + 5.590x_{it1}^r + 0.115x_{it2}^r - 0.121x_{it3}^r + 0.102x_{it4}^r + 0.984x_{it5}^r - 0.272x_{it6}^r.$$

Then the predicted value

$$\hat{y}_{it} = [\hat{y}_{it}^L, \hat{y}_{it}^U] = \begin{cases} [\hat{y}_{it}^c - \hat{y}_{it}^r, \hat{y}_{it}^c + \hat{y}_{it}^r] & \text{if } \hat{y}_{it}^r \geq 0; \\ [\hat{y}_{it}^c, \hat{y}_{it}^c] & \text{if } \hat{y}_{it}^r < 0. \end{cases}$$

4.4. The evaluation of three models

All three models are presented, but which one is the best to predict the air quality index? The fitting effect and prediction performance of these models were evaluated by calculating three error measures: mean magnitude of relative error (*MMER*), mean average absolute error (*MAE*), root mean squared error (*RMSE*). The better the fitting degree of the model is, the lower the error measure is.

To prove that our proposed models are superior to the general interval-valued regression model, this study conducts two groups of experiments below. In the first group of experiments, the data of the four cities are regarded as a whole, and the corresponding solutions were obtained by using the corresponding pooled panel interval-valued regression model (that is, the ordinary interval-valued linear regression discussed in Section 2.2), this means that such experimental data are regarded as having no individual effect. In the second set of experiments, the data of the four cities are regarded as panel interval-valued data with individual behaviour differences (individual effects), and the corresponding solutions are obtained in Sections 4.1–4.3.

For convenience, the $MMER_i$ of pooled panel interval-valued data regression model and fixed effects panel interval-valued data regression model rewrite $P-MMER_i$ and $F-MMER_i$ respectively. $MMER$, MAE_i, MAE, $RMSE_i$ and $RMSE$ are also expressed similarly.

The individual index $MMER_i$ $(i = 1, 2, \ldots, N)$ and total index $MMER$ are defined respectively:

$$MMER_i = \frac{1}{2T} \sum_{t=1}^{T} \left\{ \left| \frac{\hat{y}_{it}^L - y_{it}^L}{\hat{y}_{it}^L} \right| + \left| \frac{\hat{y}_{it}^U - y_{it}^U}{\hat{y}_{it}^U} \right| \right\},$$

$$MMER = \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \left| \frac{\hat{y}_{it}^L - y_{it}^L}{\hat{y}_{it}^L} \right| + \left| \frac{\hat{y}_{it}^U - y_{it}^U}{\hat{y}_{it}^U} \right| \right\}.$$

Table 5
Performance of the fitting errors estimate value of P-CM.

Cities	Evaluation measure (fitting)					
	$P-MMER_i$	$F-MMER_i$	$P-MAE_i$	$F-MAE_i$	$P-RMSE_i$	$F-RMSE_i$
Shanghai	0.174	0.087	7.175	4.255	9.346	5.975
Chongqing	0.064	0.042	5.130	2.659	7.743	5.611
Beijing	0.161	0.109	9.420	8.670	13.268	11.616
Tianjin	0.111	0.074	8.575	7.178	11.904	10.307
Total	0.127	0.078	7.575	5.691	11.091	8.952

Table 6
Performance of the forecasting errors estimate value of P-CM.

Cities	Evaluation measure (forecasting)					
	$P-MMER_i$	$F-MMER_i$	$P-MAE_i$	$F-MAE_i$	$P-RMSE_i$	$F-RMSE_i$
Shanghai	0.179	0.132	5.487	4.960	6.869	6.183
Chongqing	0.071	0.043	3.455	1.571	3.880	2.092
Beijing	0.340	0.189	15.129	13.232	20.071	18.208
Tianjin	0.213	0.141	14.235	12.135	20.741	19.436
Total	0.201	0.126	9.576	7.975	15.003	13.769

Table 7
Performance of the fitting errors estimate value of P-Min–Max.

Cities	Evaluation measure (fitting)					
	$P-MMER_i$	$F-MMER_i$	$P-MAE_i$	$F-MAE_i$	$P-RMSE_i$	$F-RMSE_i$
Shanghai	0.101	0.105	4.425	4.532	5.829	5.902
Chongqing	0.085	0.049	5.705	3.267	8.517	5.626
Beijing	0.101	0.097	8.206	8.067	11.432	10.918
Tianjin	0.076	0.065	6.605	6.340	9.181	8.411
Total	0.091	0.079	6.235	5.551	9.097	8.201

Table 8
Performance of the forecasting errors estimate value of P-Min–Max.

Cities	Evaluation measure (forecasting)					
	$P-MMER_i$	$F-MMER_i$	$P-MAE_i$	$F-MAE_i$	$P-RMSE_i$	$F-RMSE_i$
Shanghai	0.126	0.145	4.412	5.641	6.906	7.456
Chongqing	0.072	0.070	3.143	2.202	3.589	2.958
Beijing	0.239	0.190	14.629	13.364	20.629	19.291
Tianjin	0.178	0.137	14.035	12.564	21.846	20.342
Total	0.154	0.136	9.054	8.442	15.637	14.820

The individual index MAE_i $(1, 2, \ldots, N)$ and total index MAE of lower and upper bounds of the predictive interval are defined respectively:

$$MAE_i = \frac{1}{2T}\sum_{t=1}^{T}\left\{\left|\hat{y}_{it}^{L} - y_{it}^{L}\right| + \left|\hat{y}_{it}^{U} - y_{it}^{U}\right|\right\},$$

$$MAE = \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\left|\hat{y}_{it}^{L} - y_{it}^{L}\right| + \left|\hat{y}_{it}^{U} - y_{it}^{U}\right|\right\}.$$

The individual index $RMSE_i$ $(1, 2, \ldots, N)$ and total index $RMSE$ of lower and upper bounds of the interval are defined respectively:

$$RMSE_i = \frac{1}{2}\left\{\sqrt{\frac{\sum_{t=1}^{T}\left(\hat{y}_{it}^{L} - y_{it}^{L}\right)^2}{T}} + \sqrt{\frac{\sum_{t=1}^{T}\left(\hat{y}_{it}^{U} - y_{it}^{U}\right)^2}{T}}\right\},$$

$$RMSE = \frac{1}{2}\left\{\sqrt{\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{y}_{it}^{L} - y_{it}^{L}\right)^2}{NT}} + \sqrt{\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{y}_{it}^{U} - y_{it}^{U}\right)^2}{NT}}\right\}.$$

Based on the training data set and test data set in the attached table, the fitting errors and prediction errors of the models can be obtained. The fitting and forecasting errors for pooled panel interval-valued data regression model and three kinds of the fixed effects panel interval-valued data regression models are given in Tables 5 to 12.

As shown in Tables 5 to 12, our proposed three kinds of fixed effects panel interval-valued data regression models all have good fitting effects and prediction performance. Compared with the three models, the fitting effect and prediction performance of the P-CM model is the worst. Meanwhile, compared with the pooled panel interval-valued regression model, in general, our proposed three kinds of fixed effects panel interval-valued data regression models all have good fitting effect and prediction performance. This shows that the AQI in the four cities are heterogeneous and the AQI-related data in the four cities are not from a population, and there are individual behaviour differences in forecasting. It is best to use the panel interval-valued data model in forecasting the AQI.

In this study, our proposed panel interval-valued data regression models are applied in AQI forecasting. But according to the air monitoring data, the highest and lowest concentrations of pollutants in a day may appear at different time points, which results in a poor linear relationship. This may result in a bad forecasting effect.

5. Conclusions and further studies

This paper proposes three kinds of fixed effects panel interval-valued data regression models and presents their parameter estimation methods. Our proposed models are applied in AQI forecasting, and the experimental evaluation shows that our proposed panel interval-valued data regression models enjoy better fitting effect and predictive performance. Therefore, the three models proposed in this article can be a good choice for analysing the linear correlation between heterogeneous interval variables.

Table 9
Performance of the fitting errors estimate value of S-P-Min–Max.

Cities	Evaluation measure (fitting)					
	$P-MMER_i$	$F-MMER_i$	$P-MAE_i$	$F-MAE_i$	$P-RMSE_i$	$F-RMSE_i$
Shanghai	0.079	0.083	3.903	3.938	5.374	5.336
Chongqing	0.092	0.046	5.901	2.842	8.478	5.189
Beijing	0.118	0.119	8.956	9.292	12.347	11.912
Tianjin	0.078	0.075	7.021	7.198	10.183	9.918
Total	0.092	0.081	6.445	5.817	9.514	8.727

Table 10
Performance of the forecasting errors estimate value of S-P-Min–Max.

Cities	Evaluation measure (forecasting)					
	$P-MMER_i$	$F-MMER_i$	$P-MAE_i$	$F-MAE_i$	$P-RMSE_i$	$F-RMSE_i$
Shanghai	0.504	0.135	8.779	5.064	10.777	6.503
Chongqing	0.085	0.055	3.067	2.048	3.598	2.541
Beijing	0.675	0.195	19.712	14.017	26.182	19.637
Tianjin	0.361	0.150	17.910	12.963	26.841	20.960
Total	0.406	0.134	12.367	8.523	19.641	14.849

Table 11
Performance of the fitting errors estimate value of P-CRM.

Cities	Evaluation measure (fitting)					
	$P-MMER_i$	$F-MMER_i$	$P-MAE_i$	$F-MAE_i$	$P-RMSE_i$	$F-RMSE_i$
Shanghai	0.077	0.086	3.788	4.289	5.318	5.567
Chongqing	0.057	0.037	4.265	2.643	7.376	4.848
Beijing	0.140	0.118	8.724	8.774	11.641	11.147
Tianjin	0.077	0.070	6.823	6.746	9.580	9.318
Total	0.088	0.078	5.900	5.613	8.874	8.245

Table 12
Performance of the forecasting errors estimate value of P-CRM.

Cities	Evaluation measure (forecasting)					
	$P-MMER_i$	$F-MMER_i$	$P-MAE_i$	$F-MAE_i$	$P-RMSE_i$	$F-RMSE_i$
Shanghai	0.126	0.132	4.752	5.156	6.564	6.562
Chongqing	0.082	0.054	3.524	2.281	3.908	2.663
Beijing	0.240	0.189	15.079	13.965	21.566	20.339
Tianjin	0.156	0.129	13.097	11.936	21.753	20.802
Total	0.151	0.126	9.029	8.334	16.098	15.094

There are still some limitations on our proposed panel interval-valued data regression models, that is, the mathematical coherence of the predicted interval bounds is sometimes not guaranteed. For future studies, some constraints can be added in the least square method to ensure the mathematical coherence of the prediction interval boundary. In addition, the future study focuses on the robust panel interval-valued data models, such as adding corresponding outlier penalty terms in the models to reduce the sensitivity of the least squares method, kernel methods for panel interval-valued data. Furthermore, the panel interval-valued data nonlinear regression models are to be discussed.

CRediT authorship contribution statement

Ai-bing Ji: Development or design of models. **Jin-jin Zhang:** Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data. **Xing He:** Application of statistical, mathematical, computational, or other formal techniques to analyse or synthesize study data. **Yu-hang Zhang:** Conducting a research, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to the anonymous referees for their careful revision, valuable suggestions, and comments which improved this paper. The authors would like to thank the financial support from Hebei Key Laboratory of Machine Learning and Computational Intelligence.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2021.107798.

References

[1] M.F. Harkat, M. Mansouri, K. Abodayeh, M. Nounou, H. Nounou, New sensor fault detection and isolation strategy–based interval-valued data, J. Chemometr. 34 (2020) e3222, http://dx.doi.org/10.1002/cem.3222.
[2] E. Diday, Thinking by classes in data science: The symbolic data analysis paradigm, WIREs Comput. Stat. 8 (2016) 172–205, http://dx.doi.org/10.1002/wics.1384.
[3] E. Diday, H.-H. Bock, Analysis of Symbolic Data, Springer, Berlin, Heidelberg, ISBN: 978-3-642-57155-8, 1999.
[4] H. Park, F. Sakaori, Forecasting symbolic candle chart-valued time series, Commun. Stat. Appl. Methods 21 (2014) 471–486, http://dx.doi.org/10.5351/CSAM.2014.21.6.471.
[5] D. Wang, W. Song, W. Pedrycz, L. Cai, An integrated neural network with nonlinear output structure for interval-valued data, J. Intell. Fuzzy Systems 40 (2021) 673–683, http://dx.doi.org/10.3233/jifs-200500.

[6] Y. Zebin, D.K.J. Lin, Z. Aijun, Interval-valued data prediction via regularized artificial neural network, Neurocomputing 331 (2019) 336–345, http://dx.doi.org/10.1016/j.neucom.2018.11.063.
[7] L. Billard, E. Diday, From the statistics of data to the statistics of knowledge, J. Amer. Statist. Assoc. 98 (2003) 470–487, http://dx.doi.org/10.1198/016214503000242.
[8] L. Billard, E. Diday, Regression analysis for interval-valued data, in: Data Analysis, Classification and Related Methods, Proceedings of the Seventh Conference of the International Federation of Classification Societies, Springer, Berlin, Heidelberg, ISBN: 978-3-642-57155-8, 2000, http://dx.doi.org/10.1007/978-3-642-59789-3_58.
[9] L. Billard, E. Diday, H.-H. Bock, Symbolic regression analysis, in: K. Jajuga, A. Soko Aowski (Eds.), Classification, Clustering, and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin Heidelberg, 2002, pp. 281–288, http://dx.doi.org/10.1007/978-3-642-56181-8_31.
[10] E.d.A. Lima Neto, F.d.A.T. de Carvalho, Centre and range method for fitting a linear regression model to symbolic interval data, Comput. Statist. Data Anal. 52 (2008) 1500–1515, http://dx.doi.org/10.1016/j.csda.2009.08.010.
[11] E.d.A. Lima Neto, F.d.A.T. de Carvalho, Constrained linear regression models for symbolic interval-valued variables, Comput. Statist. Data Anal. 54 (2010) 333–347, http://dx.doi.org/10.1016/j.csda.2009.08.010.
[12] F.d.A.T. de Carvalho, E.d.A. Lima Neto, K.C.F. da Silva, A clusterwise nonlinear regression algorithm for interval-valued data, Inform. Sci. 555 (2021) 357–385, http://dx.doi.org/10.1016/j.ins.2020.10.054.
[13] L.C. Souza, R.M.C.R. Souza, G.J.A. Amaral, T.M. Silva Filho, A parametrized approach for linear regression of interval data, Knowl.-Based Syst. 131 (2017) 149–159, http://dx.doi.org/10.1016/j.knosys.2017.06.012.
[14] M. Xu, Z. Qin, A Bivariate Bayesian Method for Interval-Valued Regression Models, Knowledge-Based Systems, 2021, http://dx.doi.org/10.1016/j.knosys.2021.107396.
[15] Y. Sun, A. Han, Y. Hong, S. Wang, Threshold autoregressive models for interval-valued time series data, J. Econometrics 206 (2018) 414–446, http://dx.doi.org/10.1016/j.jeconom.2018.06.009.
[16] L. Lin, H. Chien, S. Lee, Symbolic interval-valued data analysis for time series based on auto-interval-regressive models, Stat. Methods Appl. 30 (2021) 295–315, http://dx.doi.org/10.1007/s10260-020-00525-7.
[17] G. Gonzalez-Rivera, Y. Luo, E. Ruiz, Prediction regions for interval-valued time series, J. Appl. Econometrics 35 (2020) 373–390, http://dx.doi.org/10.1002/jae.2754.
[18] S. Binbin, C. Hongmei, Y. Lei, L. Tianrui, X. Weihua, L. Chuan, Feature selection for dynamic interval-valued ordered data based on fuzzy dominance neighborhood rough set, Knowl.-Based Syst. 227 (2021) 107223, http://dx.doi.org/10.1016/j.knosys.2021.107223.
[19] C. Hsiao, Analysis of Panel Data, Cambridge University Press, ISBN: 9781139839327, 2014, http://dx.doi.org/10.1017/CBO9781139839327.
[20] S. Zhao, R. Liu, Z. Shang, Statistical inference on panel data models: A kernel ridge regression method, J. Bus. Econom. Statist. 39 (2019) 325–337, http://dx.doi.org/10.1080/07350015.2019.1660176.
[21] E. Aristodemou, Semiparametric identification in panel data discrete response models, J. Econometrics 220 (2021) 253–271, http://dx.doi.org/10.1016/j.jeconom.2020.04.002.
[22] L. Liu, H.R. Moon, F. Schorfheide, Forecasting with dynamic panel data models, Econometrica 88 (2020) 171–201, http://dx.doi.org/10.3982/ECTA14952.
[23] B.H. Beyaztas, S. Bandyopadhyay, Robust estimation for linear panel data models, Stat. Med. 39 (2020) 4421–4438, http://dx.doi.org/10.1002/sim.8732.
[24] H. Liu, Y. Pei, Q. Xu, Qunfang, estimation for varying coefficient panel data model with cross-sectional dependence, Metrika 83 (2019) 377–410, http://dx.doi.org/10.1007/s00184-019-00739-0.
[25] L. Boya, Health effects of air pollution: Evidence from China, Low Carbon Econ. 10 (2019) 81–101, http://dx.doi.org/10.4236/lce.2019.103006.
[26] H. Luo, Y. Han, X. Cheng, C. Lu, Y. Wu, Spatiotemporal variations in particulate matter and air quality over China: National, regional and urban scales, Atmosphere 12 (2020) 253, http://dx.doi.org/10.3390/atmos12010043.
[27] L. Xu, Partial correlation analysis of O3 and NO2 in Beijing area, Urban Environ. Urban Ecol. 2 (2013) 67–71.
[28] X. Xu, Combined multifractal analysis of PM2.5 trends, J. Hefei Univ. 24 (3) (2014) 26–30.
[29] M. Xu, Analysis and prediction of AQI influence factors based on partial correlation ands tepwise regression methods, Front. Environ. Prot. 3 (2017) 191–201.