Statistical analysis of Jersey City bike sharing data using spatio-temporal models

Alessandro CHAAR Lorenzo LEONI Nicola ZAMBELLI

University of Bergamo, Department of Management, Information and Production Engineering

January 11, 2023

Abstract

Is it possible to predict how many bikes to place at a rental station to optimize the bike sharing service? Starting from 2020 Jersey City (NYC) bike staring data, the aim of this study is to find an answer to this question estimating three different spatio-temporal models. The historical weather data (and not only) will try to explain the daily (and hourly) number of pickups at a station in order to help the service provider in its planning.

Keywords: bike sharing, DCM, HDGM, f-HDGM, D-STEAM.

1 Datasets description

The original bike sharing dataset[1] comes from the website Kaggle and contains the bicycle rental information in 2020 of the company Citi Bike in Jersey City (figure 1(b)). Citi Bike is a privately owned public bicycle sharing system serving the New York City boroughs of the Bronx, Brooklyn, Manhattan, and Queens, as well as Jersey City. The weather dataset, instead, contains the most significant meteorological variables' time series about NYC in 2020 and comes from the historical weather data database that the website Visual Crossing[2] makes available for the users. Moreover, we have constructed two logical dummy variables to describe the weekends and events that have occurred in New York City in 2020, i.e. the US federal holidays and the lockdown due to the COVID-19 pandemic. Through a preliminary data processing work, we have extracted from the original datasets the variables of our interest by grouping data to obtain daily and hourly information. In the following paragraphs we have summarized some information about them.

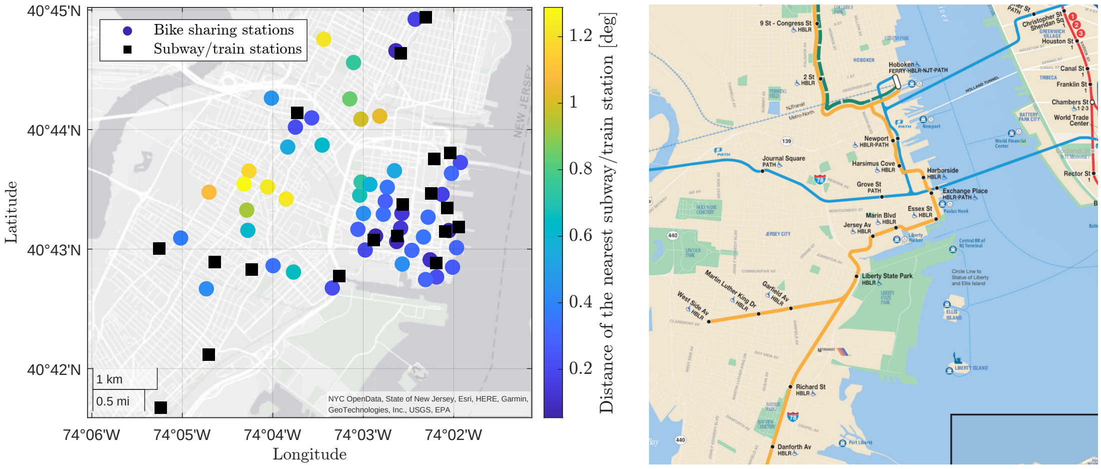
1.1 Bike sharing variables

This partition contains all the information relating to the bike rental:

- **pickups counter**: number of pickups at a rental station;
- **mean trip duration**: mean rental time of a user who picks-up a bike at a rental station;
- **mean users age**: mean age of a user who picks-up a bike at a rental station;

[1] it is possible to download it from this web page: `https://www.kaggle.com/datasets/vineethakkinapalli/citibike-bike-sharingnewyork-cityjan-to-apr-2021`.

[2] more information at this web site: `https://www.visualcrossing.com/weather-data`.



(a)

(b)

Figure 1: map of bike rental stations and of the nearest subway/train stations (a) and map of the public transport network in Jersey City (b); in particular the blue line indicates the path of the public transport service which reaches Manhattan, instead the yellow line those which extends over the entire Jersey area.

Bike sharing variable	*Unit*	*Min.*	*Max.*	*Mean*	*Median*	*Std*	*Skew.*	*Kurt.*
Mean pickups	-	0.06	58.27	18.04	17.32	10.24	0.88	4.35
Mean trip duration	min	6.60	387.21	26.76	21.30	28.92	7.08	75.13
Distance	km	0.05	1.29	0.49	0.35	0.36	0.89	2.65

Table 1: main statistics concerning bike sharing variables.

- **male counter**: number of males who pick-up a bike at a rental station;
- **female counter**: number of women who pick-up a bike at a rental station;
- **unknown gender counter**: number of people who have not specify their gender who pick-up a bike at a rental station;
- **subscribers counter**: number of customers who pick-up a bike at a rental station having an annual subscription;
- **customers counter**: number of customers who pick-up a bike at a rental station having a 4-hour pass or a 3-day pass;
- **distance**: indicates the distance between a rental station and the nearest train or subway station (figure 1(a)).

All these variables are both space-variant and time-variant, except for the distance that is a time-invariant covariate which we have obtained by carrying out an in-depth analysis of public transport network in Jersey City. In table 1 we have summarized some statistics regarding the bike sharing variables which we have decided to take into account in our analysis.

1.2 Weather variables

The weather frequency variables that Visual Crossing provides to users with daily or hourly frequency are:

Weather variable	*Unit*	*Min.*	*Max.*	*Mean*	*Median*	*Std*	*Skew.*	*Kurt.*	*R2*
Temperature	°C	-3.50	30.40	14.55	14	8.50	0.04	1.86	0.67
Feels-like temp.	°C	-7.70	33.30	13.81	13.90	9.80	0.04	1.96	0.66
Humidity	%	35.10	93.50	65.28	66.10	13.78	0.08	2.18	~ 0
Rainfall	mm	0	31.03	1.11	0	3.44	5.44	39.01	-0.06
Snowfall	cm	0	157.50	0.78	0	9.61	14.45	219.19	-0.05
Wind speed	km/h	9.90	47.80	20.64	19.90	6.89	0.98	3.90	-0.26
Cloud cover	%	0.10	100	39.45	37.15	29.79	0.35	1.95	-0.33
Visibility	km	6.50	16	15.29	16	1.54	-2.99	12.83	0.24
UV index	-	0	10	5.92	6	2.88	-0.21	1.84	0.46

Table 2: main statistics concerning weather variables. The R2 index describes the linear correlation between a meteorological covariate and the mean number of daily pickups.

- **mean temperature**;
- **mean feels-like temperature** (figure 2(a));
- **relative humidity** (figure 3(a));
- **rainfall** (figure 2(b));
- **snowfall**;
- **mean wind speed** (figure 3(b));
- **cloud cover**: percentage of covered sky (figure 3(c));
- **visibility**: maximum distance of visibility. If the field of vision is greater than 16 km, then the visibility value remains this one;
- **UV index**: indicates the energy of UV rays from 1 to 10 (figure 3(d)).

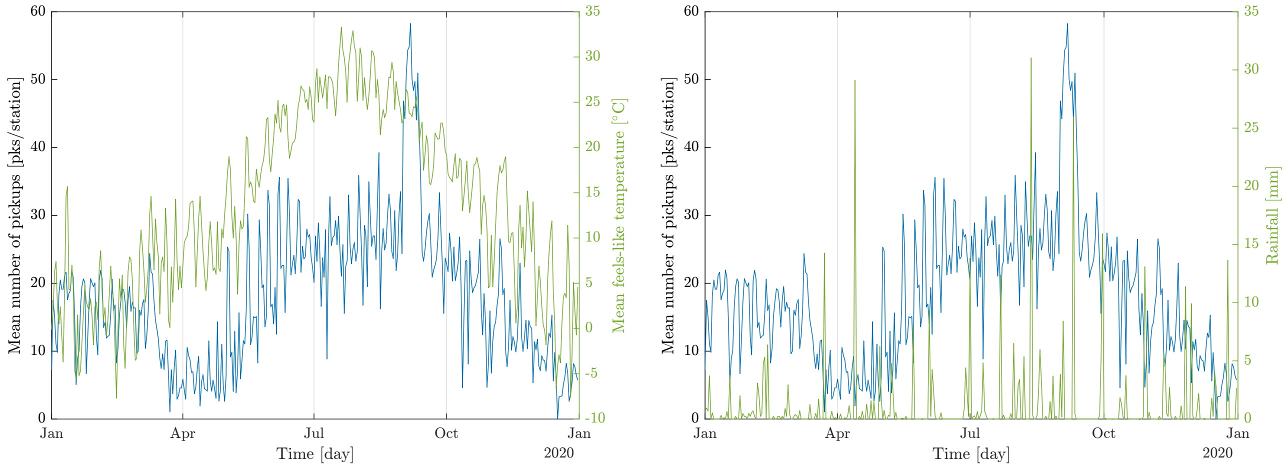
Weather variables are time-variant, but space-invariant because the 51 rental stations belong to the same geographical are. In table 2 we have summarized some of the most important statistics indicators to describe the distribution of each weather variable.

1.3 Dummy variables

We have hypothesized that not only weather may influence the demand of bicycle rentals, but also some events could provide their contribute, in particular the lockdown due to the COVID-19 pandemicand the US federal holidays. For this reason we have defined the following binary dummy variables:

- **lockdown**: this is equal to 1 if that day of 2020 there was the lockdown, 0 otherwise (figure 4(a));
- **weekends and holidays**: is equal to 1 if that day was Saturday, Sunday or a festive day, 0 otherwise (figure 4(b)).

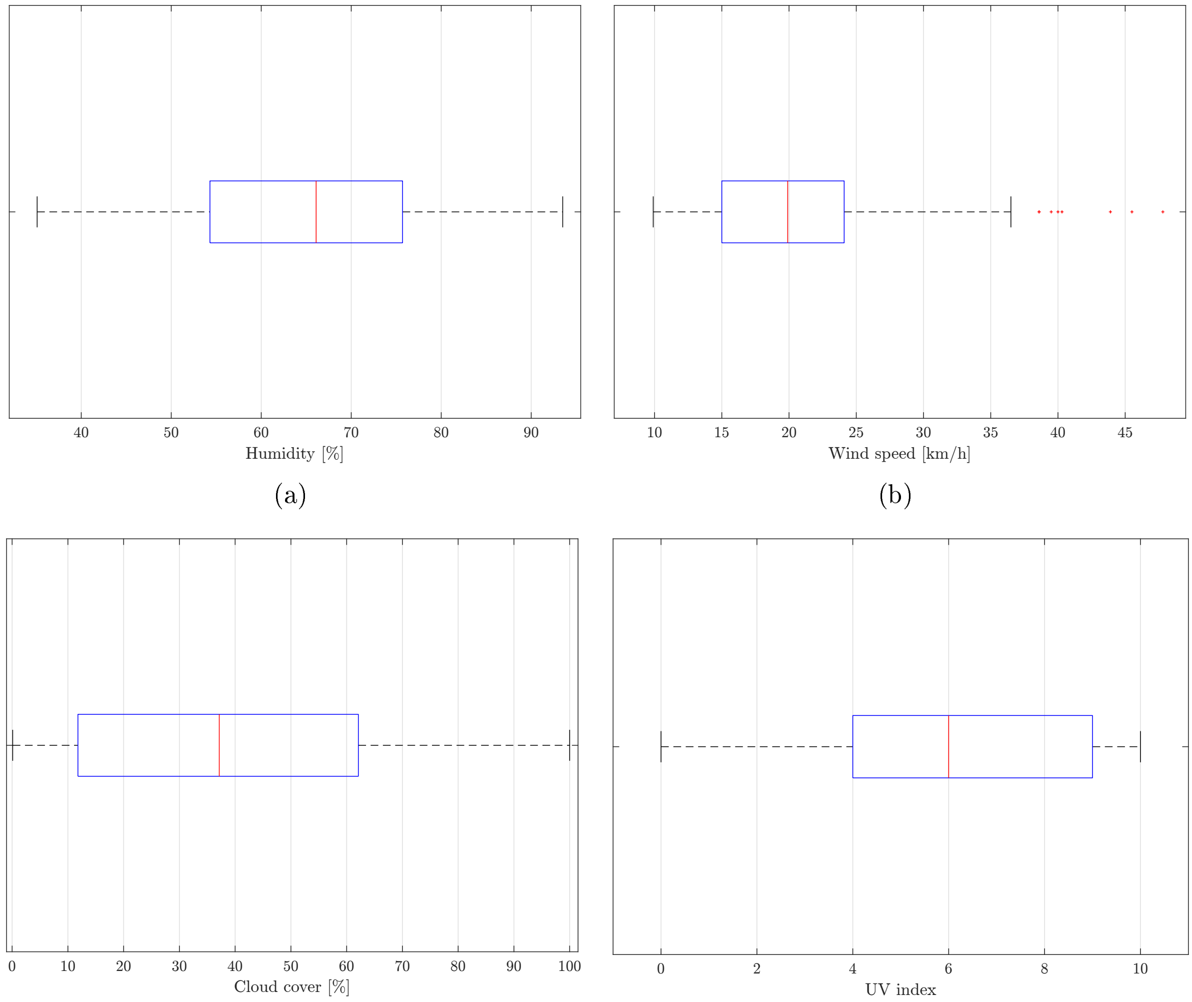
Dummy variables are time-variant, but space-invariant.



(a)

(b)

Figure 2: comparison between the trend of the mean number of daily pickups calculated on the entire bike sharing network and two weather variables, i.e. the mean daily feels-like temperature (a) and the daily rainfall (b). Through these plots we can see that temperature has a positive effects on rentals, instead rainfall a negative impact.



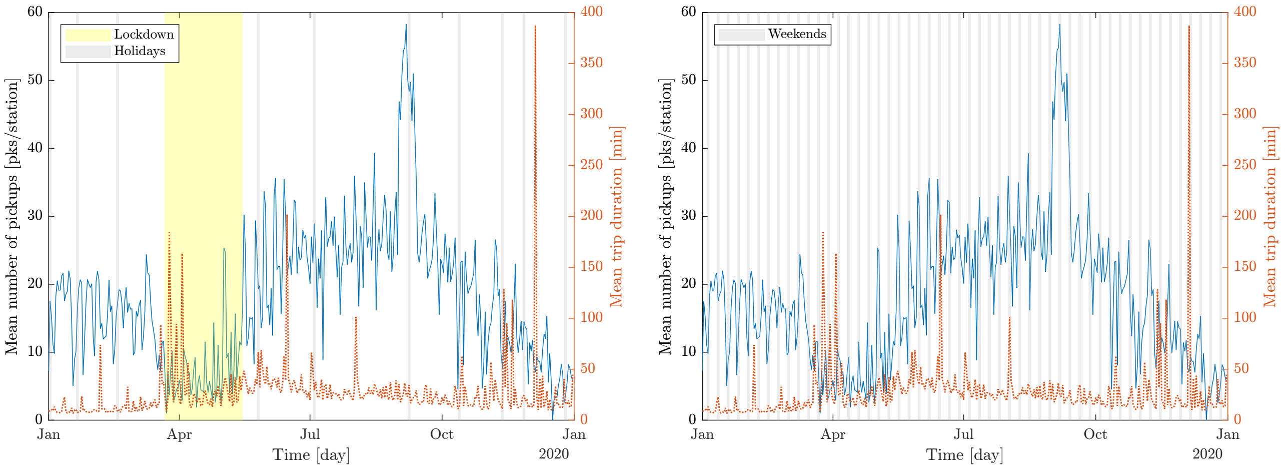
(a)

(b)

(c)

(d)

Figure 3: box-plots of some weather variables.



(a) (b)

Figure 4: in both graphs we can note as events influence the mean number of daily rentals. During the lockdown, in fact, there was not the possible to leave own home, instead in the weekends or during holidays a pickups' peak, either positive or negative, occurs.

2 Scientific questions

The aim of this study is to analyse the bike sharing phenomenon in Jersey City by using a statistical approach which involves the use of spatio-temporal models to describe the number of pickups and the mean trip duration (the observed variables) at each of 51 stations at a specific time instant. Specifically, the goals are the following:

- **inference**: how the weather variables, events and the distance between every bike sharing station and the nearest subway/train station (the covariates) can help us to describe the observed variables? Which are the most significant covariates?
- **models validation**: fixed some bike stations, which is the best spatio-temporal model among those we are taken into account in predicting the time series regarding the pickups and the mean trip duration of these test stations?

3 Methodology

To reach our aims, i.e. explaining the mean number of rentals (univariate case) and the mean trip duration (bivariate case) for each pickup station using the covariates previously described, we have decided to employ the following spatio-temporal models:

- the **Dynamic Coregionalization Model (DCM)** for daily data (Fassò et al. (2011));
- the **Hidden Dynamic Geostatistical Model (HDGM)** for daily data (Calculli et al. (2014));
- the **Functional Hidden Dynamic Geostatistical Model (f-HDGM)** for hourly data (Finazzi et al. (2021)).

After having selected the most significant covariates employing a backward approach based on the p-values of t-test (except for the f-HDGM) and estimated models through the EM algorithm, we have cross-validated them using 36 fixed pickup stations for training and 15 fixed rental stations for testing chosen casually, i.e. we have decided to use a 70-30 validation approach. In order to evaluate the performances of the estimated models we have taken into account three statistical indices:

- **RMSE$_h$** (and **MSE$_h$**) (only for f-HDG model): fixed an hour h, it expresses the mean hourly non-square validation error obtained iterating on days t and space points $\boldsymbol{s}$, i.e. on rental stations;
- **RMSE$_t$** (and **MSE$_t$**): fixed a day t, it describes the mean daily (hourly for f-HDGM) non-square validation error obtained iterating on space points $\boldsymbol{s}$ (and hour h for f-HDGM);
- **RMSE$_s$** (and **MSE$_s$**): fixed a space points $\boldsymbol{s}$, it expresses the mean daily (hourly for f-HDGM) non-square validation error obtained iterating on days t (and hour h for f-HDGM).

For doing each of these operations, i.e. model selection, estimation and validation, we have employed **D-STEM** v2, a software for MATLAB developed for modelling spatio-temporal data.

3.1 Description of the DCM

The Dynamic Coregionalization Model (DCM) is a hierarchical multivariate space-time model. For any site s $\in$ D $\sqsubset \mathbb{R}^2$ and any time t $\in \mathbb{N}$, let $y(\boldsymbol{s}, t)$ be a spatio-temporal process and its equation:

$$y(\boldsymbol{s}, t) = \boldsymbol{x}_\beta(\boldsymbol{s}, t)' \cdot \boldsymbol{\beta} + \boldsymbol{x}(\boldsymbol{s})' \cdot \boldsymbol{z}(t) + \alpha \cdot w(\boldsymbol{s}, t) + \epsilon(\boldsymbol{s}, t)$$
$$\boldsymbol{z}(t) = G \cdot \boldsymbol{z}(t-1) + \boldsymbol{\eta}(t)$$

with:

- $\boldsymbol{x}_\beta(\boldsymbol{s}, t)$ and $\boldsymbol{x}_z(\boldsymbol{s})$ are vectors of covariates. Note that $\boldsymbol{x}_z(\boldsymbol{s})$ is time invariant;
- $w(\boldsymbol{s}, t)$ $\sim$GP(0,ρ($\|\boldsymbol{s} - \boldsymbol{s}'\|$;$\boldsymbol{\theta}$) is correlated over space but IID over time;
- $\boldsymbol{z}(t)$ is $q \times 1$ dimensional with Markovian dynamics;
- G is a stable $q \times q$ transition matrix;
- $\boldsymbol{\eta}(t) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\eta)$ is the innovation with $\boldsymbol{\Sigma}_\eta$ the variance-covariance matrix;
- $\epsilon(\boldsymbol{s}, t) \sim N(0, \sigma^2_\epsilon)$ is the measurement error.

The model parameter set is Ψ={β, α, σ^2_ϵ $\boldsymbol{\theta}$, G, $\boldsymbol{\Sigma}_\epsilon$}.

3.2 Description of the HDGM

Given $y(\boldsymbol{s}, t)$ the response of a variable in position $\boldsymbol{s}$ at a discrete time t, where $\boldsymbol{s}$ belongs to a station of our geographic network, the univariate HDGM is defined according to the following formula:

$$y(\boldsymbol{s}, t) = \boldsymbol{x}_\beta(\boldsymbol{s}, t)' \cdot \boldsymbol{\beta} + a \cdot z(\boldsymbol{s}, t) + \epsilon(\boldsymbol{s}, t)$$
$$z(\boldsymbol{s}, t) = g \cdot z(\boldsymbol{s}, t-1) + \eta(\boldsymbol{s}, t)$$

Where $\boldsymbol{\beta}$ is a vector of fixed effects, a is a scale coefficient, while $\boldsymbol{x}_\beta(\boldsymbol{s}, t)'$ is the fixed effects design vector. The $\epsilon(\boldsymbol{s}, t)$ term is the univariate Gaussian measurement error independent in space and time, and $z(\boldsymbol{s}, t)$ is a latent random variable with Markovian dynamics ruled by g, the transition coefficient. The innovation $\eta(\boldsymbol{s}, t)$ is a sequence of unit variance Gaussian random fields independent in time, so is a $GP(0, \rho(\|\boldsymbol{s} - \boldsymbol{s'}\|; \boldsymbol{\theta}))$, where $\|\boldsymbol{s} - \boldsymbol{s'}\|$ is the Euclidean distance between $\boldsymbol{s}$ and $\boldsymbol{s'}$, the stations of our geographic network and $\rho(\|\boldsymbol{s} - \boldsymbol{s'}\|; \boldsymbol{\theta}))$ is a valid spatial

correlation function with range parameters $\boldsymbol{\theta}$.
The model parameter set is $\boldsymbol{\psi} = \{\boldsymbol{\beta}, a, \sigma_\epsilon^2, \boldsymbol{\theta}, g\}$, we will use the EM algorithm again to estimate $\boldsymbol{\psi}$. A benefit of the EM algorithm is that it is considered stable even when the number of parameters in $\boldsymbol{\psi}$ is not small, as in the analysis of the next chapter. The main reason is that every iteration of the EM algorithm have closed form expressions for a good part of the HDGM parameters, which improves numerical stability over Newton–Raphson and similar algorithms.
We will use the univariate HDGM to evaluate our two response variables, then we will move on to the bivariate case, evaluating both the average daily demand and the average daily rental duration. In the bivariate case the model becomes the following:

$$\boldsymbol{y}(\boldsymbol{s}, t) = \boldsymbol{X}_\beta(\boldsymbol{s}, t)' \cdot \boldsymbol{\beta} + \boldsymbol{z}(\boldsymbol{s}, t) + \boldsymbol{\epsilon}(\boldsymbol{s}, t)$$

$$\boldsymbol{z}(\boldsymbol{s}, t) = \boldsymbol{G} \cdot \boldsymbol{z}(\boldsymbol{s}, t-1) + \boldsymbol{\eta}(\boldsymbol{s}, t)$$

The structure is very similar to the univariate case, but it is important to note that the size of the terms changes. First, $\boldsymbol{y}(\boldsymbol{s}, t)$ is no longer a scalar but is a two dimensions vector, like $\boldsymbol{z}(\boldsymbol{s}, t)$. Similarly $\boldsymbol{X}_\beta(\boldsymbol{s}, t)'$ become the fixed effects design matrix. A significant difference is in $\boldsymbol{\eta}(\boldsymbol{s}, t)$ because now is a $GP(0, \boldsymbol{V} \cdot \rho(\|\boldsymbol{s} - \boldsymbol{s}'\|; \boldsymbol{\theta}))$, where V is a variance-covariance matrix of the two variable, including the scaling matrix $\boldsymbol{A}$.
The HDGM is estimated similarly to the DCM, but we only have the $\boldsymbol{z}(\boldsymbol{s}, t)$ latent variable which is estimated in the E-step by the Kalman smoother, while in DCM we use the formulas of the normal multivariate to estimate $\boldsymbol{w}(\boldsymbol{s}, t)$. Assuming that $\boldsymbol{y}(\boldsymbol{s}, t)$ is not observed at the spatial prediction locations, the Kalman smoother does also the spatial prediction. Fortunately, we have no missing values in our case study, but this function, implemented in D-STEM, will be useful in the validation phase.

3.3 Description of the f-HDGM

When we have to analyse data observed at high frequency (time domain) or resolution (spatial domain), the inferential approach on which DCM and HDGM are based may not be appropriate. In our case we have 51 stations and 1-year of data, i.e. 18.666 observations in daily data case and well 411.264 observations in hourly data case, a 2103 % increase which could cause a dramatic increase in computational time. In order to avoid this undesirable effect we have decide to resort to the *functional data analysis* (FDA): instead of estimating the time-series related to the number of hourly pickups at each rental station, a functional model estimates a linear combination of functions, common to all stations, to describe the daily trend. Given a spatial point $\boldsymbol{s}$, fixed a day t and an hour h, the values of the covariates and latent variables tell how to combine these functions to provide an estimate of $y(\boldsymbol{s}, t, h)$. **D-STEM** v2 implements the univariate functional version of the HDG model:

$$y(\boldsymbol{s}, t, h) = \boldsymbol{x}(\boldsymbol{s}, t, h)' \cdot \boldsymbol{\beta}(h) + \boldsymbol{\Phi}(h)' \cdot \boldsymbol{z}(\boldsymbol{s}, t) + \epsilon(\boldsymbol{s}, t, h)$$

$$\boldsymbol{z}(\boldsymbol{s}, t) = \boldsymbol{G} \cdot \boldsymbol{z}(\boldsymbol{s}, t-1) + \boldsymbol{\eta}(\boldsymbol{s}, t)$$

The considerations made for the HDGM are valid, however there are some differences:

- $\boldsymbol{\beta}(h)$ is a collection of splines, one for each regressor. Each spline is the linear combination of n_β basis functions; these are common to all covariates, instead the n_β combinatorial coefficients change from variable to variable in order to describe in the best way the hourly contribute of a specific regressor in explaining y;
- $\boldsymbol{\Phi}(h)$ represents a collection of n_z basis functions that properly combined with the n_z latent variables $\boldsymbol{z}(\boldsymbol{s}, t)$ describe what $\boldsymbol{x}(\boldsymbol{s}, t, h)' \cdot \boldsymbol{\beta}(h)$ is not able to explain, i.e. the spatio-temporal correlation;

	CV MSE, full model		*CV MSE, selected model*	
Model	*Pickups*	*Trip duration*	*Pickups*	*Trip duration*
2-variate	0.9870	0.9624	0.6584	0.8221
1-variate	0.4654	0.7825	0.6685	0.7868

Table 3: MSE concerning cross-validation in log-standardized scale for response variables, DCM.

Model	*Full*	*Selected*
2-variate	-6106.405	-7491.195
1-variate pickups	-3512.167	-3122.354
1-variate trip duration	-4536.846	-4309.739

Table 4: Log-likelihoods of the six models, DCM.

- the variance of ϵ is a spline obtained by linear combined n_ϵ basis functions. The aim of making σ_ϵ^2 a time signal is to find when in h domain the model performs worse.

In our specific case:

- t is the 2020 day index, instead h is the day hour index;
- $y(\boldsymbol{s}, t, h)$ indicates the number of pickups' made during the hour h of the day t at the rental station located in $\boldsymbol{s}$, instead $\boldsymbol{x}(\boldsymbol{s}, t, h)$ contains the constant, the values of the most significant weather variables, dummy variables and distance. We have also estimated two simpler models contained a reduced number of covariates in order to increase the descriptive power of the latent variables $\boldsymbol{z}(\boldsymbol{s}, t)$;
- we have chosen to use the Fourier basis seen the daily periodicity of the data;
- $n_\beta = 5$, $n_\epsilon = 5$ and $n_z = 7$; we have decide to give more resolution to the latent component.

4 Data analysis

4.1 Analysis of the DCM

4.1.1 Model selection

In order to choose parsimonious models with good performance, two different types of models are considered, namely, Full models, which contain all the available covariates (weather, distances, lockdown, holiday), and Selected models, which include only those covariates that are significant. The model selection was carried out on the univariate spatio-temporal process related to pickups, on the univariate spatio-process process related to trip duration and on the bivariate spatio-temporal process (pickup, trip duration). After obtaining the various models, we want to evaluate their performance in comparison to the full models and their predictive ability through the Cross Validation (table 3).
In table 4 we compares the models' likelihood and we can see the pickups selective univariate model have the lowest log-likelihoods while the bivariate selected model have the highest.

4.1.2 Model analysis

Figure 5 shows the trend of the latent z interacting with the covariates:

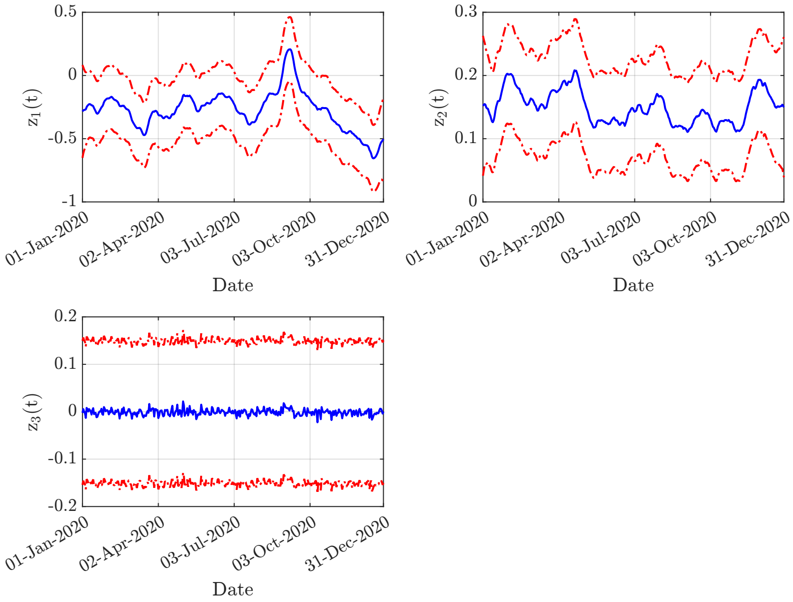


Figure 5: estimated $z_1(t)$, $z_2(t)$ and $z_3(t)$ by Kalman smoother in the univariate model for pickups.

	$\hat{\boldsymbol{\beta}}_{\boldsymbol{const}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{temp}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{rain}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{dist}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{Holiday}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{Lock}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{UV}}$
2-variate pickups	-0.109	0.264	-0.054	0.089	0.222		0.154
2-variate trip duration	-0.071	0.310		-0.064	-0.048	-0.171	0.145
1-variate pickups	0.079	0.269	-0.053	-0.206	0.001	-0.693	0.104

Table 5: estimated $\boldsymbol{\beta}$ for the bivariate model and 1-variate pickups.

- z_1, which indicates the trend of the pickups is a significant variable that has a positive effect both on the number of rentals and on the average daily rental duration.
- z_2, which indicates the distance. The trend is always positive therefore it brings a decrease in the number of pickups
- z_3 , which indicates the holidays and not affect the increase of pickups.

Figure 6 shows the trend of the latent z interacting with the covariates:

- z_1, which indicates the trend of the pickups.
- z_2, is refers to the distance. in comparison to the univariate case it can be noted an decrease in pickups.
- z_3 , which indicates the trend of the average of trip duration
- z_4 , which indicates the distance and not affect the increase of the average trip duration.

The most significant coefficients $\hat{\boldsymbol{\beta}}$ are shown in the table 5, it can be seen that both models the most significant coefficients are $\hat{\boldsymbol{\beta}}_{\boldsymbol{temp}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{UV}}$

4.2 Analysis of HDGM

4.2.1 Model selection

To select simpler but good performing models, two different types of models are considered, namely:

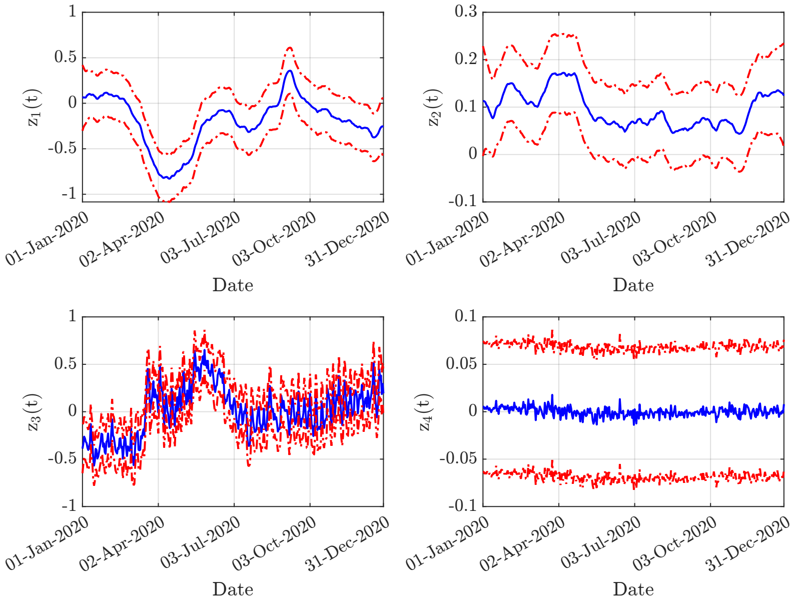


Figure 6: estimated $z_1(t)$, $z_2(t)$ and $z_3(t)$ by kalman smoother in the bivariate model.

	CV MSE, full model		*CV MSE, selected model*	
Model	*Pickups*	*Trip duration*	*Pickups*	*Trip duration*
2-variate	0.3670	0.7118	0.3609	0.7024
1-variate	0.3673	0.6889	0.3650	0.6829

Table 6: MSE concerning cross-validation in log-standardized scale for response variables, HDGM.

- **full models**: contain all the available covariates (weather, distances, lockdown and Holidays).
- **selected models**: include only the covariates that have been selected through a backward approach where at each iteration the covariate with the worst performance was excluded from the model until reaching a level significant of 5%.

Once we have obtained the select models, we want to evaluate their performance compared to the full models. Through the Cross Validation we evaluate the predictive capabilities of the models. The interesting thing is that, by making model selection, we not only obtained a simpler model to interpret, but we also managed to improve, even if only slightly, the predictive capabilities (table 6).

Another way to compare the two model categories is to take a look at the Log-likelihoods. Also in this case the selected models have a better index than the complete models, as they have slightly higher Log-likelihoods (table 7).

Model	*Full*	*Selected*
2-variate	-342.12	-268.85
1-variate pickups	3507.62	3557.33
1-variate duration	-5491.65	-5451.83

Table 7: Log-likelihoods of the six models, HDGM.

Variable	$\hat{\beta}_{const}$	$\hat{\beta}_{temp}$	$\hat{\beta}_{rain}$	$\hat{\beta}_{dist}$	$\hat{\beta}_{UV}$	$\hat{\beta}_{Holidays}$
Pickups	$-0.048_{(0.126)}$	$+0.275_{(0.035)}$	$-0.070_{(0.009)}$	$+0.039_{(0.027)}$	$+0.205_{(0.013)}$	
Duration	$+0.297_{(0.081)}$	$+0.273_{(0.044)}$			$+0.153_{(0.017)}$	$+0.202_{(0.029)}$

Table 8: estimated $\boldsymbol{\beta}$ for the multivariate model, HDGM.

Response variable	$\hat{g}_i$	$\hat{\sigma_i}$	$\hat{\theta}_i$
Pickups	$0.92_{(0)}$	$0.128_{(0.002)}$	$0.02_{(0)}$
Trip duration	$0.81_{(0.01)}$	$0.487_{(0.006)}$	$0.02_{(0)}$

Table 9: estimated parameters of the multivariate model, HDGM.

4.2.2 Model analysis

The β coefficients estimated in table 8, adequately describe the relationships of the variables with respect to our response variables:

- β_{temp}, which indicates the average perceived temperature, is a significant variable that has a positive effect both on the number of rentals and on the average daily rental duration.
- β_{rain}, causes a negative change only on the number of rentals.
- $\beta_{distance}$,as expected, the distance from the nearest public transport station, has a positive effect on the number of bike rentals.
- β_{UV}, perhaps the least expected result, which indicates that UV energy increases both response variables.
- $\beta_{Holidays}$, the dummy variable that describes whether it is a vacation or not, has a positive effect on the average rental duration, but is not significant on the number of rentals.

The other parameters which describe the selected bivariate model are shown in the table 9:

- the analysis of the $\hat{g}_i$ values suggests that the two response variables have different temporal dynamics, in particular, the number of bikes rented has higher persistence than the average rental duration;
- as expected, the $\hat{\sigma_i}$ has very distant values with respect to the two variables. In detail, the average duration has a very high $\hat{\sigma_i}$, which describes the high volatility of this variable. This value is consistent with the cross-validation MSE results of table 6, where the model that describing the average rental duration performed worse;
- the estimated parameter θ_i represents the range of the spatial correlation, common to all response variables. It amounts to just over $2km$ for this dataset, confirming a fairly regular overall spatial behaviour.

4.3 Analysis of f-HDGM

4.3.1 Analysis of relevance of the latent component $\boldsymbol{z}(\boldsymbol{s}, t)$

Due to the high computational time which the estimate of the f-HDGM requires, it was not possible to employ the stepwise approach to perform the model selection, therefore starting from the most relevant regressors identified through the estimate of an OLS model, we have built three models with a different number of covariates:

Model	*Card.* $\boldsymbol{\beta}(h)$	$V_{\boldsymbol{z}}^{1,1}$	$V_{\boldsymbol{z}}^{2,2}$	$V_{\boldsymbol{z}}^{3,3}$	$V_{\boldsymbol{z}}^{4,4}$	$V_{\boldsymbol{z}}^{5,5}$	$V_{\boldsymbol{z}}^{6,6}$	$V_{\boldsymbol{z}}^{7,7}$
Large model	10	1.71	0.75	1.21	0.36	0.65	0.12	0.17
Medium model	5	1.93	0.90	1.30	0.38	0.66	0.13	0.19
Small model	3	1.96	0.93	1.42	0.48	0.79	0.14	0.20

Table 10: variances of the latent component $\boldsymbol{z}(\boldsymbol{s}, t)$, i.e. the main diagonal of the matrix V_z, as the model complexity varies, f-HDGM.

Covariate	χ^2 *value*	*p-value*
Constant	211.16	$< 10^{-16}$
Feels-like temp.	301.72	$< 10^{-16}$
Distance	92.83	$< 10^{-16}$
Lockdown	431.19	$< 10^{-16}$
Holidays	32.71	$4.30 \cdot 10^{-6}$

Table 11: p-values of the χ^2-tests to determinate the significance of the regressors, f-HDGM.

- **large**: 9 regressors, i.e. feels-like temperature, rainfall, visibility, wind speed, cloud cover, distance and dummy variables;
- **medium**: 4 covariates, i.e. feels-like temperature, distance and dummy variables;
- **small**: 2 regressors, i.e. feels-like temperature and dummy lockdown.

In each of them we have also introduced a constant in order to allow **D-STEM** to estimate through the EM algorithm the mean number of hourly bike rentals. The aim of this study is to determinate how the relevance of the latent component $\boldsymbol{z}(\boldsymbol{s}, t)$ changes as the number of covariates, i.e. the dimension of $\boldsymbol{\beta}(h)$, varies. In table 10 are shown the n_z values of the variance of latent component; we can see that variability increases as the model complexity decreases. The reason is the following: if we decide to use a less number of covariates to describe the bike sharing phenomenon, then the descriptive power of $\boldsymbol{z}(\boldsymbol{s}, t)$ increases in order to cope with what the global component $\boldsymbol{x}(\boldsymbol{s}, t, h)' \cdot \boldsymbol{\beta}(h)$ is not able to explain due to reduction of the model complexity.

We have decided to describe in detail the *medium model*, a fair compromise between complexity and relevance of the latent component.

4.3.2 Model parameters

After having estimated the parameters of the medium model, we have performed the χ^2-test to evaluate the significance of the regressors. Results are shown in table 11: all p-values are less than 5 %, therefore we can conclude all functions $\boldsymbol{\beta}(h)$ related to the chosen covariates are significantly different from zero and consequently give their contribute to explain $y(\boldsymbol{s}, t, h)$, i.e. the number of hourly pickups at a rental station.

In the figure 7, instead, we have reported the daily trend of the estimated functions $\boldsymbol{\beta}(h)$ and the variance of ϵ, i.e. $\sigma_\epsilon(h)$. It is interesting the trend of $\beta_{const}(h)$ which expresses the mean number of daily pickups during the day; there are two peaks that probably are related with start and end work times. All others variables contribute positively in explaining $y(\boldsymbol{s}, t, h)$, except for the distance and the dummy lockdown; the negative trend of the former is correlated to the fact that an individual is less incentivized to use bike due to tiredness at the end of a

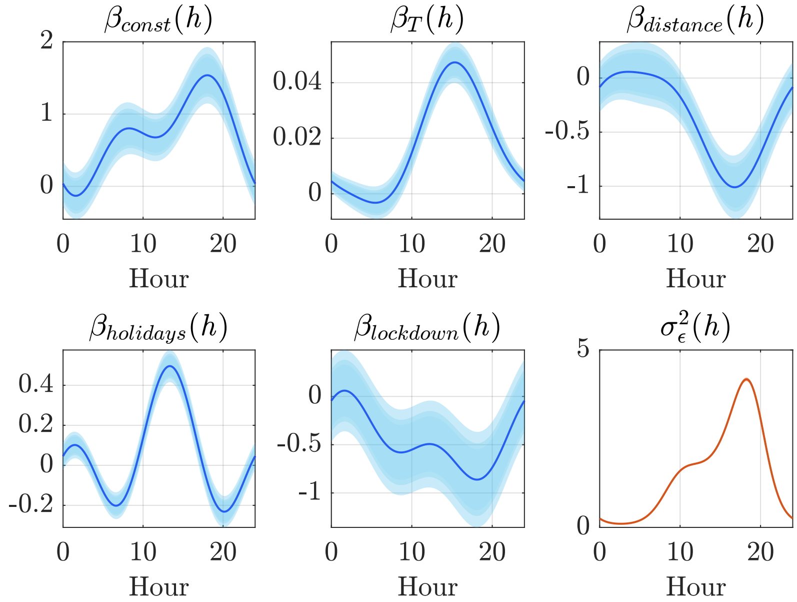


Figure 7: estimated $\beta_{const}(h)$, $\beta_T(h)$, $\beta_{distance}(h)$, $\beta_{holidays}(h)$, $\beta_{lockdown}(h)$ and $\sigma^2_\epsilon(h)$ with 90 %, 95 % and 99 % confidence bands, respectively, shown through the different shades, f-HDGM.

Index	*Min.*	*Max.*	*Mean*	*Median*	*Std*
Hourly RMSE$_h$	0.19	1.68	0.96	1.11	0.48
Hourly RMSE$_t$	0.22	2.04	1	1.04	0.36
Hourly RMSE$_s$	0.44	2.17	1	0.96	0.40

Table 12: main statistics concerning RMSE, f-HDGM.

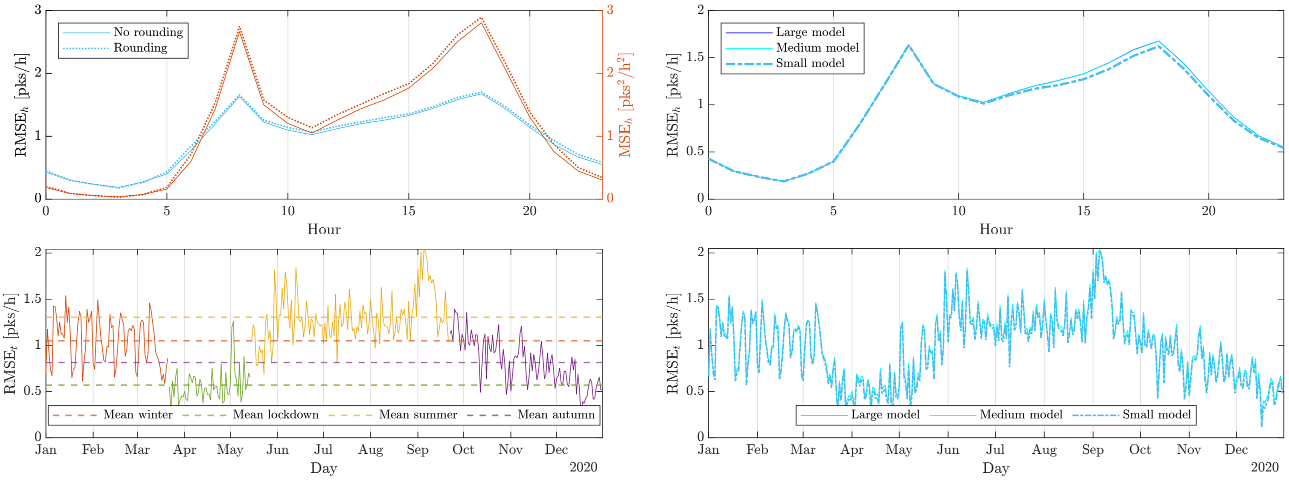
working day, that of the latter, instead, is caused to the impossibility to leave own home during the Pause Program in NYC. Moreover, $\beta_{lockdown}$ is also highly uncertain because it's different from zero for only few days. Finally, analysing the behaviour of $\sigma_\epsilon(h)$ we can note the estimated model performs worse around 20 o'clock.

4.3.3 Model validation

In figure 8(a), first plot, is shown the trend of RMSE$_h$ and RMS$_h$; thanks to the latter differences are accentuated, therefore we can note as the medium model in validation performs worse in peak hours. Dotted line describes the case in which we round $\hat{y}(\boldsymbol{s}, t, h)$ to the nearest integer. In the second plot, instead, is shown the behaviour of RMSE$_t$ during 2020; we can notice the period in which the estimated model performs worse is in summer, however the range of values assumed by this statistical index is not wide, therefore differences between means in each period are not particularly marked. Finally, in figure 8(b) is reported a comparison between the behaviours in validation of the three estimated models; albeit slightly we can note that the small model is the best because probably, thanks to his reduced complexity, doesn't go into overfitting. We please you to read table 12 for more information concerning RMSE$_h$, RMSE$_t$ and RMSE$_s$.

5 Results comparison

After having analysed the results coming from estimates of the DCM and HDGM, i.e. the unique ones for which we have employed model selection, we can summarize that most significant



(a)

(b)

Figure 8: trends of RMSE$_h$ and RMSE$_t$ concerning medium model (a) and comparison between the behaviours in validation of the three estimated models, f-HDGM.

common covariates in explaining the number of daily pickups at rental stations are feels-like temperature, rainfall, distance and UV index, instead to describe the trip duration are feels-like temperature, UV index and dummy holidays. In table 13 we can notice that, both for RMSE$_t$ and RMSE$_s$, the best model in predicting pickups is the f-HDGM in terms of both mean and standard deviation, i.e. it is not only the model which makes the smallest mean validation error, but also its RMSE is the one with the least dispersion. Moreover, we can note that HDGM performs better respect to DCM; the reason could be the following: the former contains only a single latent component, i.e. $\boldsymbol{z}(\boldsymbol{s}, t)$, instead the latter in addition has also $\boldsymbol{w}(\boldsymbol{s}, t)$, therefore it depends less on training data respect to DCM, i.e. it has less probability to fall into overfitting. Finally, regarding trip duration the best spatio-temporal model is DCM.

Index	*Variable*	*Model*	*Min.*	*Max.*	*Mean*	*Median*	*Std*
RMSE$_t$	Pickups [pks/day]	DCM	0.31	52.23	12.74	10.67	9.77
		HDGM	0	36.86	11.12	11.13	6.21
		f-HDGM	3.60	29.39	10.81	10.04	4.73
	Duration [min]	DCM	1	1001	33	11	99
		HDGM	0	2223	42	335	176
		f-HDGM	-	-	-	-	-
RMSE$_s$	Pickups [pks/day]	DCM	4.98	47.87	12.24	7.82	10.74
		HDGM	3.54	30.39	10.64	7.74	7.23
		f-HDGM	3.44	27.24	9.92	6.30	6.60
	Duration [min]	DCM	15	229	79	43	70
		HDGM	8	620	107	46	150
		f-HDGM	-	-	-	-	-

Table 13: comparison between models in validation. The best results are shown in green, the worst ones in red and intermediate values in orange.

6 Conclusions

At the end of our study we can conclude that the best spatio-temporal model to explain the bike sharing phenomenon is f-HDGM thanks to its capacity to work with data sampled in high frequency; its validation errors are more precise respect to the other models because they are built taking into account hourly data. However, the computational time requires to estimate it is greater than the other ones, therefore it is preferable to choose a not-functional statistical approach as the number of available data increases. To improve the performances of these models could be interesting to perform same the analyses on a dataset contained also bike sharing information regarding years prior the 2020; in this way there would be the possibility of taking into consideration in models estimates also the periodicity of the phenomenon. Finally, weather variables are useful to forecast if an individual will rent a bike, however they are not sufficient to explain our response variables completely due to social and individual reasons which are hard to get.

References

C. Calculli, A. Fassò, F. Finazzi, A. Pollice, and A. Turnone. Multivariate hidden dynamic geostatistical model for analysing and mapping air quality data in apulia, italy. 2014.

A. Fassò, F. Finazzi, and E. Scott. The dynamic coregionalization model in air quality risk assessment. 08 2011. URL `https://2011.isiproceedings.org/papers/950449.pdf`.

F. Finazzi, Y. Wang, and A. Fassò. D-stem v2: A software for modeling functional spatio-temporal data. *Journal of Statistical Software*, 99(10):1–29, 2021. doi: 10.18637/jss.v099.i10. URL `https://www.jstatsoft.org/index.php/jss/article/view/v099i10`.

List of Tables

1 Main statistics concerning bike sharing variables 2
2 Main statistics concerning weather variables 3
3 MSE concerning cross-validation in log-standardized scale for response variables (DCM) 8
4 Log-likelihoods of the six models (DCM) 8
5 Estimated $\boldsymbol{\beta}$ for the bivariate model and 1-variate pickups (DCM) 9
6 MSE concerning cross-validation in log-standardized scale for response variables (HDGM) 10
7 Log-likelihoods of the six models (HDGM) 10
8 Estimated $\boldsymbol{\beta}$ for the multivariate model (HDGM) 11
9 Estimated parameters of the multivariate model (HDGM) 11
10 Variances of the latent component $\boldsymbol{z}(\boldsymbol{s},t)$ as the model complexity varies (f-HDGM) 12
11 p-values of the χ^2-tests to determinate the significance of the regressors (f-HDGM) 12
12 Main statistics concerning RMSE (f-HDGM) 13
13 Comparison between models in validation 14

List of Figures

1 Map of bike rental stations and of the nearest subway/train stations and map of the public transport network in Jersey City 2
2 Comparison between the trend of the mean number of daily rentals and some weather variables 4
3 Box-plots of some weather variables 4
4 Behaviour of the mean number of daily rentals during the 2020 events 5
5 Estimated $z_1(t)$, $z_2(t)$ and $z_3(t)$ by Kalman smoother in the univariate model for pickups (DCM) 9
6 Estimated $z_1(t)$, $z_2(t)$ and $z_3(t)$ by Kalman smoother in the bivariate model (DCM) 10
7 Estimated $\beta_{const}(h)$, $\beta_T(h)$, $\beta_{distance}(h)$, $\beta_{holidays}(h)$ and $\beta_{lockdown}(h)$ (f-HDGM) . 13
8 Trends of RMSE$_h$ and RMSE$_t$ concerning medium model and comparison between the behaviours in validation of the three estimated models (f-HDGM) . . 14