Statistics in Biosciences (2024) 16:1–24
https://doi.org/10.1007/s12561-023-09393-7

On Estimation of the Effect Lag of Predictors and Prediction in a Functional Linear Model

Haiyan Liu[1] · Georgios Aivaliotis[1] · Vijay Kumar[2] · Jeanine Houwing-Duistermaat[1,3]

Received: 24 May 2022 / Revised: 8 September 2023 / Accepted: 15 September 2023 /
Published online: 8 November 2023

Abstract
We propose a functional linear model to predict a functional response using multiple functional and longitudinal predictors and to estimate the effect lags of predictors. The coefficient functions are written as the expansion of a basis system (e.g. functional principal components, splines), and the coefficients of the basis functions are estimated via optimizing a penalization criterion. Then effect lags are determined by simultaneously searching on a prior designed grid mesh based on minimization of a proposed prediction error criterion. Mathematical properties of the estimated regression functions and predicted responses are studied. The performance of the method is evaluated by extensive simulations and a real data analysis application on chronic obstructive pulmonary disease (COPD).

Keywords Functional data analysis · Effect lag functional linear model · Functional principal component analysis · COPD

1 Introduction

Due to advancements in science and technology, the amount of available, often open-source, data is growing exponentially. New opportunities for statistical research arise by formulating relevant and interesting questions which can be answered with these data. This work is motivated by questions in urban analytics and three temporal datasets, namely COPD hospital admissions, NO2 measurements and visits to gyms in Leeds from 2013 to 2018. The NO2 data are measured at ten locations

✉ Haiyan Liu
h.liu1@leeds.ac.uk

[1] Department of Statistics, University of Leeds, Leeds, UK

[2] Department of Physics and Astronomy, University of Sussex, Brighton, UK

[3] Department of Mathematics, Radboud University Nijmegen, Nijmegen, The Netherlands

and extrapolated to postcode level. The visits to gyms and the COPD admissions are aggregated to postcode level. The health and pollution data are sampled densely and regularly in time (we call this dense functional data) and the physical exercise data are obtained at irregular time intervals with few measurements (we call this sparse longitudinal data). The questions we aim to answer with these data are: Do life style and pollution affect COPD hospital admissions at postcode level? Does a rise in pollution have a direct effect on COPD hospital admissions or is there a delay? A first step to answer these questions is to model health outcomes over time as function of pollution and physical exercise over time.

Relationships between temporal data are often not synchronous and involve a delay in the effects. Exposure of a person to high pollution might show to affect this person's health with some delay and the effect might fade away after some time. Other examples are, a treatment with a medicine might take some time to start affecting the patient and after the end of the treatment it might still be having an effect for limited amount of time. Historical exposure to high temperatures might not have an effect on the growth of trees anymore after a certain period and it may also take some time before high temperatures result in lower growth rates.

Motivated by the problem of dense temporal and sparse longitudinal predictors, we consider estimation and prediction for a functional regression model [7]. We estimate the intervals through the corresponding lags of the effect of predictors on response. Specifically for our motivating example, we estimate the influence of dense functional pollutants (nitrogen dioxide (NO2) concentrations) and sparse longitudinal predictor (physical activities) on dense COPD hospital admissions. Moreover, we want to estimate the effect lags of pollutants on COPD hospital admission, i.e. from when the predictors have an influence on the response and until when this influence disappears.

The classical function-on-function linear model, which was first formulated by Ramsay and Dalzell [17], reads as follows:

$$Y(t) = \beta_0(t) + \int_0^T \beta_1(s,t)X(s)ds + \epsilon(t),\ t \in [0,T]$$

where $Y(t)$ is the response trajectory, $X(s)$ is the predictor trajectory, $\epsilon(t)$ is the error process, $\beta_0(t)$ is the intercept process, $\beta_1(s,t)$ is the two-dimensional regression coefficient function which shows the influence of X on Y. A drawback of this model is that the entire predictor trajectory $X(s)$ including the future values, i.e. when $s > t$, is assumed to influence the current value of response trajectory Y at time t. Clearly this is not appropriate in many applications. Malfait and Ramsay [5, 8, 12] proposed historical functional linear models, where only the past of the predictor trajectory influences the response at the current time:

$$Y(t) = \beta_0(t) + \int_{t-\delta_2}^{t-\delta_1} \beta_1(s,t)X(s)ds + \epsilon(s),\ t \in [0,T]$$

where δ_1 and δ_2 $(0 < \delta_1 < \delta_2 < T)$ are the lags for the influence of predictor trajectory on response trajectory. For one dense functional predictor, Malfait and Ramsay

[12] consider the triangular basis expansion of the coefficient function which is estimated at each observation point. A penalized approach which allows varying lags for the historical functional linear model has been developed by [5]. Kim et al. [8] considers the situation that both predictor process and response process are sparsely and irregularly observed. Pomann et al. [16] has extended the historical functional linear model to multiple homogeneous predictors, and the response is influenced by the predictors from a fixed starting effect time to current time. In this paper we extend this work to a model with both dense functional and sparse longitudinal predictors.

Our contribution of this paper is threefold: firstly, a model with multiple heterogeneous (sparse longitudinal and dense functional) predictors subject to time lags (both starting and ending points) that are fixed but unknown is proposed. We propose estimators for the coefficient regression functions. The effect lags (i.e. the fixed starting and ending effect times) are determined by minimizing the prediction error. Secondly, the asymptotic behavior of the estimated coefficient functions, and the predicted response curve is investigated. Thirdly, the three temporal datasets are integrated and the relationships between COPD hospital admissions and lifestyle and pollution are modelled.

The paper is organized as follows: In Section 2, the historical function-on-function linear model for multiple heterogeneous predictors is introduced. In Section 3, we consider the estimation of the coefficient functions and the uniform consistency of the estimators is established. In Section 4, the prediction of the response trajectories is proposed and the asymptotic property of the predicted trajectories is established. The determination of the lags is proposed in Section 5. Extensive numerical simulations are considered in Section 6 to show the asymptotic properties of our proposed estimators. In Section 7, the COPD and NO2 datasets are analysed and the lags of the influence of NO2 on COPD are determined. We finish by drawing some conclusions and further discussion.

2 Model

Suppose the observations are $\{Y_{ij}, t_{ij} : i = 1, ..., n, j = 1, ..., m_{Yi}\}$, $\{W_{1ij}, s_{1ij}: i = 1, ..., n, j = 1, ..., m_{1i}\}$ and $\{W_{2ij}, s_{2ij} : i = 1, ..., n, j = 1, ..., m_{2i}\}$, where Y_{ij} is the j-th observation of response for the i-th subject (at time $t_{ij} \in [0, 1]$) and m_{Yi} is the number of observations of response for the i-th subject, similarly, W_{1ij}, W_{2ij} are the j-th observation of subject i for two predictors respectively and m_{1i} and m_{2i} are the numbers of observations of the two predictors (at time $s_{1ij}, s_{2ij} \in [0, 1]$) respectively. For example, the response Y_{ij} corresponds to the standardised COPD hospital admissions in the i-th postcode district at time t_{ij}. The predictor W_{1ij} corresponds to the NO2 concentrations in the i-th postcode district at time s_{1ij}. The predictor W_{2ij} corresponds to the standardised number of physical activities in the i-th postcode district at time s_{2ij}.

Let $W_{1ij} = X_{1i}(s_{1ij}) + \epsilon_{1ij}$, $W_{2ij} = X_{2i}(s_{2ij}) + \epsilon_{2ij}$ and $X_{1i}(t)$, $X_{2i}(t)$ be independent copies of underlying square-integrable random functions $X_1(t)$ and $X_2(t)$ over

[0, 1] respectively. Without loss of generality, we assume $\mu_{X_1}(t) = E[X_1(t)] = 0$ and $\mu_{X_2}(t) = E[X_2(t)] = 0$. See Remark 2 for guidance on how to estimate the means in real data applications. We denote by $C_{X_1}(s,t) = cov(X_1(s), X_1(t))$ the covariance of $X_1(t)$ and $C_{X_2}(s,t) = cov(X_2(s), X_2(t))$ the covariance of $X_2(t)$. We assume that the first predictor curves $X_{1i}(t)$ are observed on a dense and regular grid of points, i.e. $m_{11} = ... = m_{1n} := m_1$, while the second predictor curves $X_{2i}(t)$ are observed on a sparse and irregular grid of points s_{2ij}. The observations W_{1ij} and W_{2ij} are the discrete versions of $X_{1i}(t)$ and $X_{2i}(t)$ with iid mean-zero and variance-finite noise ϵ_{1ij} and ϵ_{2ij} respectively. The noise ϵ_{1ij} and ϵ_{2ij} are independent of $X_{1i}(t)$ and $X_{2i}(t)$ respectively. For the responses Y_{ij}, they are observed either on a sparse and irregular grid or dense grid of t_{ij}.

We define the lag historical functional linear model with two heterogeneous predictors $X_1(t)$ and $X_2(t)$ and the response $Y(t)$ as

$$\begin{aligned} Y_{ij} &= \beta_0(t_{ij}) + \int_{\delta_{11}}^{\delta_{12}} \beta_1(s, t_{ij}) X_{1i}(t_{ij} - s) ds \\ &\quad + \int_{\delta_{21}}^{\delta_{22}} \beta_2(s, t_{ij}) X_{2i}(t_{ij} - s) ds + e_{ij} \text{ for } t_{ij} \geq \max(\delta_{12}, \delta_{22}) \end{aligned} \tag{1}$$

where $i \in \{1, ..., n\}$, $j \in \{1, ..., m_{Yi}\}$, $\beta_0 : [0,1] \to \mathbb{R}$, $\Delta_1 = [\delta_{11}, \delta_{12}]$, $\Delta_2 = [\delta_{21}, \delta_{22}]$, $\beta_1 : \Delta_1 \times [0,1] \to \mathbb{R}$ and $\beta_2 : \Delta_2 \times [0,1] \to \mathbb{R}$ are continuous two-dimensional coefficient functions, and e_{ij} are independent measurement errors with mean zero and finite variance σ_e^2. Errors e_{ij} are assumed to be independent of $X_{1i}(t)$ and $X_{2i}(t)$.

In Equation (1), we assume that the δ's are fixed, similar to [8]. Given the δ 's, the observations Y_{ij} are modelled at times $t_{ij} \geq \max\{\delta_{12}, \delta_{22}\}$. Considering a subset only of the observations based on the lags is not unusual and is similar to the approach for AR(p) models, where model selection (determine p) comes first and then parameter estimation where observations at time $t < p$ are not available.

Notice that (1) is equivalent to

$$\begin{aligned} Y_{ij} = \beta_0(t_{ij}) &+ \int_{t_{ij}-\delta_{12}}^{t_{ij}-\delta_{11}} \beta_1(t_{ij} - s, t_{ij}) X_{1i}(s) ds \\ &+ \int_{t_{ij}-\delta_{22}}^{t_{ij}-\delta_{21}} \beta_2(t_{ij} - s, t_{ij}) X_{2i}(s) ds + e_{ij} \end{aligned} \tag{2}$$

then the model (1) means that given the entire predictor curves X_{1i} and X_{2i}, the response for the i-th subject at time $t_{ij} \geq \max\{\delta_{12}, \delta_{22}\}$ is only affected by the values of X_{1i} over time-window $[t_{ij} - \delta_{12}, t_{ij} - \delta_{11}]$ and by the values of X_{2i} over time-window $[t_{ij} - \delta_{22}, t_{ij} - \delta_{21}]$.

Whether $X_{1i}(s)$ for s in the time-window $[t_{ij} - \delta_{12}, t_{ij} - \delta_{11}]$ and $X_{2i}(s)$ for s the time-window $[t_{ij} - \delta_{22}, t_{ij} - \delta_{21}]$ have an effect on the response Y_{ij} also depends on the coefficient functions $\beta_1(t_{ij} - s, t_{ij})$ and $\beta_2(t_{ij} - s, t_{ij})$ respectively. These functions represent the effect of the predictors on the outcome variable and can be equal to zero.

3 Estimation of the model parameters

Let $\{B_{11}(s), ..., B_{1K_1}(s)\}$ and $\{B_{21}(s), ..., B_{2K_2}(s)\}$ be two pre-specified basis functions on Δ_1 and Δ_2. Then the two-dimensional coefficient functions $\beta_1(s,t)$ and $\beta_2(s,t)$ are assumed to be represented as

$$\beta_1(s,t) = \sum_{k=1}^{K_1} B_{1k}(s) b_{1k}(t),\ s \in \Delta_1,\ t \in [0,1] \tag{3}$$

and

$$\beta_2(s,t) = \sum_{k=1}^{K_2} B_{2k}(s) b_{2k}(t),\ s \in \Delta_2,\ t \in [0,1] \tag{4}$$

respectively, where K_1 and K_2 capture the resolution of the fit and should be chosen accordingly and $b_{1k}(t)$ and $b_{2k}(t)$ are the unknown time-varying coefficient functions defined on [0, 1]. Various basis functions such as Fourier, B-spline, wavelet basis can be used depending on the specific features of the coefficient functions.

Substituting (3) and (4) into equation (1), we have

$$\begin{aligned} Y_{ij} =& \beta_0(t_{ij}) + \sum_{k=1}^{K_1} b_{1k}(t_{ij}) \int_{\delta_{11}}^{\delta_{12}} B_{1k}(s) X_{1i}(t_{ij}-s) ds \\ &+ \sum_{k=1}^{K_2} b_{2k}(t_{ij}) \int_{\delta_{21}}^{\delta_{22}} B_{2k}(s) X_{2i}(t_{ij}-s) ds + e_{ij} \\ =:& \beta_0(t_{ij}) + \sum_{k=1}^{K_1} b_{1k}(t_{ij}) \tilde{X}_{1ik}(t_{ij}) + \sum_{k=1}^{K_2} b_{2k}(t_{ij}) \tilde{X}_{2ik}(t_{ij}) + e_{ij} \\ =& \beta_0(t_{ij}) + \mathbf{b}_1^T(t_{ij}) \tilde{\mathbf{X}}_{1i}(t_{ij}) + \mathbf{b}_2^T(t_{ij}) \tilde{\mathbf{X}}_{2i}(t_{ij}) + e_{ij}, \end{aligned} \tag{5}$$

where $\tilde{X}_{1ik}(t_{ij}) = \int_{\delta_{11}}^{\delta_{12}} B_{1k}(s) X_{1i}(t_{ij}-s) ds$, $\tilde{X}_{2ik}(t_{ij}) = \int_{\delta_{21}}^{\delta_{22}} B_{2k}(s) X_{2i}(t_{ij}-s) ds$, $\mathbf{b}_1(t_{ij}) = (b_{11}(t_{ij}), ..., b_{1K_1}(t_{ij}))^T$, $\mathbf{b}_2(t_{ij}) = (b_{21}(t_{ij}), ..., b_{2K_2}(t_{ij}))^T$, $\tilde{\mathbf{X}}_{1i}(t_{ij}) = (\tilde{X}_{1i1}(t_{ij}), ..., \tilde{X}_{1iK_1}(t_{ij}))^T$, and $\tilde{\mathbf{X}}_{2i}(t_{ij}) = (\tilde{X}_{2i1}(t_{ij}), ..., \tilde{X}_{2iK_2}(t_{ij}))^T$. Note that the observed times t_{ij} depend on subject i. Model (1) reduces to a varying coefficient model with K_1 induced predictors $\tilde{X}_{1ik}(t_{ij})$ and K_2 induced predictors $\tilde{X}_{2ik}(t_{ij})$ which can be regarded as realizations of $\tilde{X}_{1k}(t)$ and $\tilde{X}_{1k}(t)$ at t_{ij} respectively.

At first, notice that $\mu_{X_1}(t) = \mu_{X_2}(t) = 0$ implies $\beta_0(t_{ij}) = E[Y_{ij}]$, so $\beta_0(t)$ can be estimated by smoothing Y_{ij} via local smoothing method based on the pooled data, see for example Yao et al. [19]. We denote $Y_{ij} - \hat{\beta}_0(t_{ij})$ by Y_{ij}, where $\hat{\beta}_0(t_{ij})$ is an estimator of $\beta_0(t)$ evaluated at time t_{ij}.

In order to derive the estimator of $\{b_{11}(t), ..., b_{1K_1}(t)\}$ and $\{b_{21}(t), ..., b_{2K_2}(t)\}$, we assume $t_{ij} = t_j^0$ to simplify the notation in this paragraph, i.e. the observation times for different subjects are the same. We then estimate $b_{1k}(t_j^0)$ and $b_{2k}(t_j^0)$ by minimizing the penalized sum of squared errors (PSSE):

$$
\begin{aligned}
PSSE_{b_1,b_2} &= \sum_{i=1}^{n} e_{ij}^2 + \rho_1 \|\mathbf{b}_1(t_j^0)\|^2 + \rho_2 \|\mathbf{b}_2(t_j^0)\|^2 \\
&= \|\mathbf{Y}_j - \tilde{\mathbf{X}}_1(t_j^0)\mathbf{b}_1(t_j^0) - \tilde{\mathbf{X}}_2(t_j^0)\mathbf{b}_2(t_j^0)\|^2 \\
&\quad + \rho_1 \|\mathbf{b}_1(t_j^0)\|^2 + \rho_2 \|\mathbf{b}_2(t_j^0)\|^2,
\end{aligned} \tag{6}
$$

where $\|\cdot\|$ is the Euclidean norm of a vector, $\mathbf{Y}_j = (Y_{1j}, ..., Y_{nj})^T$, $\tilde{\mathbf{X}}_1(t_j^0) = \left(\tilde{\mathbf{X}}_{11}(t_j^0), ..., \tilde{\mathbf{X}}_{1n}(t_j^0)\right)^T_{n\times K_1}$, $\tilde{\mathbf{X}}_2(t_j^0) = \left(\tilde{\mathbf{X}}_{21}(t_j^0), ..., \tilde{\mathbf{X}}_{2n}(t_j^0)\right)^T_{n\times K_2}$, $\rho_1 > 0$ and $\rho_2 > 0$ are the regularization parameters which are assumed to be time independent in order to reduce the high variability. The penalization does not only prevent overfitting but also guarantees the invertibility of matrix while solving the minimization problem. Then the minimizer of (6) is

$$
\begin{bmatrix} \hat{\mathbf{b}}_1(t_j^0) \\ \hat{\mathbf{b}}_2(t_j^0) \end{bmatrix} = \left(\mathbf{Z}_j^T\mathbf{Z}_j + \begin{bmatrix} \rho_1 I_{K_1} & 0 \\ 0 & \rho_2 I_{K_2} \end{bmatrix}\right)^{-1} \left(\mathbf{Z}_j^T\mathbf{Y}_j\right),
$$

where I_K is the $K \times K$ identity matrix and

$$
\mathbf{Z}_j = \begin{bmatrix} \tilde{X}_{111}(t_j^0) & \cdots & \tilde{X}_{11K_1}(t_j^0) & \tilde{X}_{211}(t_j^0) & \cdots & \tilde{X}_{21K_2}(t_j^0) \\ \vdots & & \vdots & \vdots & & \vdots \\ \tilde{X}_{1n1}(t_j^0) & \cdots & \tilde{X}_{1nK_1}(t_j^0) & \tilde{X}_{2n1}(t_j^0) & \cdots & \tilde{X}_{2nK_2}(t_j^0) \end{bmatrix}.
$$

Notice that when $t_{ij} \neq t_j^0$, $\hat{\mathbf{b}}_1(t_j^0)$ needs to be replaced by $\hat{\mathbf{b}}_1(t_{ij})$ and $\mathbf{Z}_j$ can be adapted correspondingly by replacing t_j^0 by t_{ij} and the formulas in this paragraph hold.

Therefore, for arbitrary $t \in [0, 1]$, we have

$$
\begin{bmatrix} \hat{\mathbf{b}}_1(t) \\ \hat{\mathbf{b}}_2(t) \end{bmatrix} = \left(\begin{bmatrix} \hat{\mathbf{C}}_{11}(t) & \hat{\mathbf{C}}_{12}(t) \\ \hat{\mathbf{C}}_{21}(t) & \hat{\mathbf{C}}_{22}(t) \end{bmatrix} + \begin{bmatrix} \frac{\rho_1}{n} I_{K_1} & 0 \\ 0 & \frac{\rho_2}{n} I_{K_2} \end{bmatrix}\right)^{-1} \begin{bmatrix} \hat{\mathbf{C}}_{1Y}(t) \\ \hat{\mathbf{C}}_{2Y}(t) \end{bmatrix} \tag{7}
$$

where $\hat{\mathbf{C}}_{11}(t) = \left[\hat{C}_{\tilde{X}_{1k},\tilde{X}_{1l}}(t)\right]_{kl}$ is a $K_1 \times K_1$ matrix with $\hat{C}_{\tilde{X}_{1k},\tilde{X}_{1l}}(t)$ an estimator of $C_{\tilde{X}_{1k},\tilde{X}_{1l}}(t) = cov\left(\tilde{X}_{1k}(t), \tilde{X}_{1l}(t)\right)$, $\hat{\mathbf{C}}_{12}(t) = \left[\hat{C}_{\tilde{X}_{1k},\tilde{X}_{2l}}(t)\right]_{kl}$ is a $K_1 \times K_2$ matrix with $\hat{C}_{\tilde{X}_{1k},\tilde{X}_{2l}}(t)$ an estimator of $C_{\tilde{X}_{1k},\tilde{X}_{2l}}(t) = cov\left(\tilde{X}_{1k}(t), \tilde{X}_{2l}(t)\right)$, $\hat{\mathbf{C}}_{21}(t) = \left[\hat{C}_{\tilde{X}_{2k},\tilde{X}_{1l}}(t)\right]_{kl}$ is a $K_2 \times K_1$ matrix with $\hat{C}_{\tilde{X}_{2k},\tilde{X}_{1l}}(t)$ an estimator of $C_{\tilde{X}_{2k},\tilde{X}_{1l}}(t) = cov\left(\tilde{X}_{2k}(t), \tilde{X}_{1l}(t)\right)$, $\hat{\mathbf{C}}_{22}(t) = \left[\hat{C}_{\tilde{X}_{2k},\tilde{X}_{2l}}(t)\right]_{kl}$ is a $K_2 \times K_2$ matrix with $\hat{C}_{\tilde{X}_{2k},\tilde{X}_{2l}}(t)$ an estimator of $C_{\tilde{X}_{2k},\tilde{X}_{2l}}(t) = cov\left(\tilde{X}_{2k}(t), \tilde{X}_{2l}(t)\right)$, $\hat{\mathbf{C}}_{1Y}(t) = \left[\hat{C}_{\tilde{X}_{11},Y}(t), ..., \hat{C}_{\tilde{X}_{1K_1},Y}(t)\right]^T$ is a vector and $\hat{C}_{\tilde{X}_{1l},Y}(t)$ is an estimator of $C_{\tilde{X}_{1l},Y}(t) = cov\left(\tilde{X}_{1l}(t), Y(t)\right)$, and $\hat{\mathbf{C}}_{2Y}(t) = \left[C_{\tilde{X}_{21},Y}(t), ..., C_{\tilde{X}_{2K_2},Y}(t)\right]^T$ is a vector and $\hat{C}_{\tilde{X}_{2l},Y}$ is an estimator of $C_{\tilde{X}_{2l},Y}(t) = cov\left(\tilde{X}_{2l}(t), Y(t)\right)$.

To obtain the necessary quantities in (7), we have to consider the embedding covariance functions and their corresponding estimations. We only give the

details for $C_{\tilde{X}_{1k},\tilde{X}_{1l}}(t)$ here and for the other covariance functions see Appendix A. For $C_{\tilde{X}_{1k},\tilde{X}_{1l}}(t)$, we have

$$\begin{aligned} C_{\tilde{X}_{1k},\tilde{X}_{1l}}(t) &= cov\big(\tilde{X}_{1k}(t), \tilde{X}_{1l}(t)\big) \\ &= \int_{\delta_{11}}^{\delta_{12}} \int_{\delta_{11}}^{\delta_{12}} B_{1k}(s) B_{1l}(u) E[X_1(t-s) X_1(t-u)] du ds \\ &= \int_{\delta_{11}}^{\delta_{12}} \int_{\delta_{11}}^{\delta_{12}} B_{1k}(s) B_{1l}(u) C_{X_1}(t-s, t-u) du ds, \end{aligned}$$

where $C_{X_1}(s,u)$ is the covariance between $X_1(s)$ and $X_1(u)$. Since predictor X_1 is densely observed, $C_{X_1}(s,u)$ can be estimated by bivariate kernel smoothing, see [2]:

$$\hat{C}_{X_1}(s,u) = \frac{1}{(m_1 b)^2} \sum_{j,k=1}^{m_1} K\left(\frac{s - s_{1j}}{b}, \frac{u - s_{1k}}{b}\right) \frac{1}{n} \sum_{i=1}^{n} W_{1ij} W_{1ik},$$

where b is a bandwidth, m_1 is the number of observations of the dense and regular predictor X_1 for each subject, and K is a bivariate kernel function.

Once $\hat{\mathbf{b}}_1(t)$ and $\hat{\mathbf{b}}_2(t)$ are obtained (for given lags δ's and regularization parameters ρ's), we can estimate coefficient functions by

$$\hat{\beta}_1(s,t) = \sum_{k=1}^{K_1} B_{1k}(s) \hat{b}_{1k}(t), \ s \in \Delta_1, \ t \in [0,1]$$

and

$$\hat{\beta}_2(s,t) = \sum_{k=1}^{K_2} B_{2k}(s) \hat{b}_{2k}(t), \ s \in \Delta_2, \ t \in [0,1].$$

Theorem 1 *Under assumptions* (*A*1–*A*5) *and* (*B*1–*B*3) *specified in Appendix B, denote* $I_t = [\max\{\delta_{12}, \delta_{22}\}, 1]$,

$$\lim_{n\to\infty} \sup_{s,t \in \Delta_1 \times I_t} |\hat{\beta}_1(s,t) - \beta_1(s,t)| = 0 \quad \text{in probability.}$$

$$\lim_{n\to\infty} \sup_{s,t \in \Delta_2 \times I_t} |\hat{\beta}_2(s,t) - \beta_2(s,t)| = 0 \quad \text{in probability.}$$

Remark 1 Here the convergence rate depends on the tuning parameters, e.g. the bandwidth b, used in estimating the corresponding covariance structures (including auto-covariance structure of the predictor processes and cross-covariance structure between the predictor processes and response process). Note that the bandwidth b should be chosen in such a way that there are a sufficient number of observations in the interval to estimate the covariance. For details of the proof, see Appendix C.

Remark 2 If the mean $\mu_{X_1}(t)$ and $\mu_{X_2}(t)$ of the predictors $X_1(t)$ and $X_2(t)$ are not equal to 0, they can be estimated nonparametrically. Specifically, for the sparse and

irregular longitudinal predictor, one can use the method proposed by Yao et al. [19, 20] or by [9, 14]. For the dense and regular functional predictor, methods such as kernel estimation proposed by [2] can be used.

4 Prediction

Suppose we observe a new discrete response curve $\mathbf{Y}_j^* = (Y^*(t_1^*), ...Y^*(t_{m^*}^*))$, discrete dense predictor trajectory $\mathbf{W}_1^* = (W_1^*(s_{11}), ..., W_1^*(s_{1m_1}))^T$ and discrete sparse predictor trajectory $\mathbf{W}_2^* = (W_2^*(s_{21}^*), ...W_2^*(s_{2m_2}^*))^T$. From the original model (1), the predicted response curve is

$$E\left[Y^*(t)|X_1^*, X_2^*\right] = \beta_0(t) + \int_{\delta_{11}}^{\delta_{12}} \beta_1(s,t)X_1^*(t-s)ds + \int_{\delta_{21}}^{\delta_{22}} \beta_2(s,t)X_2^*(t-s)ds. \tag{8}$$

However, the lags $\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$ and regularization parameters ρ_1 and ρ_2 have to be determined and the functional representation of the predictor trajectories $X_1^*(s)$ and $X_2^*(s)$ have to be recovered from data.

For $X_1^*(s)$, it can be easily recovered by kernel smoothing, since the sampling is dense. However for $X_2^*(s)$, since the sampling is sparse and irregular, we use functional principal component analysis (FPCA) and approximate the curve by a limited number of functional principal components. We assume $X_2^*(s) \sim X_2(s) \in L^2[0,1]$ and $E[X_2(s)] = 0$. Denote the covariance of $X_2(s)$ by $C_{X_2}(s,u) = cov(X_2(s), X_2(u))$, then the Mercer's theorem gives the following spectral decomposition of the covariance

$$C_{X_2}(s,t) = \sum_{l=1}^{\infty} \lambda_l \phi_l(s)\phi_l(u)$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ are eigenvalues and ϕ_l are orthonormal eigenfunctions. By Karhunen-Loève (KL) expansion, $X_2^*(s)$ can be represented as

$$X_2^*(s) = \sum_{l=1}^{\infty} \xi_l^* \phi_l(s)$$

where $\xi_l^* = \int_0^1 X_2^*(s)\phi_l(s)ds$ are the functional principal component scores and are uncorrelated random variables with mean 0 and variance λ_l. The truncated formula for $X_2^*(s)$ is:

$$X_2^{*L}(s) = \sum_{l=1}^{L} \xi_l^* \phi_l(s).$$

The covariance $C_{X_2}(s,t)$ can be estimated as we discussed in last section and the eigenfunctions ϕ_l can be estimated following the spectral decomposition of the estimated covariance. However, the scores ξ_l^* cannot be approximated by numerical integration which is used for dense functional data. In fact, under the Gaussian assumption, denote $\boldsymbol{\phi}_l = (\phi_l(s_{21}^*), ..., \phi_l(s_{2m_{X_2}^*}^*))^T$, the best linear predictor for ξ_l^* is (see [11], Yao et al. [19, 20] or see the application in [10])):

$$\tilde{\xi}_l^* = \lambda_l \boldsymbol{\phi}_l^T \Sigma^{-1} \mathbf{W}_2^*$$

where $\Sigma = var(\mathbf{W}_2^*)$. Then the estimate of ξ_l^* can be defined as

$$\hat{\xi}_l^* = \hat{\lambda}_l \hat{\boldsymbol{\phi}}_l^T \hat{\Sigma}^{-1} \mathbf{W}_2^*.$$

The number of eigenfunctions L can be selected to be the number of eigenfunctions that explain 99% of the variation. Once obtaining the estimation of eigenfunctions ϕ_l, scores ξ_{il} and L, $X_2^*(s)$ can be recovered as

$$\hat{X}_2^*(s) = \sum_{l=1}^{L} \hat{\xi}_l^* \hat{\phi}_l^*(s).$$

After plugging the functional representation of the predictor curves $\hat{X}_1^*(s)$ and $\hat{X}_2^*(s)$ into (8), we have

$$\begin{aligned} \hat{Y}_L^*(t) &= \int_{\delta_{11}}^{\delta_{12}} \hat{\beta}_1(s,t)\hat{X}_1^*(t-s)ds + \int_{\delta_{21}}^{\delta_{22}} \hat{\beta}_2(s,t)\hat{X}_2^*(t-s)ds \\ &= \int_{\delta_{11}}^{\delta_{12}} \hat{\beta}_1(s,t)\hat{X}_1^*(t-s)ds + \int_{\delta_{21}}^{\delta_{22}} \hat{\beta}_2(s,t) \sum_{l=1}^{L} \hat{\xi}_l^* \hat{\phi}_l(t-s)ds. \end{aligned} \tag{9}$$

Define

$$\tilde{Y}^*(t) = \int_{\delta_{11}}^{\delta_{12}} \beta_1(s,t)X_1^*(t-s)ds + \int_{\delta_{21}}^{\delta_{22}} \beta_2(s,t) \sum_{l=1}^{\infty} \tilde{\xi}_l^* \phi_l(t-s)ds.$$

and

$$\tilde{Y}_L^*(t) = \int_{\delta_{11}}^{\delta_{12}} \beta_1(s,t)X_1^*(t-s)ds + \int_{\delta_{21}}^{\delta_{22}} \beta_2(s,t) \sum_{l=1}^{L} \tilde{\xi}_l^* \phi_l(t-s)ds.$$

Theorem 2 *Under assumptions* (A1–A5) *and* (B1–B3) *in Appendix B, denote* $I_t = [\max\{\delta_{12}, \delta_{22}\}, 1]$, *for all* $t \in I_t$, *we have*

$$\lim_{n\to\infty} \hat{Y}_L^*(t) = \tilde{Y}^*(t) \quad \text{in probability.}$$

Remark 3 Note that the number of eigenfunctions L used in the KL expansion of the sparse and irregular predictor process is a function of sample size n and goes to infinity as n goes to infinity. For details of the proof, see Appendix C.

5 Computation of the Lags

The final task is to estimate the time lag δ's. For selecting δ's and ρ's, we consider the normalized prediction error (NPE) criterion and the K-fold cross validation criterion. Specifically, for pre-specified δ and ρ, NPE in this situation is defined as

$$NPE\{(\delta,\rho)\} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_{Yi}}\frac{\left|\hat{Y}_{ij} - Y_{ij}\right|}{|Y_{ij}|} \tag{10}$$

where $\hat{Y}_{ij}$ is the predicted value for the jth measurement on the ith response trajectory $Y(t)$ obtained based on the pre-specified δ and ρ, $N = \sum_{i=1}^{n} m_{Y_i}$. Similarly, for pre-specified δ and ρ, we define K-fold cross validation criterion as follows. We divide the dataset into K equal parts. For each $k = 1, ..., K$, we fit the model to the other $K-1$ parts, which gives the estimation of coefficient functions, and further gives the prediction $\hat{Y}_{ij}^{-k}$ in the kth part. Then the K-fold cross validation score is defined as,

$$CV\{(\delta,\rho)\} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i\in k\text{th part}}\frac{1}{m_{Y_i}}\sum_{j=1}^{m_{Y_i}}\left(\hat{Y}_{ij}^{-k} - Y_{ij}\right)^2. \tag{11}$$

Similar criteria are considered in [8] and Pomann et al. [16].

Then δ's and ρ's are chosen in a hierarchical manner. Let D_1 and D_2 be the sets of potential lags for the first and second predictor, i.e. $\{(\delta_{11}, \delta_{12})\}$ and $\{(\delta_{21}, \delta_{22})\}$, respectively. Let D_ρ be the set of potential regularization parameters $\{(\rho_1, \rho_2)\}$. Firstly, for a fixed point of lags $\delta^0 = \left((\delta_{11}^0, \delta_{12}^0), (\delta_{21}^0, \delta_{22}^0)\right) \in D_1 \times D_2$, we calculate the NPE values for all $\rho \in D_\rho = \{(\rho_1, \rho_2)\}$, i.e. $NPE\{(\delta^0, \rho)\}_{\rho\in D_\rho}$. Then the ρ that achieves the smallest NPE value, i.e. $\rho_{opt}(\delta^0) = \min_{\rho\in D_\rho} NPE\{(\delta^0, \rho)\}$, is chosen as the optimal ρ for the given fixed point of lags δ^0. Secondly, we calculate the cross validation score $CV\{(\delta^0, \rho_{opt}(\delta^0))\}$ for the given δ^0 based on the $\rho_{opt}(\delta^0)$. At last, we repeat the above steps for all $\delta \in D_1 \times D_2$ and we obtain the cross validation score for all $\delta \in D_1 \times D_2$, $CV\{(\delta, \rho)\}_{\delta\in D_1\times D_2, \rho_\delta \in D_\rho}$. Then, the optimal δ is chosen to be the one with the smallest cross validation score.

Remark 4 Since for our simulation study and real data analysis, Y_{ij}, for j such that $t_j > \max\{\delta_{12}, \delta_{22}\}$, is larger than 0.5, we define NPE as in formula (9). If $\min\{|Y_{ij}|\}$ is very close to zero, one may define NPE as

$$NPE\{(\delta,\rho)\} = \frac{1}{N}\sum_{i=1}^{n}\frac{\sum_{j=1}^{m_{Yi}}\left|\hat{Y}_{ij}-Y_{ij}\right|}{\sum_{j=1}^{m_{Yi}}|Y_{ij}|}.$$

In our simulation and real data analysis, the two NPE criteria give the same results.

Remark 5 Here, for pre-specified δ, we first chose ρ from a set of ρ values using the NPE criterion. Then we chose δ based on the CV criterion from a set of δ values. The advantage is that the computationally faster NPE criterion for choosing ρ and the more refined CV criterion for selecting the δ are used.

6 Simulations

We study efficiency of the NPE criterion for selecting the time lags δ's and regularization parameters ρ's.

For $n = 50,\ 100,\ 150,\ 200$ subjects, we first generate two predictor curves $X_1(t)$ and $X_2(t)$ on dense and equally spaced time points over [0, 1], i.e. $\{j/99, j = 0, ..., 99\}$ and then accordingly generate the response curve $Y(t)$ at time points $\{j/99, j = 0, ..., 99\}$. The number of measurements made on the ith response m_{Yi} is randomly selected from 20 to 50, the number of measurements made on the ith predictor m_{1i} is 100 and the number of measurements made on the ith predictor m_{2i} is randomly selected from 30 to 50.

Define $X_{1i}(t) = \xi_{i1}\sin(2\pi t) + \xi_{i2}t^2$ with $\xi_{i1} \overset{iid}{\sim} N(0,1)$ and $\xi_{i2} \overset{iid}{\sim} N(0,1)$, $X_{2i}(t) = \zeta_i\cos(2\pi t)$ with $\zeta_i \overset{iid}{\sim} N(0,1)$. We take the same time lags for both X_1 and X_2, i.e. $\delta_{11} = \delta_{21} = 0.1$, $\delta_{12} = \delta_{22} = 0.4$. For coefficient functions, we choose $\beta_0(t) = t + t^{1/5}$, $\beta_1(s,t) = \sin(2\pi s)\cos(\pi t)$, $t \in [0,1]$, $s \in [0.1, 0.4]$, $\beta_2(s,t) = \sin(4\pi s)\cos(2\pi t)$, $t \in [0,1]$, $s \in [0.1, 0.4]$. The measurement errors are taken to be independent normal with signal to noise ratio 20 for the predictors and response. Figure 1 shows the simulated data for the first replicate with $n = 100$.

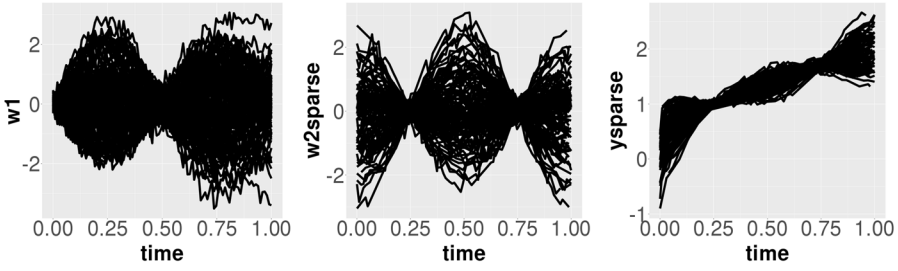

Fig. 1 Simulated data of first replicate: The left plot shows the discrete noisy observations of the first predictor which are observed for a dense and regular grid. The middle plot shows the discrete noisy observation of the second predictor which are observed for a sparse and irregular grid. The right plot shows the discrete noisy observations of the response which are sparsely and irregularly observed

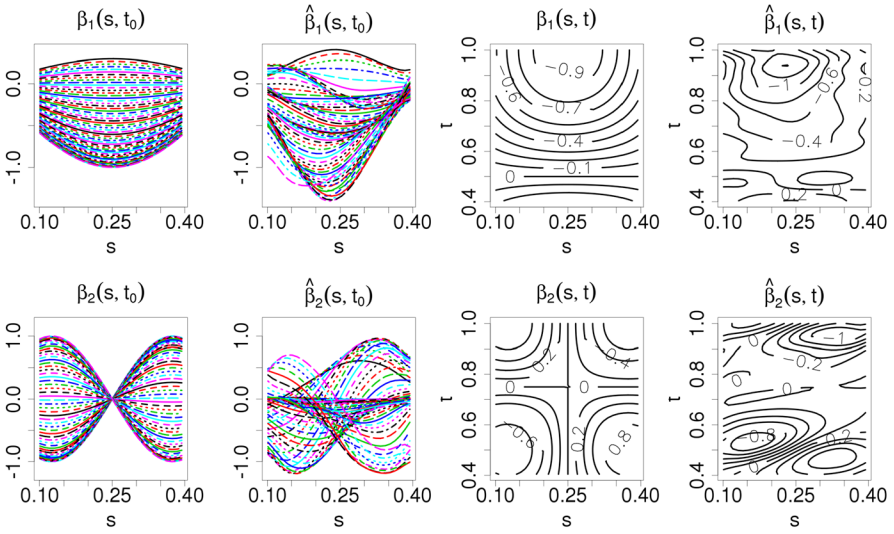

Fig. 2 The estimated functions for the first replicate: The left upper corner plot is the true β_1: abscissa is s with the domain [0.1, 0.4], ordinate is the values of β_1, there are 60 curves and they are $\beta_1(s, t_j)$ for $t_j = j/99$, $j = 40, ..., 99$. The second left upper plot is the estimation of β_1. The third left upper plot is the contour line of the true β_1. The fourth left upper plot is the contour line of the estimated $\hat{\beta}_1$. The bottom panel shows the true and estimated β_2

Table 1 NPEs based on correct lags

n	50	100	150	200
NPE×100	2.08	1.95	1.86	1.79

Since we could not assume any prior on the coefficients and B-spline basis are computationally fast and have good properties, we use B-spline functions for the estimation and prediction. With respect to K_1 and K_2, we use B-spline functions of degree 4 with 10 equally spaced interior knots over Δ_1 and Δ_2 (number of bases is 14). For details on B-spline basis, see for example [4]. The number of functional principal components is chosen such that 99% of variation is kept. The penalized parameters ρ_1 and ρ_2 are chosen on the dense grid of ρ_1, $\rho_2 \in [10^{-5}, 10^{-2}]$ with 20 equally-spaced points in both directions. We use NPE criterion and 10-fold cross validation criterion to determine the regularization parameters and the lags. Notice that in order to check the estimation performance, the estimation procedure is done under the correct lags, i.e. $\delta_{11} = \delta_{21} = 0.1$, $\delta_{12} = \delta_{22} = 0.4$. Figure 2 shows the result of one simulation, where ρ_1 is chosen as 4.28×10^{-4}, ρ_2 is chosen as 8.86×10^{-4} and the corresponding NPE is 1.95×10^{-2}. From Fig. 2, we conclude that our model successfully reveals the structure of coefficient functions.

Table 1 shows the asymptotic properties of our estimation. For different number of observations $n = 50, 100, 150, 200$, the NPE values are shown and also the estimation is based on the correct lags. As we can see, the NPE decreases as n increases which is expected based on Theorem 1.

For evaluating the performance of our model on selecting the effect lags, the ρ's are determined based on the NPE criterion and the δ's are determined based on 10-fold cross-validation score. Since the true $\delta_{11} = \delta_{21} = 0.1$ and $\delta_{12} = \delta_{22} = 0.4$, in order to save computational time, we fix the ending point i.e. $\delta_{11} = \delta_{21} = 0.1$, use the same starting point, i.e. $\delta_{12} = \delta_{22}$ and search over $\{0.3, 0.4, 0.5\}$. That is we have three combinations but there is only one correct combination. Our model has 65 correct choices out of 100 simulations.

7 Data Analysis

Chronic Obstructive Pulmonary Disease (COPD) continues to be one of the leading causes of morbidity and mortality in the world and a burden on many national health systems [13]. Many of the risk factors associated with the disease are controlled by each patient's personal decisions, an example of such a risk factor is how much a patient smokes (see [1]). Other risk factors are more complicated and are determined by socioeconomic factors such as where a patient lives (for example the presence of sport facilities and the amount of air pollution in the neighbourhood) and their access to healthcare. These factors are more likely to be effected by government and policy. In 2003, the US National Heart, Lung, and Blood Institute estimated that the total costs (direct and indirect) of COPD was approximately $32.1 billion for the year (see [13]). Modelling COPD in a predictive way could be a useful asset in allocating health resources to better deal with predicted spikes in COPD exacerbations and can be used to identify areas in which preventative measures can be taken to better COPD health outcomes. Both of these approaches could potentially make significant headway in reducing the morbidity related to COPD and the economic costs of the disease.

Air quality is a known factor that affects a person's health and quality of life. Previous research suggests a link between air quality and COPD hospitalisation (see [21]). More specifically, a study looking at the effect of pollutants on 94 COPD sufferers living in London found that a rise in NO2 concentration accounted for a 6% increase in the odds of a symptomatic fall in peak flow rate (see [15]). We aim to link heterogeneous data sets regarding NO2 concentration and exercise intensity to model COPD hospital admissions for the city of Leeds.

Healthcare, air quality and lifestyle datasets are typically a mixture of static, temporal, dense and sparse data. The pollution data is usually fairly regular over time (the data we present is collected every 15 min), but there is only a limited number of monitoring sites in a city meaning that the data are spatially sparse. Physical activity data can come from a variety of sources stretching from a small group of volunteers who allow their daily movements to be tracked, to a large sample based on information from logged visits to local gym facilities.

7.1 Pollution Data

Pollution data was taken from 10 automatic monitoring points from across Leeds taking measurements of NO2 concentration every 15 min. The dataset from 2013 to 2018 can be found Ratified air quality - nitrogen dioxide. The location of the 10 monitors is shown in Fig. 6 and a snapshot of the data is shown in Table 7 in Appendix D. A mean value at each collection point was calculated daily for all years. Some collection points reported N/As for long periods of time making the data sparse at some points in time.

The 10 collection points were mapped onto a 861 point grid of Leeds. The location of the grid is shown in Fig. 8 in Appendix D. Using the Krige function in R, an interpolation over this grid was carried out using the inverse distance weighted interpolation method with a power of 2 (default). N/As were excluded. The points of the grid were then allocated to postcode district by matching them to the closest postcode (Euclidean distance). This was done using the website http://www.geodojo.net/uk/converter/. The data was then aggregated down to district level by taking the mean of all data points within a district. From this the daily NO2 for each postcode district in Leeds is calculated along with the standard deviation between all the points that lie within the district boundaries, see Table 9 in Appendix D.

The daily mean NO2 concentrations from 2013 to 2018 for the 18 postcode districts in Leeds and their kernel smoothing with data-adaptive local plug-in bandwidth selection curves are shown in the left and right panel of Fig. 3 respectively. As expected a yearly seasonal effect for daily mean NO2 is observed with peaks in the winter.

7.2 Physical Activity Data

Raw physical activity data was obtained from a local authority concerning use of their gym and sports facilities for the period 2013 to 2018. These data were given per postcode sector level and we further aggregated these to postcode district level (e.g. LS1).

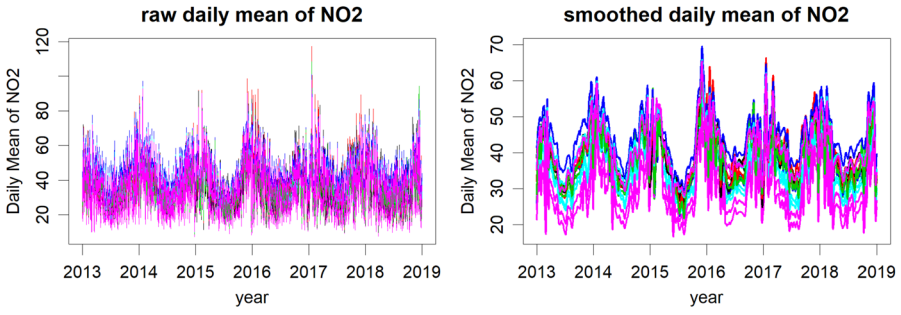

Fig. 3 Raw and smoothed daily mean NO2 concentration of 18 postcode districts

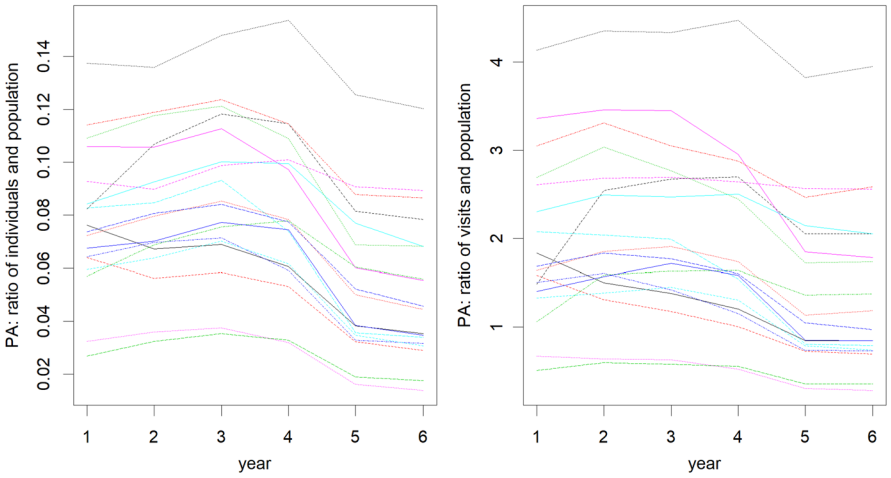

Fig. 4 Raw physical activity data for 18 postcode districts around Leeds. Left panel shows the standardised number of individuals using gym facilities for 18 postcode districts across Leeds from 2013 to 2018. Right panel shows the number of visits to gym facilities per person per week for 18 postcode districts across Leeds from 2013 to 2018

To standardise the physical activity data, the population size per district was used, i.e. the counts per postcode district were divided by the population size of the district. For the population size, we used mid year population estimates per lower layer super output area (LSOA) which were obtained from the Office for National Statistics (ONS). These were converted to postcode district by using a postcode lookup which maps postcodes onto a LSOA. When the LSOA covers multiple postcode districts, the LSOA population was split equally between them.

The physical activity curves for the 18 postcode districts in Leeds are given in Fig. 4. These curves show little variation in time.

7.3 COPD Data

Temporal data on hospital admissions due to COPD in Leeds was provided by the Leeds Teaching Hospitals NHS Trust. Raw data include admission date, age, disease, unique ID, sex, and area by postcode information. We removed records of subjects with ages that were not numbers. We restricted the dataset to ages above 20 years and to the time window of 2013 to 2018. The obtained dataset comprised 7944 COPD hospital admissions.

The data was then grouped by day. To standardise, the expected count of hospital admissions E_i was calculated. We divided the total number of counts over the whole 6 year period by the number of days in this period. Then to take the size of the population into account, again, the mid year population estimates obtained from the ONS at LSOA level were used (see above). Thus, for each district i, $i \in \{1, 2, ..., 18\}$, with

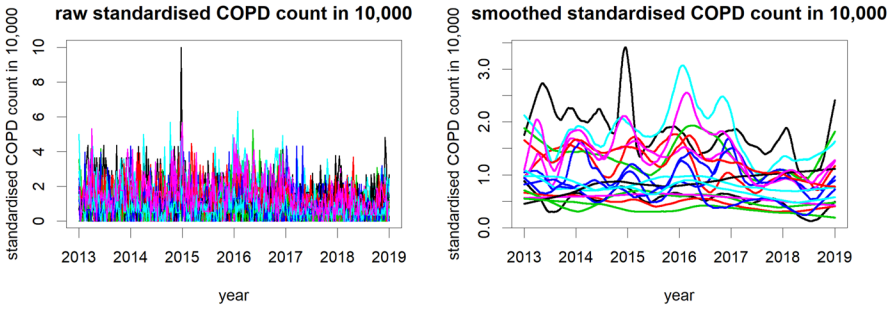

Fig. 5 Raw and smoothed weekly standardised COPD hospital admissions in 10,000. The left panel shows the raw weekly standarised COPD hospital admissions in 18 postcode districts in Leeds from 1st January 2013 to 31st December 2018. The right panel shows the corresponding smoothing curves that are estimated through kernel smoothing with data-adaptive local plug-in bandwidth selection

a population size P_i, and an expected admission count E_i, the standardised count is defined as

$$Standardised\ Count_i = \frac{Count_i}{P_i E_i},$$

where $Count_i$ is the weekly admission count in the ith district. The weekly admissions after standardisation based on district population (per 10,000) are shown in the left panel of Fig. 5 and the corresponding kernel smoothing with data-adaptive local plug-in bandwidth selection curves [3, 6] are given in the right figure. It appears that there is a yearly seasonality phenomenon with peaks every late winter and early spring.

7.4 Data Analysis Results

As a preliminary analysis, we first fitted a standard linear regression model with average standardised COPD hospital admissions Y as response variable and average standardised weekly number of gym visits and average daily NO2 value as covariates. Both covariates appeared to have a statistically significant effect on the outcome. The estimate of β_1 is -2.86 with p value of 0.003 and the estimates of β_2 of 0.007 with p value of 0.003.

Next we are interested in the time aspect. Figure 3 and Fig. 5 show temporal fluctuations and a potential effect lag (delay) of the influence of NO2 concentrations on COPD hospital admissions. On the other hand, Fig. 4 shows that physical activity curves for 18 postcode districts across Leeds have little variation in time. Therefore, we did not consider physical activity as a covariate in a functional regression model.

In order to obtain a rough idea of the effect lag and duration of influence, we compared the position of local maximum of NO2 concentrations and hospital admissions. Specifically, based on the background, we restrict the starting point and ending point within one period (one year). And then we designed quite dense searching grid (with respect to the smoothness of the curves) for

ending point as $\{7, 14, 21, 24.5, 28, 31.5, 35\}$ (days) and that for starting point as $\{84, 98, 112, 126, 140, 154, 168, 182, 196, 199.5, 203, 206.5, 210, 213.5, 217, 220.5, 224, 238, 252, 266, 280\}$ (days). The mean functions of both predictor and outcome are estimated using kernel smoothing method as mentioned in Remark 2 on page 9. We used the same basis system and number of basis for estimation and prediction as in the simulation section, i.e. we use B-spline functions of degree 4 with 10 equally spaced interior knots (number of basis is 14). Here, we did 7*21 searches and the computation time is 4.6 h (on a standard laptop using Rstudio) which is affordable.

The optimal combination of ending point and starting point appears to be 24.5 and 203 days, i.e. 3.5 and 29 weeks based on leave-one-curve-out cross validation squared prediction error criterion. This means that $t - 203$ is the starting effective time and $t - 24.5$ is the ending effective time for NO2 concentration to have effect on COPD hospital admissions at time t. In other words, it takes around 3.5 weeks for NO2 concentration to have effect on COPD hospital admissions and this effect lasts around 25.5 weeks.

We did another 7*21 searches on a slightly different grid, i.e. ending points are $\{7, 14, 21, 22.75, 26.25, 29.75, 33.25\}$ (days) and starting points are $\{84, 98, 112, 126, 140, 154, 168, 182, 196, 197.75, 201.25, 204.75, 208.25, 211.75, 215.25, 218.75, 224, 238, 252, 266, 280\}$ (days). The results are only slightly different: the optimal combination of ending point and starting point is 26.25 and 210 days, i.e. 3.75 and 28.25 weeks based on leave-one-curve-out cross validation squared prediction error criterion.

8 Discussion

In this paper, we have developed a functional regression model that combines heterogeneous predictors, in particular sparse and dense functional predictors, with their effects on the response being restricted by time lags that in effect create an interval where the predictors affect the response. We prove the consistency of the coefficient functions of both the dense and sparse predictors. For prediction, we recovered new dense predictors using kernel smoothing, whereas we used FPCA for the sparse ones and proved the asymptotic property of the predictions. Finally, we minimized the normalized prediction error in order to find the optimal lags. Simulation studies are conducted to evaluate the model, the coefficient functions are estimated and the lags are determined with high accuracy.

The lags are estimated by using a grid search. For the simulation study, a grid search is adopted over several points that are close to the true effect lags. For real applications, we do not know the true effect lags and a larger grid is necessary or preliminary information can be used. One can also consider to run a preliminary search for each predictor separately in order to get a rough estimate of the lags and then design the grid accordingly. In our data application, we did 7*21 searches, and the computation time is 4.6 h (on a standard laptop using Rstudio). We did another search and obtained almost the same estimates. For these searches, the assumption is made that the true lag is included in the grid or close to one if the elements in the

grid. Preliminary runs can provide some evidence that the true lags are included in the run. Moreover, if the final estimated lags are near the end points of the intervals/grid it is recommended that the grid is extended and a new fit is implemented.

Our model assumes that the lags do not change over time. Since the coefficient functions changes over time, the effective length of the historical interval of covariate values having an effect on the outcome can still be smaller for some periods. It might be interesting to investigate whether the length of the historical interval changes over time or is periodic. However such a model will be computationally intensive. Another extension is to use multivariate dense functional and sparse and irregular longitudinal predictors. As the estimator is derived from formula (6), if more than two predictors exist, one can estimate their covariance structure based on the type of the predictors (dense, sparse and irregular) then the corresponding $\hat{b}(t)$ can be obtained from formula (6). Note that this does not result in an over-parameterisation.

Our model is used to analyse the influence of daily mean NO2 concentrations on the COPD hospital admissions in 18 postcode districts in Leeds. It appeared that NO2 concentrations of 25.5 weeks ago still have an effect on COPD hospital admissions and that the NO2 concentrations of the last 3.5 weeks have not yet an effect. While it makes sense that historical NO2 concentrations have an effect and that recent values have not yet an effect on COPD hospital admissions, the estimates of the lags themselves need careful interpretation. First we do not provide standard errors. One way to obtain standard errors might be a bootstrap method, however such a method is very intensive computationally. Also the lags have to be interpreted together with the β functions which can vary over a time. Thus for a part of the time interval between lags, the effect of the covariate can be zero if the coefficient function is zero in this interval. Further, the models are fitted on aggregated data.

COPD hospital admissions are essentially Poisson distributed in each day, however in order to validate our methodology we aggregate weekly and standardised based on district populations and then consider these weekly aggregated data as continuous. Therefore, future research on general response is of great importance both in theory development and applications. Another extension is to model the spatial correlation of the COPD admissions. More efficient parameter estimates can be obtained by using this information.

Our model can be used to manage resources of hospitals, since it predicts the delay of hospital admissions after NO2 pollution as well as the duration. Another potential application is decision making around policies for improving the health in a city. For example should a city invest in more sport facilities or in a greener transport system? To answer such questions, the model should probably be extended with more factors. Moreover one might want to address possible missingness processes, for example the observed use of sport facilities is an underestimation since people might do sports using other facilities or in other districts. Recently, we developed methods for appropriate estimation of the mean function for temporal data subject to a detection limit [9].

To conclude we presented methods for functional analysis of temporal data where the effect of the temporal covariate might be delayed. We used this method for an interesting problem from urban analytics. We identified a delayed effect of 3.5 weeks for NO2 air pollution on COPD hospital admissions.

Appendix

Appendix A: Covariance Functions and Corresponding Estimation in Section 3

- For $C_{\tilde{X}_{2k},\tilde{X}_{2l}}(t)$, we have

$$\begin{aligned} C_{\tilde{X}_{2k},\tilde{X}_{2l}}(t) &= cov\big(\tilde{X}_{2k}(t), \tilde{X}_{2l}(t)\big) \\ &= \int_{\delta_{21}}^{\delta_{22}} \int_{\delta_{21}}^{\delta_{22}} B_{2k}(s)B_{2l}(u)E[X_2(t-s)X_2(t-u)]duds \\ &= \int_{\delta_{21}}^{\delta_{22}} \int_{\delta_{21}}^{\delta_{22}} B_{2k}(s)B_{2l}(u)C_{X_2}(t-s,t-u)duds \end{aligned}$$

  where $C_{X_2}(s,u)$ is the covariance between $X_2(s)$ and $X_2(u)$. Since the predictor X_2 is sparsely observed, $C_{X_2}(s,u)$ can be estimated by local linear surface smoother [19] which is defined through minimizing

$$\sum_{i=1}^{n} \frac{1}{(m_{2i}b)^2} \sum_{j\neq k=1}^{m_{2i}} K\left(\frac{s-s_{2ij}}{b}, \frac{u-s_{2ik}}{b}\right) \times (W_{2ij}W_{2ik} - \alpha_0 - \alpha_1(s-s_{2ij}) - \alpha_2(u-s_{2ik}))^2$$

  with respect to α_0, α_1, α_2, where b is a bandwidth, m_{2i} is the number of observations of X_2 for the subject i and K is a bivariate kernel function. And $\hat{C}_{X_2}(s,u) = \hat{\alpha}_0$.
- For $C_{\tilde{X}_{1k},\tilde{X}_{2l}}(t)$, we have

$$\begin{aligned} C_{\tilde{X}_{1k},\tilde{X}_{2l}}(t) &= cov\big(\tilde{X}_{1k}(t), \tilde{X}_{2l}(t)\big) \\ &= \int_{\delta_{11}}^{\delta_{12}} \int_{\delta_{21}}^{\delta_{22}} B_{1k}(s)B_{2l}(u)E[X_1(t-s)X_2(t-u)]duds \\ &= \int_{\delta_{11}}^{\delta_{12}} \int_{\delta_{21}}^{\delta_{22}} B_{1k}(s)B_{2l}(u)C_{X_1,X_2}(t-s,t-u)duds \end{aligned}$$

  where $C_{X_1,X_2}(s,u)$ is the covariance between $X_1(s)$ and $X_2(u)$. Since the predictor X_1 is densely observed and X_2 is sparsely observed, $C_{X_1}(s,u)$ can be estimated by local surface smoothing.
- For $C_{\tilde{X}_{2k},\tilde{X}_{1l}}(t)$, it is similar to $C_{\tilde{X}_{1k},\tilde{X}_{2l}}(t)$.
- For $C_{\tilde{X}_{1l},Y}(t)$, we have

$$\begin{aligned} C_{\tilde{X}_{1l},Y}(t) &= cov\big(\tilde{X}_{1l}(t), Y(t)\big) \\ &= \int_{\delta_{11}}^{\delta_{12}} B_{1l}(s)E[X_1(t-s)Y(t)]ds \\ &= \int_{\delta_{11}}^{\delta_{12}} B_{1l}(s)C_{X_1,Y}(t-s,t)ds \end{aligned}$$

where $C_{X_1,Y}(s,u)$ is the covariance between $X_1(s)$ and $Y(u)$. Since X_1 is densely observed and Y is sparsely or densely observed, $C_{X_1,Y}(s,u)$ can be estimated by local linear surface smoothing.

- For $C_{\tilde{X}_{2l},Y}(t)$, it is similar to $C_{\tilde{X}_{1l},Y}(t)$.

Appendix B: Assumptions

We first give the assumptions (A) which are needed for both Theorem 1 and Theorem 2.

We assume the data of the sparse predictor $\{W_{2ij}, s_{2ij} : i = 1, ..., n,\ j = 1, ..., m_{2i}\}$ and the sparse response $\{Y_{ij}, t_{ij} : i = 1, ..., n,\ j = 1, ..., m_{Y_i}\}$ to be iid samples from the joint densities, $g_{X_2}(x,s)$ and $g_Y(y,t)$. The observations of dense predictor $\{W_{1ij}, s_{1ij} : i = 1, ..., n,\ j = 1, ..., m_1\}$ are also assumed to be iid for different i. We assume t_{ij} and s_{2ij} are iid with marginal densities $f_t(t)$ and $f_s(s)$, while $s_{1ij} - s_{1i(j-1)}$ are small and fixed for any i and j. We assume $(X_{2ij}, X_{2il}, s_{2ij}, s_{2il})$ and $(Y_{ij}, Y_{il}, t_{ij}, t_{il})$ are identically distributed with joint density functions $g_{X_2X_2}(x_1, x_2, s_1, s_2)$ and $g_{YY}(y_1, y_2, t_1, t_2)$ respectively. Let $p_1,\ p_2 \in \mathbb{N}$ such that $0 \le p_1 + p_2 \le 4$.

(A1) For $p_1 + p_2 = p,\ 0 \le p_1, p_2 \le p$. The derivatives $\frac{d^p f_t(t)}{d^p t}$ and $\frac{d^p f_s(s)}{d^p s}$ exist and are continuous on [0, 1] with $f_t(t) > 0$ and $f_s(s) > 0$ on [0, 1]. The derivatives $\frac{d^p g_{X_2}(x,s)}{d^p s}$ and $\frac{d^p g_Y(y,t)}{d^p t}$ exist and are continuous on $\mathbb{R} \times [0,1]$. The derivatives $\frac{d^p g_{X_2X_2}(x_1,x_2,s_1,s_2)}{d^{p_1}s_1 d^{p_2}s_2}$ and $\frac{d^p g_{YY}(y_1,y_2,t_1,t_2)}{d^{p_1}t_1 d^{p_2}t_2}$ exist and are continuous on $\mathbb{R}^2 \times [0,1]^2$.

(A2) The number of observations for sparse and irregular predictor m_{2i} is a random variable such that $m_{2i} \sim_{iid} M_2$, where $m_2 > 0$ is a discrete random variable with $P(M_1 > 1) > 0$. The number of observations sparse and irregular response m_{Y_i} for the i-th subject is a random variable such that $m_{Y_i} \sim_{iid} M_Y$, where $M_Y > 0$ is a discrete random variable with $P(M_Y > 1) > 0$. We assume M_2 and M_Y are independent. The observation times and measurements are assumed to be independent of the number of observations for any subjects and for any subset of any subjects, which means $\{W_{2ij}, Y_{ik}, s_{2ij}, t_{ik} : j \in J_i, k \in K_i\}$ is independent of M_2 and M_Y where $J_i \in \{1, ..., M_2\}$ and $K_i \in \{1, ..., M_Y\}$.

Let $K(\cdot)$ and $K(\cdot,\cdot)$ be the nonnegative univariate and bivariate kernel functions for smoothing mean functions and auto-covariance and cross-covariance functions (surfaces).

(A3) The bivariate kernel function $K(\cdot,\cdot)$ is assumed to be a product kernel of univariate kernel $K(\cdot)$, i.e, $K(\cdot,\cdot) = K(\cdot)K(\cdot)$. The univariate kernel $K(\cdot)$ is assumed to be a symmetric probability density function with support $[-1, 1]$ such that $0 < \int K(u)u^2 du < \infty$. The boundary kernels to are used.

Let b_s and b_d be the bandwidths used for estimating the mean functions of the sparse and irregular processes Y and X_2 and dense and regular processes X_1 respectively. Let b_{C_s} and b_{C_d} be the bandwidths used for estimating the auto-covariance

and cross-covariance of the process with sparse and irregular processes involved in, i.e. Y and X_2 and auto-covariance of dense and regular process X_1.

(A4) As the sample size $n \to \infty$ and $m_1 \to \infty$, we assume $b_s \to 0$, $nb_s^4 \to \infty$, $nb_s^6 < \infty$; $b_d = b_d(n, m_1) \to 0$, $cm_1^{-1/3} \le b_d \ll n^{-1/4}$; $b_{C_s} \to 0$, $nb_{C_s}^6 \to \infty$, $nb_{C_s}^8 < \infty$; $b_{C_d} = b_{C_d}(n, m_1) \to 0$, $cm_1^{-1/3} \le b_{C_d} \ll n^{-1/4}$ where c is a positive number.

(A5) Assume the fourth moments of X_2 and Y are finite. Assume the mean and auto-covariance functions of X_1 are twice differentiable on [0, 1] and $[0, 1]^2$.

We give the assumptions (B) which are only needed for Theorem 2.

(B1) The number of eigenfunctions $L = L(n)$ in the KL expansion, which depends on the sample size n, satisfies the rate conditions given in assumption (B5) of Yao et al. [20].

(B2) The FPC scores ξ and measurement errors ϵ in predictors observations are independent of each and are Gaussian.

(B3) The number, location, and values of measurements for a given subject remain unaltered as the sample size $n \to \infty$.

Appendix C: Proofs of Theorem 1 and Theorem 2

Proof (of Theorem 1) Uniform consistency of $\hat{C}_{X_1}(s,u)$ is given in Theorem 4 of [2], uniform consistency of $\hat{C}_{X_1,X_2}$, $\hat{C}_{X_2,X_1}$, $\hat{C}_{X_1,Y}$, $\hat{C}_{X_2,Y}$ is given in Lemma 1 of Yao et al. [20], uniform consistency of $\hat{C}_{X_2}(s,u)$ is given in Theorem 1 of Yao et al. [19]. Then the uniform consistency of $\hat{\mathbf{C}}_{11}(t)$, $\hat{\mathbf{C}}_{12}(t)$, $\hat{\mathbf{C}}_{21}(t)$, $\hat{\mathbf{C}}_{22}(t)$, $\hat{\mathbf{C}}_{1Y}(t)$, $\hat{\mathbf{C}}_{2Y}(t)$ can be obtained. Therefore the uniform consistency of $\hat{\mathbf{b}}_1(t)$ and $\hat{\mathbf{b}}_2(t)$ follows and thus that of $\hat{\beta}_1(s,t)$ and $\hat{\beta}_2(s,t)$ can be obtained. □

Proof (of Theorem 2) For fixed L, we have

$$
\begin{aligned}
&|\hat{Y}_L^*(t) - \tilde{Y}^*(t)| \\
&\le |\hat{Y}_L^*(t) - \tilde{Y}_L^*(t)| + |\tilde{Y}_L^*(t) - \tilde{Y}^*(t)| \\
&\le \left| \int_{\delta_{11}}^{\delta_{12}} \hat{\beta}_1(s,t)\hat{X}_1^*(t-s)ds - \int_{\delta_{11}}^{\delta_{12}} \beta_1(s,t)X_1^*(t-s)ds \right| \\
&\quad + \left| \int_{\delta_{21}}^{\delta_{22}} \hat{\beta}_2(s,t) \sum_{l=1}^{L} \hat{\xi}_l^* \hat{\phi}_l(t-s)ds - \int_{\delta_{21}}^{\delta_{22}} \beta_2(s,t) \sum_{l=1}^{L} \tilde{\xi}_l^* \phi_l(t-s)ds \right| \\
&\quad + \left| \int_{\delta_{21}}^{\delta_{22}} \beta_2(s,t) \sum_{l=1}^{L} \tilde{\xi}_l^* \phi_l(t-s)ds - \int_{\delta_{21}}^{\delta_{22}} \beta_2(s,t) \sum_{l=1}^{\infty} \tilde{\xi}_l^* \phi_l(t-s)ds \right| \\
&= I_1 + I_2 + I_3
\end{aligned}
$$

For I_1, from the uniform consistency of $\hat{\beta}_1(s,t)$ established in Theorem 1 and the uniform consistency of kernel smoother, we have $I_1 \to 0$ as $n \to \infty$.

For I_2, from the uniform consistency of $\hat{\beta}_2(s,t)$ established in Theorem 1, the uniform consistency of $\hat{\xi}_l^*$ for $\tilde{\xi}_l^*$ from Theorem 3 in Yao et al. [19], and the uniform consistency of $\hat{\phi}_l$ from Theorem 2 in Yao et al. [19], we have $I_2 \to 0$ as $n \to \infty$.

For I_3, following Lemma A.3 in Yao et al. [19], we have $I_3 \to 0$ as $n \to \infty$.

Therefore, Theorem 2 follows.

Appendix D: Figures and Tables

See Figs. 6, 7, 8 and 9

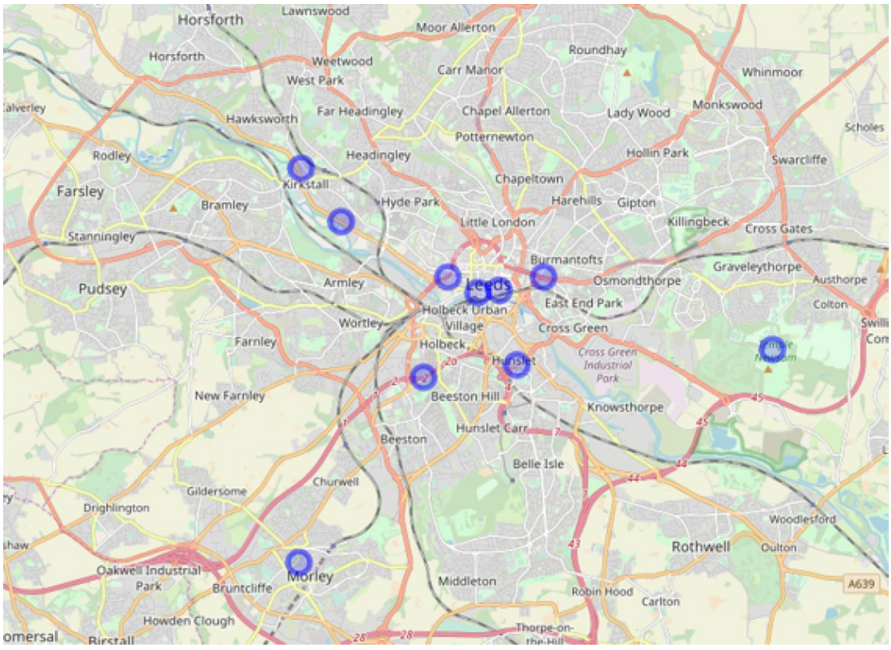

Fig. 6 Map of automatic monitoring points across Leeds

	A	B	C	D	E	F	G	H	I	J
1	Site Name	X	Y	0	1	2	3	4	5	6
2	Corn Exchange	53.79623	-1.54061							
3	Haslewood Close	53.79868	-1.52677	29.48958	44.26042	30.47312	45.25	62.375	31.73958	33.32292
4	Jack Lane, Hunslet	53.78262	-1.5351	26.69792	43.55405	32.17708	45.8125	87.57292	47.11458	52.21875
5	Kirkstall Rd	53.80869	-1.58924	15.0625	31.5625	10.51042	27.73958	49.625	25.01042	41.23958
6	Temple Newsam	53.78553	-1.45601	11.20213	20.62105	9.159574	18.87368	34.31579	9.368421	11.65263
7	Tilbury Terrace	53.78045	-1.56397	21.14894	31.12766	15.45745	30.02128	60.71277	23.16129	46.82979
8	Bishopgate Street	53.79578	-1.54708	51.22917	61.32292	49.76842	60.69792	86.3871	63.01042	59.15625
9	Abbey Street	53.81768	-1.60223	24.45833	35.39583	17	24.32258	61.95833	42.09375	47.16842

Fig. 7 Raw pollution data for different monitoring points around Leeds

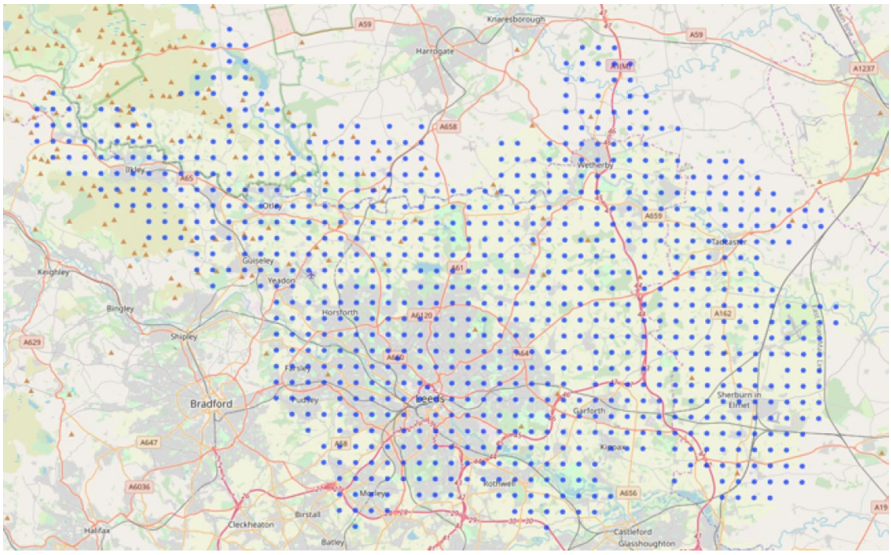

Fig. 8 861 point interpolation grid of Leeds

	A	B	C	D
1		district	mean NO2 Âµg m-3	standard deviation
2	0	LS21	40.1199444	12.21198349
3	1	LS22	38.5843241	11.813595
4	2	HG3	40.20073581	12.12746236
5	3	HG5	39.07152924	11.87532649
6	4	LS29	40.01509971	12.21634599
7	5	LS17	40.43790384	12.05346486

Fig. 9 Pollution by postcode district with standard deviation

Acknowledgements We would like to thank the referees for very useful constructive remarks. We thank the a local authority for providing physical activity data. We gratefully acknowledge the services of the Research Data and Informatics Team at Leeds Teaching Hospitals NHS Trust, in particular Atif Rabani, for extracting routinely-collected clinical data (COPD hospital admissions) and de-identifying it for use in this research. The research leading to these results has received funding from the Alan Turing Institute, the European Union's Seventh Framework Programme (FP7-Health-F5-2012) under grant agreement number 305280 (MIMOmics), and the Yujie Talent Project of North China University of Technology No. 107051360023XN075-04.

References

1. Bartal M (2005) COPD and tobacco smoke. Monaldi Arch Chest Dis 63(4)
2. Beran J, Liu H (2014) On estimation of mean and covariance functions in repeated time series with long-memory errors. Lithuanian Math J 54(1):8–34
3. Brockmann M, Gasser T, Herrmann E (1993) Locally adaptive bandwidth choice for kernel regression estimators. J Am Stat Assoc 88(424):1302–1309
4. Fan J, Gijbels I (1996) Local Polynomial Model Appl. CRC Press, Boca Raton
5. Harezlak J, Coull BA, Laird NM, Magari SR, Christiani DC (2007) Penalized solutions to functional regression problems. Comput Stat Data Anal 51(10):4911–4925
6. Herrmann E (1997) Local bandwidth choice in kernel regression estimation. J Comput Graph Stat 6:35–54
7. Horvath L, Kokoszka P (2012) Inference for functional data with applications. Springer, Berlin
8. Kim K, Sentürk D, Li R (2011) Recent history functional linear models for sparse longitudinal data. J Stat Plan Inference 141(4):1554–1566
9. Liu H, Houwing-Duistermaat J (2022) Fast estimators for the mean function for functional data with detection limits. Stat 11(1):e467
10. Liu H, Del Galdo F, Houwing-Duistermaat J (2018) Prediction and forecasting models based on patient's history and biomarkers with application to Scleroderma disease. arXiv:1811.04290
11. Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis. Academic Press, Cambridge
12. Malfait N, Ramsay JO (2003) The historical functional linear model. Canadian J Stat 31(2):115–128
13. Mannino DM, Buist AS (2007) Global burden of COPD: risk factors, prevalence, and future trends. The Lancet 370(9589):765–773
14. Peng J, Paul D (2009) A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. J Comput Graph Stat 18(4):995–1015
15. Peacock JL, Anderson HR, Bremner SA, Marston L, Seemungal TA, Strachan DP, Wedzicha JA (2011) 333Outdoor air pollution and respiratory health in patients with COPD. Thorax 66(7):591–596
16. Pomann GM, Staicu AM, Lobaton EJ, Mejia AF, Dewey BE, Reich DS, Shinohara RT (2016) A lag functional linear model for prediction of magnetization transfer ratio in multiple sclerosis lesions. Annals Appl Stat 10(4):2325–2348
17. Ramsay JO, Dalzell CJ (1991) Some tools for functional data analysis. J Royal Stat Soc 53(3):539–572
18. Ramsay JO, Silverman BW (2005) Functional data analysis. Springer Series in Statistics, 2nd edn. Springer New York, NY
19. Yao F, Müller HG, Wang JL (2005) Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100(470):577–590
20. Yao F, Müller HG, Wang JL (2005) Functional linear regression analysis for longitudinal data. Annals of Stat 33(6):2873–2903
21. Zhang Z, Wang J, Lu W (2018) Exposure to nitrogen dioxide and chronic obstructive pulmonary disease (COPD) in adults: a systematic review and meta-analysis. Environ Sci Pollut Res 25(15):15133–15145