MOX-Report No. 19/2019

Modelling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan.

Torti, A.; Pini, A.; Vantini, S.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Modelling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan

Agostino Torti[1] | **Alessia Pini**[2] | **Simone Vantini**[3]

[1]MOX- Department of Mathematics, Politecnico di Milano, Pizza Leonardo da Vinci 32, 20133, Milan, Italy.
E-mail: agostino.torti@polimi.it

[2]MOX- Department of Mathematics, Politecnico di Milano, Pizza Leonardo da Vinci 32, 20133, Milan, Italy.
E-mail: simone.vantini@polimi.it

[3]Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo A. Gemelli 1, 20123, Milan, Italy.
E-mail: alessia.pini@unicat.it

In today's world bike sharing systems are becoming increasingly common in all main cities around the world. To understand the spatio-temporal patterns of how people move by bike through the city of Milan, we apply functional data analysis to study the flows of a bike sharing mobility network. We introduce a complete pipeline to properly analyse and model functional data through a concurrent functional-on-functional model taking into account the effects of weather conditions and calendar on the bike flows.

KEYWORDS
Bike Sharing System, FDA, functional-on-functional model, Milan

Abbreviations: FDA, functional data analysis; BSS, bike sharing system; IWT, interval wise testing; NIL, Nucleo di identità locale.

1 | INTRODUCTION

In the last years, due to urbanization and globalization, the demand for transportation increased like never before (Martellato (2017)). The growth of urban population combined with the increase of traffic congestion, environmental pollution and fuel prices have driven urban developers and city councillors to experiment new sustainable mobility systems. To tackle these problems a paradigm shift in the field of mobility has been necessary, which has lead to a gradually establishing of a model based on sharing mobility. In particular, it is of particular interest the continuous growth of bike sharing systems (BSSs) in all major cities, which is replacing how people make short trips in highly urbanized areas (Fishman (2016)). In Italy bike sharing mobility is growing up very fast and only in 2017 the bike sharing usage has increased by 147%. Nowadays, in Italy, there are 265 cities with at least one BSS, for a total of 39500 available bikes. Milan, in particular, is the most advanced reality in the Italian sharing mobility and holds the 44% of all available bikes in the country, i.e. 17380.+ (Gentili et al. (2017)). The growth of this service has allowed more Italians to move by bike and has revealed how important it is to study this phenomenon in order to handle with mobility management and plan the city of the future.

The aim of this work is to use bike sharing data to study mobility in the city of Milan providing useful information to the municipality both for the management of urban mobility network and for the management of the bike sharing service. We are interested in understanding the global behaviour of the city and its spatio-temporal patterns: we would like to know how people move by bike, departure and arrival times and venues, studying the variability within and between days. We also aim to quantify how external factors, such as weather conditions (rain, temperature,wind...) or particular events, influence people's mobility behaviour. In the end, we aim to define a procedure able to model the future bike trips between the districts of the city knowing the values of the external factors.

There is extensive literature analysing and modelling BSSs in different cities around the world. To have a good idea of the evolution of works on BSSs, one can look at Fishman (2016). The spatio-temporal patterns are one of the most studied aspects (e.g., Froehlich et al. (2009); Borgnat et al. (2013); Zhou (2015); Faghih-Imani and Eluru (2016); Saberi et al. (2018)). Indeed, understanding patterns of usage is fundamental for fleet management of BSSs and it can give insights on the social behaviour of the users. Each one of the cited works analyses a BSS in a different city, respectively, Barcellona, Lione, Chicago, New York and London, but all of them find out similar results: differences in the usage between weekdays and weekends are identified; moreover, it is also revealed the presence of three main peaks of usage during weekdays, respectively, in the morning, at lunchtime and in the evening. The limit of all these works on BSSs is that time is treated as a discrete variable and the data are observed looking at their average behaviour in a time interval. In the reality, instead, a mobility datum is continuously dependent on time and a thus a natural way to analyse the nature of these data is to see them as continuous functions of time. For this reason, in order to develop our analyses, we make use of tools from Functional Data Analysis (FDA), the branch of statistics dealing with curves, surfaces or anything else varying over a continuum (e.g., Ramsay and Silverman (2005)). As far as we know, there are only few works which apply FDA to study BSSs (e.g., Han et al. (2018); Gervini and Khanal (2018)) or, more generally, to any mobility sharing system. In Han et al. (2018) a scalar-on-functional linear model is implemented to understand how the total daily number of rentals is affected by the bike sharing activity in previous days. In Gervini and Khanal (2018), instead, the daily demand distribution at every station in the city of Chicago is treated as a functional datum and a hierarchical clustering technique is applied to find common patterns. To our knowledge, instead, FDA has never been employed to the problem of analyzing and modelling flows in a mobility network.

Our analyses focus on BikeMi, the major and older BSS in the city of Milan. We model it as a weighted time-varying network, in which the districts of the city are seen as nodes, while bikes moving from one district to another one represent the weight of each edge (Borgnat et al. (2013)). For each day we define the corresponding functional flow

from district *A* to district *B* as a function representing the rate of bikes that are leaving from district *A* to go towards district *B* at time t. In conclusion, we model a BSS as a complex network with functional flows on its edges. Applying FDA tools to this framework, we introduce a new methodology to study not only a shared mobility system but also, more in general, any complex time evolving network whose edges can be modelled as functional data (e.g., telecommunications networks). There is an important observation that has to be made on the number of flows we are going to analyse: the bike stations are concentrated into 39 of the 88 NILs, the official districts in which Milan is divided, and this means 39^2 possible paths. So, in our work, we are going to analyse flow data on 1521 different paths. This fact urged us in developing an interactive interface to explore the results of the analyses. An example of the functional flows are illustrated in Figure 1 for two randomly chosen NILs in the city of Milan.

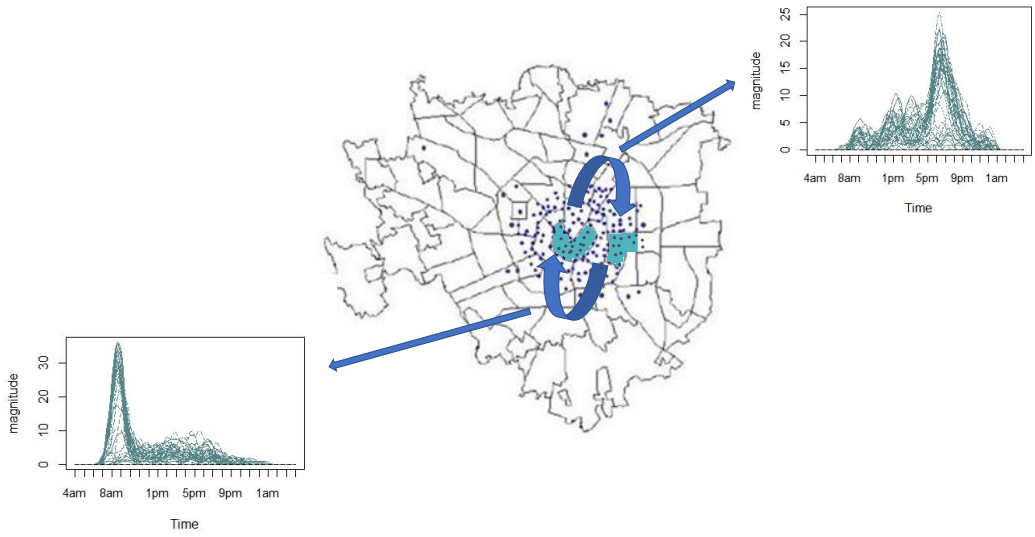


FIGURE 1 City of Milan is divided into its 88 NILs highlighting the position of all the 263 working bike stations (blue dots) in the analysed period, from the 25th of January to the 6th of March 2016. Moreover, two functional flows samples from Duomo (NIL 1) to 22 Marzo (NIL 26) and vice versa are displayed.

To develop our analyses, for each flow we built a concurrent functional-on-functional model (e.g., Ramsay and Silverman (2005); Kim et al. (2016)) taking into account the effects of weather conditions and calendar on the bike flows. A concurrent model is simultaneous, meaning that the functional response and the functional covariates are defined on the same domain and furthermore that the expected value of the functional response given the covariates is a function of the value that the covariates assume at the same point of the domain. Specifically, the bike flows constitute the functional response variables, while the weather conditions are the functional covariates and the type of the day is a scalar dummy covariate.

Of course this is a strong approximation of the reality in which the expected value of the response variable at each point of the domain is a function of the value that the covariates assume in all the domain. Nevertheless, we propose a modelling based on the concurrent model for at least three reasons: first, being plausible that weather conditions have only a short range effect, a concurrent model is a good approximation; second, using a concurrent model there is

no need for a penalization (no extra tuning parameters); third, a concurrent model is easier to be interpreted by our stakeholder.

The rest of this paper is organized as follows. In Section 2 we describe the concurret functional-on-functional model, discussing the methodology proposed for parameter estimation, inference, diagnostic and prediction. In Section 3 our analyses are shown step by step: we present the used datasets, we estimate the functional data through a smoothing procedure, we apply the developed methodology and we report the main results. In Section 4 a Shiny app to dynamically show the results for all the network is presented. In Section 5 conclusions are presented and discussed. In the end, in Appendix A and Appendix B we report some details related to the implemented procedure.

2 | METHODOLOGY

2.1 | Model Estimation and Inference

The Functional-on-Functional Linear Model

We introduce a concurrent functional-on-functional linear model with interaction. Despite the large number of works on functional models, the majority of them do not readily accommodate the existence of an interaction among covariates, even if some exceptions can be found in few works dealing with functional models with scalar responses (e.g., Li et al. (2010); Usset et al. (2016)). In our work we extend the literature related to this topic by adding a two-way scalar-functional interaction to a functional response model. Suppose to have a sample of n continuous squared-integrable random functions $y_i(t)$, s.t $y_i(t) \in L^2[a,b] \cap C^0[a,b]$, $\forall i \in \{1,...,n\}$ and $\forall t \in [a,b]$. We assume that $y_i(t)$ follow the concurrent functional-on-functional linear model

$$y_i(t) = \beta_0(t) + \sum_{k_s=1}^{K_s} \beta_{k_s \cdot}(t) x_{k_s i} + \sum_{k_f=1}^{K_f} \beta_{\cdot f_k}(t) z_{k_f i}(t) + \sum_{k_s=1=1}^{K_s} \sum_{k_f=1=1}^{K_f 1} \beta_{k_s k_f}(t) x_{k_s i} z_{k_f i}(t) + \epsilon_i(t), \tag{1}$$

where $\forall i \in \{1,...,n\}, t \in [a,b]$:

- $x_{k_s i} \in \Re$ for $k_s = 1,...,K_s$ are known scalar covariates;
- $z_{k_f i}(t) \in L^2[a,b] \cap C^0[a,b]$ for $k_f = 1,...,K_f$ are known functional covariates;
- $\beta_0(t)$ is the unknown fixed functional intercept;
- $\beta_{k_s \cdot}(t)$ and $\beta_{\cdot k_f}(t)$ for $k_s = 1,...,K_s$, $k_f = 1,...,K_f$ are the unknown fixed functional regression coefficients, respectively, for scalar and functional covariates;
- $\beta_{k_s k_f}(t)$ for $k_s = 1,...,K_s$, $k_f = 1,...,K_f$ are the unknown fixed functional regression coefficients for interaction terms;
- $\epsilon_i(t)$ are independent and identically distributed random functions with zero-mean and finite total variance, i.e. $E(\|\epsilon_i(t)\|^2_{L^2}) < \infty$.

The above model can also be expressed in the following more compact way:

$$y(t) = Z(t)\beta(t) + \epsilon(t),\ t \in [a,b].$$

In this notation $y(t)$ and $\epsilon(t)$ are, respectively, a n-vector of functional responses and of functional residuals, $\beta(t) =$

$(\beta_0(t), ..., \beta_K(t))'$ is a $(K+1)$-vector of functional regression coefficients with $K = K_s + K_f + K_s K_f$ and $Z(t)$ is a $nx(K+1)$ design matrix of ones, scalar covariates, functional covariates and interactions.

Model Estimation

Once the model has been formulated, we have to estimate its functional parameters: this step is handled by least squares estimation, as suggested, for instance, by Ramsay and Silverman (2005). The ordinary least squares (OLS) estimators of the functional coefficients $\beta_k(t)$, with $k = 0, ..., K$, is found by minimizing the sum over units of the squared L^2 distances between the observed functional dependent variables and those predicted by the linear model (1), hence minimizing $\sum_{i=1}^{n} \int_a^b \left[y_i(t) - \left\{ \beta_0(t) + \sum_{k_s=1}^{K_s} \beta_{k_s \cdot}(t) x_{k_s i} + \sum_{k_f=1}^{K_f} \beta_{\cdot f_k}(t) z_{k_f i}(t) + \sum_{k_s=1=1}^{K_s} \sum_{k_f=1=1}^{K_f 1} \beta_{k_s k_f}(t) x_{k_s i} z_{k_f i}(t) \right\} \right]^2 dt$. Because of the interchangeability of integration and summation, this boils down to a point-wise minimization of $\sum_{i=1}^{n} \left[y_i(t) - \left\{ \beta_0(t) + \sum_{k_s=1}^{K_s} \beta_{k_s \cdot}(t) x_{k_s i} + \sum_{k_f=1}^{K_f} \beta_{\cdot f_k}(t) z_{k_f i}(t) + \sum_{k_s=1=1}^{K_s} \sum_{k_f=1=1}^{K_f 1} \beta_{k_s k_f}(t) x_{k_s i} z_{k_f i}(t) \right\} \right]^2$ for each $t \in [a, b]$. Thus, the OLS estimate $\hat{\beta}(t) = (\hat{\beta}_0(t), ..., \hat{\beta}_K(t))'$, with $t \in [a, b]$, coincides with the OLS estimator of the corresponding ordinary linear regression model at point t.

Model Inference and Interval Wise Testing Procedure

In this subsection we describe the adopted procedure to perform valid tests of various hypotheses on the functional regression coefficients of model (1). We want to test if a covariate has a significant effect on the functional response providing also the related portions of the domain that are responsible for rejecting a null hypothesis (domain selection). More precisely, we do not simply want a infinite repetition of a point-wise test along all the domain, which would be meaningless to our purpose, but we want to globally asses the validity of our test for each interval of the domain. For all these reasons, we make use of the interval-wise testing (IWT) procedure, which has been introduced by Pini and Vantini (2017) in the framework of testing differences between two populations. This procedure relies on the definition of an adjusted p-value function, provided with a control of the interval-wise error rate, to select intervals of the domain where the null hypothesis is rejected. This method has been applied and extended both to multi-way functional analysis of variance (ANOVA) (e.g,. Pini et al. (2017)) and to a functional-on-scalar linear model (e.g., Abramowicz et al. (2018)). In our work we make a further step ahead using the same procedure to make inference on the functional regression coefficients of a concurrent functional-on-functional linear model.

In detail, let $C \in R^{(q \times (K+1))}$ be any real-valued full rank matrix, where q denotes the number of hypotheses on linear combinations of functional regression coefficients to be jointly tested, with $1 \leq q \leq K+1$, and let $\mathbf{c}_0(t) = (c_{01}(t), ..., c_{0q}(t))'$ be a vector of fixed functions in $L^2[a, b] \cap C^0[a, b]$ representing the value under the null hypothesis of the tested linear combination. Then, we can test hypotheses on one or more linear combinations of the functional coefficients as follows:

$$\begin{cases} H_{0,C} : C\beta(t) = \mathbf{c}_0(t) & \forall t \in [a, b] \\ H_{1,C} : C\beta(t) \neq \mathbf{c}_0(t) & \text{for at least one} \quad t \in [a, b]. \end{cases} \tag{2}$$

Choosing properly C and $\mathbf{c}_0(t)$, we can both develop a global test for our model, the analog of the F-test for the ordinary linear regression model, and a single test for the significance of each functional regression coefficient, the analog of the t-test. The IWT procedure produces an adjusted p-value function $\tilde{p}_c(t)$ which can be thresholded at level α to select the portions of the domain imputable for the rejection of the null hypothesis $H_{0,C}$ (domain selection). Moreover, the used IWT procedure globally asses the validity of our test along all the domain providing a (asymptotic) control of the Interval Wise Error Rate (IWER). This type of control implies that the probability of detecting false positive intervals is (asymptotically) controlled at level α (for more details and demonstrations see Abramowicz et al. (2018)). In Appendix A and Appendix B, a brief explanation of this procedure and some details on the implementation are included. The IWT

procedure, in addition, can be used for reducing the functional model (1) through a backward elimination in order to select only the significant covariates (Abramowicz et al. (2018)).

2.2 | Model Diagnostic

In this section we deal with the diagnostic of a concurrent functional-on-functional linear model. Although there is a large literature on scalar-on-scalar linear regression diagnostic assessing the assumptions of the model and detecting influential observation (e.g., Cook and Weisberg (1982)), little has been done on developing diagnostic measures for functional regression models. In the last years, a few diagnostic measures have been developed for regression models with functional responses (e.g., Shen and Xu (2007); Chiou and Müller (2007); Gao et al. (2015)). Both Shen and Xu (2007) and Chiou and Müller (2007) propose some diagnostic measures including residuals and defining a scalar single-case Cook's distance for linear models with functional responses. In Gao et al. (2015) one step ahead is done. Authors define both a global and local Cook's distance for detecting multiple curves in a functional-on-scalar linear model: the global Cook's distance returns a scalar value for each set of curves; the local Cook's distance, instead, is evaluated point-wise in the domain of the functional response variable. In our work we strengthen the literature of model diagnostic in the functional scenario by defining a procedure able to select influential observations in a concurrent functional-on-functional model. Firstly, we check the estimated functional residuals, $\hat{\epsilon}_i(t) \quad \forall i \in 1, \ldots, n \quad$ and $\quad \forall t \in [a, b]$. Secondly, we observe the functional errors under a leave one out cross validation analysis (LOOCV), $y_i(t) - \hat{y}_i(t) \quad \forall i \in 1, \ldots, n \quad$ and $\forall t \in [a, b]$, where $\hat{y}_i(t)$ denotes the predicted i-observation obtained by fitting the model without the i-observation in the data and predicting the i-observation as explained in Sect. 2.3 . Successively, similarly to the local Cook's distance introduced in Gao et al. (2015), we define a Cook's distance function, a curve which estimates the scalar Cook's distance - as introduced Cook (1979) - for each functional response variable in each point of the domain. This approach does not only allow to detect influential observations, but it is also able to explain why: in line with the idea of the domain selection, the introduced Cook's distance function is able to select the portions of the domain that bring to identify an observation as an influential one. In the end, in addition to the detection of influential data, we define a proper statistic, called DFBETA function, to study the difference in each coefficient estimation with and without an influential observation.

Cook's distance function

Cook's distance is a commonly used estimate of the influence of a data point in ordinary linear regression model. Given a functional-on-functional model of the type (1), we need to estimate the related Cook's distance functions for every i-observation along the whole domain. To evaluate these curves, we make a point-wise estimation using the ordinary scalar Cook's distance for each given t on the corresponding ordinary linear regression model.
Consequentially, $\forall i \in \{1, \ldots, n\}$ and $\forall t \in [a, b]$, we define:

$$D_i(t) = \frac{\sum_{j=1}^{n} (\hat{y}_j(t) - \hat{y}_{j(-i)}(t))^2}{(K+1)S^2(t)}, \quad \textit{with} \quad S^2(t) = \frac{\hat{\epsilon}^T(t)\hat{\epsilon}(t)}{K+1}, \quad \textit{and} \quad \hat{\epsilon}(t) = y(t) - \hat{y}(t). \tag{3}$$

At this point, we have to define when and where an observation is influential according to its Cook's distance function. In the ordinary regression analysis there are different opinions regarding which cut-off value to use for spotting highly influential points, but the simpler suggested operational guideline is $D_i > 1$. Transposing it to the functional case, we say that a i-observation is influential if there is at least one point of the domain in which $D_i(t) > 1$. Moreover, we say that the i-observation is influential in the interval $I \subseteq [a, b]$ of the domain if $D_i(t) > 1 \; \forall t \in I \subseteq [a, b]$.

DFBETA Function

Given a functional model of the type (1), in order to study the difference in each coefficient estimation with and without an observation, we define a measure of influence called DFBETA function. After having computed $\forall i \in \{1, ..., n\}$ the OLS estimators $\hat{\beta}(t) = (\hat{\beta}_0(t), ..., \hat{\beta}_K(t))'$ and $\hat{\beta}_{(-i)}(t) = (\hat{\beta}_{0(-i)}(t), ..., \hat{\beta}_{K(-i)}(t))'$, the $(K+1)$-vector of DFBETA functions for each observation is defined as follows:

$$\text{DFBETA}_i(t) = \hat{\beta}(t) - \hat{\beta}_{(-i)}(t) \quad \forall t \in [a, b], \forall i \in \{1, ..., n\}. \tag{4}$$

2.3 | Prediction of Functional Data

The prediction of functional data and the evaluation of its uncertainty is a open-question in literature. Some works propose to perform prediction based on functional time series (e.g., Hörmann and Kokoszka (2012); Canale and Vantini (2016)), others through functional linear models (e.g., Antoch et al. (2010)) and others using bootstrap (e.g., Goldsmith et al. (2013)). Another interesting open question is how to define the prediction bands for the estimated curves. Despite different works on both simultaneous and point-wise confidence bands for functional data (e.g., Goldsmith et al. (2013); Degras (2017); Chang et al. (2017)), in the context of regression models with functional responses, as far as we know, there are no works which try to compute functional simultaneous prediction bands for the predicted curves. In our work we do not cover this lack in literature, but we propose a first attempt to predict future observations with related point-wise prediction bands.

Once the final functional-on-functional model (1) has been estimated, it can be straightforwardly applied to give a prediction of the functional response in the following way: $\hat{y}_{new}(t) = z_{new}(t)\hat{\beta}(t)$, where $z_{new}(t)$ contains both scalar and functional new covariates. To obtain, instead, the prediction bands for the predicted curve, we use the empirical distribution function of the estimated residuals $\{\hat{\epsilon}_i(t)\}_{i=1,...,n}$. Since we are working with a n-sample of functional data, we define the point-wise functional empirical distribution function in the following way:

$$\hat{F}_t(s) = \frac{1}{n}\sum_{i=1}^{n} 1_{\hat{\epsilon}_i(t) \leq s}, \quad \textit{with} \quad t \in [a, b], \quad s \in \mathbb{R}. \tag{5}$$

To obtain the prediction bands, we calculate, for each t, the $\alpha/2$ and the $(1-\alpha/2)$ percentiles of the empirical distribution. Then, we add them to the the predicted value for each point t.

3 | DATA ANALYSIS

3.1 | Preprocessing

In this section we introduce the data analysed in our work to study mobility in the city of Milan. To achieve this aim we use two datasets: the first one related to BikeMi, the analysed BSS, and the second one containing weather information. BikeMi is a fixed-stations service, meaning that there are docking stations located through the city that subscribers have to use to pick up and drop off bikes. It started in Milan in November 2008 and during these years it has evolved considerably: at the beginning there were only 200 stations while nowadays there are more then 400 with almost 5000 available bikes. Our analyses cover a period of 42 days, from the 25$^{\text{th}}$ of January 2016 to the 6$^{\text{th}}$ of March 2016, in which there have been 350093 bike trips. The dataset consists of a log of all the rentals, with the station and the time of both

departure and arrival. In order to focus on the global behaviour of the city and subsequently to better visualize results, we decide to aggregate the closer stations. For the aggregation, we consider NILs, the 88 official districts in which the city of Milan is divided, and we assign each station to the corresponding NIL. We focus on flows between NILs on the basis that it is more interesting from a urban planning perspective. Nevertheless, an analysis of flows between stations would be straightforward. In the analysed period, there were 263 operating docking stations, which are concentrated only into 39 of the 88 NILs (Figure 1). So, in our work, we study flows on 39^2 possible paths, i.e. 1521. On each selected path, the route from NIL A to NIL B, we define a different function for each day of the dataset. More precisely, since in the analysed period the service was operative from 7:00am to 1:00am, we decide to redefine the day such that it begins at 4:00am and ends at 4:00am of the following day in such a way that the bike activity is null at the edges of the domain. At this point we are ready to compute the functional data. In order to turn the available raw data into smooth functions, we use a kernel density estimation smoothing method (e.g., Hastie et al. (2001)). This method, as well all the literature regarding smoothing techniques for functional data, has received much attention in last decades and has been well-established over the years. For each day, we define a functional flow from NIL A to NIL B as a function representing the rate of bikes per hour that are leaving from NIL A to go towards NIL B at time t along a single day. As a consequence, the obtained functional flows represent at each time t the hourly rate of departure at time t. However, since bike trips are of short duration and they generally take less than 20 minutes, the estimated functional flows can be used to extract information both on departure rate and arrival rate. In Figure 2 we report, as example, the obtained 42 functional flows from Duomo (NIL 1) to itself, which is the most travelled path in the period of interest. Moreover, in the plot the flows are coloured according to week and weekend days. It is immediate to observe a completely different behaviour in the two cases. With regard to the notation, we call $y_{ijk}(t)$ the function representing the bikes going from NIL j to NIL k on day i, with $i \in \{1, \ldots, n\}$ and $t \in (4am, 4am)$.

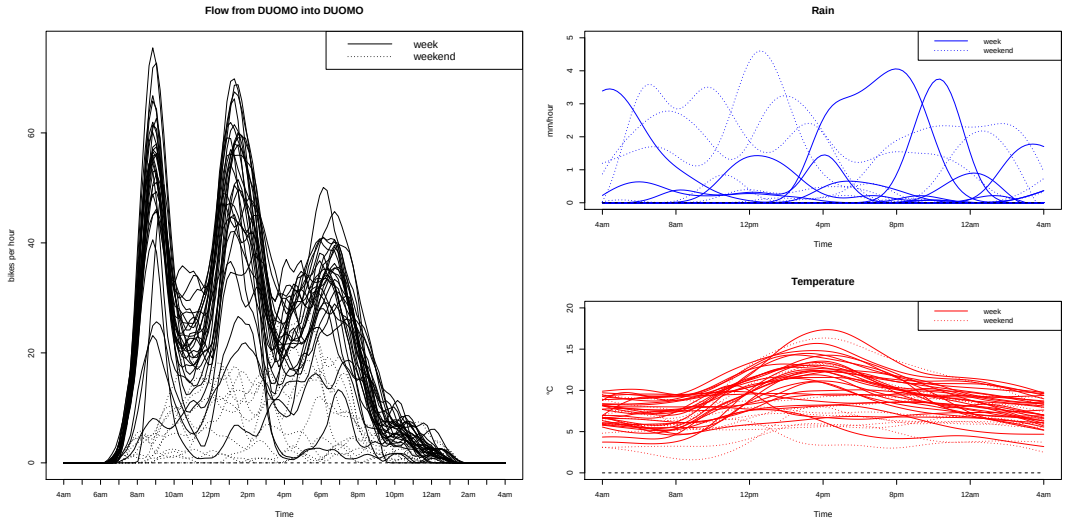


FIGURE 2 From left to right: functional flows from Duomo into Duomo, functional rain data and functional temperature data. In all figures weekdays are represented by a solid line while weekends by a dotted line.

Once the functional flows have been obtained, the step ahead is to aggregate them with external factors, so that we can understand their behaviour. For this reason we download weather data from ARPA Lombardia (www.arpalombardia.it)

related to rain and temperature. Since these data have generally a different shape within the same day, to directly compare the functional flows with weather information, we look at both rain and temperature as daily functions of time. Starting from the rain (mm/hour) and the temperature (C), for each day, we turn these raw data into smooth functions. The implemented method is the Local Weighted Polynomial Regression (e.g., Hastie et al. (2001)). Figure 2 shows the results of this procedure on both rain, which reveals the precipitations for each day, and temperature, which represents the temperature in degrees Celsius at each instant for each day of the analysed period. Moreover, like before, the functions are coloured according to week and weekend days.

3.2 | Applied Methodology

In this section we report the specific methods that we used to analyse and model the bike flows between districts of Milan. It is clearly infeasible to show the procedure and the related results on all the available paths of the BikeMi network, thus we decide to report here the general methodology and comment on a specific path, i.e. the flows from NIL 1 to itself, that, for simplicity, we will just call, in the reminder of this section, $y_i(t)$.

3.2.1 | Model Estimation and Inference

Initially, we tried to apply a concurrent functional-on-functional linear model of the form (1) to our data. However, since a linear model allows for negative values of the response variable and we want to fit and successively predict non negative functions, it turned out to be unsuitable to our purpose. This can be easily checked, for example, looking at the errors under a LOOCV analyses proposed in Sect. 2.2. To overpass this problem, we use a concurrent functional-on-functional log-linear model, which is able to preserve positivity of the analysed functional flows. The proposed approach is easy and make use of a simple transformation: we introduce $\widetilde{y}_i(t) = ln(y_i(t) + \delta)$, with δ close to zero but strictly positive, and we fit the model (1) using the transformed $\widetilde{y}_i(t)$ as response variables. Then, we can obtain information on the original flows through the exponential transformation $y_i(t) = e^{\widetilde{y}_i(t)} - \delta$. The parameter δ is necessary because the flows $y_i(t)$ can also be equal to zero and so $ln(y_i(t))$ can be undefined. To select the most appropriate δ, we perform a sensitivity analysis with $\delta \in (0, 2)$ and we show the results for $\delta = 1$.

In the initial full model we use six scalar covariates and two functional covariates. The scalar covariates are dummy variables indicating the days of the week. We use Monday as baseline and we define a different dummy variable for each day of the week from Tuesday until Sunday. The functional covariates are related to rain and temperature. We introduce for each day-i the function $z_{f_{rain}i}(t)$, which is equal to the amount of rain (mm/hour) through time, and the function $z_{f_{dtemp}i}(t)$, which is the difference through time with respect to the average daily temperature function of the period in degrees Celsius (C), $z_{f_{temp}i}(t) - \bar{z}_{f_{temp}i}(t)$. As in classical multiple regression, we start from an initial full model with all the available covariates and we apply a backward elimination procedure. Using the IWT procedure proposed in Sect. 2.1, we perform a t-test (2) on each regression coefficient. At each step we remove the covariate with the largest minimum value of the adjusted p-value function until only the significant coefficients remain in the model. Specifically, we say that a coefficient is significant if the adjusted p-value function of the related t-test is smaller than a threshold value α in at least one point of the domain. In our analyses we fix $\alpha = 0.05$ as significance level. Because of the high number of covariates, six scalar and two functional, we initially fit the model without interaction. With regards to the scalar covariates, two dummy variables related to Saturday and Sunday are significant. However, using properly test (2), we discover that their regression coefficients are not statistically different. Thus, we further simplify the model introducing a unique dummy covariate $x_{s_{we}i}$, which is equal to one if day-i is a Saturday or a Sunday, zero otherwise. Moving to the functional covariates, both rain and temperature are significant. At this point we add all the possible

scalar-functional interactions among covariates in the model and we test their significance. In the end, we obtain the following final reduced model:

$$log(y_i(t)+1) = \beta_0(t) + \beta_{we}(t)x_{s_{wei}} + \beta_{rain}(t)z_{f_{rain}i}(t) + \beta_{temp}(t)z_{f_{dtemp}i}(t) + \beta_{we\_rain}(t)x_{s_{wei}}z_{f_{rain}i}(t) + \epsilon_i(t), \quad (6)$$

with $i \in \{1, ..., n\}$ and $t \in (4am, 4am)$.

In the above final model there are three remaining covariates: the scalar covariate $x_{s_{wei}}$ due to the weekend-effect, the functional covariate $z_{f_{rain}i}(t)$ related to rain and the functional covariate $z_{f_{dtemp}i}(t)$ related to temperature. Moreover, it is quite interesting the presence of a scalar-functional interaction between $x_{s_{wei}}$ and $z_{f_{rain}i}(t)$.
In Figure 3 all the estimated functional regression coefficients, together with their adjusted p-value functions, are displayed. The blue rectangles indicate the portions of domain where the p-value is lower than $\alpha = 0.05$. Since we are working with a log-linear model, but we are interest in the absolute variability of $y(t)$, the effect of the obtained coefficients on $y(t)$ is not linear but it is multiplicative. Indeed, each change in a covariate linearly leads to a percentage change in the response variable $y(t)$. More precisely, given a log-linear model $ln(y) = \beta_0 + \beta_1 x + \epsilon$, then, the exact percentage change in y depending on x is $\%\Delta y = (e^{\Delta x \beta_1} - 1) \cdot 100$. The coefficient $\beta_0(t)$, which is significant in all the domain, represents the average behaviour during week ($x_{s_{wei}} = 0$) with no rain ($z_{f_{rain}i}(t) = 0$) and temperature equal to its average daily behavior ($z_{f_{dtemp}i}(t) = 0$). However, it gives information on $ln(y(t)+1)$, so, to be interpreted, we need to apply the exponential transformation. The weekend coefficient $\beta_{we}(t)$ is always negative and has a significant p-value in all the domain, meaning that during weekends the number of bike trips is smaller than during weekdays. This reduction is more evident in the morning where the difference between week and weekend is more marked: in this time interval $\beta_{we}(t) = -2.5$, meaning a reduction of almost 90%. The rain coefficient $\beta_{rain}(t)$ is negative and with a significant p-value in every moment of the day, meaning that an increase in the amount of precipitations (in terms of mm/h of rain) implies a percentage decrease in the number of trips. On weekdays this reduction is not constant during the day but it is more evident in the morning: here $\beta_{rain}(t) = -3$, thus just one extra millimeter of rain implies a decrease of almost 95% of the bike flows. This effect is counterbalanced during weekends by the interaction term $\beta_{we\_rain}(t)$ which has a significant p-value only in the morning around 10:00am. In this time interval $\beta_{we\_rain}(t)$ is positive and $\beta_{rain}(t) + \beta_{we\_rain}(t)$ is about -1, meaning that the rain has a minor influence during weekends: one millimeter of rain implies a decrease of just about 65%. Looking at the daily distributions of rain in the analysed period (Figure 2), it is plausible that this effect is due to the limited availability of rainy data and not to a different attitude of users according to the day of the week. Indeed, in the 42 analysed days only 15 are rainy days and in the morning it generally rains heavily during weekends. In general, we can say that rain has a massive impact on the number of bike trips all hours of the day: even a small amount of rain leads people not to move by bike and over a rain threshold people do not use bikes. So, to estimate more accurately the effect of rain on bike trips, we would need a larger dataset. The temperature coefficient $\beta_{temp}(t)$ is statistically significant only in the afternoon after 4:00pm: here it is about 0.08, meaning that one degree above (below) the average daily temperature at a given time of the day increases (decreases) bike flows by about 8%. This means that if the temperature is higher than the average ($z_{f_{dtemp}i}(t) \geq 0$), people are more motivated to travel by bike; if instead the temperature is lower than the average ($z_{f_{dtemp}i}(t) \leq 0$), people are more reluctant to pick-up bikes.

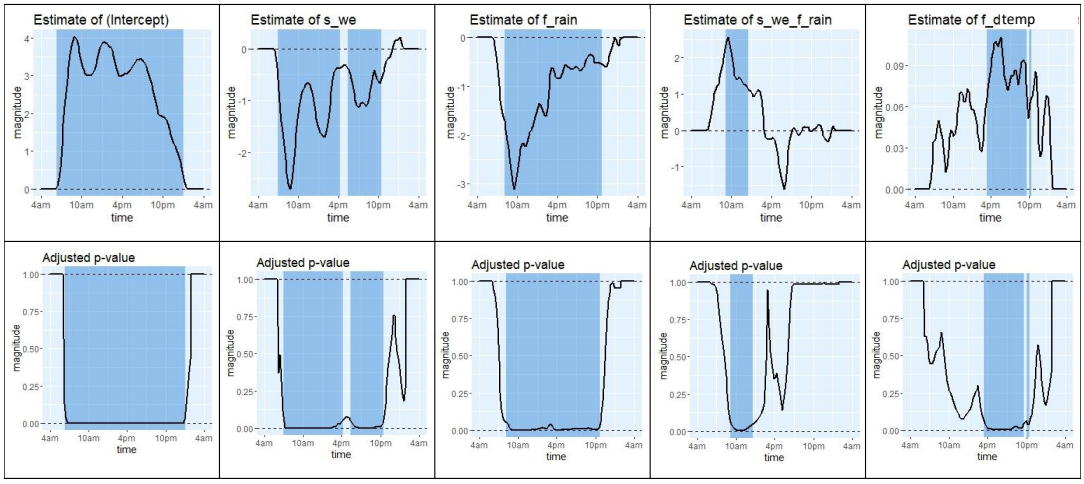


FIGURE 3 Functional regression coefficients of the reduced log model and their adjusted p-value functions. From left to right: $\beta_0(t)$, $\beta_{we}(t)$, $\beta_{rain}(t)$, $\beta_{we\_rain}(t)$ and $\beta_{temp}(t)$.

3.2.2 | Model Diagnostic

Regarding model diagnostic of the final reduced log-linear model, we firstly check the residuals and the errors under a LOOCV analysis. Not finding any strange behaviour, we secondly focus on the identification of possible influential observations. For this purpose, we use the Cook's distance function introduced in Sect. 2.2. However, since we are working with a model of the form $\widetilde{y}(t) = In(y(t)+1) = Z(t)\beta(t) + \epsilon(t)$, but we are interested in the variability of the original functional flows $y_i(t)$, the formula defined in (3) is not suitable. Indeed, we want to identify observations $y_i(t)$ which influence the estimate of $y(t)$ and not $In(y(t)+1)$. For this reason, we readjust (3) using the exponential transformation as follows:

$$\widetilde{D}_i(t) = \frac{\sum_{j=1}^{n}(\hat{y}_j(t) - \hat{y}_{j(-i)}(t))^2}{(K+1)S^2(t)}, \quad with \quad \hat{y}_i(t) = e^{\hat{\widetilde{y}}_i(t)} - 1, \quad S^2(t) = \frac{\hat{\epsilon}^T(t)\hat{\epsilon}(t)}{(K+1)} \quad and \quad \hat{\epsilon}(t) = y(t) - \hat{y}(t).$$

In Figure 4, the estimated readjusted Cook's distance functions for the final log model (6) and a dotted black horizontal line equal to one (the fixed cut-off value) are plotted together. We notice four influential observations: the 9th, the 16th, the 25th and the 27th of February. Three of them, the 09/02, the 16/02 and the 27/02, are rainy days, each one with a high level of rain in the time interval where the related Cook's distance function is bigger than one; the 25/02, instead, is a not rainy day. It is a Thursday and it has a high Cook's distance function in the morning. To better understand why these observations appear to be influential, we plot the DFBETA functions, obtained using (5), for each regression coefficient and highlighting the four discovered days (Figure 5). The three rainy days, as expected, affect the estimate of the two rainy coefficients $\beta_{rain}(t)$ and $\beta_{we\_rain}(t)$. This implies that rainy days have a significant impact in estimating the coefficients of the model and that our model is sensible to the loss of rainy observations. This can be due to the fact that

the number of rainy days in analysed dataset is very low. Focusing instead on the 25th of February, highlighted in light blue, it affects the estimate of the intercept around 8:00am, in the same time interval where its related Cook's distance is bigger than one. These results are due to the fact that this day has an atypical behaviour and that there are no bike trips before 8:00am (Figure 6).

Regarding the three rainy days, we decide not to remove them from the dataset, because necessary to estimate the effect of rain on bike trips. Indeed, our dataset has the problem not to have enough rainy days to estimate precisely their behaviour. Moving towards the 25th of February, instead, it seems to be due to an error in the registration of data, indeed, on this day, there are never bike trips before 8:00am. Because of this atypical behaviour, we decide to remove it from our data and we refit model (6).

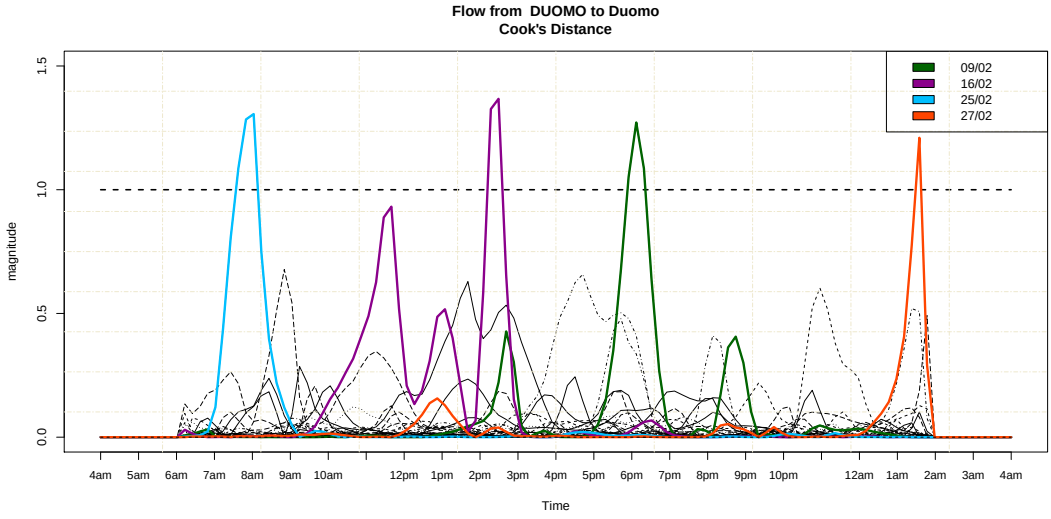
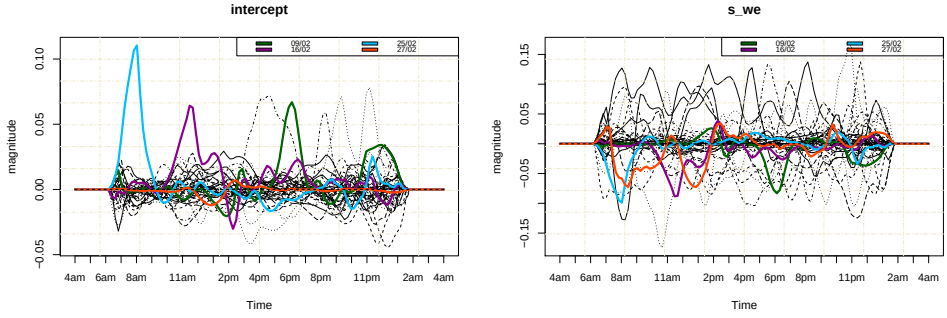


FIGURE 4 Readjusted Cook's distance functions of the log model on the path from Duomo into itself.



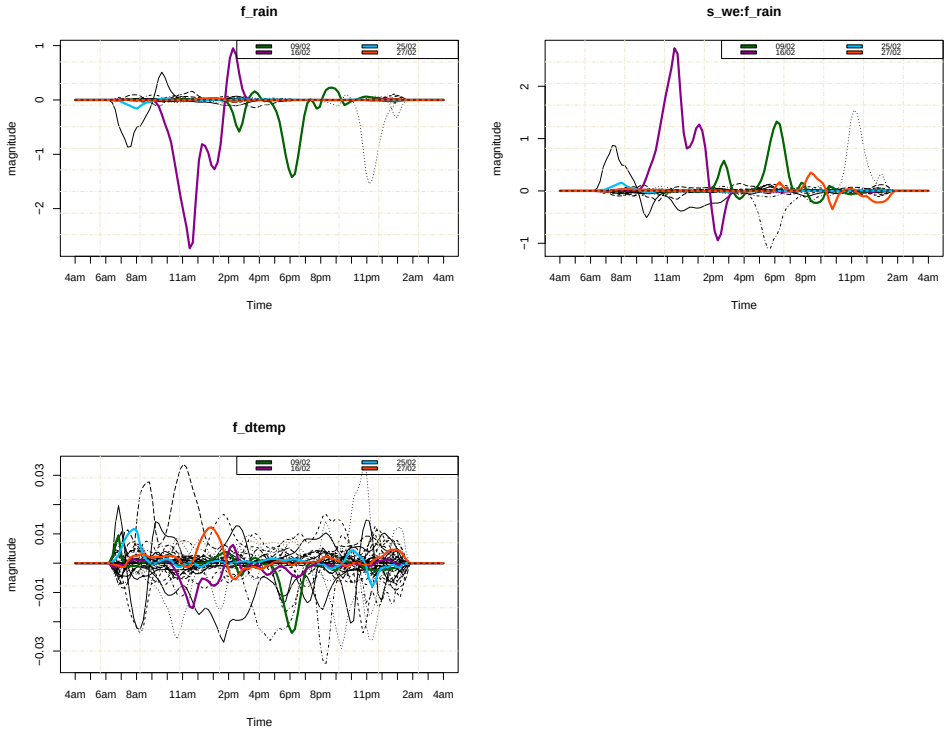


FIGURE 5 DFBETA functions of the log model on the path from Duomo into itself. From the top left to the bottom right: $\beta_0(t)$, $\beta_{we}(t)$, $\beta_{rain}(t)$, $\beta_{we\_rain}(t)$ and $\beta_{temp}(t)$.

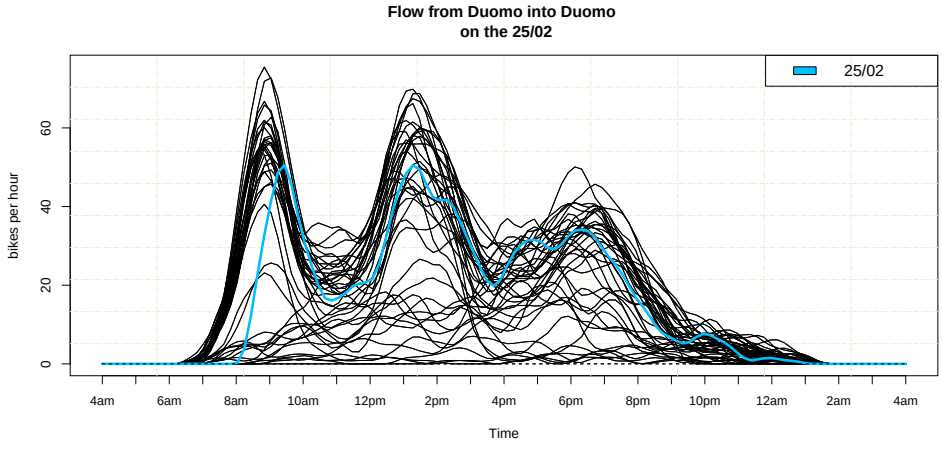


FIGURE 6 Flows from Duomo into Duomo highlighting the 25th of February.

3.2.3 | Prediction

Once the final model has been estimated and validated, it is ready to be used to predict future observations. In our case study it can be useful both for short-term predictions, using the reliable weather forecast of the following hours or days, and to investigate the network behavior in a business what-if scenario perspective.

We show the potential of our model trying to predict a hypothetical day on the path from Duomo to itself. Using the log model fitted without the 25th of February, we apply the methodology presented in Sect. 2.3 to predict the transformed log variable and its prediction bands. Then, we use the exponential transformation $\hat{y}_{new}(t) = e^{\hat{\tilde{y}}_{new}(t)} - 1$ to obtain our true curve with its related prediction bands. We predict a flow during a weekday ($x_{s_{wei}} = 0$) with some random values of rain and temperature (Figure 7). The weather conditions are displayed in comparison to the data of the analysed period: we observe that there is a little bit of rain in the morning (around one mm/hour) and that the temperature is under its average daily profile (around five C). The predicted curve is highlighted in orange and the related point-wise prediction bands at level 95% in light blue. Observing the results, we can notice the combined effect of both rain and temperature: the rain, which is present in the morning, implies that there are almost no bike trips in the same time interval; the temperature, which has a significant effect only after 4:00pm, instead, implies that the number of bike trips is less than usual.

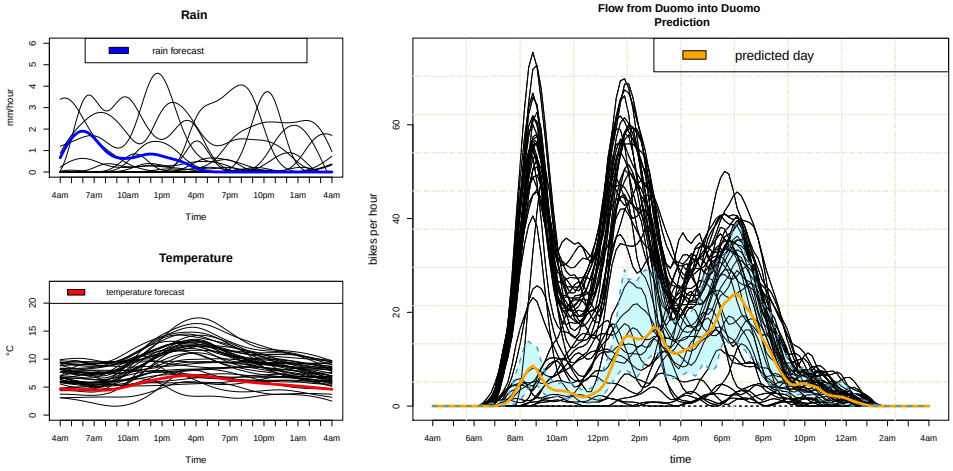


FIGURE 7 Left: selected value of rain (in blue) and temperature (in red) compared to the observed weather data (in black) in the period of interest. Right: predicted curve (in orange) with the point-wise prediction bands at level 95% (in light blue) compared to the observed flows (in black) in the period of interest.

3.3 | Results for the Entire Network

In this section we present some general results on the city of Milan obtained by applying the procedure shown in Sect. 3.2 on all the flows of the BikeMi network. Since the BikeMi service has been modelled like a network with 39 nodes, theoretically there are 39^2 available paths, i.e. 1521. So, to study all the flows between the districts of the city, we applied the log-linear model on each path of the network. In reality the network is very sparse in the sense that about 88% of the available paths have in average less than ten bikes per day. There are only 185 paths with at least ten bikes per day, but about 77% of all the trips in the analysed period is concentrated on these paths. For this reason

we applied our model only on the 185 paths having an average daily number of trips greater or equal to ten. For all the paths with a sufficient rate of bikes per day, we have been able to study both the effects of calendar and weather condition by looking at the functional coefficients of the final models.

Firstly, observing the coefficients related to the calendar dummy variables, we found out that the bike sharing activity has generally a well defined behavior depending on the day of the week and the selected path. The vast majority of the models found out statistical evidence to divide the days of the week into weekdays and weekends. Looking at the intercepts, which reveal the average behaviour during weekdays with no rain and temperature in the average, the bike sharing service seems mainly used by workers and during weekdays: in the morning people use to go in the city center, while in the late afternoon they seem to come back from where they leaved at the beginning of the day. This trend is easily observable by looking at the flows going into Duomo and leaving from it. On this purpose, in figure 8 we report the results for the 19 NILs nearer to Duomo, Duomo included, which involve almost the 35% of the overall trips on their own. We show both the average incoming (on the left) and outcoming (on the right) trips in Duomo estimated using the intercepts of the model on that paths: we observe an opposite behaviour in the two situations, indeed, the great number of people going into Duomo in the morning (8:00am-9:00am) is balanced from an equal number of people leaving from Duomo in the afternoon (6:00pm-8:00pm). The path with the highest number of trips in a day is the one from Duomo on itself (red curves), but the most outstanding connection happens to be between Magenta and Duomo (green curves). Observing always these paths, during weekends, instead, there is not anymore the presence of these evident peaks in the rush hours. Indeed, the coefficients related to the corrective weekend term, $\beta_{we}(t)$, have a similar shape to the respective intercepts but with an opposite sign. These weekend coefficients imply a reduction which is different according to the selected path and the time of the day, varying from no difference to a negative impact of the 90% on the most travelled paths. Looking at their adjusted p-value functions, we notice that they are significant primarily in the rush hours and some times at lunch break. Beyond this distinction between days during week and days during weekend, in some cases it has been observed a different behaviour on Friday. This happens, for example, on the path from Duomo to Magenta and from Duomo to Brera. On these two paths, the corrective regression coefficients related to Friday are significant around 8:00pm and they are negative, meaning that on Friday night people travelling by bike are less then during other weekdays. This effect is minimal, indeed, the coefficients in this time interval imply a reduction by about 4%.

Moving towards the effect of weather conditions, some conclusions can be drawn by looking at the coefficients of the functional log model applied on the different bike flows inside the city of Milan. In Sect. 3.2.1 we have already commented the effect of both rain and temperature on the flows from Duomo on itself, and looking at the other flows, as expected, the results are similar. With regards to the rain, it has a massive impact on the number of bike trips all hours of the day: even a small amount of rain leads people not to move by bike and over a rain threshold people do not use bikes. The interaction term *Rain-Weekend* appears to be significant only in the morning of the most travelled paths. In this time interval, as in the case of the flows from Duomo to itself, it is positive to counterbalance the effect of the rain on Saturdays and Sundays. Turning to the temperature, we find out that a temperature higher than its average implies an increase of the number of bike trips, while a colder temperature makes people more reluctant to travel by bike. This effect, however, is significant only on the most travelled paths around 4:00pm, and it affects the bike trips only in a small proportion (an increase/decrease of one degree changes the final results of almost 6%).

In the end, we point out some of the main results of the model diagnostic procedure applied on the different bike flows inside the city of Milan. In Sect. 3.2.1, estimating the Cook's distance function for the flows from Duomo on itself, we discovered four influential observations: three rainy days (the 09/02, the 16/02 and the 27/02) and a not rainy Thursday (the 25/02). Reapplying the same procedure on the other paths of the network, the same observations or just some of them are found out. Observing the DFBETA functions related to these days, it can be noticed that the first three days

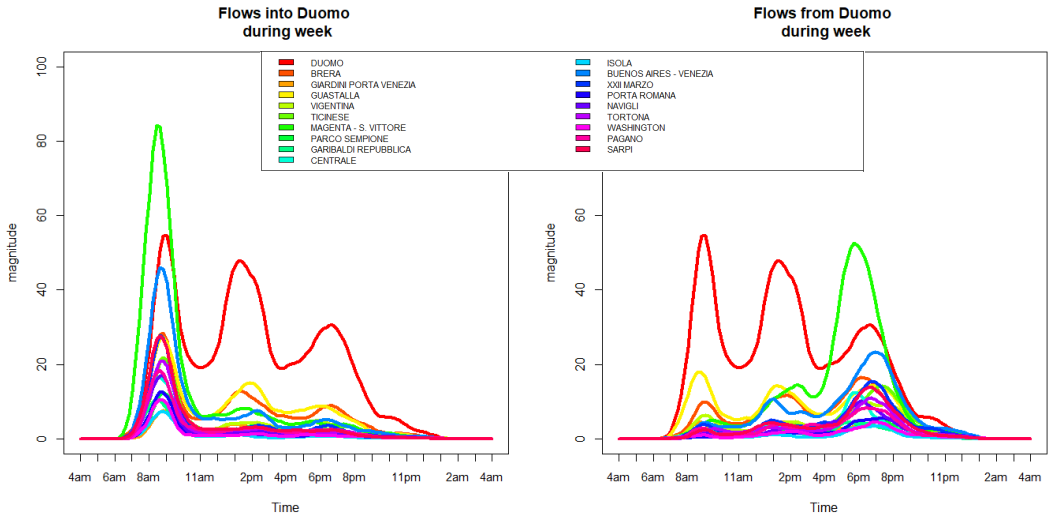


FIGURE 8 Left: average behavior of the flows going into Duomo during week with no rain and temperature in the average. Right: average behavior of the flows leaving from Duomo during week during week with no rain and temperature in the average. The 19 selected NILs are: Duomo (NIL 1), Brera (NIL 2), Giardini di Porta Venezia (NIL 3), Guastalla (NIL 4), Vigentina (NIL 5), Ticinese (NIL 6), Magenta S. Vittore (NIL 7), Parco Sempione (NIL 8), Garibaldi Repubblica (NIL 9), Centrale (NIL 10), Isola (NIL 11), Buenos-Aires Venezia (NIL 21), 22 Marzo (NIL 26), Porta Romana (NIL 27), Navigli (NIL 44), Tortona (NIL 50), Washington (NIL 51), Pagano (NIL 69) and Sarpi (NIL 69).

affect significantly the estimate of the two rainy coefficients, while the last day has a negative impact on the estimate of the intercept in the early morning. The conclusions taken on these days for all the paths of the network are the same done in the specif case of the path from Duomo on itself. We decide to not remove the rainy days from the dataset, because necessary to estimate the effect of rain on bike trips. We instead decide to remove the 25th of February from all our data. Indeed, on this day there are never bike trips before 8:00am and it seems to be due to an error in the registration of data.

4 | SHINY APP

Recalling the fact that we are studying the flows in a network made of 39 nodes (the NILs of Milan with at least one bike station), it is clearly infeasible to report all the results. To overcome this problem, we create an app by which it is possible to dynamically modify the model's inputs in order to choose the scenario of interest and to visualize the associated data and the subsequent results, as commented below. By doing this, we would like to both visualize the statistically significant spatio-temporal patterns of the users and to model each possible functional flow in the BikeMi network. Figure 9 shows how the app appears. Some green numbers and letters are added in the picture in order to explain how the app works. The app is composed from three main parts, identified by the numbers one, two and three.

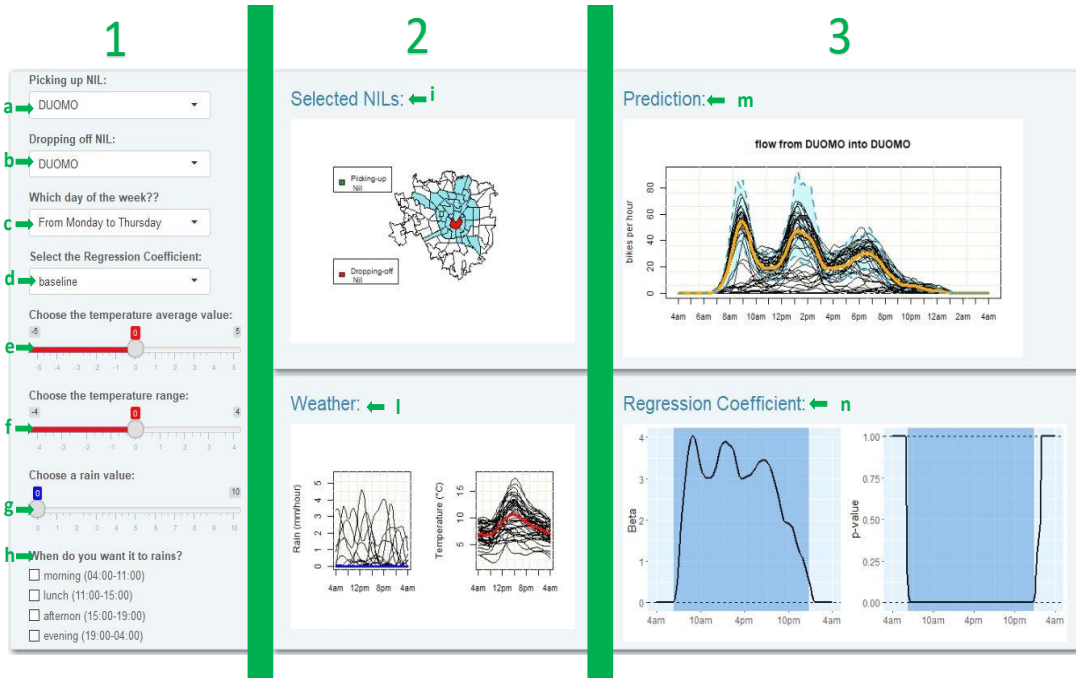


FIGURE 9 Shiny App.

1. On the left-hand side of the screen there is the "Input panel". This is the place where the user can select the scenario of interest thanks to seven drop-down boxes (labelled in the picture as "a", "b", "c", "d", "e", "f", "g" and "h"). Firstly, in "a" and "b", the user can choose a specific path selecting both the departure NIL (in "a") and the arrival NIL (in "b"). Then, it is possible to select the day of the week during which we want to predict the flow (in "c"). The possible options are: from Monday to Thursday, Friday or Weekend. In "e", "f", "g" and "h" it is possible to choose the weather conditions for the day we want to predict by setting both temperature (in "e" and "f") and rain (in "g" and "h"). The initial values of the temperature, which are zero both in "e" and "f", give exactly the average daily temperature function. Acting on "e" it is operated a vertical translation, while acting on "f" it is defined its range. The rain, instead, can be selected according to the "time" (in "h") and to the "strength" (in "g"). The "time" drop-down box allows to choose the intervals of the day during which it is raining. The day is split in four sections: the morning (4am-11am), the lunch-break (11am-13pm), the afternoon (15pm-19pm) and the evening (19pm-4am). Choosing the "strength" there are eleven possibilities which give a value between zero and 5mm/hour (the maximum rain value in the observed data). In the end, in "d", it is possible to select one functional regression coefficient of the model. The choice is among the followings: *Baseline* ($\beta_0(t)$), *Weekend* ($\beta_{we}(t)$), *Friday* ($\beta_{fri}(t)$), *Rain* ($\beta_{rain}(t)$), *Temperature* ($\beta_{dtemp}(t)$) and *Weekend-Rain* ($\beta_{we\_rain}(t)$).

2. In the central part of the screen it is possible to visualize the data related to the selected scenario: the two NILs (labelled in "i") and the weather conditions (labelled in "l"). Both the picking-up NIL and the dropping-off NIL are plotted on the map of Milan, respectively, in green and in red. All the remaining NILs in which the BikeMi service is operative are instead highlighted in light blue. The choosen rain and temperature, instead, are plotted in comparison to all the weather data of the analysed period.

3. On the right-hand side of the screen are shown the estimated results: the predicted flows (labelled in "m") ad the regression coefficients (labelled in "n"). The "Prediction panel" shows all the functional flows in our dataset on the selected path and highlights in orange the predicted curve. Moreover, the empirical prediction bands (with $\alpha = 5\%$) are added to the plot and the region that they define is coloured in light blue. The "Regression Coefficient panel" shows the selected functional regression coefficient and its adjusted p-value function. The blue rectangles indicate the portions of domain where the p-value is below the selected threshold $\alpha = 0.05$. Furthermore, if the coefficient is not significant for the model, $\alpha > 0.05$, it appears the sentence: "*the selected coefficient is not significant for the model*".

5 | CONCLUSIONS

The aim of our work was to use bike sharing data to study mobility in the city of Milan providing useful information both to the municipality and urban planners and the fleet managers. From this perspective, we have investigated the presence of spatio-temporal patterns with the purpose of understanding how people move by bike, departure and arrival and venues, studying the variability within and between days and taking into account the effects of weather conditions.

To build our analyses we have applied, for the first time, FDA to study the flows of a bike sharing mobility network. Moreover, we have developed a complete pipeline to properly analyse and model functional data through a concurrent functional-on-functional model. Focusing on models with functional responses, many topics - e.g., regarding model inference, model diagnostic and model prediction - should still be improved. In this framework our work has addressed some interesting issues on different aspects. Firstly, we have incorporated a two-way scalar-functional interaction in our model, that has allowed to explain better the variability of the response variable without worsen significantly the computational time needed and without overcomplicating the interpretation of the model. Secondly, we have applied for the first time the IWT procedure, as introduced by Pini and Vantini (2017), to test hypotheses on the functional regression coefficients of a functional-on-functional model. Then, and most importantly, we dealt with the diagnostic of a functional model in a innovative way. We have defined a Cook's distance function able not only to detect the influential observations, but also able to select the portions of the domain that bring to identify an observation as an influential one. According to this idea, we have also defined another statistic, called DFBETA function, able to highlight the portions of the domain where the removal of an observation affects the estimation of the functional regression coefficients.

From a practical point of view, the developed approach have been applied to study mobility in the city of Milan. The data used in our analyses are about BikeMi, the main and older BSS of Milan, in the period that goes from the 25^{th} of January 2016 to the 6^{th} of March 2016. Since we have studied the flows of a network of 39 nodes, meaning a total of 39^2 potential paths, i.e. 1521, we developed an app (created with Shiny, Studio (2014)) thanks to which it is possible to visualize the results for each path in an interactive way. This app allows us both to visualise the spatio-temporal patterns of the users and to model future bikes flows between the districts of the city. Regarding the spatio-temporal patterns, we have found out that BikeMi is mostly used by workers at a specific time of the day and on some specif paths. This information can be used by the municipality of Milan and by urban planners in improving the mobility management in different ways: first, designing new and safer bike paths on the most travelled routes; second, bringing other workers to prefer this mean of transport instead of their own motor vehicle in order to reduce traffic congestion and air pollution. Regarding the modelling of future bike flows, this tool can be useful to the fleet managers for at least two reasons: first, managing the short term rebalancing procedure by predicting future bike flows for the following hours or days; second, investigating the network behavior in a business what-if scenario perspective.

The main limitation of our work is that the available data cover a period of six weeks during winter. Moreover, among these days, only 15 of them are rainy days, and due to this fact, the effects of the rain are more difficult to estimate and the obtained results are more challenging to interpret. Despite all of this, the developed approach is ready to be used to analyse larger datasets and different shared mobility systems in other cities. As a matter of fact, it can be applied to any time evolving mobility network which can be modelled with functional data on its edges.

REFERENCES

Abramowicz, K., Häger, C. K., Pini, A., Schelin, L., Sjöstedt de Luna, S. and Vantini, S. (2018) Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics*.

Antoch, J., Prchal, L., Rosaria De Rosa, M. and Sarda, P. (2010) Electricity consumption prediction with functional linear regression using spline estimators. *Journal of Applied Statistics*, **37**, 2027–2041.

Borgnat, P., Robardet, C., Abry, P., Flandrin, P., Rouquier, J.-B. and Tremblay, N. (2013) A dynamical network view of lyon's vélo'v shared bicycle system. 267–284.

Canale, A. and Vantini, S. (2016) Constrained functional time series: applications to the italian gas market. *International Journal of Forecasting*, **32**, 1340–1351.

Chang, C., Lin, X. and Ogden, R. T. (2017) Simultaneous confidence bands for functional regression models. *Journal of Statistical Planning and Inference*, **188**, 67–81.

Chiou, J.-M. and Müller, H.-G. (2007) Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis*, **51**, 4849–4863.

Cook, R. D. (1979) Influential observations in linear regression. *Journal of the American Statistical Association*, **74**, 169–174.

Cook, R. D. and Weisberg, S. (1982) Residuals and influence in regression.

Degras, D. (2017) Simultaneous confidence bands for the mean of functional data. *Wiley Interdisciplinary Reviews: Computational Statistics*, **9**, e1397.

Faghih-Imani, A. and Eluru, N. (2016) Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of new york citibike system. *Journal of Transport Geography*, **54**, 218–227.

Fishman, E. (2016) Bikeshare: A review of recent literature. *Transport Reviews*, **36**, 92–113. URL: `http://dx.doi.org/10.1080/01441647.2015.1033036`.

Freedman, D. and Lane, D. (1983) A nonstochastic interpretation of reported significance levels. *Journal of Business Economic Statistics*, **1**, 292–98. URL: `https://EconPapers.repec.org/RePEc:bes:jnlbes:v:1:y:1983:i:4:p:292-98`.

Froehlich, J., Neumann, J., Oliver, N. et al. (2009) Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI*, vol. 9, 1420–1426.

Gao, Q., Ahn, M. and Zhu, H. (2015) Cook's distance measures for varying coefficient models with functional responses. *Technometrics*, **57**, 268–280.

Gentili, V., Milioni, D., Refrigeri, L., Rossi, G., Soprano, L. and Squitieri, F. (2017) *SECONDO RAPPORTO NAZIONALE SULLA SHARING MOBILITY*.

Gervini, D. and Khanal, M. (2018) Exploring patterns of demand in bike sharing systems via replicated point process models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Goldsmith, J., Greven, S. and Crainiceanu, C. (2013) Corrected confidence bands for functional data using principal components. *Biometrics*, **69**, 41–51.

Han, K., Müller, H.-G., Park, B. U. et al. (2018) Smooth backfitting for additive modeling with small errors-in-variables, with an application to additive functional regression for multiple predictor functions. *Bernoulli*, **24**, 1233–1265.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hörmann, S. and Kokoszka, P. (2012) Functional time series. **30**, 157–186.

Kim, J. S., Maity, A. and Staicu, A.-M. (2016) General functional concurrent model.

Li, Y., Wang, N. and Carroll, R. J. (2010) Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association*, **105**, 621–633.

Martellato, G. (2017) *Sharing mobility management: Fornire alle persone servizi di mobilità sostenibile in forma collaborativa*.

Pesarin, F. and Salmaso, L. (2010) *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons Inc.

Pini, A. and Vantini, S. (2017) Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, **29**, 407–424. URL: `http://dx.doi.org/10.1080/10485252.2017.1306627`.

Pini, A., Vantini, S., Grasso, M. and Colosimo, B. (2017) Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**, 55–81.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional data analysis*. Springer, New York.

Saberi, M., Ghamami, M., Gu, Y., Shojaei, M. H. S. and Fishman, E. (2018) Understanding the impacts of a public transit disruption on bicycle sharing mobility patterns: A case of tube strike in london. *Journal of Transport Geography*, **66**, 154–166.

Shen, Q. and Xu, H. (2007) Diagnostics for linear models with functional responses. *Technometrics*, **49**, 26–33.

Studio, R. (2014) Inc. shiny: Web application framework for r. *R package version 0.10. 0*.

Usset, J., Staicu, A.-M. and Maity, A. (2016) Interaction models for functional regression. *Computational statistics & data analysis*, **94**, 317–329.

www.arpalombardia.it () URL: `http://www.arpalombardia.it/Pages/ARPA_Home_Page.aspx`.

Zhou, X. (2015) Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in chicago. *PloS one*, **10**, e0137922.

ACKNOWLEDGEMENT

We are deeply grateful to Clear Channel for having shared with us the data used in our work.

A | IWT PROCEDURE

The applied IWT procedure is based on three steps presented in detail in Abramowicz et al. (2018). We briefly report them here:

1. **Interval-wise testing**
   Given any closed interval $I \subseteq [a, b]$, we want to test:

$$\begin{cases} H^I_{0,C} : C\beta(t) = \mathbf{c}_0(t) & \forall t \in I \\ H^I_{1,C} : C\beta(t) \neq \mathbf{c}_0(t) & \text{for some} \quad t \in I. \end{cases} \tag{7}$$

   To perform tests of linear hypotheses (7) we use the statics $T^I_c = \int_I T_c(t)dt$, where $T_c(t) = \left(C\hat{\beta}(t) - \mathbf{c}_0(t)\right)' \left(C\hat{\beta}(t) - \mathbf{c}_0(t)\right)$ and $\hat{\beta}(t)$ is the OLS estimator of the functional regression coefficients (see Sect. 2.2).

2. **Adjusted p-value functions**
   Let p^I_C denote the p-value of test (7) obtained using functional permutation tests based on the Freedman and Lane permutation scheme (Freedman and Lane (1983)). It is obtained by calculating the portion of permutations leading to a larger value of the test static than the test static on the observed data. In order to identify sub-domains where the null hypothesis of test (7) is rejected, we make use of an adjusted p-value function that is defined, for each point t, as the supremum p-value of all interval-wise tests on intervals containing t:

$$\tilde{p}_C(t) = \sup_{I \ni t} p^I_C, t \in [a, b]$$

3. **Domain selection**
   The intervals of the domain presenting a rejection of the null hypothesis (7) are obtained by thresholding the corresponding adjusted p-value functions at level α.

B | DETAILS ON THE IMPLEMENTATION

In this section we report the implemented procedure to obtain the adjusted adjusted p-value function. Since this function is defined $\forall t \in [a, b]$ as a supremum over uncountably many subintervals covering point t, it is practically infeasible to compute it. We therefore use a numerical approximation procedure, as already done in previous works applying IWT (e.g., Pini and Vantini (2017) or Abramowicz et al. (2018)). The following steps describe this procedure for testing linear hypotheses of the form (7) on the all domain $[a, b]$.

1. Partition of the domain into K equally sized atomic sub-intervals I_j of length $\delta t = (b - a)/K$ with $j = 1, ..., K$. Let S be the family (of size $K/(K+1)/2$) of all possible intervals $J \in S$ obtained as unions of the K atomic sub-intervals I_j.

2. For all intervals $J \in S$, approximate the value of the t-static T^J_c under the full model by a rectangle quadrature rule with step δt applied to the point-wise test static $T_c(t) = \left(C\hat{\beta}(t) - \mathbf{c}_0(t)\right)' \left(C\hat{\beta}(t) - \mathbf{c}_0(t)\right)$.

3. Estimate the p-value p_C^J of the tests on each $J \in S$ using the freedman and Lane permutation scheme (Freedman and Lane (1983)) with B $\in N$ random chosen permutations (conditional Monte Carlo algorithm, Pesarin and Salmaso (2010)).

4. Estimate the adjusted p-vale function $\tilde{p}_C(t)$ at point t as the maximum of the corresponding p-values of all intervals $J \in S$ containing t.

5. Select the significant intervals of the domain by tresholding at level α the obtained approximated adjusted p-value function.

This procedure relies on the two parameters K and B. It is import to point out that, as K tends to infinity, the error arising from the discretization of the domain becomes negligible. Moreover, increasing the number of random permutations B, the Monte Carlo error tends to zero. In our work we fix K=119, indeed, all our functional data are evaluated on a equally spread grid of 120 points. The choice of K is due to a balance between computation time needed and results. It has been shown that, by increasing the number of points of the grid, the obtained adjusted p-value functions do not change significantly, but the required computation time increases considerably. With regards to the number of permutations, instead, we fix B=1000, that is the most used parameter in literature. However, our results are robust even to a smaller B.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

18/2019 Delpopolo Carciopolo, L.; Cusini, M.; Formaggia, L.; Hajibeygi, H.
Algebraic dynamic multilevel method with local time-stepping (ADM-LTS) for sequentially coupled porous media flow simulation

17/2019 Antonietti,P.F.; De Ponti, J.; Formaggia, L.; Scotti, A.
Preconditioning techniques for the numerical solution of flow in fractured porous media

14/2019 Antonietti, P.F.; Facciolà, C; Verani, M.
Mixed-primal Discontinuous Galerkin approximation of flows in fractured porous media on polygonal and polyhedral grids

15/2019 Brandes Costa Barbosa, Y. A.; Perotto, S.
Hierarchically reduced models for the Stokes problem in patient-specific artery segments

16/2019 Antonietti, P.F.; Houston, P.; Pennesi, G.; Suli, E.
An agglomeration-based massively parallel non-overlapping additive Schwarz preconditioner for high-order discontinuous Galerkin methods on polytopic grids

12/2019 Capezza, C.; Lepore, A.; Menafoglio, A.; Palumbo, B.; Vantini, S.
Control charts for monitoring ship operating conditions and CO2 emissions based on scalar-on-function regression

11/2019 Benacchio, T.; Klein, R.
A semi-implicit compressible model for atmospheric flows with seamless access to soundproof and hydrostatic dynamics

10/2019 Abramowicz, K.; Pini, A.; Schelin, L.; Sjostedt de Luna, S.; Stamm, A.; Vantini, S.
Domain selection and family-wise error rate for functional data: a unified framework

09/2019 Antonietti, P.F.; Facciola', C.; Verani, M.
Unified analysis of Discontinuous Galerkin approximations of flows in fractured porous media on polygonal and polyhedral grids

13/2019 Manzoni, A; Quarteroni, A.; Salsa, S.
A saddle point approach to an optimal boundary control problem for steady Navier-Stokes equations