Supplementary materials for this article are available at https://doi.org/10.1007/s13253-023-00561-2.

A Nonparametric Bootstrap Method for Heteroscedastic Functional Data

Rubén FERNÁNDEZ-CASAL, Sergio CASTILLO-PÁEZ, and Miguel FLORES

The objective is to provide a nonparametric bootstrap method for functional data that consists of independent realizations of a continuous one-dimensional process. The process is assumed to be nonstationary, with a functional mean and a functional variance, and dependent. The resampling method is based on nonparametric estimates of the model components. Numerical studies were conducted to check the performance of the proposed procedure, by approximating the bias and the standard error of two estimators. A practical application of the proposed approach to pollution data has also been included. Specifically, it is employed to make inference about the annual trend of ground-level ozone concentration at Yarner Wood monitoring station in the United Kingdom.

Supplementary material to this paper is provided online.

Key Words: Functional data analysis; Resampling methods; Local linear estimation; Variogram.

1. INTRODUCTION

Air pollution is considered one of the biggest health challenges worldwide in urban environments. There are a wide variety of urban air pollutants, such as carbon monoxide (CO), nitrogen oxides (NO_x), sulphur dioxide (SO_2), particulate matter ($PM_{2.5}$ and PM_{10}) and ozone (O_3). In this work, we will focus on ground-level ozone, as it has been shown to have serious health effects on humans and can also damage plants and trees (e.g. Karlsson et al. 2017), but the proposed methodology could be also applied to other pollutants.

Rubén Fernández-Casal, Sergio Castillo-Páez and Miguel Flores contributed equally to this work.

R. Fernández-Casal (✉), Departamento de Matemáticas, Universidade da Coruña, Facultad de Informática, Campus de Elviña s/n, 15071 A Coruña, Spain. Centro de Investigación TIC (CITIC), Universidade da Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain . (E-mail: *ruben.fcasal@udc.es*).
S. Castillo-Páez, Departamento de Ciencias Exactas, Universidad de las Fuerzas Armadas ESPE, Av. General Rumiñahui s/n, 171103 Sangolquí, Ecuador . (E-mail: *sacastillo@espe.edu.ec*).
M. Flores, Department of Mathematics, Faculty of Sciences, Escuela Politécnica Nacional, Ladrón de Guevara E11–253, 170525 Quito, Pichincha, Ecuador. SIGTI Group, Faculty of Administrative Sciences, Escuela Politécnica Nacional, Ladrón de Guevara E11–253, 170525 Quito, Pichincha, Ecuador .
(E-mail: *miguel.flores@epn.edu.ec*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 29, Number 1, Pages 169–184
https://doi.org/10.1007/s13253-023-00561-2

Nowadays, technological developments related to sensor technology and IoT have made it possible to have data sources where air pollutants and related variables are continuously monitored. For instance, the UK Automatic Urban and Rural Network (AURN, https://uk-air.defra.gov.uk/networks/network-info?view=aurn) records large volumes of information, including the pollutants cited above. As the observed values can be considered realizations of a functional process, the application of functional data analysis (FDA) techniques may be useful on the assessment of air pollution impact on human health and ecosystems (see e.g. Ramsay and Silverman 2005; Ferraty and Vieu 2006; Manteiga and Vieu 2007; Ullah and Finch 2013 for a general view of this methodology). There are several studies available in the literature on functional methods applied to environmental data (e.g. Febrero et al. 2008; Delicado et al. 2010; Giraldo et al. 2010; Embling et al. 2012; Sancho et al. 2014; Xiao and Hu 2018). Many of these developments used the tools implemented in different R packages. Among them, we may highlight the packages `fda` (Ramsay et al. 2020), `rainbow` (Shang and Hyndman 2019), and `fda.usc` (Febrero and Oviedo 2012), which allow the application of descriptive, outlier detection, regression, classification, clustering, dimension reduction, variance analysis and bootstrap methods, among others.

Bootstrap methods for functional data can be of great interest for problem solving in many fields, including air quality data analysis (see e.g. McMurry and Politis 2011). They are often used to approximate characteristics of the distribution of statistics related to the process under study. For instance, among many other applications, this includes estimating the probability that a pollutant exceeds a certain threshold value (e.g. UK air quality guidelines state that eight-hour average of ozone should not exceeded $100\,\mu\mathrm{g\,m}^{-3}$ more than 10 times a year). Classical bootstrap procedures have been employed for functional data analysis, including naive, parametric, and block bootstrap methods. For example, Ferraty et al. (2010) studied the asymptotic validity of naive and wild bootstrap methods for inference on a nonparametric functional regression model. Several resampling procedures specifically designed for functional data have also been proposed (de Castro et al. 2005; Politis and Romano 2010). Among them, we may highlight the smoothed bootstrap method proposed in Cuevas et al. (2006), where they compare its performance with those of the naive and parametric bootstrap methods. However, for the results obtained with a bootstrap procedure to be reliable, it must adequately reproduce the variability of the underlying process.

The proposed bootstrap procedure is an adaptation of the method developed by Castillo-Páez et al. (2019) for spatial data. The idea would be to consider the functional process as a spatial process of dimension one so that repeated (independent) measurements are observed at some discretization points. This method requires the modelling of the variability of the process, which is done employing nonparametric techniques. Following the usual procedure in geostatistics, the modelling of the dependence is done through the semivariogram. For this purpose, a new package `npfda` (Fernandez-Casal et al. 2023) has been developed, adapting the tools implemented in the `npsp` package (Fernandez-Casal 2023) for this particular case (see the supplementary material for more details).

This methodology was applied to ground-level ozone data. The data set consist of daily averages of ozone concentration ($\mu\mathrm{g\,m}^{-3}$) recorded over the period from 1988 to 2020 at the Yarner Wood monitoring site in the UK (available at https://uk-air.defra.gov.uk/data). These data were pre-processed, applying the usual outlier detection and data imputation

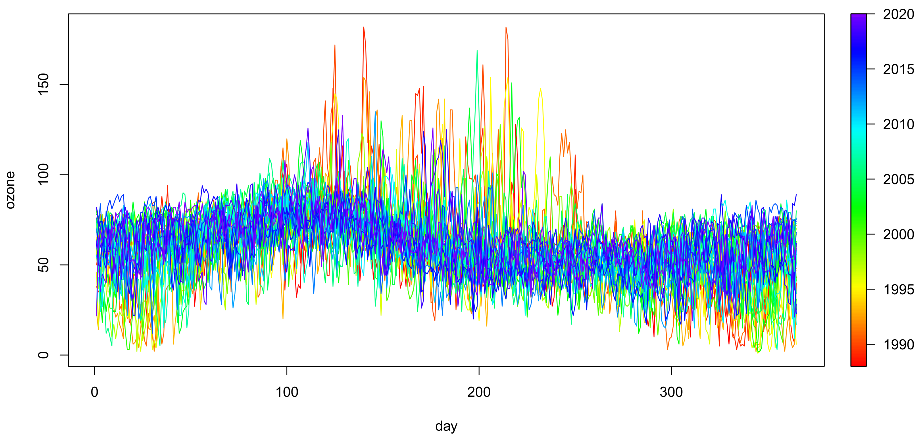

Figure 1. Annual ozone curves (μg m^{-3}), from 1988 to 2020, at the Yarner Wood site in the UK (using a rainbow colour scale for the year) .

methods, using the package `climatol` (Guijarro 2019). It is assumed that the observations corresponding to each year are (partial) realizations of a functional process, so the data consist of 33 curves observed at 365 discretization points. This curves are shown in Fig. 1. As an initial objective, we will assume that we intend to make inferences about the annual trend of the ozone level. Specifically, the estimation of the mean curve and the construction of confidence intervals (Sect. 4). However, this methodology can be used for a large number of problems including the analysis of other pollution related variables.

The remainder of the paper is organized as follows. The general model, the nonparametric estimators and the proposed bootstrap method, are presented in Sect. 2. The performance of this procedure is illustrated through numerical studies in Sect. 3, where the results are compared with those derived from the naive and smoothed bootstrap approaches. In Sect. 4, we describe an application of the proposed methodology to the ozone data. Finally, Sect. 5 contains a summary of the main conclusions and some finals remarks.

2. METHODOLOGY

Suppose that $\mathcal{S}_n = \{Y_i(t)\}_{i=1}^n$, for $t \in [a, b] \subset \mathbb{R}$, is a set of n independent observations of a functional variable $Y(t)$ defined over $\mathbb{R}$, verifying:

$$Y_i(t) = \mu(t) + \sigma(t)\varepsilon_i(t), \tag{1}$$

being $\mu(t)$ the functional trend, $\sigma^2(t)$ the functional variance, and $\varepsilon_i(t)$ a random error process with zero mean, unit variance and correlations

$$\operatorname{Cov}\left(\varepsilon_i(t), \varepsilon_{i'}(t')\right) = \delta_{ii'}\rho\left(\left|t - t'\right|\right),$$

for $1 \leq i, i' \leq n$ and $a \leq t, t' \leq b$, where $\delta_{ii'} = 1$ if $i = i'$, $\delta_{ii'} = 0$ if $i \neq i'$ and $\rho(\cdot)$ is the correlogram function.

In practice, each $Y_i(t)$ is observed at a discrete set of points $t_j \in [a, b] \subset \mathbb{R}$, with $j = 1, \ldots, p$. This set of observations can be expressed as a matrix $\mathbf{Y}$ of order $n \times p$, with $\mathbf{Y}_{ij} = Y_i(t_j)$. Furthermore, if $\mathbf{y}_i = \left(Y_i(t_1), \ldots, Y_i(t_p)\right)^\top$ is the vector corresponding to the i-th row of $\mathbf{Y}$, the elements of its covariance matrix $\operatorname{Cov}(\mathbf{y}_i) = \mathbf{\Sigma}_0$ (within-curve covariance matrix) are

$$(\mathbf{\Sigma}_0)_{jj'} = \sigma(t_j)\sigma(t_{j'})\rho\left(\left|t_j - t_{j'}\right|\right),$$

for $i = 1, \ldots, n$. Consequently, $\mathbf{\Sigma}_0 = \mathbf{D}\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{D}$, where $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ (within-curve correlation matrix) is the covariance matrix of $\boldsymbol{\varepsilon}_i = \left(\varepsilon_i(t_1), \ldots, \varepsilon_i(t_p)\right)^\top$, for $i = 1, \ldots, n$, being $\mathbf{D} = \operatorname{diag}(\sigma(t_1), \ldots, \sigma(t_p))$. Nevertheless, the dependence structure is estimated through the error semivariogram:

$$\gamma(u) = \frac{1}{2}\operatorname{Var}(\varepsilon(t) - \varepsilon(t+u)) = 1 - \rho(u).$$

2.1. Nonparametric Estimation

The proposed procedure starts with the nonparametric estimation of the trend, the conditional variance and the dependence, following an iterative algorithm similar to the one described in Fernández-Casal et al. (2017). However, in this case, since multiple realizations of the process are available, it has been observed that a bias correction in the estimation of the small-scale variability seems to be not necessary.

The trend is estimated by linear smoothing of

$$\left\{(t_j, Y_i(t_j)) : 1 \le i \le n, 1 \le j \le p\right\}.$$

This estimator can be written explicitly in terms of the sample means $\bar{Y}(t) = \frac{1}{n}\sum_i Y_i(t)$:

$$\hat{\mu}(t) = \mathbf{e}_1^\top \left(\mathbf{X}_t^\top \mathbf{W}_t \mathbf{X}_t\right)^{-1} \mathbf{X}_t^\top \mathbf{W}_t \bar{\mathbf{y}} = \mathbf{s}_t^\top \bar{\mathbf{y}} \tag{2}$$

where $\bar{\mathbf{y}} = \left(\bar{Y}(t_1), \ldots, \bar{Y}(t_p)\right)^\top$, $\mathbf{e}_1 = (1, 0)^\top$, $\mathbf{X}_t$ is a matrix with the j-th row equal to $\left(1, t_j - t\right)$, $\mathbf{W}_t = \operatorname{diag}\{K_h(t_1 - t), \ldots, K_h(t_p - t)\}$, $K_h(u) = \frac{1}{h}K(\frac{u}{h})$, K is a kernel function and h is the bandwidth parameter.

The small-scale variability of the process, determined by the conditional variance and the temporal dependence of the error process, is estimated from the residuals $r_{ij} = Y_i(t_j) - \hat{\mu}(t_j)$. An estimate of the conditional variance $\hat{\sigma^2}(\cdot)$ is obtained by linear smoothing of:

$$\{(t_j, r_{ij}^2) : 1 \le i \le n, 1 \le j \le p\},$$

analogously to the trend estimate, using a bandwidth h_2.

A pilot local linear estimate of the error semivariogram $\hat{\gamma}(\cdot)$ is obtained by the linear smoothing of the semivariances,

$$\left\{\left(t_j - t_{j'}, \tfrac{1}{2}(\hat{\varepsilon_{ij}} - \hat{\varepsilon_{ij'}})^2\right) : 1 \le i \le n, 1 \le j < j' \le p\right\},$$

of the standardized residuals $\hat{\varepsilon_{ij}} = r_{ij}/\hat{\sigma}(t_j)$. The corresponding bandwidth parameter will be denoted by h_3. Additionally, as this estimator is not necessarily conditionally negative definite (it cannot be used directly for prediction or simulation), a flexible Shapiro–Botha variogram model (Shapiro and Botha 1991) is fitted to the pilot estimates to obtain the final variogram estimate $\bar{\gamma}(\cdot)$.

Although the choice of the kernel function is of secondary importance, the bandwidth parameters play an important role in the performance of the local linear estimators described above, since they control the shape and size of the local neighbourhoods used for computing the corresponding estimates, determining their smoothness. However, when the data are correlated, traditional smoothing parameter selection methods for nonparametric regression will often fail to provide useful results (Opsomer et al. 2001). To take the dependence into account, we recommend the use of the "bias-corrected and estimated" generalized cross-validation criterion (CGCV) proposed in Francisco-Fernández and Opsomer (2005). In the case of the trend estimator $\hat{\mu}(\cdot)$, this method consists in selecting the bandwidth h that minimizes:

$$\mathrm{CGCV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\bar{Y}(t_i) - \hat{\mu}(t_i)}{1 - \frac{1}{n}\mathrm{tr}(\mathbf{S}\hat{\mathbf{R}}_{\bar{\mathbf{y}}})}\right)^2,$$

where $\mathrm{tr}(\mathbf{A})$ stands for the trace of a square matrix $\mathbf{A}$, $\mathbf{S}$ is the smoothing matrix, a square matrix whose ith row is equal to $\mathbf{s}_{t_i}$ (the smoother vector for $t = t_i$), and $\hat{\mathbf{R}}_{\bar{\mathbf{y}}}$ is an estimate of the correlation matrix of the sample means $\bar{\mathbf{y}}$. This matrix can be easily obtained bearing in mind that:

$$\mathrm{Cov}\left(\bar{Y}(t_j), \bar{Y}(t_{j'})\right) = \tfrac{1}{n}\sigma(t_j)\sigma(t_{j'})\rho\left(\left|t_j - t_{j'}\right|\right).$$

An analogous procedure can be used to select the bandwidth h_2 for the variance estimation. Nevertheless, this method will require an estimate of the correlation matrix of the squared residuals (or of their sample means, if we use the previous approximation). Under the assumptions of normality and zero mean for the residuals, the covariance matrix of the squared residuals admits the following expression:

$$\mathbf{\Sigma}_{\mathbf{r}^2} = 2\mathbf{\Sigma}_{\mathbf{r}} \odot \mathbf{\Sigma}_{\mathbf{r}}, \tag{3}$$

where $\odot$ represents the Hadamard product and $\mathbf{\Sigma}_{\mathbf{r}}$ the covariance matrix of the residuals (Ruppert et al. 1997), from which it is simpler to approximate the required correlations. The bandwidth parameter h_3 for the estimation of the variogram could be selected, for instance, by minimizing the cross-validation relative squared error of the semivariogram estimates (see e.g. Fernández-Casal and Francisco-Fernández 2014). Although, as this criterion does not take into account the dependence between the sample semivariances, the resulting bandwidth should be increased (for example by multiplying it by a factor between 1.5 and 2) to avoid under-smoothing the variogram estimates.

The above criteria, for the selection of optimal bandwidths for trend and variance approximation, require estimation of the small-scale variability of the process, leading to a circular problem. To avoid it, an iterative algorithm is used. Starting with an initial h and h_1 bandwidths (e.g. obtained by any of the available methods for independent data). At each iteration, the bandwidths are selected using the variance and variogram estimates computed in the previous iteration, and the model components are re-estimated. The algorithm is considered to have converged when there are no significant changes in the selected bandwidths, indicating similar small-scale variability estimates. Typically, a single iteration of this algorithm is sufficient in practice. This procedure is implemented in the `npf.fit()` function of the `npfda` package (Fernandez-Casal et al. 2023). More details are provided in the supplementary material.

2.2. Nonparametric Bootstrap

Using the nonparametric estimates of the trend $\hat{\mu}(\cdot)$, the variance $\hat{\sigma^2}(\cdot)$ and the semivariogram $\bar{\gamma}(\cdot)$ obtained with the procedure described in previous section, the proposed bootstrap algorithm is as follows:

1. Form the standardized residuals matrix $\hat{\mathbf{E}}$, whose ith row is equal to $\hat{\boldsymbol{\varepsilon}}_i = \hat{\mathbf{D}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})$, where $\hat{\mathbf{D}} = \operatorname{diag}(\hat{\sigma}^2(t_1), \ldots, \hat{\sigma}^2(t_p))$ and $\hat{\boldsymbol{\mu}} = \left(\hat{\mu}(t_1), \ldots, \hat{\mu}(t_p)\right)^\top$.
2. Construct an estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ of the within-curve correlation matrix from $\bar{\gamma}(\cdot)$, and compute its Cholesky decomposition $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}} = \mathbf{U}^\top\mathbf{U}$.
3. Compute the uncorrelated standardized residuals $\mathbf{E} = \hat{\mathbf{E}}\mathbf{U}^{-1}$ and scale them (jointly, by subtracting the overall sample mean and dividing by the overall sample standard deviation).
4. Use the scaled values to derive an independent bootstrap sample $\mathbf{E}^*$ (by resampling the rows and columns of $\mathbf{E}$).
5. Compute the bootstrap errors $\boldsymbol{\varepsilon}^* = \mathbf{E}^*\mathbf{U}$.
6. Obtain the bootstrap sample $\mathbf{Y}^*$, with

$$\mathbf{y}_i^* = \hat{\boldsymbol{\mu}} + \hat{\mathbf{D}}\boldsymbol{\varepsilon}_i^*,$$

   for $i = 1, \ldots, n$.
7. Repeat B times steps 4–6 to obtain the B bootstrap replicates $\left\{\mathbf{Y}_1^*, \ldots, \mathbf{Y}_B^*\right\}$.

As stated in the Introduction, the replicates derived from this algorithm can be used to approximate characteristics of the distribution of a statistic under study. For example, they can be used to approximate the standard error and bias of an estimator (as illustrated in Sects. 3 and 4), as well as to compute confidence intervals (Sect. 4), among many other potential applications.

3. SIMULATION RESULTS

This section presents various studies comparing the performance of the proposed nonparametric bootstrap method (NPB) with the smoothed bootstrap (SB) method proposed by Cuevas et al. (2006) and the naive bootstrap (NB) method. The SB algorithm is implemented in the `fdata.bootstrap()` function of the `fda.usc` package and can be summarized as follows:

1. Draw a standard bootstrap replicate $\mathbf{Y}_0^*$ from $\mathbf{Y}$, by uniform resampling of the rows $\mathbf{y}_1, \ldots, \mathbf{y}_n$.
2. Generate $\mathbf{Z}$, such that each row $\mathbf{z}_i = (Z_i(t_1), \ldots, Z_i(t_p)))^\top$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\alpha\hat{\Sigma}_{\mathbf{Y}}$, where $\hat{\Sigma}_{\mathbf{Y}}$ is the sample covariance matrix of the observed values $\mathbf{Y}$ (an estimate of Σ_0) and α is a smoothing parameter (controlling the amount of additional variability), and such that $\mathbf{z}_i$ is independent of $\mathbf{z}_{i'}$ if $i \neq i'$ $(\mathrm{Cov}\left(Z_i(t_j), Z_{i'}(t_{j'})\right) = 0)$.
3. Compute the bootstrap sample as $\mathbf{Y}^* = \mathbf{Y}_0^* + \mathbf{Z}$.
4. Repeat B times steps 1–3 to obtain the B bootstrap replicates $\left\{\mathbf{Y}_1^*, \ldots, \mathbf{Y}_B^*\right\}$.

The difficulty in applying this method in practice is the proper selection of the α parameter. However, in the results shown below, we set $\alpha = 0.05$ following the authors' recommendation.

Note that the naive bootstrap (NB) can be obtained as a particular case when $\alpha = 0$. In this case, steps 2 and 3 in the previous algorithm can be skipped, resulting in the naive bootstrap replicates $\mathbf{Y}^* = \mathbf{Y}_0^*$.

Numerical studies were carried out to study the behaviour of the three bootstrap procedures (NPB, SB, NB) under different scenarios. In each case, $N = 2000$ curve samples of sizes $n = 25$, 50 and 100, with $p = 101$ regular discretization points in the interval [0, 1], following the model (1) were generated. In order to take into account the effect of different functional forms of the trend and variance, the following theoretical functions were considered: $\mu_1(t) = 2.5 + \sin(2\pi t)$ (nonlinear trend), $\mu_2(t) = 10t(1-t)$ (polynomial trend), $\mu_3(t) = 2$ (constant trend), $\sigma_1^2(t) = (\frac{15}{16})^2[1-(2t-1)^2]^2 + 0.1$ (nonlinear variance), $\sigma_2^2(t) = 0.5(1+t)$ (linear variance) and $\sigma_3^2(t) = 1$ (constant variance, i.e. homoscedastic case). The random errors ε_i were normally distributed with zero mean, unit variance and isotropic exponential variogram:

$$\gamma_\varepsilon(u) = c_0 + (1-c_0)\left(1 - \exp\left(-3\frac{|u|}{a}\right)\right),$$

(for $u \neq 0$), where c_0 is the nugget effect ($1 - c_0$ is the partial sill) and a is the practical range. The values considered in the simulations were $a = 0.3, 0.6, 0.9$, and $c_0 = 0, 0.2, 0.5$. For instance, Fig. 2 provides an idea of the shape of the simulated samples in two of the studied scenarios.

In each scenario, $B = 1000$ bootstraps replicates were obtained using both the SB and NPB methods. The performance of both methods was analysed by comparing the results in

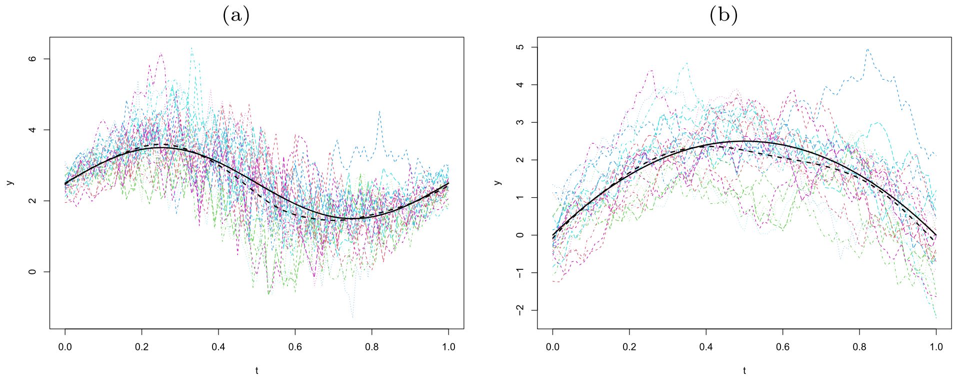

Figure 2. Simulated samples of size $n = 25$ with μ_1 (nonlinear), σ_1^2 (nonlinear), $c_0 = 0.2$ and $a = 0.6$ (**a**), and with μ_2 (polynomial), σ_2^2 (linear), $c_0 = 0$ and $a = 0.9$ (**b**). The theoretical trends are shown in solid lines and the nonparametric estimates in black dashed lines .

the approximation of characteristics of two estimators. More specifically, we will consider the approximation of the bias and the standard error (se) of the nonparametric trend $\hat{\mu}(t)$ and conditional variance $\widehat{\sigma^2}(t)$ estimators described in Sect. 2.1. The general procedure to approximate the bias and the standard error of an estimator $\hat{\theta}(t)$ from bootstrap resamples is as follows:

1. Derive B replicates $\left\{\mathbf{Y}_1^*, \ldots, \mathbf{Y}_B^*\right\}$ from the original data.
2. Compute B estimates of $\theta(t)$ from the B replicates, which will be denoted by $\left\{\hat{\theta}_1^*(t), \ldots, \hat{\theta}_B^*(t)\right\}$.
3. Approximate the bootstrap version of $\sigma(\hat{\theta}(t))$ as follows:

$$\widehat{se}^*(\hat{\theta}^*(t)) = \left\{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^*(t) - \bar{\hat{\theta}}^*(t)\right)^2\right\}^{\frac{1}{2}}, \tag{4}$$

   where $\bar{\hat{\theta}}^*(t) = \sum_{b=1}^{B} \hat{\theta}_b^*(t)/B$.

4. In a similar way, obtain the bootstrap counterpart of $\text{Bias}(\hat{\theta}(t))$ through

$$\widehat{\text{Bias}}^*(\hat{\theta}^*(t)) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^*(t) - \hat{\theta}(t)\right). \tag{5}$$

To avoid the effect that the bandwidth selection criteria might have on the results, the local linear trend and variance estimators were computed using the bandwidths that minimized the corresponding (theoretical) mean average squared errors (MASE). For the trend estimator, this criterion can be expressed as follows:

$$\text{MASE}(h) = \frac{1}{p}(\mathbf{S}\boldsymbol{\mu} - \boldsymbol{\mu})^t(\mathbf{S}\boldsymbol{\mu} - \boldsymbol{\mu}) + \frac{1}{np}\text{tr}(\mathbf{S}\boldsymbol{\Sigma}_0\mathbf{S}^t),$$

where $\boldsymbol{\mu} = \left[\mu(t_1), \ldots, \mu(t_p)\right]^t$. An analogous approach was used in the case of the variance estimator, by using (3) to approximate the corresponding covariance matrix.

At each simulation, the bias and variance of the two estimators were approximated through (5) and (4). To measure the accuracy of these bootstrap estimates, mean squared (MSE) errors were computed, using theoretical values, $\text{Bias}(\hat{\theta}(t))$ and $\sigma(\hat{\theta}(t))$, approximated by simulation. For example, in the case of the approximation of the bias of the trend estimator:

$$\text{MSE}(t) = E\left\{\left[\widehat{\text{Bias}}^*(\hat{\mu}^*(t)) - \text{Bias}(\hat{\mu}(t))\right]^2\right\}.$$

The averages of these errors over the discretization points will be denoted by AMSE.

Similar results were observed across the simulation scenarios, although only a few representative outcomes are presented here for brevity. Overall, the proposed method showed superior performance in approximating the bias of both estimators. The bias approximations obtained with the SB and NB methods were closer to zero, particularly for the trend estimator.

In addition, the results obtained with the SB and NB methods were more similar than expected, since the replicates with the SB method have more variability. Only slight differences between these two methods were observed when approximating the bias of the variance estimator. For example, Fig. 3 compares the theoretical values with the bootstrap approximations of the bias and the standard error of both estimators for μ_1 (nonlinear), σ_1^2 (nonlinear), $n = 50$, $c_0 = 0.2$ and $a = 0.6$.

Unexpectedly, the standard error approximations obtained with the SB and NB methods turned out to be slightly better than those obtained with the NBP method, especially when the sample size is small. For instance, Table 1 summarizes the errors obtained in the approximation of the bias and standard error of the estimators with both bootstrap procedures considering the different sample sizes, for μ_1 (nonlinear), σ_1^2 (nonlinear), $c_0 = 0.2$ and $a = 0.6$. It can be observed that as the sample size n increases, the squared errors decrease, suggesting the consistency of the approximations obtained with both methods. A clear improvement is observed when using the SB or the NB methods to approximate the standard error of the variance estimator with the smallest sample size, obtaining very similar results with both methods as the number of observations increases. However, the NPB method outperforms the other methods at approximating the bias in all cases, especially when the variance estimator is considered.

The influence of the temporal dependence on the bootstrap approximations was also studied. For instance, Table 2 shows the results obtained for the trend estimator $\hat{\mu}(t)$ considering the different nugget (c_0) and practical range (a) values, for μ_1 (nonlinear), σ_1^2 (nonlinear) and $n = 100$. In these cases, the errors corresponding to the standard error approximations are quite similar for all three methods. As for the bootstrap estimates of biases, as expected, it is generally observed that the errors decrease as the nugget increases (which corresponds to lower temporal dependency). This effect is particularly pronounced when the SB or NB method is used. A similar behaviour is observed when the practical range increases.

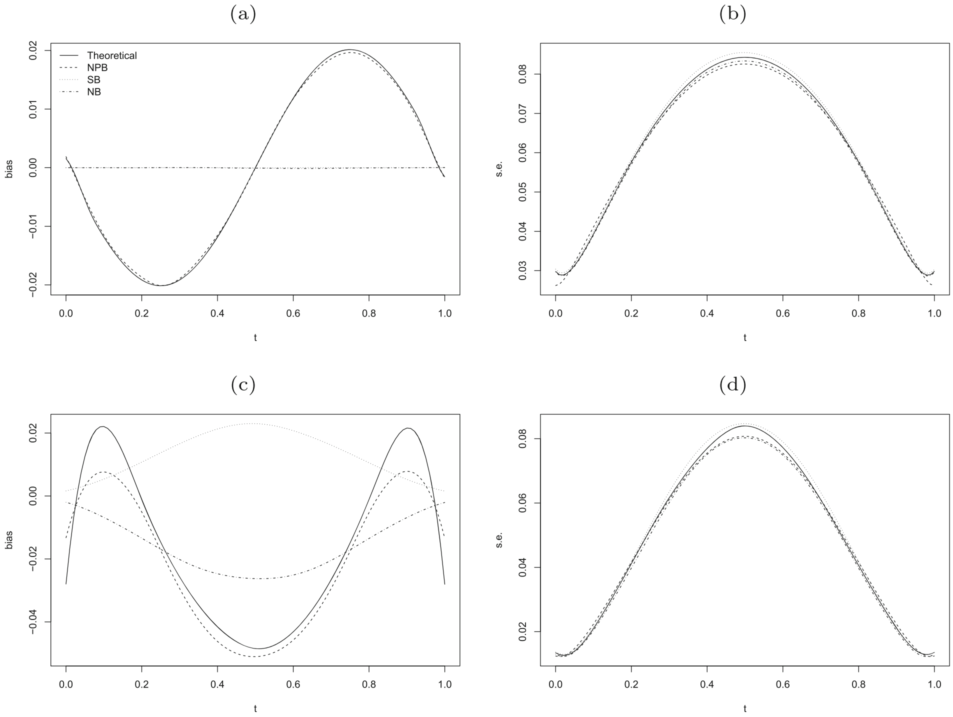

Figure 3. Comparison of the theoretical bias (left) and standard error (right) with their bootstrap approximations, for the local linear trend (top) and the variance (bottom) estimators, considering μ_1 (nonlinear), σ_1^2 (nonlinear), $n = 50, c_0 = 0.2$ and $a = 0.6$. The theoretical values are shown in solid lines, the NPB, SB and NB approximations in dashed, dotted and dot-dashed lines, respectively .

Table 1. Monte Carlo approximations of the AMASE ($\times 10^2$) of the bias and standard error bootstrap estimates, for the local linear trend $\hat{\mu}(t)$ and variance $\hat{\sigma^2}(t)$ estimators, considering μ_1 (nonlinear), σ_1^2 (nonlinear), $n = 100$ $c_0 = 0.2$ and $a = 0.6$

		$n = 25$		$n = 50$		$n = 100$	
Estimator	Method	Bias	se	Bias	se	Bias	se
$\hat{\mu}(t)$	SB	0.104	0.031	0.046	0.009	0.020	0.002
	NB	0.104	0.031	0.046	0.009	0.021	0.002
	NPB	0.020	0.036	0.009	0.010	0.004	0.003
$\hat{\sigma^2}(t)$	SB	1.433	0.076	0.767	0.028	0.531	0.009
	NB	1.594	0.078	0.688	0.029	0.384	0.009
	NPB	0.097	0.123	0.038	0.029	0.018	0.009

Finally, Table 3 illustrates the effect of the assumed theoretical functional forms in model (1) on the errors in bias and standard error approximations of the variance estimator (for $n = 100, c_0 = 0.2$, and $a = 0.6$). Once again, the NPB method consistently outperforms the other methods in approximating biases across the different scenarios. When the variance model remains fixed, similar results are obtained with all methods when the trend varies. However, for the same theoretical trend, different behaviours are observed when the functional form of

Table 2. Monte Carlo approximations of the AMASE ($\times 10^2$) of the bootstrap estimates of the bias and standard error of $\hat{\mu}(t)$, considering the different c_0 and a values, with μ_1 (nonlinear), σ_1^2 (nonlinear) and $n = 100$

		$a = 0.3$		$a = 0.6$		$a = 0.9$	
c_0	Method	Bias	*se*	Bias	*se*	Bias	*se*
0	SB	0.036	0.002	0.023	0.003	0.016	0.003
	NB	0.036	0.002	0.023	0.003	0.016	0.003
	NPB	0.007	0.002	0.004	0.004	0.003	0.005
0.2	SB	0.030	0.002	0.020	0.002	0.015	0.003
	NB	0.030	0.002	0.021	0.002	0.016	0.002
	NPB	0.006	0.002	0.004	0.003	0.003	0.004
0.5	SB	0.023	0.001	0.018	0.002	0.015	0.002
	NB	0.023	0.001	0.018	0.002	0.015	0.002
	NPB	0.005	0.001	0.004	0.002	0.003	0.002

Table 3. Monte Carlo approximations of the AMASE ($\times 10^2$) of the bootstrap estimates of the bias and standard error of $\widehat{\sigma^2}(t)$, considering the different theoretical trend and variance functions, with $n = 100$, $c_0 = 0.2$, and $a = 0.6$

Theoretical		μ_1 (nonlinear)		μ_2 (polynomial)		μ_3 (constant)	
	Method	Bias	*se*	Bias	*se*	Bias	*se*
σ_1 (nonlinear)	SB	0.531	0.010	0.536	0.009	0.541	0.009
	NB	0.384	0.009	0.383	0.009	0.379	0.009
	NPB	0.018	0.009	0.018	0.009	0.017	0.009
σ_2^2 (linear)	SB	0.654	0.006	0.654	0.006	0.645	0.005
	NB	0.694	0.005	0.690	0.005	0.673	0.006
	NPB	0.004	0.013	0.004	0.013	0.004	0.013
σ_3^2 (constant)	SB	1.104	0.009	1.105	0.010	1.092	0.010
	NB	1.170	0.009	1.163	0.009	1.136	0.009
	NPB	0.008	0.022	0.008	0.022	0.008	0.022

the theoretical variance changes. While the error in bias approximations increases notably with the SB and NB methods when simpler variance models are considered, a similar effect is observed with the NPB method in standard error approximations. This may be attributed to the slight underestimation of variance by the local linear estimator $\widehat{\sigma^2}(t)$ in these cases, resulting in a small negative bias that the SB and NB methods approximate with values close to zero, and producing slightly lower variability in the NPB method.

4. APPLICATION TO POLLUTION DATA

In this section, the practical performance of proposed methodology is illustrated through its application to the data set of ground-level ozone concentrations briefly mentioned in the Introduction ($n = 33$ and $p = 365$).

The iterative process described at the end of Sect. 2.1 was used to estimate the model components. As a stopping criterion, an absolute percentage difference of less than 10%

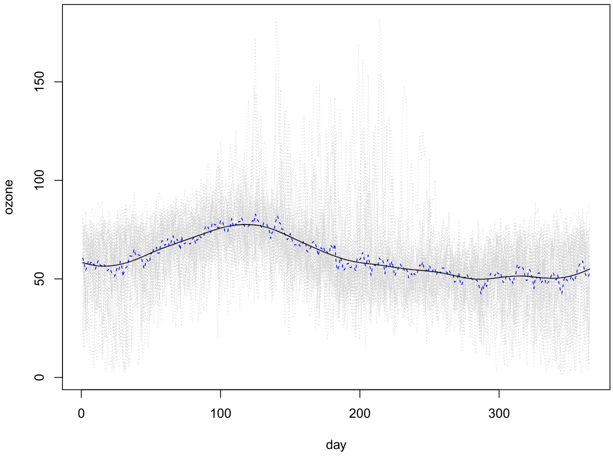

Figure 4. Sample mean (dashed line) and nonparametric trend estimates (solid line), of the ozone data (grey dotted lines) .

between bandwidths was used. Two iterations were performed in this case. (Although only one would have been necessary since the selected bandwidths for trend and variance estimation were nearly identical to those of the second iteration, further details can be found in the supplementary material.) The final trend estimate is shown in Fig. 4, computed with a bandwidth $h = 36.077$ selected by the CGCV criterion, where an increase in mean ozone levels is observed during springtime.

Then, from the final residuals, the variance estimate $\hat{\sigma^2}(\cdot)$ (with a bandwidth $h_2 = 33.106$ selected by the CGCV criterion), the pilot semivariogram estimates $\hat{\gamma}(\cdot)$ (with a bandwidth $h_3 = 3.713$ selected by minimizing the CV relative squared error) and its Shapiro-Botha fit $\bar{\gamma}(\cdot)$ were computed. Figure 5a shows the standard deviation estimate, where an increase in the variability in ozone concentration at the beginning of summer and in winter. The variogram estimates are shown in Fig. 5b. The final variogram has a nugget effect of $\hat{c}_0 = 0.307$ (which may be interpreted as the proportion of independent variability) and a practical range $\hat{a} \approx 32.7$ (a distance beyond which the temporal correlation can be considered negligible).

With these nonparametric estimates, the NPB approach was applied to make inference about the trend of the functional process. Thus, the bias and standard error of the local linear trend estimator were approximated with $B = 2000$ replicates. Figure 6 shows an example of the results obtained, the bias-corrected trend estimates (solid line) and pointwise confidence intervals (point lines), computed adding and subtracting two standard errors to the corrected trend estimate. The NPB method also allows the construction of pointwise confidence intervals using the basic percentile method (see e.g. Davison and Hinkley, 1997, Section 5.2), obtaining practically identical results. (The basic bootstrap replicas are shown in dotted grey lines; see the supplementary material for further details.)

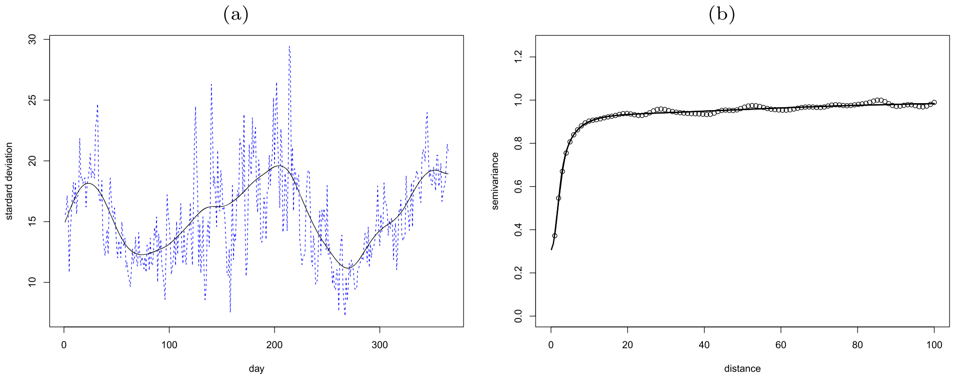

Figure 5. Sample variance and nonparametric variance estimates (**a**), dashed and solid lines, respectively, and semivariogram estimates (**b**) of the ozone data .

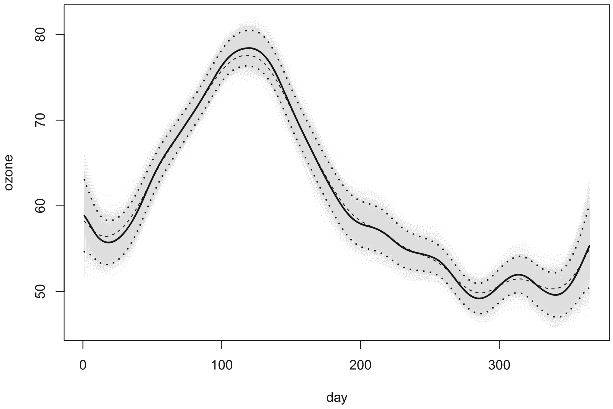

Figure 6. Bias-corrected (solid line) and uncorrected (dashed line) trend estimates, pointwise confidence intervals (point lines) and basic bootstrap replicas (dotted grey lines) obtained as a result of the application of the NPB method to the ozone data .

5. CONCLUSION

The performance of the proposed methodology was validated by a simulation study, showing its good behaviour under different scenarios, considering distinct theoretical trend and variance functions and including several degrees of temporal dependence. The results were compared to those obtained with the SB and NB approaches, showing that the new method seems to be better at reproducing the process variability. Specifically, the NPB method proved to be much better at approximating the bias of the estimators considered, as the SB or NB methods tend to produce bias approximations close to zero. Although, unexpectedly, the standard error approximations obtained with the SB and NB methods turned out to be slightly better than those obtained with the NBP method when the sample size is small.

To improve performance in the case of small samples, a correction for the bias due to the direct use of the residuals in the estimation of the small-scale variability, similar to that proposed in Fernández-Casal et al. (2017) for the spatial case, could be investigated.

The NPB method proposed in this study is designed for nonstationary heteroscedastic processes. However, it can be easily adapted to cases where either the mean or variance is assumed to be constant, such as when using residuals from a functional regression model. If any of these assumptions is reasonable, the procedure could be simplified, and even better results could be expected. On the other hand, the proposed functional model may not be appropriate in certain cases. For example, in the ozone dataset, it might be reasonable to assume that there is a yearly effect in the functional mean or in the variance. In such cases, more sophisticated estimators, such as semiparametric ones, could be considered for these components. However, the bootstrap procedure would remain analogous. Whereas if it is not appropriate to assume that the distribution of the standardized errors is homogeneous, it would be necessary to modify the resampling procedure. These aspects could be the subject of future researches, including the presence of dependence between curves.

The NPB technique was used for approximating characteristics of estimators and for the construction of confidence intervals. Moreover, it can also be employed in other inference problems, including hypothesis testing (e.g. related to the trend or variance functions), estimation of the probability that a pollutant concentration level exceed air quality guidelines, outlier detection (e.g. due to pollution episodes or sensor failures), among many others.

ACKNOWLEDGEMENTS

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work has been supported by grant PID2020-113578RB-I00, funded by MCIN/AEI/10.13039/501100011033. The research of Rubén Fernández-Casal has also been supported by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020/14) and by CITIC that is supported by Xunta de Galicia, convenio de colaboración entre la Consellería de Cultura, Educación, Formación Profesional e Universidades y las universidades gallegas para el refuerzo de los centros de investigación del Sistema Universitario de Galicia (CIGUS). The research of Sergio Castillo Páez has also been supported by the Universidad de las Fuerzas Armadas ESPE, from Ecuador.
Author Contribution All authors contributed equally to this work.

Data Availability The pre-processed data are supplied with the `R` package `npfda`, as the `ozone` data set. The original data can be downloaded from https://uk-air.defra.gov.uk/data. The code used to apply the proposed methodology to the pollution data and the results generated are included in the supplementary material.

Declarations
Competing Interests The authors have no relevant financial or non-financial interests to disclose.

[Received October 2022. Revised June 2023. Accepted June 2023. Published Online August 2023.]

REFERENCES

Castillo-Páez S, Fernández-Casal R, García-Soidán P (2019) A nonparametric bootstrap method for spatial data. Comput Stat Data Anal 137:1–15

Cuevas A, Febrero M, Fraiman R (2006) On the use of the bootstrap for estimating functions with functional data. Comput Stat Data Anal 51(2):1063–1074

de Castro BF, Guillas S, González Manteiga W (2005) Functional samples and bootstrap for predicting sulfur dioxide levels. Technometrics

Delicado P, Giraldo R, Comas C et al (2010) Statistics for spatial functional data: some recent contributions. Environmetrics Off J Int Environmetrics Soc 21(3–4):224–239

Embling CB, Illian J, Armstrong E et al (2012) Investigating fine-scale spatio-temporal predator-prey patterns in dynamic marine ecosystems: a functional data analysis approach. J Appl Ecol 49(2):481–492

Febrero Bande M, Oviedo de la Fuente M (2012) Statistical computing in functional data analysis: the R package fda.usc. J Stat Softw 51(4):3–20. https://doi.org/10.18637/jss.v051.i04

Febrero M, Galeano P, González-Manteiga W (2008) Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. Environmetrics Off J Int Environmetrics Soc 19(4):331–345

Fernandez-Casal R (2023) npsp: Nonparametric spatial (geo)statistics. R package version 0.7-11. https://rubenfcasal.github.io/npsp

Fernández-Casal R, Francisco-Fernández M (2014) Nonparametric bias-corrected variogram estimation under non-constant trend. Stoch Environ Res Risk Assess 28(5):1247–1259

Fernández-Casal R, Castillo-Paez S, Garcia-Soidan P (2017) Nonparametric estimation of the small-scale variability of heteroscedastic spatial processes. Spat Stat 22:358–370

Fernandez-Casal R, Castillo-Paez S, Flores M (2023) npfda: Nonparametric functional data analysis. R package version 0.1-4. https://rubenfcasal.github.io/npfda

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice, vol 76. Springer, Berlin

Ferraty F, Van Keilegom I, Vieu P (2010) On the validity of the bootstrap in non-parametric functional regression. Scand J Stat 37(2):286–306

Francisco-Fernández M, Opsomer J (2005) Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. Can J Stat 33:279–295

Giraldo R, Delicado P, Mateu J (2010) Continuous time-varying kriging for spatial prediction of functional data: an environmental application. J Agric Biol Environ Stat 15(1):66–82

Guijarro J (2019) Climatol: Climate Tools (Series Homogenization and Derived Products). R package version 3.1.2. https://CRAN.R-project.org/package=climatol

Karlsson PE, Klingberg J, Engardt M et al (2017) Past, present and future concentrations of ground-level ozone and potential impacts on ecosystems and human health in Northern Europe. Sci Total Environ 576:22–35. https://doi.org/10.1016/j.scitotenv.2016.10.061

Manteiga WG, Vieu P (2007) Statistics for functional data. Comput Stat Data Anal 51(10):4788–4792. https://doi.org/10.1016/j.csda.2006.10.017

McMurry T, Politis D (2011) Resampling methods for functional data. The Oxford handbook of functional data analysis. Oxford Univ. Press, Oxford, pp 189–209

Opsomer JD, Wang Y, Yang Y (2001) Nonparametric regression with correlated errors. Stat Sci 16:134–153

Politis DN, Romano JP (2010) K-sample subsampling in general spaces: the case of independent time series. J Multivar Anal 101(2):316–326

Ramsay JO, Graves S, Hooker G (2020) fda: Functional data analysis. https://CRAN.R-project.org/package=fda, r package version 5.1.9

Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, New York. https://doi.org/10.1007/b98888

Ruppert D, Wand MP, Holst U et al (1997) Local polynomial variance-function estimation. Technometrics 39(3):262–273

Sancho J, MartÃnez J, Pastor J et al (2014) New methodology to determine air quality in urban areas based on runs rules for functional data. Atmos Environ 83:185–192. https://doi.org/10.1016/j.atmosenv.2013.11.010

Shang HL, Hyndman R (2019) rainbow: Bagplots, boxplots and rainbow plots for functional data. https://CRAN.R-project.org/package=rainbow, r package version 3.6

Shapiro A, Botha JD (1991) Variogram fitting with a general class of conditionally nonnegative definite functions. Comput Stat Data Anal 11(1):87–96

Ullah S, Finch CF (2013) Applications of functional data analysis: a systematic review. BMC Med Res Methodol 13(1):1–12

Xiao W, Hu Y (2018) Functional data analysis of air pollution in six major cities. J Phys Conf Ser 1053(012):131. https://doi.org/10.1088/1742-6596/1053/1/012131

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