Functional Data Analysis of London Bike Sharing Data

STE

Abstract. This project applies Functional Data Analysis (FDA) to London's bike-sharing dataset available in kaggle[1] to uncover rental data patterns and their relationship with categorical factors, namely, seasons, and weather conditions. By leveraging Fourier smoothing, Functional Principal Component Analysis (FPCA), Functional ANOVA (FANOVA), Functional Regression, Depth Analysis, and clustering based on FPCA scores, certain characteristics, trends, group differences, outliers, representative cases, in the dataset are revealed. The findings evidence the utility of FDA in understanding subset of urban mobility patterns and could provide the basis for the further exploration of similar temporal data, utilizing relevant methodologies other than non-functional methods like time series analysis, multivariate regression, etc.

Keywords: Functional Data Analysis, Fourier Smoothing, FPCA, FANOVA, Functional Regression, Depth Analysis, London Bike Sharing Data

1 Introduction

Bike-sharing systems have become an indispensable component of urban mobility. Understanding their dynamics is useful for optimizing operations, planning, and policy-making, either for the service provider or urban planner. Traditional linear regression, time-series methods often treat data as discrete points, without capturing the smoothness and continuity inherent in temporal usage patterns.

Functional Data Analysis (FDA) [2], [3], addresses these challenges by treating data as smooth functions over a continuum (e.g., hours in a day). In this paper, FDA is utilized to London's bike-sharing data to analyze hourly rental counts and uncover underlying structures shaped by seasonality, holidays, and weather conditions.

The `fda` package in R [8], and its manual along with FDA section in R project website [6, 5], are the important tools utilized in functional data analysis in this paper, which are accompanied with additional examples, methods and code provided in the Functional and Topological Data Analysis course at Università degli Studi di Milano [7], by professor Micheletti Alessandra to build a comprehensive FDA workflow.

The analysis in this paper includes:

– Data preparation, exploratory analysis.
– Functional data creation and fourier smoothing.
– FPCA to identify main modes of variation.
– FANOVA to assess the impact of categorical factors (season, holiday, weather) on functional profiles.
– Functional regression to model rental behavior.
– Depth analysis and functional boxplots to identify atypical curves or outliers, with wilcoxon tests, deep curve analysis.
– Clustering based on FPCA scores to visualize depth distribution patterns.

It was expected that the results would reveal distinct daily patterns, seasonal variations, holiday effects, and weather-related particularities.

2 Mathematical Foundations

2.1 Functional Representation of Data

Given n observations of time-series data $\{y_i(t)\}_{i=1}^n$ over a domain $\mathcal{T}$ (e.g., 24 hours), FDA represents each observation as a smooth function:

$$y_i(t) \approx f_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t),$$

where $\{\phi_k(t)\}_{k=1}^K$ are basis functions (e.g., Fourier or B-splines), and c_{ik} are coefficients [2].

2.2 Fourier Smoothing

For periodic data, a Fourier basis is often suitable:

$$\phi_k(t) = \begin{cases} 1, & k = 0, \\ \sqrt{2}\cos(k\omega t), & k \text{ even}, \\ \sqrt{2}\sin(k\omega t), & k \text{ odd}, \end{cases}$$

with $\omega = \frac{2\pi}{T}$ and $T = 24$. Smoothing chooses λ via Generalized Cross-Validation (GCV) to balance fit and smoothness:

$$\mathcal{L}(f) = \int_{\mathcal{T}} (y(t) - f(t))^2 dt + \lambda \int_{\mathcal{T}} (f''(t))^2 dt.$$

[9]

2.3 Functional Principal Component Analysis (FPCA)

FPCA decomposes functions into orthogonal modes of variation:

$$f_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t),$$

where $\mu(t)$ is the mean function, ξ_{ik} are scores, and $\phi_k(t)$ are eigenfunctions of the covariance operator. Eigenvalues λ_k measure variance explained[2].

2.4 Functional ANOVA (FANOVA)

FANOVA tests differences among groups (e.g., seasons) at the functional level:

$$f_i(t) = \mu(t) + \sum_{g=1}^{G} \beta_g(t) z_{ig} + \epsilon_i(t).$$

Significance is assessed via $F(t)$-like statistics comparing between-group and within-group variation [8].

2.5 Depth Analysis

Functional depth quantifies how central a curve is within a sample. The Modified Band Depth (MBD) is:

$$\text{MBD}(f) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} I(f(t) \in [f_i(t), f_j(t)], \forall t),$$

where $I(\cdot)$ is an indicator [10].

3 Data Preparation and Initial Checking

The dataset includes hourly rental counts, temperatures, humidity, wind speed, weather codes, holiday indicators, and seasonal labels. After parsing timestamps, initial check was performed to handle missing values, compute summary statistics, and engineer features (hour, day, month, year, day-of-week, etc) for further analysis.

Various aggreggations or grouping were performed to examine the general patterns. Initial exploratory analyses included TSA style line plots and heatmaps at various temporal scales to identify preliminary patterns.

4 Exploratory Analysis

4.1 Aggregate Patterns

The below plots show total rental counts data organized in different time scale, showing hourly peaks, weekly cycles, monthly trends, and yearly patterns (Figure 1).

- **Daily Rentals by Hour of Day**: Peak usage occurs during commuting hours (8 AM and 5-6 PM), with minimal activity at night.
- **Weekly Rentals by Day of Week**: Rentals are generally higher on weekdays compared to weekends, reflecting work-related usage.
- **Monthly Rentals by Day of Month**: Rentals are relatively stable throughout each month, with occasional spikes indicating events or holidays.
- **Yearly Rentals by Month of Year**: Rental activity peaks during summer months (June to August) and decreases significantly in winter (December to February). Trends are consistent across years.

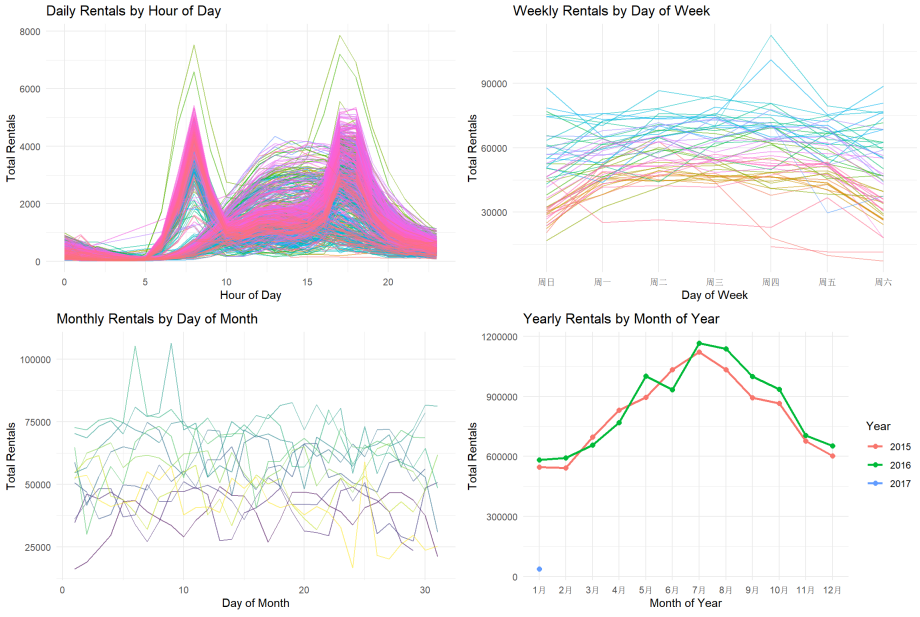

Fig. 1. Total Rental Plots by hour of day, day of week, day of month, and month of year.

4.2 Heatmaps

Heatmaps (Figure 2) offer a way to visualize intensity of rentals(total rental counts) over daily-hourly, weekly-daily, monthly-daily, and yearly-monthly grids. They highlight regular peaks and seasonal shifts.

- **Daily-Hourly Rentals**: Rentals peak during morning and evening commute hours, as seen in denser color blocks.
- **Weekly-Daily Rentals**: Weekdays show more activity compared to weekends, with mid-week having the highest concentration.
- **Monthly-Daily Rentals**: Peak rentals occur during summer months (July and August), with steady demand throughout the months.
- **Yearly-Monthly Rentals**: Rentals are highest during summer, especially in July, while winter months experience the least activity. Trends are consistent over years.

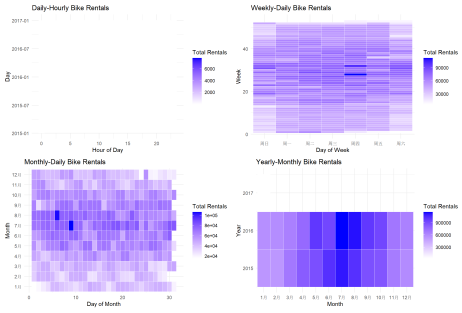

Fig. 2. Heatmaps of Daily-Hourly, Weekly-Daily, Monthly-Daily, and Yearly-Monthly Bike Rentals.

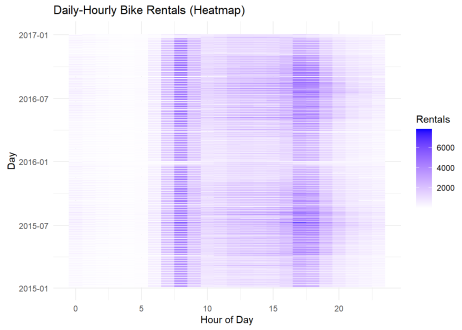

Fig. 3. Heatmaps of Daily-Hourly

To better visualize daily rental patterns, below heat map (Figure 4) could offer better intuition, displaying hourly bike rental trends across different temporal dimensions.

– **Hourly Rentals by Day of Week**: Rentals peak on weekdays during commute hours (8–10 AM, 5–7 PM), with lighter activity on weekends.
– **Hourly Rentals by Day of Month**: Consistent peaks during similar time slots (morning and evening hours) throughout the month, reflecting habitual commuting behavior.
– **Hourly Rentals by Month of Year**: Rentals surge during summer months, particularly mid-year, while winter months have fewer rentals. Peak hours remain aligned with workday schedules.

These visualizations reveal distinct commuter and certain seasonal trends in bike-sharing usage.

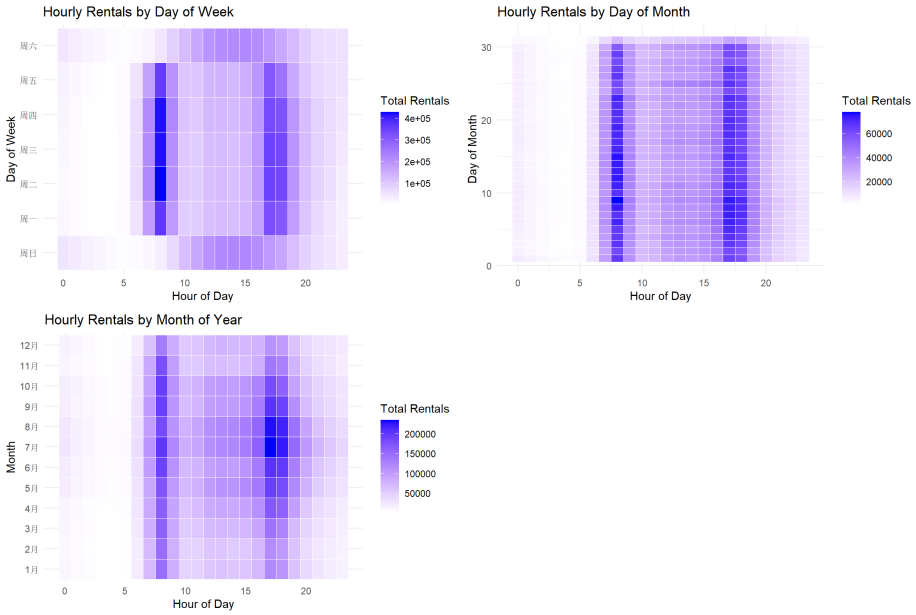

Fig. 4. Hourly Rentals stratified by Day of Week, Day of Month, and Month of Year.

4.3 Hourly Trends by Season, Month, and Holiday

The below (figure 5) shows hourly trends(total rental counts) by season, month, and holiday vs. non-holiday.

– Hourly Trends by Season: Spring, summer, and fall show significant morning (7–9 AM) and evening (5–7 PM) peaks, reflecting commuter usage. Winter displays a subdued trend with lower rentals overall.
– Hourly Trends by Month: Rentals align with commuting hours across months, peaking during warmer months (May to September). Winter months (December to February) experience a drop in activity.

– Holiday vs. Non-Holiday Trends: Non-holiday trends exhibit sharp commuting peaks, while holidays show a flatter curve with more mid-day activity, indicating leisure-oriented usage.

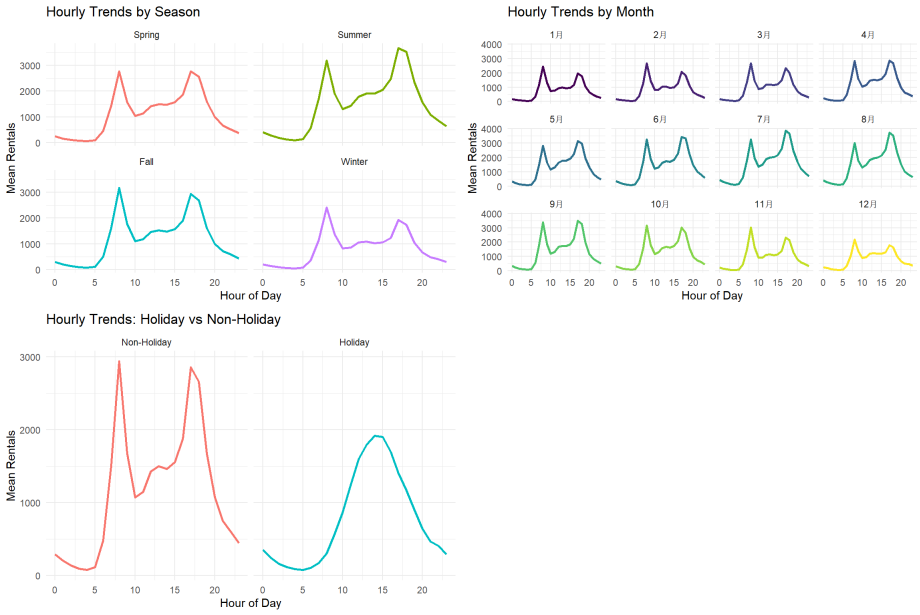

Fig. 5. Hourly Bike Rental Patterns by Season, Month, and Holiday Status

These preliminary explorations would provide basis for assumptions or expectations of further analytical results, where additional research could focus on particular patterns other than the obvious trends or characteristics.

5 Functional Data Analysis

5.1 Fourier Smoothing

In functional data analysis, Fourier smoothing represents data using Fourier basis functions to effectively capture and smooth periodic patterns.[2]

In this step, function data is created before smoothing. The visualization in below (figure 6) compares original and Fourier-smoothed patterns of bike rentals and key weather variables across hours of a day, 24 hours. The rental counts (cnt) shows distinct peaks during morning and evening rush hours, reflecting commuter behavior. Temperature metrics (t1 and t2) exhibit a gradual rise and fall. Humidity (hum) inversely correlates with temperature, peaking at early morning and late night. Wind speed fluctuates minimally but shows slight daytime increases. Smoothing enhances trend continuity, removing noise while

preserving underlying patterns, aiding in understanding rental and weather dynamics across hourly timeframes.

Generalized Cross-Validation (GCV) was implemented to determine the optimal regularization parameter for smoothing each dataseries.

The Fourier basis is particularly suited for cyclic or periodic data, such as hourly time series, ensuring smooth, continuous curves. Although b-spline basis along with registration were tested, such approach is not shown here for brevity and consideration of the nature of the dataset.

Note: Number of basis functions for Fourier expansions and is set in the beginning.

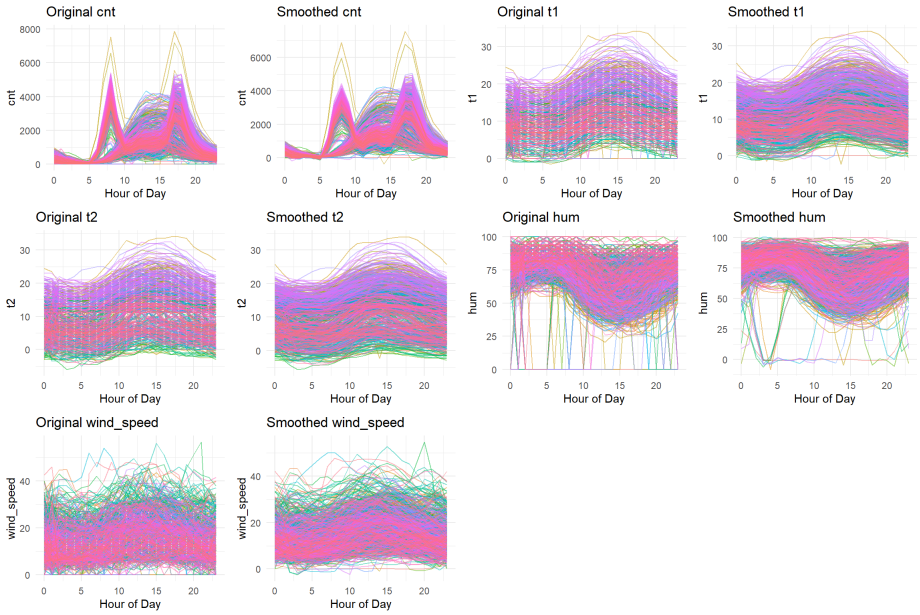

Fig. 6. Original vs. Smoothed Hourly Trends for Bike Rentals and Weather Variables (cnt, t1, t2, hum, wind_speed).

5.2 Functional Principal Component Analysis Results

Functional principal component analysis (FPCA) is a statistical technique that extends principal component analysis to functional data, enabling the investigation of dominant modes of variation in datasets where each observation is a function or curve. This method is particularly useful for dimension reduction and feature extraction in functional data analysis. [2]

The functional principal component analysis (FPCA) conducted on the variables of hourly bike rental data (e.g., count, temperature t1 and t2, humidity, wind speed) has yielded certain insights into patterns and variance distribution:

PCA results and Variance Explanation

In this analysis, FPCA was performed on smoothed functional data for each variable using a Fourier basis. Initially, a maximum number of principal components `max_nharm` was specified. The variance proportion `varprop` for each component was calculated, and the cumulative variance `cum_var` was computed. A dynamic threshold of `90%` cumulative variance was set to determine the optimal number of principal components `dynamic_nharm`. This approach ensures that the selected components capture the majority of data variability. In addition, the variance explained plots were generated, and Varimax rotation was applied to facilitate clearer component interpretation.The results of current step are displayed in tabular(table 1),plots(figure 7) forms below.

- **Variance Explained for cnt**: The bar plot shows the proportion of variance explained by each principal component (PC) for bike counts (`cnt`). Two PCs were sufficient to explain 94% of the variance, evident from the steep cumulative variance line. This indicates that the primary dynamics of bike rentals can be captured with just two components.
- **Variance Explained for t1**: For the temperature variable `t1` (e.g., real-time temperature), a single PC explains 92% of the variance. The single peak suggests a dominant, consistent diurnal pattern across all days.
- **Variance Explained for t2**: Similar to `t1`, one PC explains 92.7% of the variance for `t2` (e.g., temperature feel). The variance structure reflects the strong influence of a single, smooth day time trend.
- **Variance Explained for hum**: For humidity (`hum`), three PCs were required to explain 90% of the variance. This suggests more variability in the humidity patterns that cannot be captured with fewer components. The underlying complexity requires further analysis.
- **Variance Explained for windspeed**: Wind speed also required three PCs to capture 91.6% of its variance, indicating complex hourly variations that contribute to the dynamics of the variable.

Table 1. Table 1 Metrics for Functional Principal Component Analysis (FPCA)

Variable	Chosen PCs	Cumulative Variance Explained (%)
Count (cnt)	2	94.00
Temperature (t1)	1	92.07
Feels-like Temperature (t2)	1	92.72
Humidity (hum)	3	90.22
Wind Speed	3	91.59

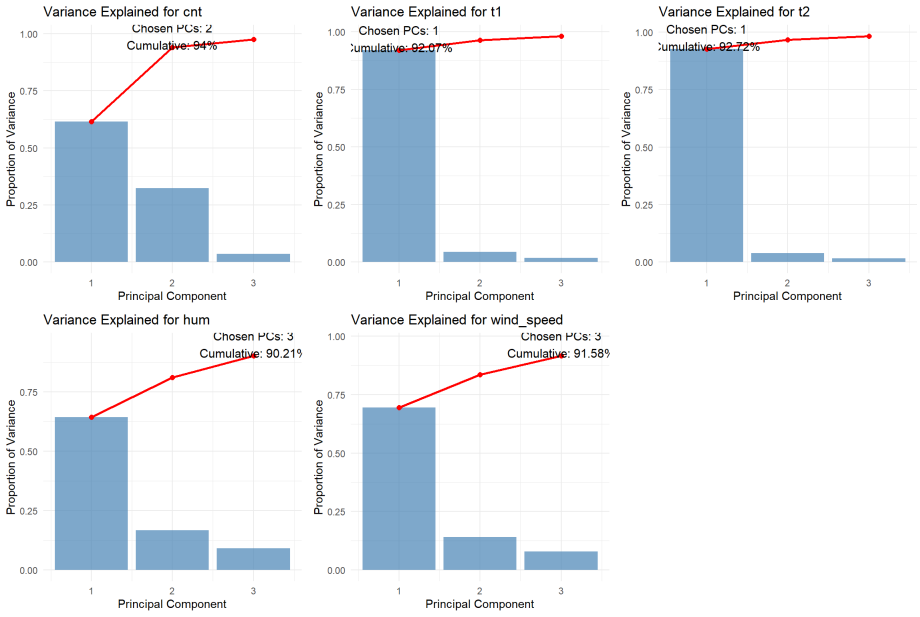

Fig. 7. Variance explained by principal components for each variable. The cumulative variance threshold of 90% determines the number of components selected dynamically for FPCA

Principal Component Functions

Principal Component Functions help to further visualize the main patterns and sources of variability in the data, where dominant trends can be identified.

- **PC Functions for cnt**: Two PCs for `cnt` capture the morning and evening rental peaks. The first PC reflects general trends, while the second PC distinguishes between different rental intensities during these periods.
- **PC Functions for t1 and t2**: Both temperature variables (`t1` and `t2`) rely on a single PC, with smooth harmonic curves that peak in the early afternoon, reflecting typical temperature patterns.
- **PC Functions for hum**: The three PCs for humidity illustrate a more complex behavior, with distinct amplitude changes over the day. The second and third PCs highlight opposing patterns, such as early morning versus late afternoon variability.
- **PC Functions for windspeed**: Similarly, the wind speed PCs reflect variability across different times of the day. The three PCs capture fluctuations, such as mid-morning peaks or afternoon troughs.

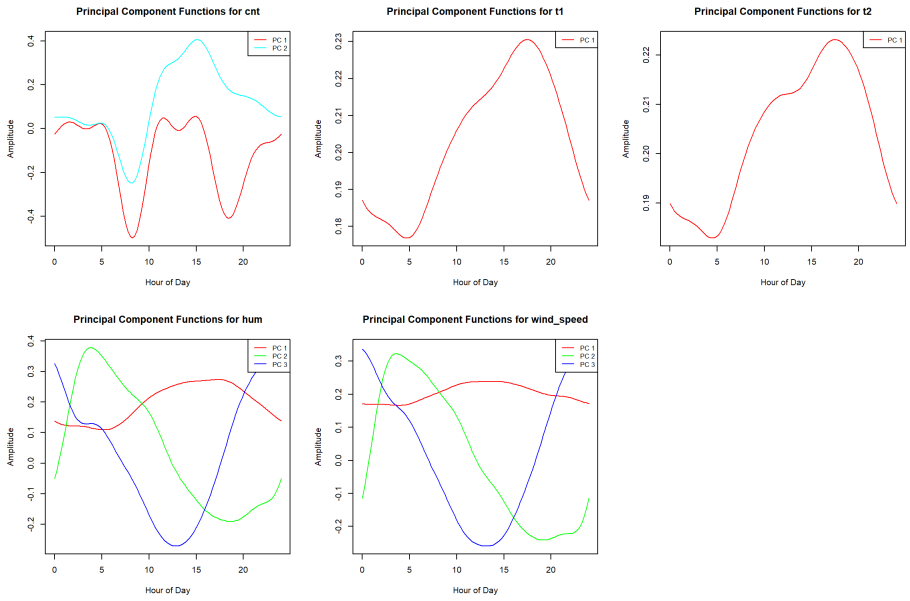

Fig. 8. Principal component functions for each variable, showing the harmonic amplitudes across the hours of the day. These illustrate the key temporal patterns captured by the functional principal components.

Varimax-Rotated PCs

Varimax-rotated FPCA function plots enhance interpretability by simplifying the structure of principal components. This rotation maximizes the variance of loadings within each component, making each factor more distinct and easier to visualize.

- **Rotated PCs for cnt**: The varimax-rotated PCs refine the interpretation, separating morning and evening rental peaks. The second PC highlights differences in peak intensities.
- **Rotated PCs for t1 and t2**: For temperature, rotation does not significantly alter the single PC since the patterns are already well-defined.
- **Rotated PCs for hum and windspeed**: After rotation, the PCs for humidity and wind speed are more distinct, showing clearer temporal patterns. For instance, in humidity, PCs now emphasize specific times of rising or falling trends(possibly related to additional weahter conditions), while wind speed PCs show similar dynamics.

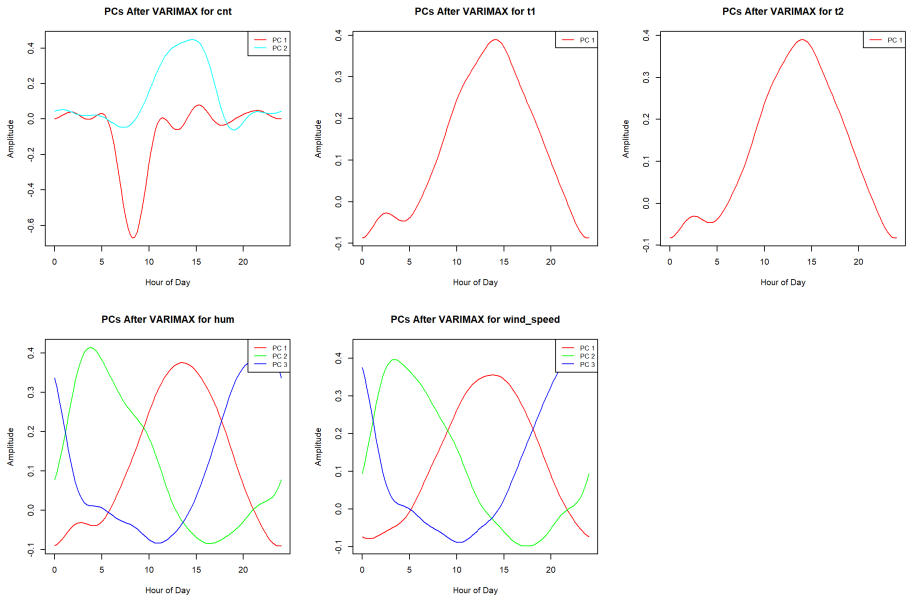

Fig. 9. Principal component functions after VARIMAX rotation for each variable, enhancing interpretability by emphasizing orthogonality and temporal separation of patterns across the day

5.3 Functional ANOVA (FANOVA)

Functional analysis of variance (FANOVA) extends traditional ANOVA techniques to functional data, enabling the assessment of differences among groups where each observation is a function or curve. This approach is particularly useful in fields such as ergonomics and medicine, where data are often inherently functional. [2] Here FANOVA was performed to assess effects of season, holiday, and weather factors.

Table 2. FANOVA Metrics

Grouping	SSE	SST	R-Squared
Season	6,523,967,477	42,499,620,416	0.8465
Holiday	7,752,890,769	42,499,620,416	0.8176
Weather	7,736,004,488	42,499,620,416	0.8180

Key Metrics :

– **SSE (Sum of Squared Errors):** Measures unexplained variability for each grouping factor. Lower values indicate better fit.
– **SST (Total Sum of Squares):** Measures total variability in the data.

– **R-Squared:** Proportion of variance explained by the factor. Higher values (close to 1) indicate stronger explanatory power.

Observations :
– **Season** explains 84.65% of the variance, indicating strong temporal patterns in bike rentals based on the season.
– **Holiday** explains 81.76% of the variance, showing holidays influence demand but less strongly than seasons.
– **Weather** explains 81.80% of the variance, suggesting significant but slightly weaker impacts compared to seasons.

Below plots display the factor effects on total bike rental counts.

Seasonal Effects :
– Each plot represents the effect of a season (Spring, Summer, Fall, Winter) on hourly bike rentals.
– **Summer** shows the highest peak around mid-day and early evening, indicating increased demand during warmer weather.
– **Winter** has lower overall effects, with smaller peaks during commuting hours (morning and evening).

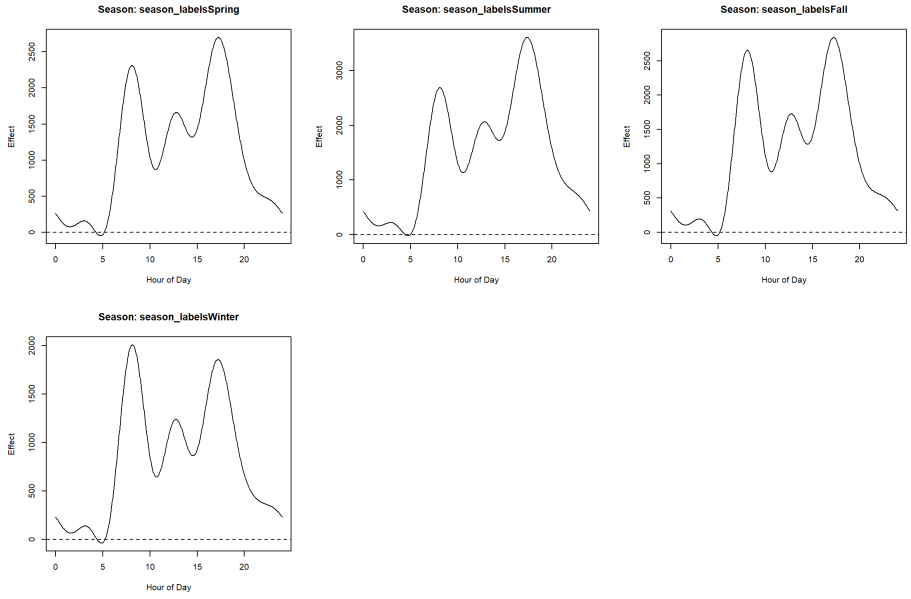

Fig. 10. Seasonal Effects on Bike Rentals - Functional ANOVA results showing the impact of different seasons on hourly bike rental patterns

Holiday Effects :
– Two plots show the effect of holidays and non-holidays on rentals.

– On holidays, bike rentals peak in the afternoon, reflecting leisure use. Non-holiday usage peaks during commute hours.

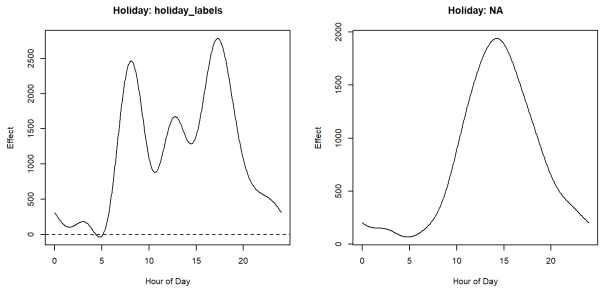

Fig. 11. Holiday Effects on Bike Rentals - Functional ANOVA results depicting the contrasting impact of holidays versus non-holidays on hourly rentals

Weather Effects :

– Weather condition plots display rental variations under different weather scenarios (Clear, Rain, Thunderstorms, etc.).
– Clear weather shows the highest effect, while rain and thunderstorms reduce demand significantly.

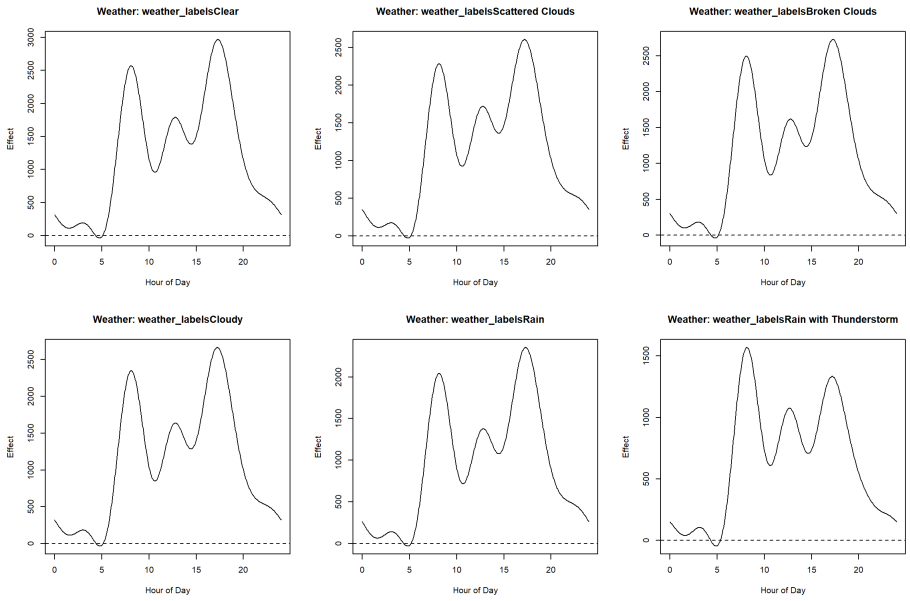

Fig. 12. Weather Effects on Bike Rentals - Functional ANOVA results illustrating the influence of various weather conditions on hourly bike rentals

Conclusion:

FANOVA identifies and quantifies the impact of season, holidays, and weather on bike rentals. With further analysis, the impact of different factors vary, but in general, commuter patter in non holiday, warmer weather is related to higher rental activity, extreme weather could reduce the rental significantly.

5.4 Functional Regression

Functional regression extends traditional regression techniques to scenarios where either predictors or responses are functions, enabling the modeling of relationships between functional and scalar variables. This approach is particularly useful in fields such as biostatistics and environmental science, where data are often inherently functional.[2] Incorporating temperature, humidity, wind speed, and holiday indicators into a functional regression model would be the first step to generalize variable relationships with such model(additional factors are not considered in this analysis).

Model Setup

The functional regression model implemented can be expressed as:

$$Y(t) = \beta_0(t) + \int X_1(s)\beta_1(s,t)\,ds + \int X_2(s)\beta_2(s,t)\,ds + \int X_3(s)\beta_3(s,t)\,ds + \beta_4(t)\cdot\text{Holiday} + \epsilon(t)$$

where:

- $Y(t)$: Response functional variable (hourly bike rental counts as a function of time t).
- $X_1(s), X_2(s), X_3(s)$: Functional predictors (e.g., t_2: temperature, hum: humidity, $wind$: wind speed) smoothed over time.
- $\beta_0(t)$: Functional intercept, representing the baseline trend in bike rentals over time.
- $\beta_1(s,t), \beta_2(s,t), \beta_3(s,t)$: Functional coefficients for predictors X_1, X_2, X_3, capturing the time-varying influence of predictors on $Y(t)$.
- $\beta_4(t)$: Scalar functional coefficient for the categorical predictor Holiday, indicating holiday-specific adjustments in bike rentals.
- $\epsilon(t)$: Residual error term, capturing the unexplained variance in bike rentals.

Model Variations

1. With Constant: - Includes the intercept term $\beta_0(t)$ to model the baseline effect of bike rentals independent of predictors. - Formula:

$$Y(t) = \beta_0(t) + \int X_1(s)\beta_1(s,t)\,ds + \cdots + \beta_4(t)\cdot\text{Holiday} + \epsilon(t)$$

2. Without Constant: - Excludes the intercept term $\beta_0(t)$, assuming the baseline is captured solely by the predictors. - Formula:

$$Y(t) = \int X_1(s)\beta_1(s,t)\,ds + \cdots + \beta_4(t)\cdot\text{Holiday} + \epsilon(t)$$

Coefficient Estimation: Functional coefficients $\beta_j(t)$ $(j = 1, 2, 3, 4)$ are estimated using the fRegress function. - Smoothing is applied using Fourier basis with a predefined number of basis functions ($nbasis$).

The time-varying coefficients $\beta_j(t)$ reveal how the influence of each predictor changes across the hours of the day.

Model	**SSE**	**SST**	**R^2**
With Constant	6,335,076,696	42,498,816,057	0.851
Without Constant	6,715,499,466	42,498,816,057	0.842

Table 3. Summary of Regression Model Metrics

Metric	With Constant	Without Constant
MSE	640,359.5	678,813.2
RMSE	800.2247	823.9012
MAE	596.343	607.9362
MAPE	18,002,644%	18,088,914%

Table 4. Additional metrics

Metric/Plot	**With Constant**	**Without Constant**
Residual Spread	Centered around 0; slightly tighter.	Centered around 0; slightly larger spread.
Extreme Residuals	Few outliers above 6000.	More pronounced outliers above 8000.
Residual Histogram	Symmetric, tightly clustered around 0.	Symmetric but slightly heavier tails.
Q-Q Plot	Deviates at tails; closer fit to normality.	Larger deviations at tails.
Coefficient Magnitudes	Smaller, as baseline captured by constant.	Larger, as coefficients absorb baseline.
Predictor Effects	Time-varying; peak around midday (holidays).	Similar time-varying effects as with constant.

Table 5. Comparison of Models With and Without Constant

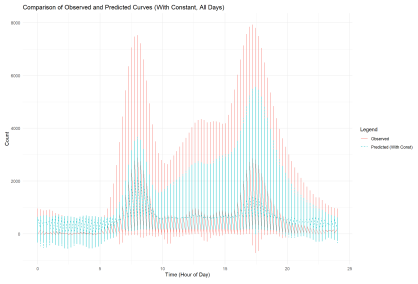

Fig. 13. Observed vs Predicted cnt functional data with Model with Constant

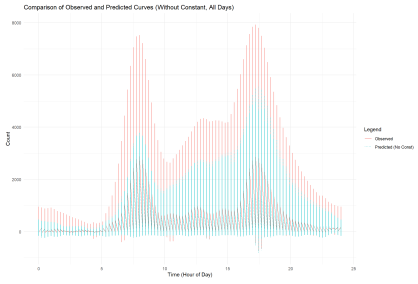

Fig. 14. Observed vs Predicted cnt functional datawith Model without Constant

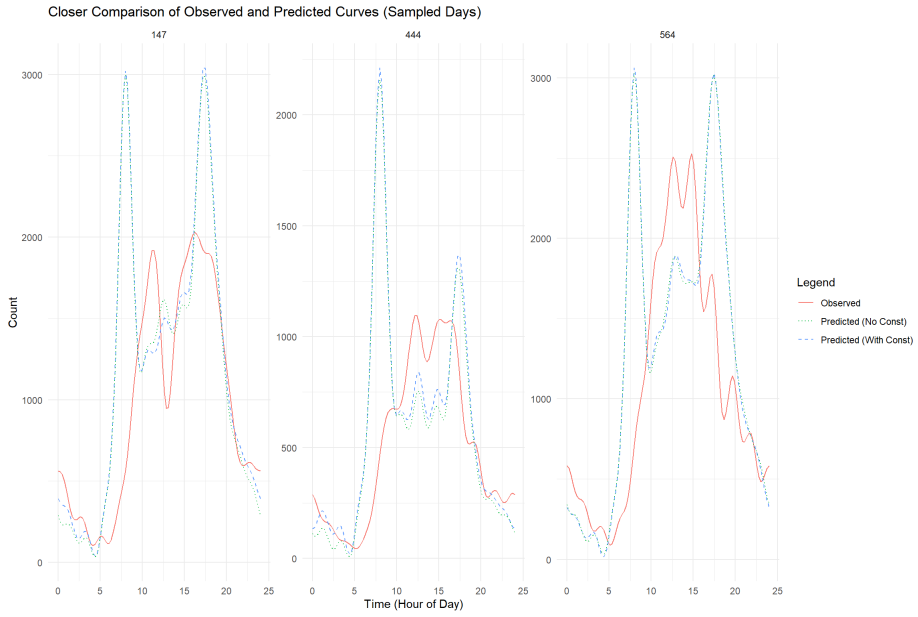

Fig. 15. Comparison of Observed and Predicted Curves (Sampled Days)

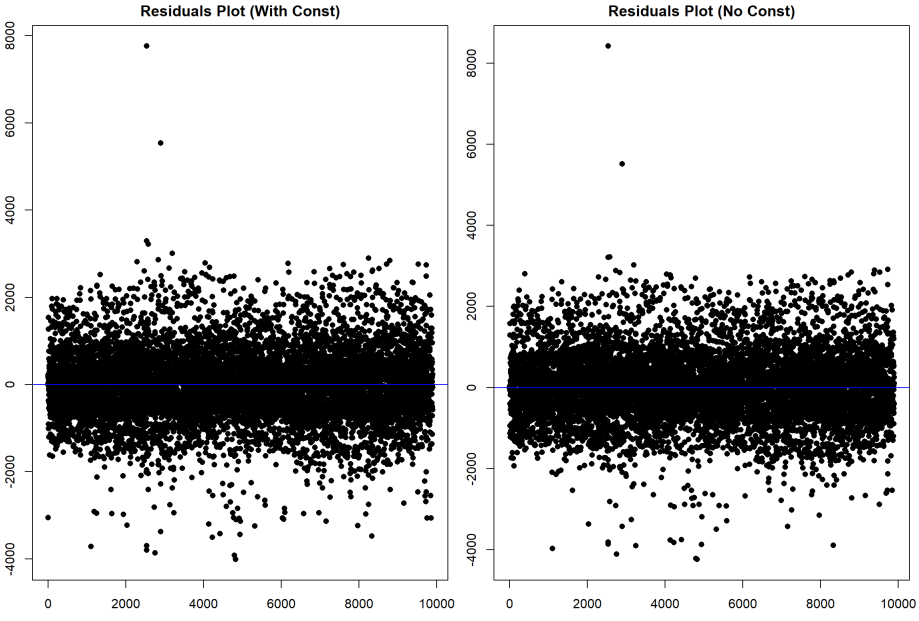

Fig. 16. plots of residuals with and without constant

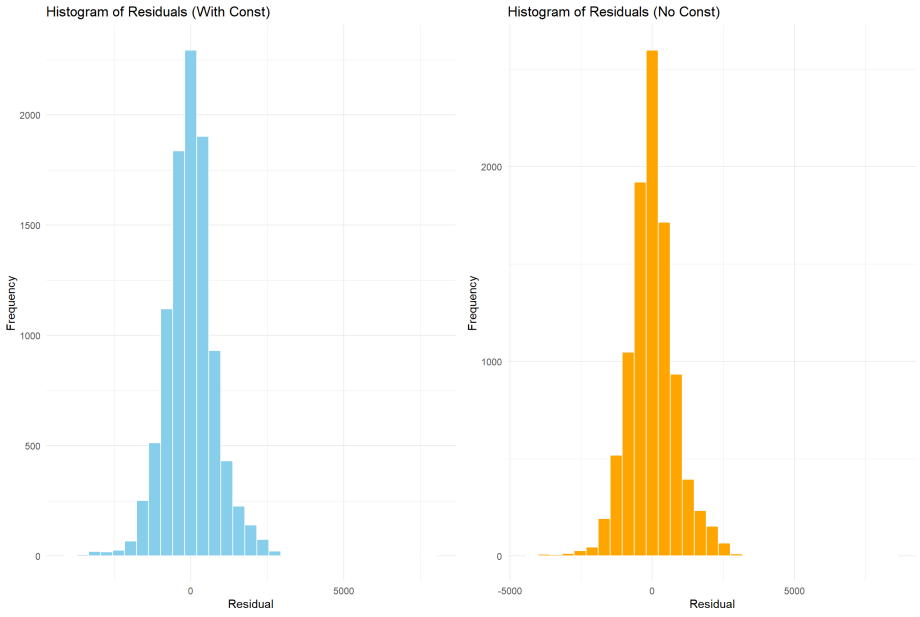

Fig. 17. histogram of residuals with and without constant

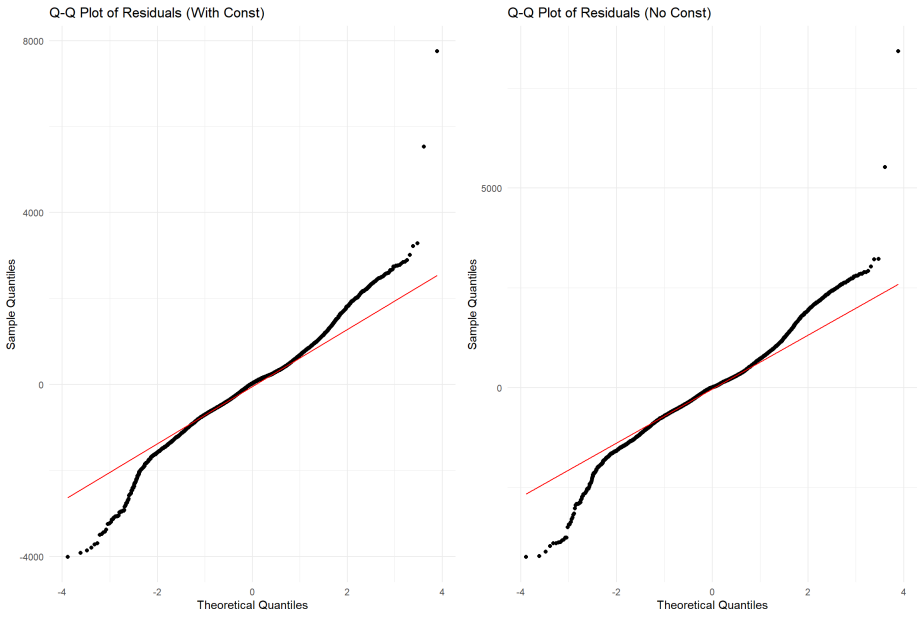

Fig. 18. Q-Q Plot Comparison with and without constant

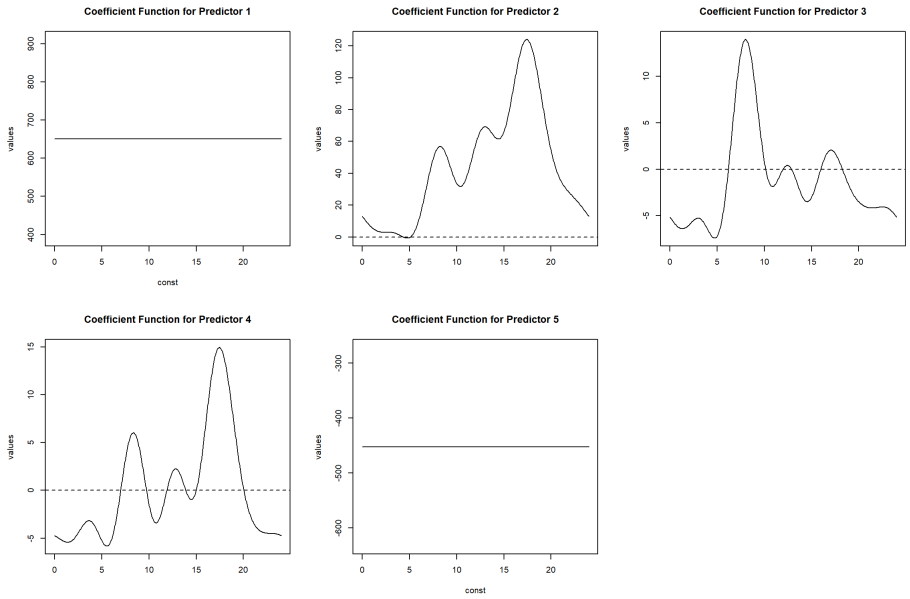

Fig. 19. coefficients for model with constant

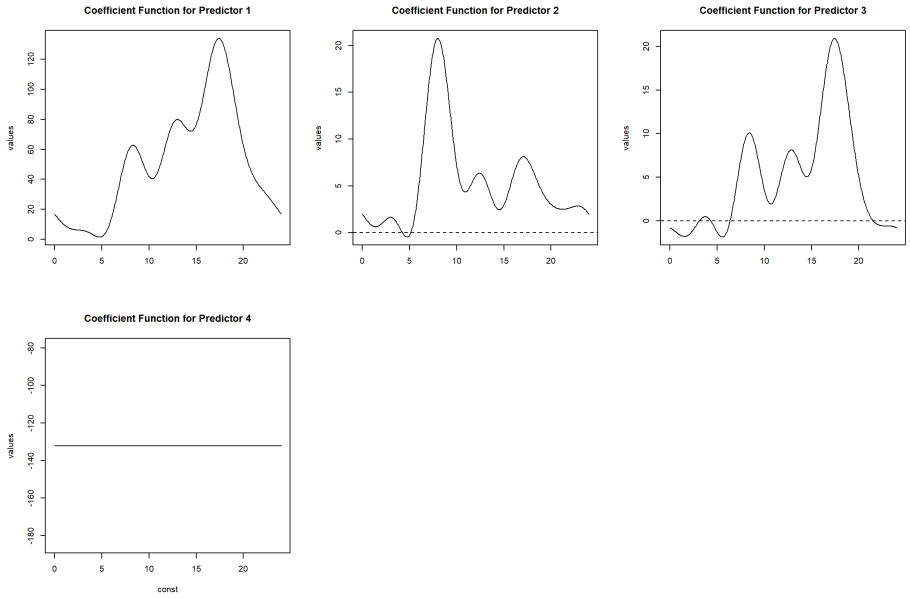

Fig. 20. coefficients for model without constant

General observations

- Model Fit:
  - Including the constant improves residual distribution by reducing extreme values and tightening the spread.
  - The constant term at certain extent effectively captures the baseline trend in bike rentals.
- Predictor Effects: Both models show similar time-varying effects for predictors, with midday peaks for holidays and temperature.
- Normality of Residuals: The model with a constant adheres more closely to normality, as evident in the Q-Q plot and histogram.
- Extreme Residuals: Outliers in both models need to be investigated, as they may indicate data issues or unmodeled phenomena.
- Model fit: Although both models have over `80%` accuracy, the model with a constant provides a slightly better fit and should be preferred.
- Variable(function) and regression strategy choices: Introducing additional variable(functional data) could be useful to increase accuracy but would also increase the difficulty of interpretations. Additional scalar to function and function to function regression strategies can also be further explored.

5.5 Depth Analysis and Functional Boxplots

Depth analysis is useful to identify outliers and anomalies. Functional boxplots are visual tools of depth analysis for summarizing the variability and identifying outliers in functional datasets. The plots display the functional boxplots for five variables: `cnt`, `t1`, `t2`, `hum`, and `wind_speed`. These boxplots(Figure 21) are constructed using the Modified Band Depth (MBD) method applied to smoothed functional data(Each variable is smoothed using Fourier basis functions to create a functional data representation over the 24-hour period).

Depth Calculation

Modified Band Depth (MBD): Measures the centrality of each functional curve relative to the overall data. Higher depth values indicate that a curve is closer to the central region, while lower values suggest potential outliers.

Metrics in Functional Boxplots

- **Median Curve (Black Line):** Represents the central trend of the functional data.
- **Envelope (Shaded Region):** Captures 50% of the data, highlighting typical variations around the median.
- **Whiskers (Blue Lines):** Represent the range of the inner 75% of the data, excluding potential outliers.

– **Outliers (Red Dashed Lines):** Identify curves that deviate significantly from the main dataset, based on their depth values.

Variable Analysis

– **cnt (Bike Counts):**
  • Displays two prominent peaks during the day (likely corresponding to morning and evening rush hours).
  • High variability during peak hours, with several outliers (red dashed lines) likely representing anomalously high or low rental counts.
– **t1 (Temperature 1) and t2 (Temperature 2):**
  • Both variables show relatively stable trends throughout the day.
  • Narrower envelopes suggest lower variability in temperature compared to bike counts.
  • Median curves for both variables are close to the center of the envelope, indicating consistent patterns.
– **hum (Humidity):**
  • Significant variability, with a wide envelope across the 24-hour period.
  • Red dashed lines indicate a substantial number of outliers, reflecting fluctuations in humidity that could affect bike rental behavior.
– **windspeed:**
  • Displays relatively stable trends with a few prominent outliers.
  • The median curve indicates typical wind speeds, while variability is less evident compared to counts and humidity.

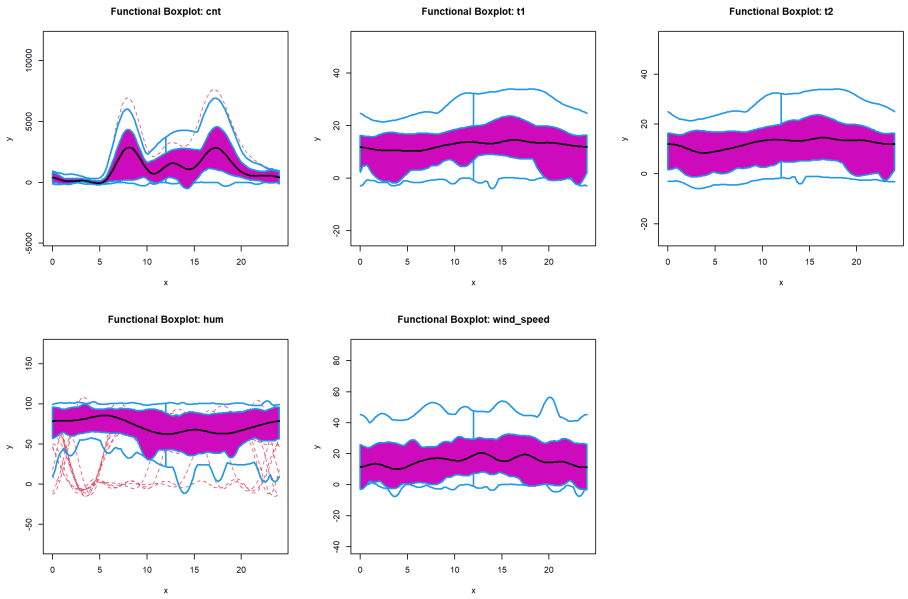

Fig. 21. Functional Boxplots of Key Variables

Depth distribution and Wilcoxon tests

The below plots provide additional insights into the relationships between depth and mean variable values, along side with Wilcoxon tests to compare groups (e.g., holidays vs. non-holidays and seasonal depth variations).

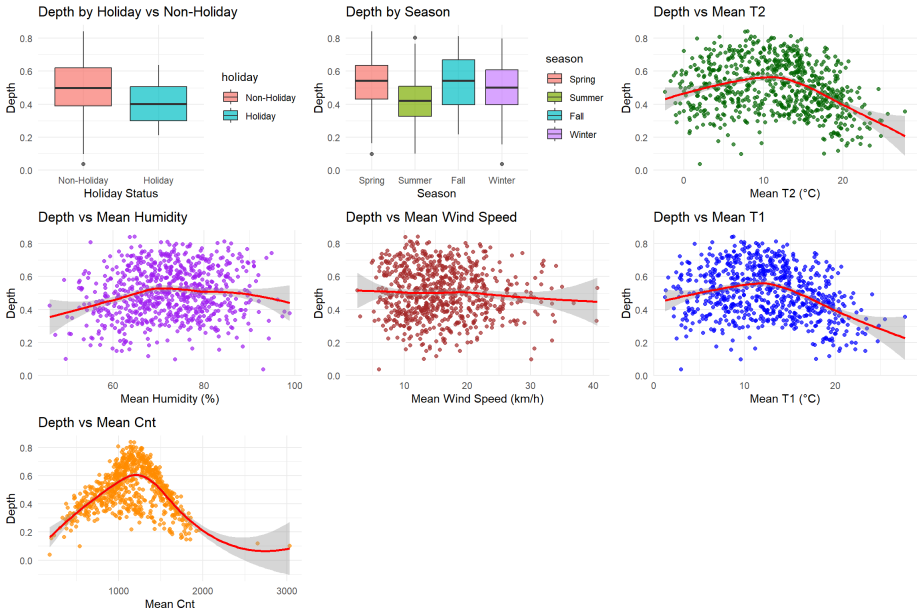

Fig. 22. Depth Relationships

Comparison	**Test Statistic (W)**	**p-value**	**Alternative Hypothesis**
Holidays vs Non-Holidays	3507	0.008084	True location shift is not equal to 0
Summer vs Winter	11845	1.296e-06	True location shift is not equal to 0

Table 6. Wilcoxon test results

Metrics and Comparisons

- **Holiday vs. Non-Holiday Analysis(Boxplot (Top-Left)):**Shows the depth distribution of cnt functional data for holidays versus non-holidays. The Wilcoxon test reveals a significant difference (p=0.008) between the two groups, suggesting depth differs between holiday and non-holiday periods.
- **Seasonal Analysis(Boxplot (Top-Middle)):** Compares depth distributions of cnt functional data across seasons. A Wilcoxon test comparing summer and winter demonstrates a highly significant difference, indicating seasonality influences depth(additional pair comparison is not performed but could be further analyzed).

Mean Variable Values with Depth Value of cnt funcional data

- **Mean Count (Cnt):** Represents user activity (e.g., bike rentals) and its relationship to depth. Depth peaks at moderate activity levels, suggesting that typical days have balanced activity, while very high or low activity levels are less central or more atypical.
- **Mean Temperature 1 (T1):** Shows a clear non-linear trend with depth. Moderate temperatures have higher depth values, aligning with more typical days, while extreme temperatures (either low or high) correspond to lower depth with higher uncertainty.
- **Mean Temperature 2 (T2):** Similar to T1.
- **Mean Humidity:** Displays a curvature trend, where moderate humidity levels are associated with higher depth values, making these days more central. Extremely low or high humidity values are linked with lower depth and uncertainty, indicating atypical conditions.
- **Mean Wind Speed:** Exhibits a slight negative relationship with depth, where higher wind speeds are less typical (lower depth), while lower wind speeds are more central (higher depth).

Deepest Curves Analysis

In functional data analysis, depth measures provide a center-outward ranking of curves, facilitating robust statistical analyses such as identifying medians and detecting outliers. This approach allows for the ordering of functional data from the center outward, enabling the definition of robust statistics like medians and trimmed means, and serves as a foundation for outlier detection and classification methods. [12]

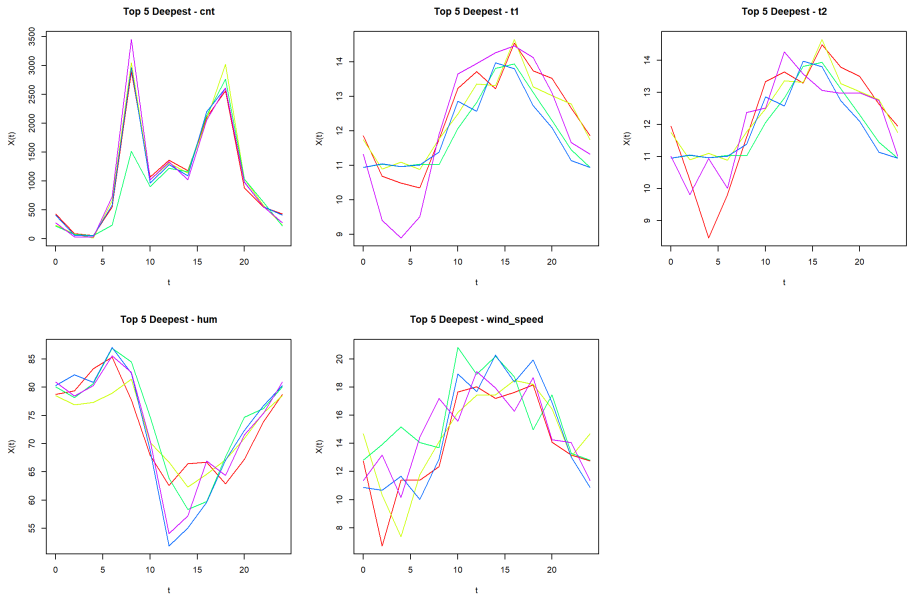

Fig. 23. Deepest Curves Analysis

The plots display the top 5 deepest (most central) curves for variables as "cnt," "t1," "t2," "humidity," and "wind speed." These curves represent the most typical patterns for each variable across hours of the day, encapsulating the core structure and distant from outlier influence.

5.6 Clustering Based on FPCA Scores

Applying k-means clustering to Functional Principal Component Analysis (FPCA) scores effectively groups days into clusters with distinct characteristics.

This approach combines the strengths of functional data analysis and clustering techniques to identify patterns in time series data. By representing complex temporal data through FPCA scores, which capture the main modes of variation, k-means clustering can partition the data into meaningful groups. This methodology has been successfully applied in various studies, such as the clustering of multidimensional curves with spatio-temporal structures.[13]

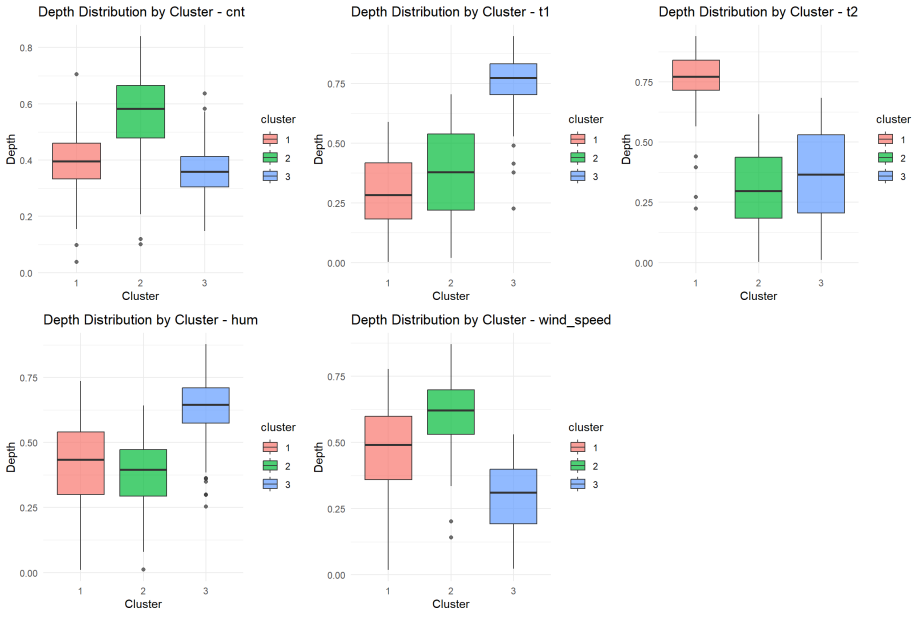

Fig. 24. Clustering Based on FPCA Scores

- **Depth Distribution by Cluster - cnt:** The boxplot shows the variability in depth within each cluster, with Cluster 2 showing the highest median depth and a broader spread compared to Cluster 1 and Cluster 3.
- **Depth Distribution by Cluster - t1:** The depth distribution for temperature ("t1") across clusters highlights distinct patterns. Cluster 3 has the highest median depth and lower spread, while Cluster 1 shows lower median compared to the others. Cluster 2 has slightly higher median value than cluster 1 and most spread among all.
- **Depth Distribution by Cluster - t2:** The depth variability for temperature ("t2") across clusters indicates that Cluster 1 exhibits the highest median depth, while Cluster 3 shows a wider range in values, reflecting greater variability.
- **Depth Distribution by Cluster - hum:** Clusters for humidity ("hum") display distinct depth characteristics, with Cluster 2 having the lowest median depth and Cluster 3 showing highest median narrower spread of depth values.
- **Depth Distribution by Cluster - wind_speed:** The depth distribution for wind speed reveals significant clustering, with Cluster 2 showing higher median depth values and narrower range, while Cluster 3 exhibits lowest median compared to the others.

General Conclusion: The depth distributions across clusters for each variable reveal unique clustering patterns, highlighting distinct characteristics for

user activity, temperature, humidity, and wind speed. These insights indicate meaningful variability in the data, suggesting that FPCA-based clustering captures certain central patterns and variations across functional profiles. Further indepth analysis could be performed to understand the inner group dynamics.

6 Conclusion

This study utilized Functional Data Analysis (FDA) to uncover and confirm temporal patterns and the influence of external factors on the hourly rental count variable in London's bike-sharing data. The analysis revealed distinct daily periodicities, significant seasonal and weather effects, and behavioral changes on holidays. Below are some points to highlight:

- Strong daily rental peaks during commuting hours, with variations influenced by seasons and holidays.
- Significant seasonal differences in general, with higher usage in warmer periods.
- Depth analysis identified anomalous patterns, offering a tool for detecting outliers, especially on count and humidity variables.
- Clustering identified distinct subgroups based on FPCA scores, which could be further analyzed to derive patterns.
- Deeper analysis on regression outliers, depth analysis outliers, could offer further insights on urban mobility planning and rental resource allocations.

These findings can inform or support decisions on operational strategies (e.g., bike redistribution, predicting seasonal demands, understanding holiday or weather-driven shifts), infrastructure planning, and policy decisions for different stake holders. The methods and results demonstrate the power of FDA in analyzing urban mobility patterns.

7 Further reasearch

In general, FDA could have various approaches despite the nature difference of underlying dataset. For example, dynamically choosing best basis function number for fourier smoothing, choosing b spline smoothing and registration, various choice of maximum number of principal components, choosing different function(changing variable and factor sets, choosing function on scalar or scalar on function regressions, etc) to perform regression, thouroughly compare all factor variable pairs' effect(winter vs fall, etc). This project could be expanded to many different combination of techniques and choices of tools used. The current project has been focusing on the completeness of FDA workflow instead of comprehensive analysis and ehausting all analytical techniques. FDA could

be utilized in many different pratical analysis cases, and its potential could be further explored.

Appendix

R code as in attchment and also github repo with additional plots. Metadata:

- `timestamp`: Hourly timestamp.
- `cnt`: Hourly bike rental count.
- `t1`, `t2`: Real and "feels like" temperatures (°C).
- `hum`: Humidity (%).
- `windspeed`: Wind speed (km/h).
- `weathercode`: Weather categories (1:Clear,2:Scattered Clouds,3:Broken Clouds, 4:Cloudy,7:Rain,10:Thunderstorm, 26:Snowfall,94:Freezing Fog).
- `isholiday`: Indicator if the day is a holiday.
- `isweekend`: Indicator if the day is a weekend.
- `season`: 0:Spring,1:Summer,2:Fall,3:Winter.

References

1. https://www.kaggle.com/datasets/hmavrodiev/london-bike-sharing-dataset
2. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer (2005).
3. Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis. Springer (2002).
4. The R Project for Statistical Computing, `https://cran.r-project.org`.
5. CRAN Task View: Functional Data Analysis, `https://cran.r-project.org/web/views/FunctionalData.html`.
6. Ramsay, J.O., Hooker, G., Graves, S.: `fda` package on CRAN, `https://cran.r-project.org/web/packages/fda/index.html`.
7. Università degli Studi di Milano, Functional and Topological Data Analysis course materials, `https://www.unimi.it/en/education/degree-programme-courses/2025/functional-and-topological-data-analysis`.
8. Ramsay, J.O., Hooker, G., Graves, S.: Functional Data Analysis with R and Matlab. Springer (2009).
9. Silverman, B.W.: Smoothing Techniques for Curve Estimation. Springer (1996).
10. Cuevas, A., Febrero, M., Fraiman, R.: Robust Estimation and Classification for Functional Data via Projections. Computational Statistics (2007).
11. Piter, C., Särkkä, S., Kaski, S.: Functional Spatiotemporal Models for Urban Mobility. *arXiv:2003.12041* (2020).
12. Sguera, C., López-Pintado, S. (2021). A notion of depth for sparse functional data. *TEST*, 30, 630–649.
13. Adelfio, G., Di Salvo, F., Chiodi, M. (2018). Space-Time FPCA Clustering of Multidimensional Curves. In: Perna, C., Pratesi, M., Ruiz-Gazen, A. (eds) Studies in Theoretical and Applied Statistics. SIS 2016. Springer Proceedings in Mathematics & Statistics, vol 227. Springer, Cham.