Statistics Webinar Series #1

# Functional Data Analysis for Energy Consumption Modelling

Friday, 12 June 2020

Novri Suhermi (n.suhermi@lancaster.ac.uk)
PhD in Applied Social Statistics, Department of Mathematics and Statistics

In collaboration with: Juhyun Park, Jan Bastiaans, Dénes Csala

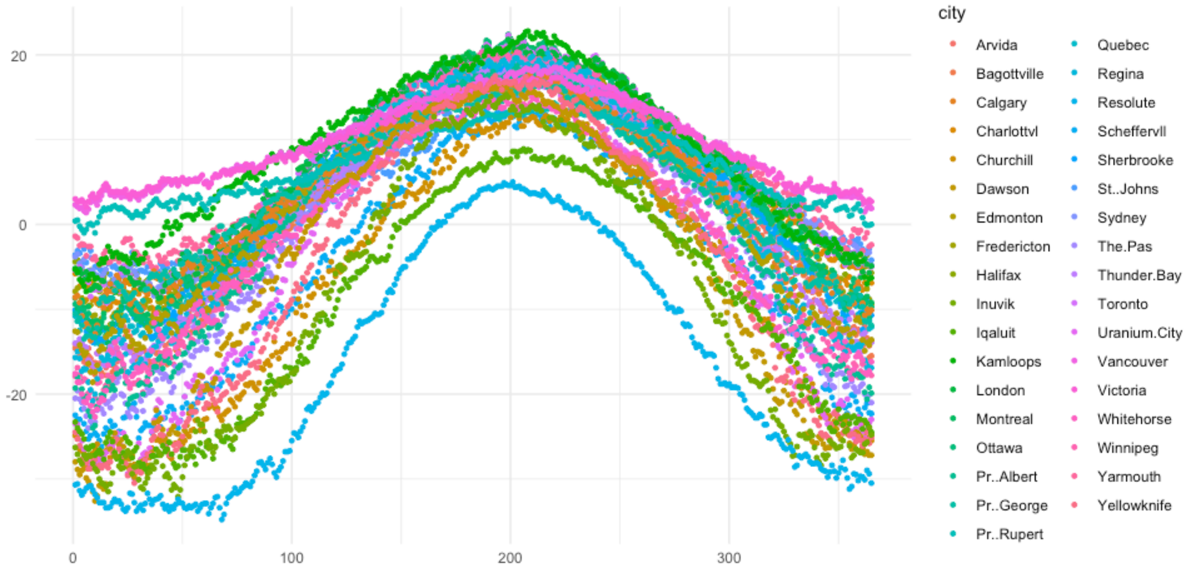
# Outline

- What is functional data?
- Representing functional data
- Exploring functional data
- Functional regression model
- Application to energy
- Challenges and other topics

# What is functional data?

# Functional data

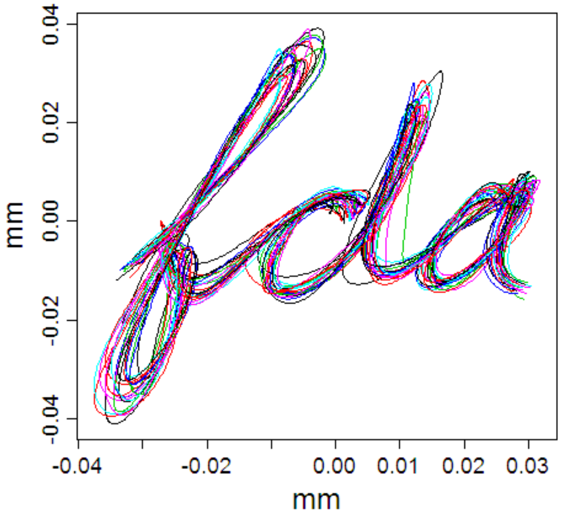
- Quantity
- Frequency
- Similarity
- Smoothness



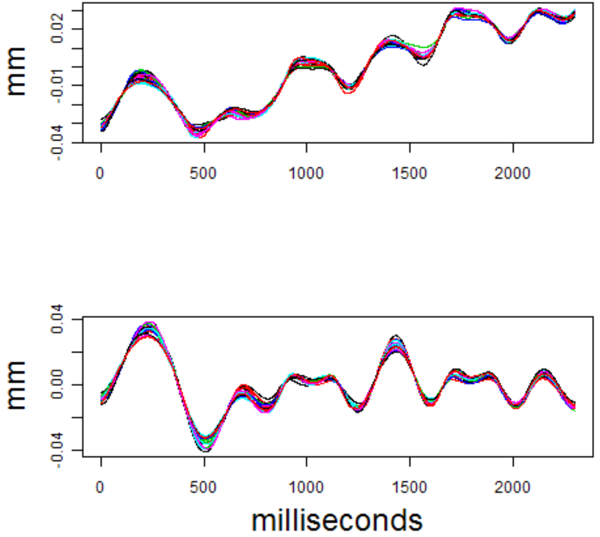

Daily temperatures in Canada

# Functional data






2-dimensional handwriting data




1-dimensional handwriting data

- 20 replications
- 1401 observations within replications
- High-frequency measurements
- Smooth, but complex processes
- Repeated observations
- Multiple dimensions

# Functional data

Functional data is multivariate data with an ordering on the dimensions (Müller, 2006).

Key assumption is *smoothness*:

$$x_{ij} = x_i(t) + \varepsilon_{ij}$$

with $t$ is continuous time, $x_i(t)$ are smooth.

Functional data = the functions $x_i(t)$.

Functional data analysis (FDA) = analysis of data that are functions.

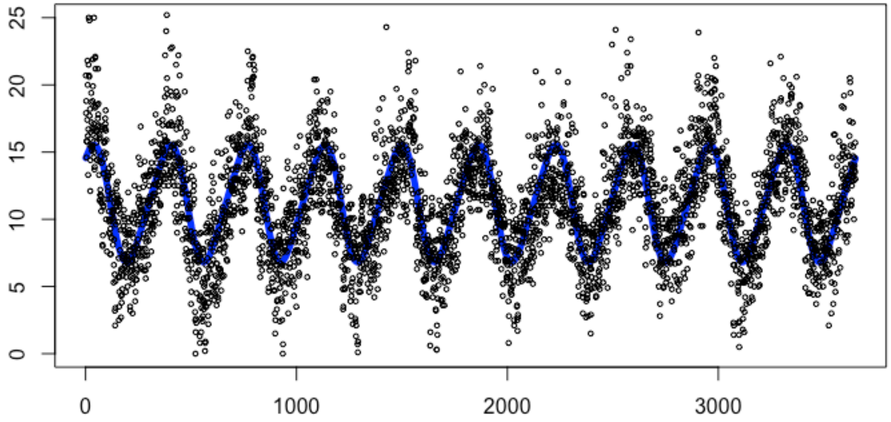
# Necessities for Functional Data

- Must believably derive from a smooth process.
- Process should not be easily parameterizable (should not be able to write down a formula).
- Enough data to resolve the essential features of the process.
- Some repetition in the process.
- Do not need equally-spaced or perfect measurements.

# Representing functional data

# From discrete to functional data

Represent data recorded at discrete times as continuous function in order to




Daily temperatures over 10 years in Melbourne

- Allow evaluation of record at any time point
- Evaluate rates of change
- Reduce noise
- Allow registration onto a common time-scale

This process is called smoothing.

# From discrete to functional data

Simplest dataset form in FDA

$$x_n(t_{j,n}) \in \mathbb{R}, \quad t_{j,n} \in [T_1, T_2], \quad n = 1, 2, \ldots, N, \quad j = 1, \ldots, J_n$$

- $N$ curves are observed on a common interval $[T_1, T_2]$
- The values of the curves are never known at all point $t \in [T_1, T_2]$
- They are available only at some specific points $t_{j,n}$

# From discrete to functional data

The object in FDA are smooth curves

$$\{x_n(t) : t \in [T_1, T_2], n = 1, 2, \ldots, N\}$$

- The values $x_n(t)$ exist at any point $t$, but observed only at selected points $t_{j,n}$
- Typically expressed in basis expansion

# From discrete to functional data

Consider only one record

$$y_j = x(t_j) + \varepsilon_j$$

represent $x(t)$ as

$$x(t) = \sum_{k}^{K} c_k \phi_k(t) = \Phi(t)\mathbf{c}$$

where $\Phi(t)$ is called basis system and $\mathbf{c}$ is a coefficient vector.

Kokoszka & Reimherr (2017)

# Smoothing functional data by least squares

Recall

$$x(t) = \sum_{k}^{K} c_j \phi_k(t) = \Phi(t)\mathbf{c}$$

Minimize

$$\text{SMSSE}(\mathbf{y}|\mathbf{c}) = \sum_{j=1}^{n} \left[ y_j - \sum_{k}^{K} c_k \phi_k\left(t_j\right) \right]^2 = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c})$$

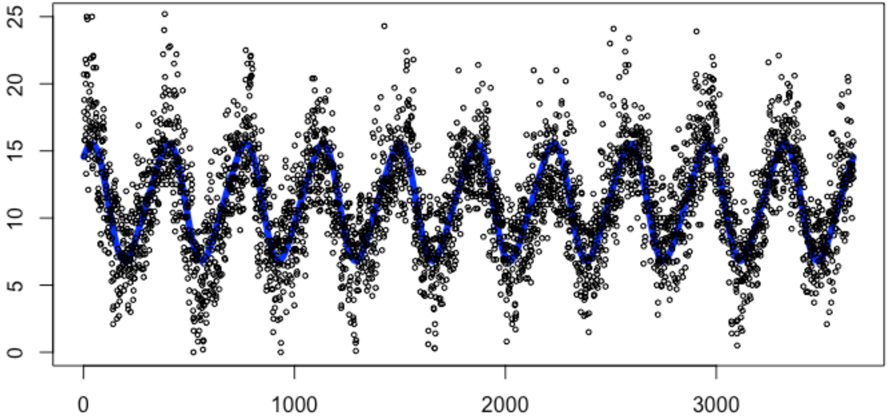
Solution

$$\hat{\mathbf{c}} = \left(\mathbf{\Phi}'\mathbf{\Phi}\right)^{-1} \mathbf{\Phi}'\mathbf{y}$$

# Basis expansions

## Fourier basis (for periodic data)

$$x(t) = c_1 + c_2 sin(\omega t) + c_3 cos(\omega t) + c_4 sin(2\omega t) + c_5 cos(2\omega t) + \ldots$$

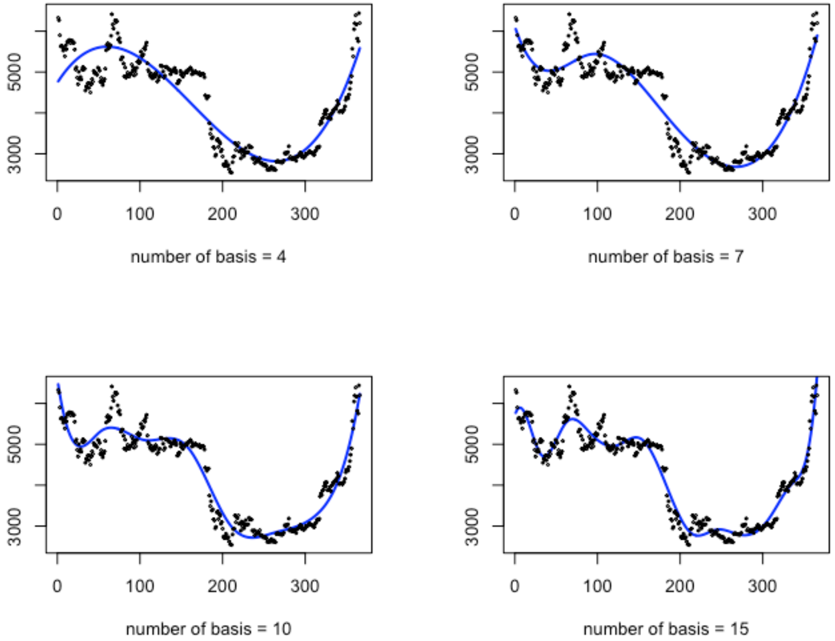



Smoothing using Fourier Basis

# Basis expansions

## B-Spline basis

$$x(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau)$$




Smoothing using B-Spline

# Basis expansions

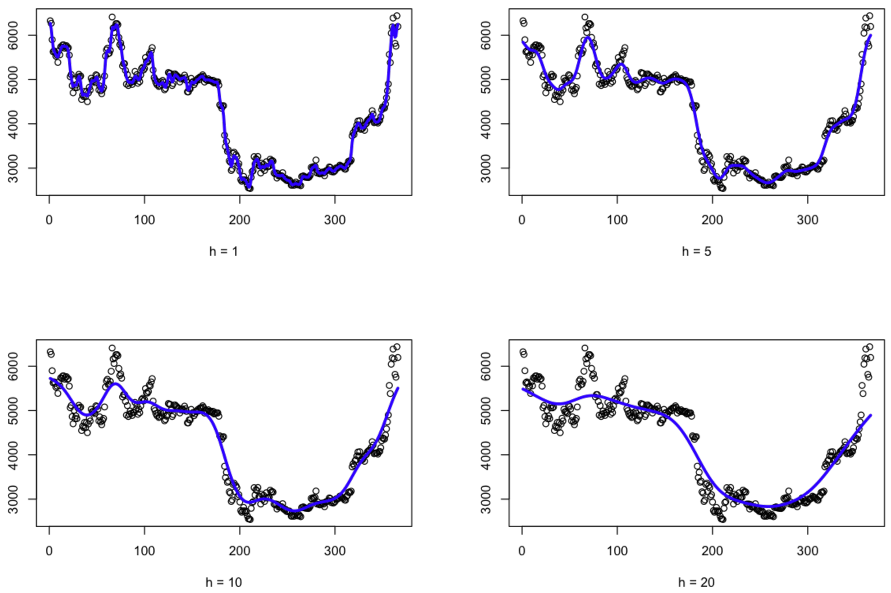
## Kernel smoothing (local estimator)

$$x(t) = \sum_{j=1}^{n} S_j(t) y_j$$

where $S(t)$ is a Nadaraya-Watson kernel estimator given by

$$S_j(t) = \frac{\text{Kern}[(t_j - t)/h]}{\sum_r \text{Kern}[(t_r - t)/h]}$$

and $\text{Kern}(.)$ is a kernel function.




Kernel smoothing for Bitcoin price

# Exploring functional data

# Summary statistics for functional data

## Functional mean

$$\bar{x}(t) = N^{-1} \sum_{i=1}^{N} x_i(t)$$

## Functional covariance

$$\text{cov}_X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t_1) - \bar{x}(t_1)][x_i(t_2) - \bar{x}(t_2)]$$

## Functional variance

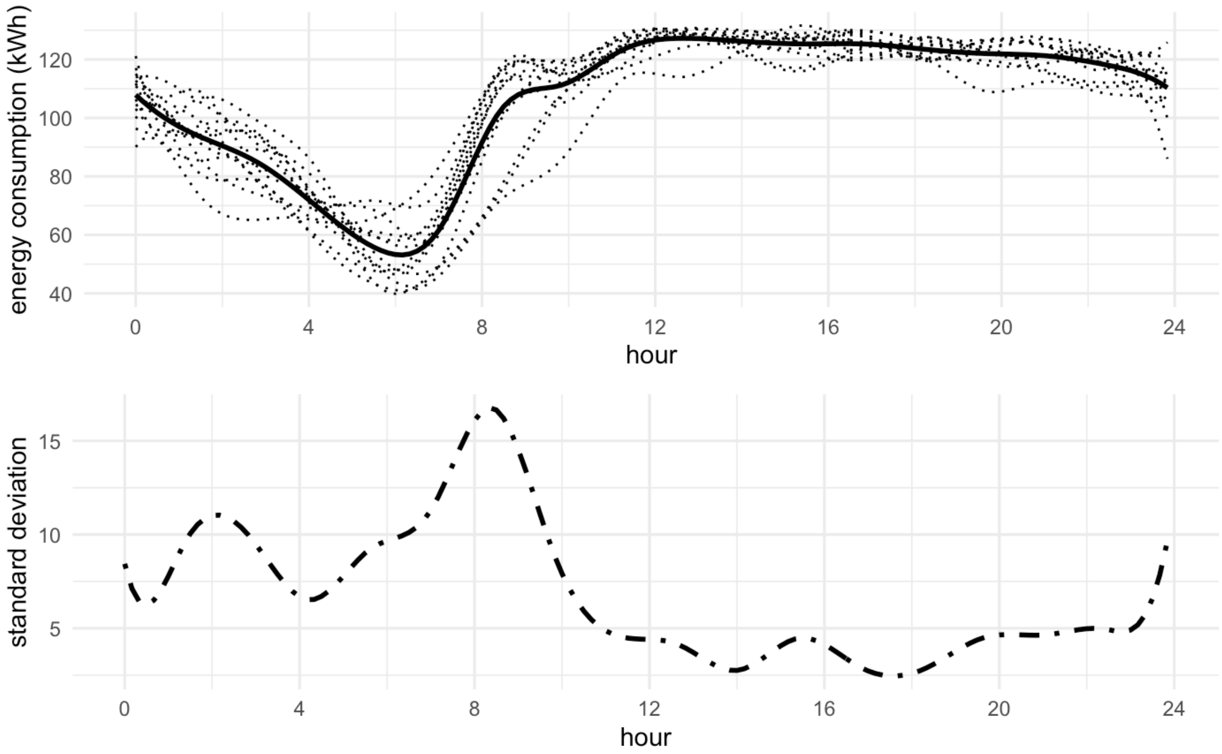
$$\text{var}_X(t) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t) - \bar{x}(t)]^2$$

## Functional correlation

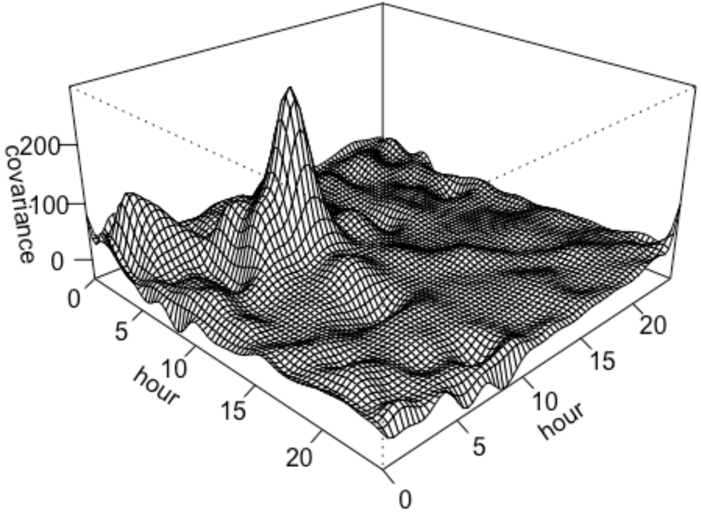
$$\text{cov}_X(t_1, t_2) = \frac{\text{cov}_X(t_1, t_2)}{\sqrt{\text{var}_X(t_1)\text{var}_X(t_2)}}$$

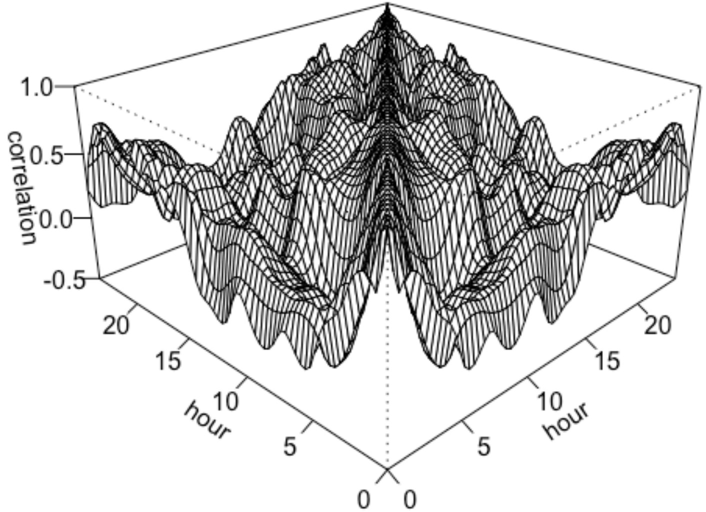
Ramsay & Silverman (2005)

# Functional mean and variance example



# Functional covariance and correlation example



# Functional regression model

# Functional regression model

**Simple linear regression model for non-functional data**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where $\beta_0$ and $\beta_1$ are unknown parameters.

**Functional concurrent regression model for functional data**

$$Y_i(t) = \beta_0(t) + \beta_1(t) X_i(t) + \varepsilon_i(t)$$

The coefficients $\beta_0$ and $\beta_1$ vary over time. This model is also called time-varying coefficient model, introduced by Hastie and Tibshirani (1993).

# Estimation using local kernel method

Minimize

$$\arg \min_{\boldsymbol{\beta}(t),\boldsymbol{\beta}(t)^{(1)}} \sum_{i=1}^{T} \left[ y(t) - \mathbf{X}(t)'\boldsymbol{\beta}(t) - (t_i - t)\mathbf{X}(t)'\boldsymbol{\beta}(t)^{(1)} \right]^2 K_b(t_i - t)$$

Solution

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}(t) \\ \hat{\boldsymbol{\beta}}(t)^{(1)} \end{pmatrix} = \begin{pmatrix} S_{T,0}(t) & S_{T,1}^{\top}(t) \\ S_{T,1}(t) & S_{T,2}(t) \end{pmatrix}^{-1} \begin{pmatrix} T_{T,0}(t) \\ T_{T,1}(t) \end{pmatrix}$$

$$S_{T,s}(t) = \frac{1}{T} \sum_{i=1}^{T} \mathbf{X}(t_i)'\mathbf{X}(t_i)\,(t_i - t)^s\, K_b(t_i - t)$$

$$T_{T,s}(t) = \frac{1}{T} \sum_{i=1}^{T} \mathbf{X}(t_i)'\mathbf{X}(t_i)\,(t_i - t)^s\, K_b(t_i - t)\, y(t_i)$$

Wu & Zhang (2006)

# Other functional regression models

## Total effect

$$Y_i(t) = \beta_0(t) + \int_T X_i(s)\beta_1(s,t)ds + \varepsilon_i(t)$$

## Cumulative effect

$$Y_i(t) = \beta_0(t) + \int_0^t X_i(s)\beta_1(s,t)ds + \varepsilon_i(t)$$

## Short-term feed-forward effect

$$Y_i(t) = \beta_0(t) + \int_{t-\delta}^t X_i(s)\beta_1(s,t)ds + \varepsilon_i(t)$$

Ramsay & Silverman (2005)

# Diagnostics checking

Goodness-of-fit test

$$\mathrm{H}_0 : \mathbb{E}\left(A_{ek} | A_{X_j}\right) = \mathbb{E}\left(A_{ek}\right) = 0$$ (residual mean is zero)

Test statistics

$$T_0 = \sum_{k=1}^{K_e} \lambda_{ek} T_k / \sum_{k=1}^{K_e} \lambda_{ek}$$

Empirical p-value

$$\tilde{p} = \frac{1}{M} \sum_{m=1}^{M} I\left\{T_0^{(m)*} \geq \hat{T}_0\right\}$$

Chiou & Müller (2007)

# Application to energy

# Background

*This research is focused on establishing the influence of stakeholder groups on building energy consumption. These groups, including building occupants, designers, operators and maintainers, each have different levels of influence over the amount of energy consumed in a building.*

*Understanding the degree and nature of that influence will lead to improved methods for reducing energy consumption in buildings. It will provide vitally important new insights which will help us achieve our aims of reducing consumption and carbon emissions.*

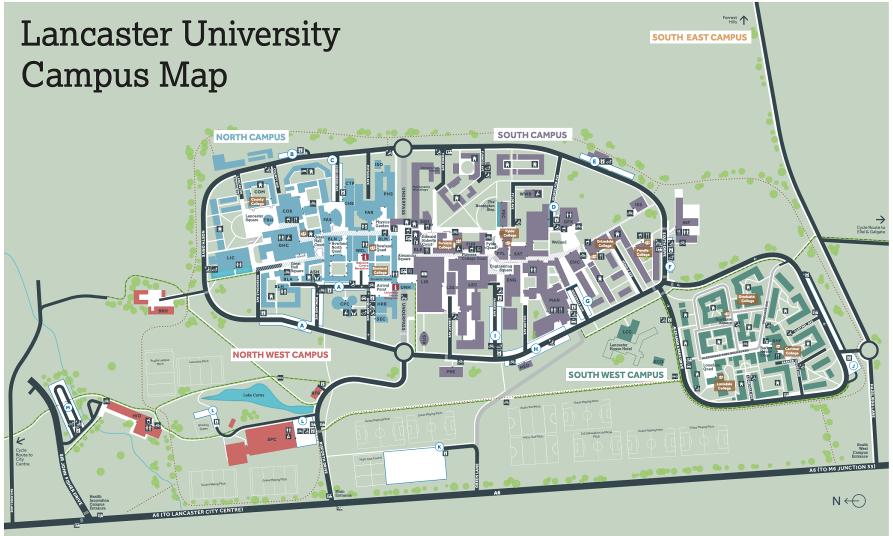
Jan Bastiaans,

Managing Director

Desco Analytics Ltd.

# Available datasets

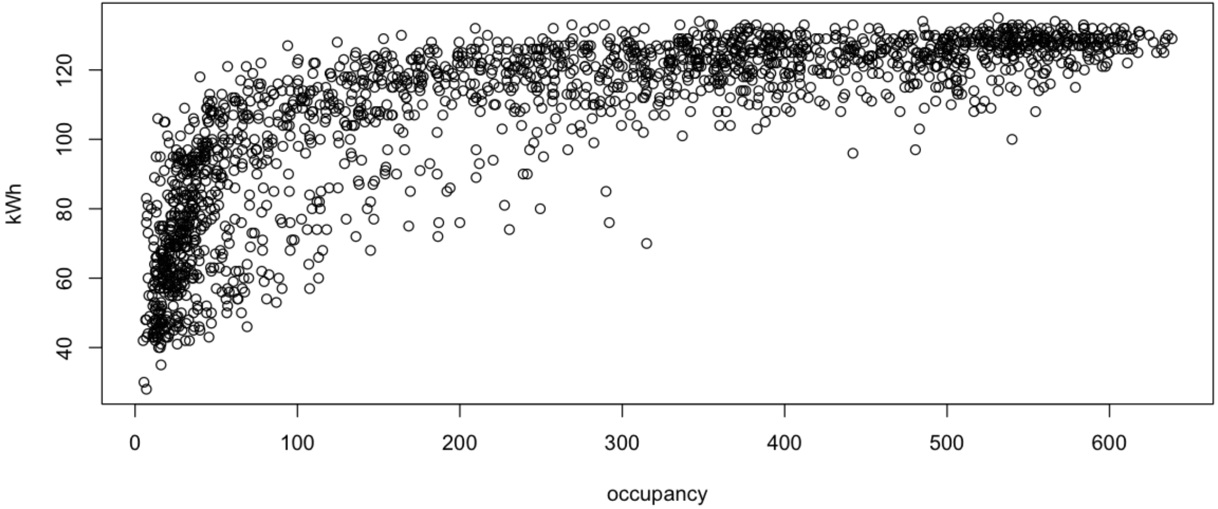
- Over 800 energy meters, measuring energy consumption in over 100 buildings
- Over 30,000 sensors in the Building Management System, including internal temperatures, ventilation and heating.
- Weather data from an on-site meteorology station
- Building occupancy data using WiFi data and room booking data
- Room and building data, using the asset management system



# Dataset and variables (current study)

- Data: lighting consumption (kWh) and building occupancy (person) of university library
- 13 days period (18-30 May 2018)
- Recorded every 10 minutes
- $13 \times 24 \times 6 = 1872$ observations in total
- Response variable ($Y$): lighting consumption
- Predictor variable ($X$): building occupancy

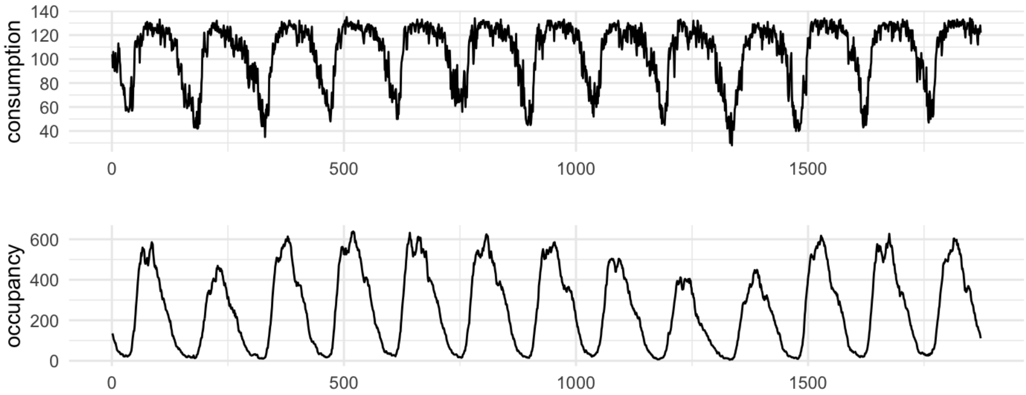
# Scatter plot: ligthing consumption vs occupancy




Scatter plot of lighting consumption vs occupancy

The plot shows that the two variables have nonlinear relationship

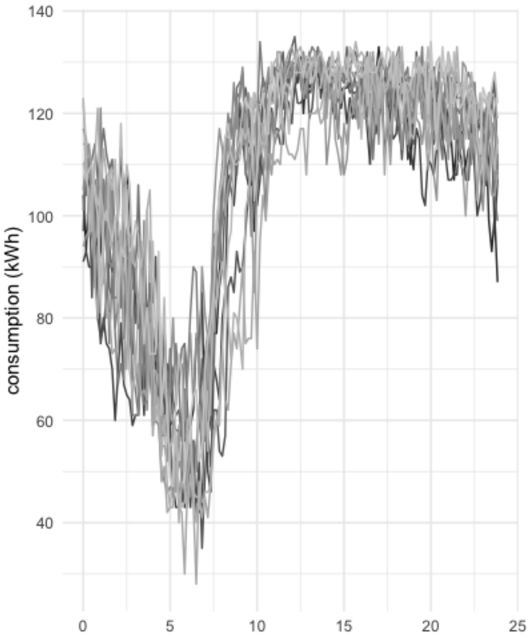
# Time series plot

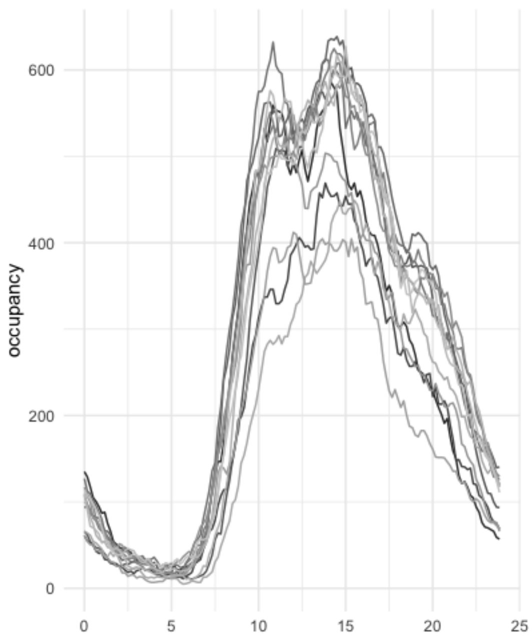



Time series plots of energy consumption and occupancy

- Data has daily seasonal pattern
- The effect of occupancy might vary over time

# Time series plot



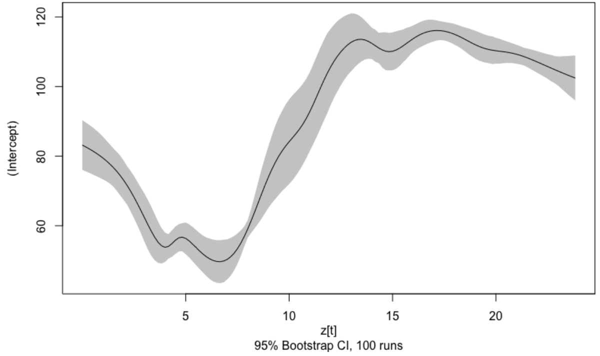



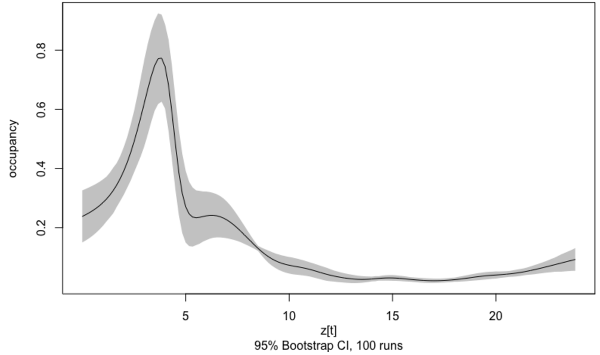

Time series plots of daily data (24 hour scale)

# Functional regression

## Results

```
Call:
NULL

Class: tvlm

Summary of time-varying estimated coefficients:
================================================
        (Intercept) occupancy
Min.          49.67   0.02062
1st Qu.       67.78   0.03017
Median       102.64   0.06756
Mean          89.85   0.15824
3rd Qu.      110.90   0.23767
Max.         116.11   0.77298

Bandwidth:  1
Pseudo R-squared:  0.8845
```



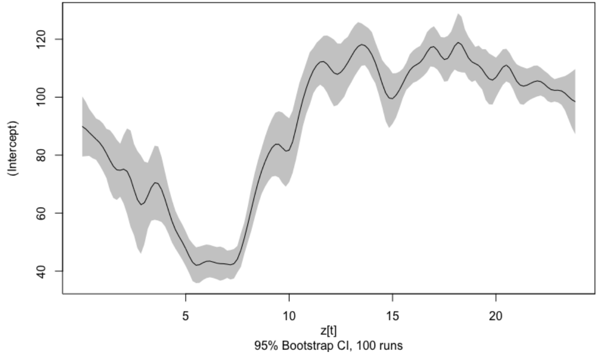



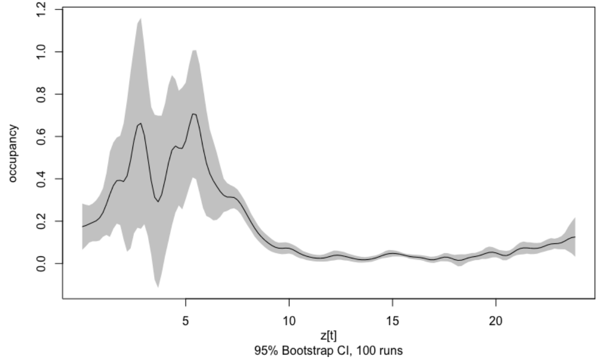

Coefficient plots

# Functional regression

## Results (optimal bandwith)

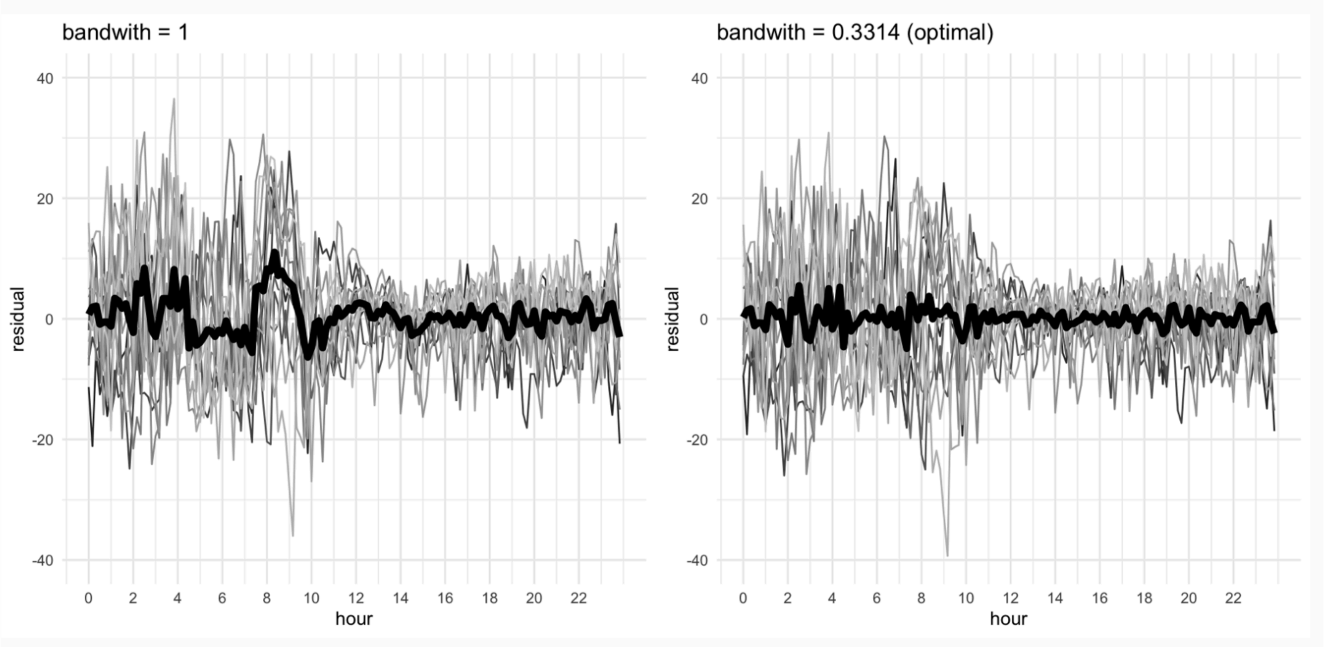
```
Call:
NULL

Class: tvlm

Summary of time-varying estimated coefficients:
================================================
        (Intercept) occupancy
Min.          42.03   0.01477
1st Qu.       70.59   0.03388
Median       101.44   0.07222
Mean          89.66   0.17355
3rd Qu.      110.03   0.30370
Max.         118.90   0.70642

Bandwidth:  0.3314
Pseudo R-squared:  0.9003
```



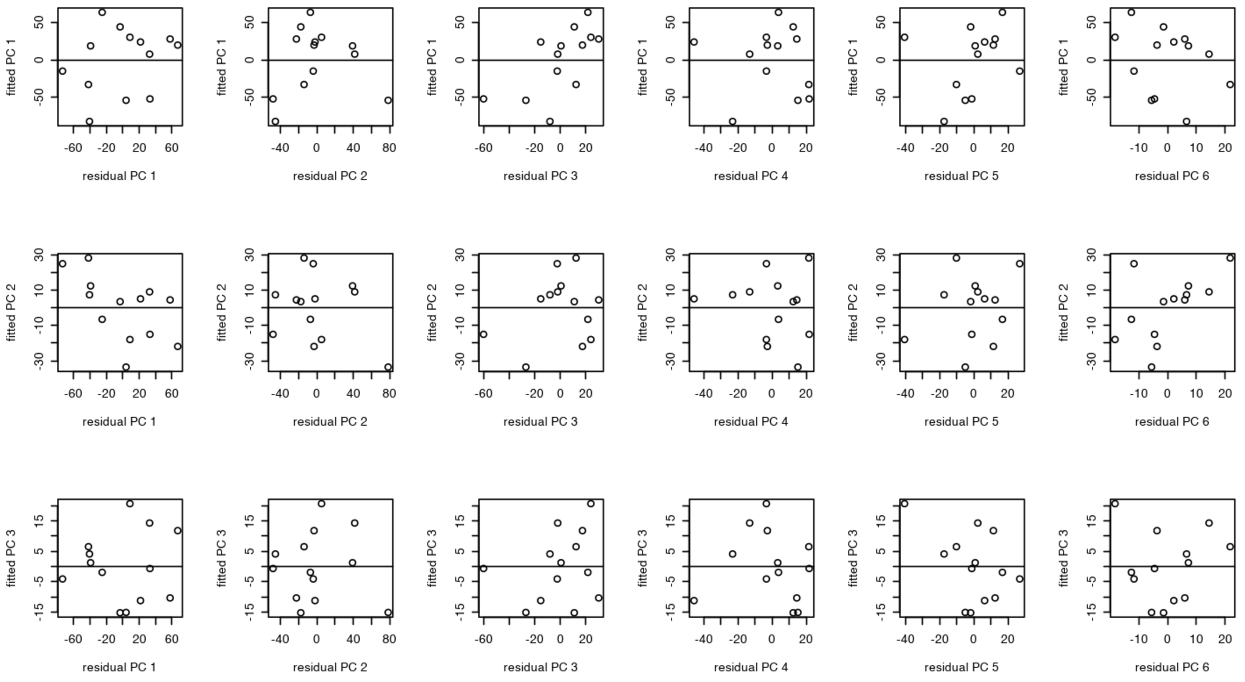

Coefficient plots

# Residual analysis



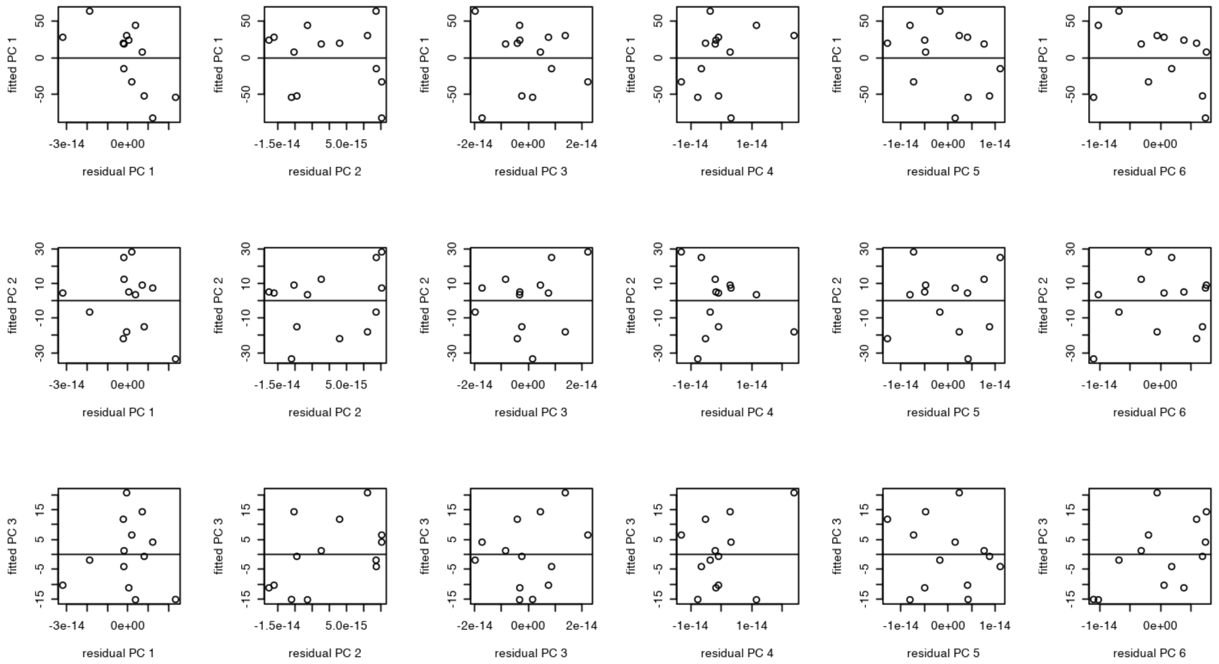

Residual plot

# Residual analysis



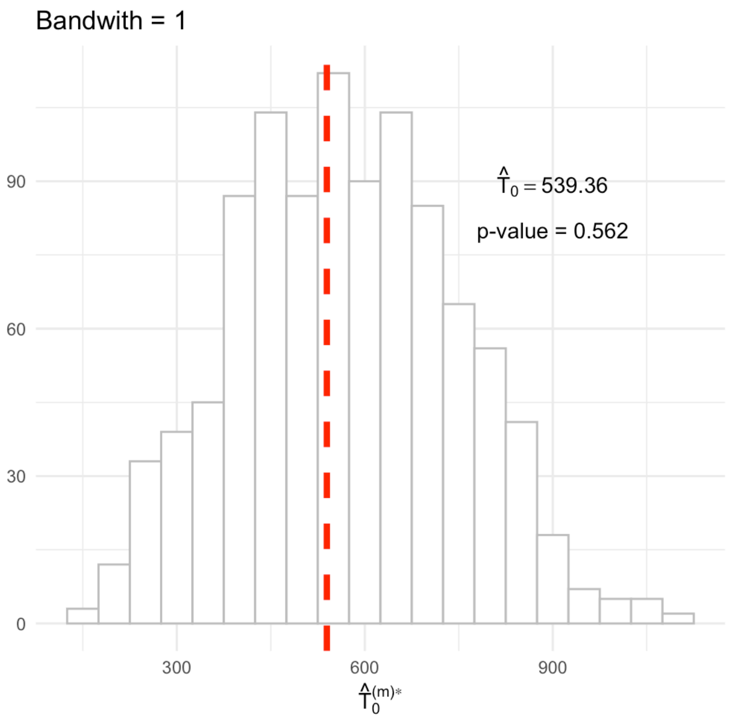

Residual FPC score vs Fitted FPC score (Model bandwidth 1)

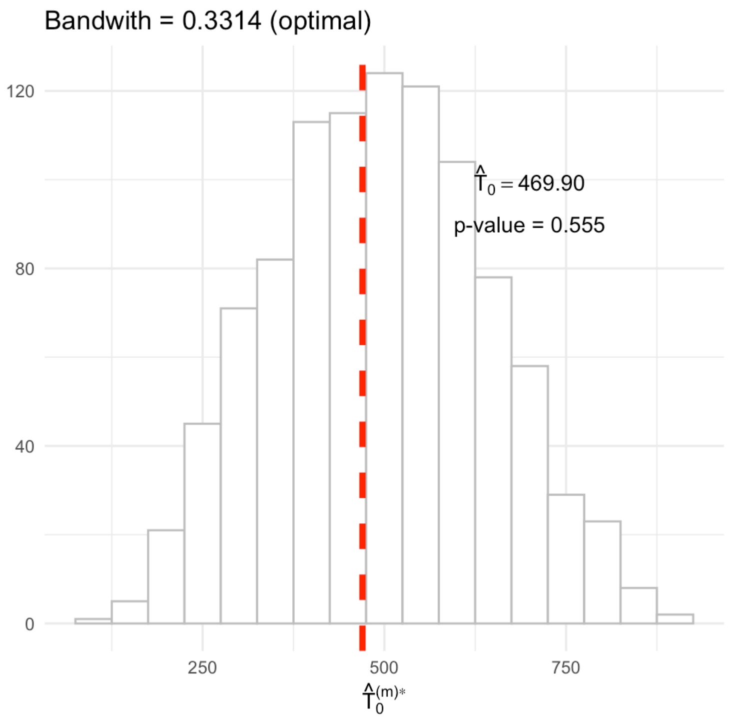
# Residual analysis




Residual FPC score vs Fitted FPC score (Model optimal bandwidth)

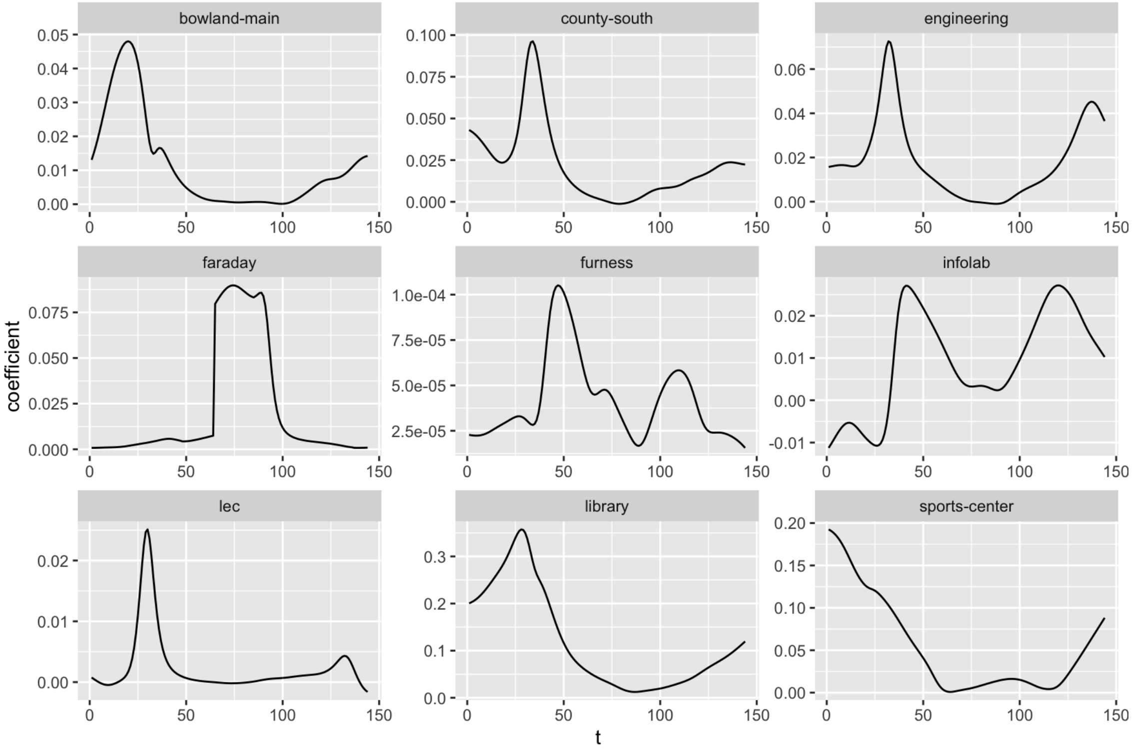
# Goodness-of-fit test




Distribution under null hypothesis

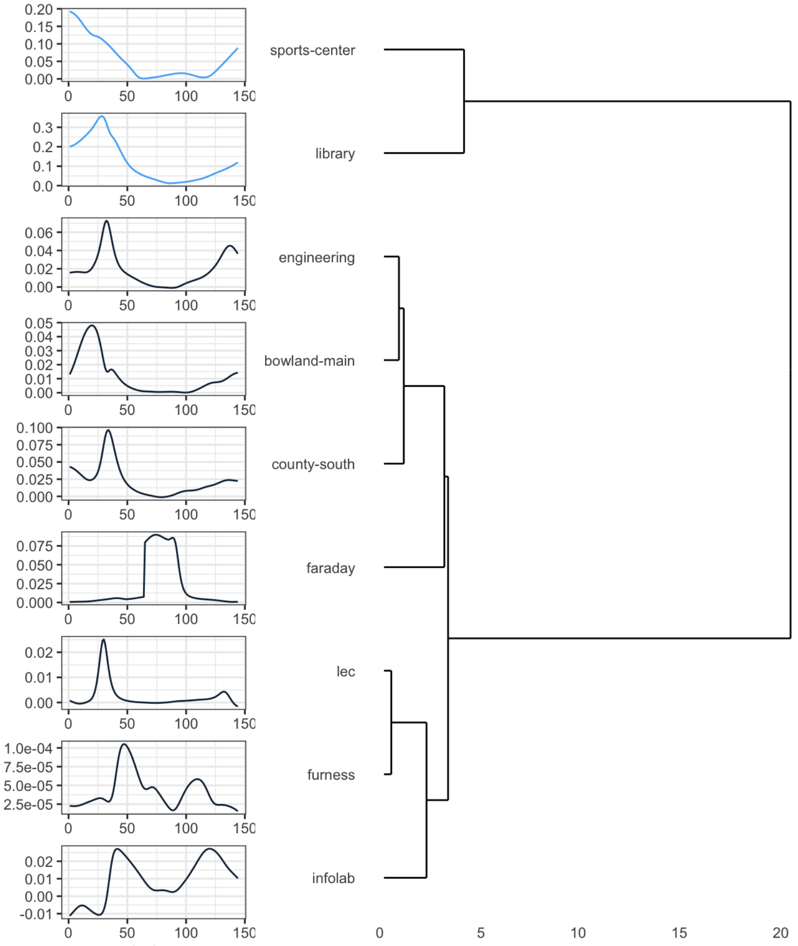
# Results from other buildings




Slope functions (occupancy effect)

Comparison of the regression coefficient function between buildings

# Grouping the buildings

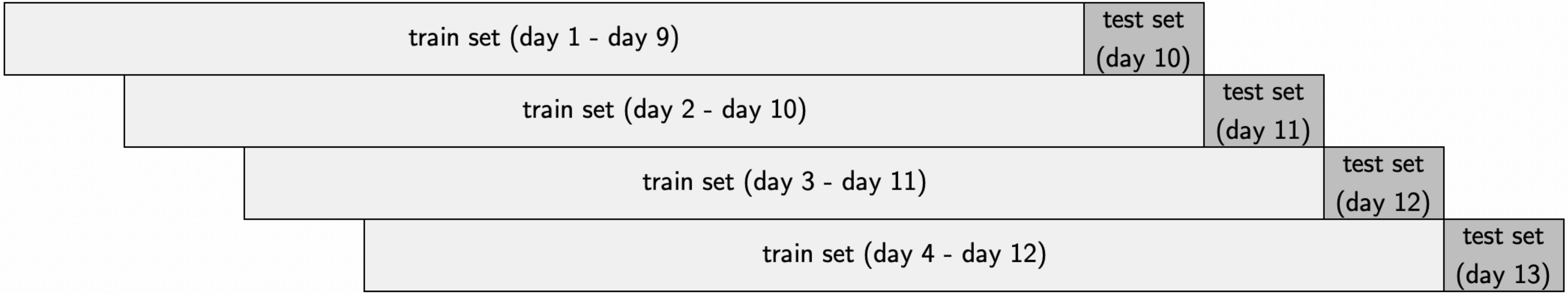


# Forecasting

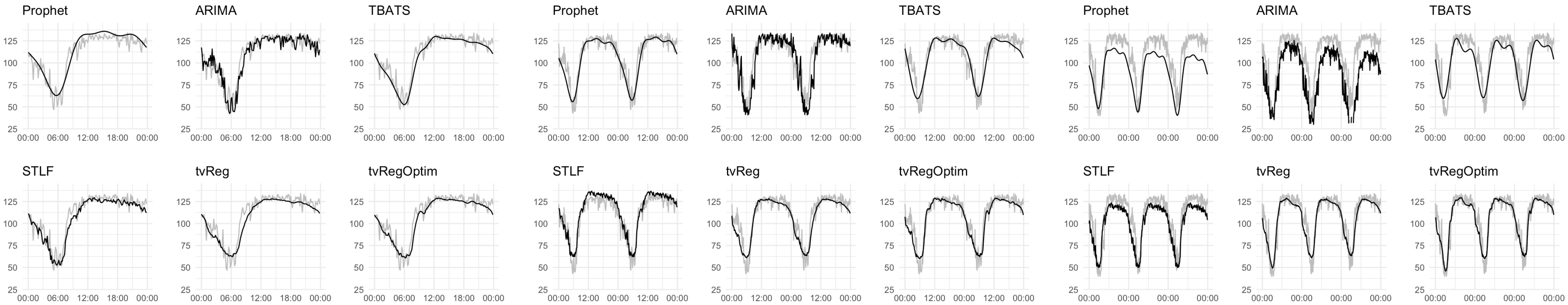
- We perform forecasting and compare the results to several popular time series models
  - Prophet
  - STLF
  - ARIMA
  - TBATS
- Forecast comparison is done by out-of-sample forecast evaluation

# Forecasting

We also perform time series cross-validation with rolling window



# Forecast results




| | Prophet | TBATS | ARIMA | STLF | tvReg | tvRegOptim |
|---|---|---|---|---|---|---|
| **1-day forecast** | | | | | | |
| RMSE | 8.2479 | 7.5574 | 9.8762 | 7.9633 | 8.4773 | 7.5942 |
| MAPE | 0.0653 | 0.0583 | 0.0766 | 0.0611 | 0.0650 | 0.0595 |
| sMAPE | 0.0630 | 0.0604 | 0.0789 | 0.0635 | 0.0659 | 0.0599 |
| **2-day forecast** | | | | | | |
| | Prophet | TBATS | ARIMA | STLF | tvReg | tvRegOptim |
| RMSE | 8.6500 | 10.3034 | 13.6979 | 8.4563 | 8.9744 | 8.3337 |
| MAPE | 0.0667 | 0.0813 | 0.0988 | 0.0735 | 0.0718 | 0.0666 |
| sMAPE | 0.0681 | 0.0814 | 0.1060 | 0.0696 | 0.0721 | 0.0666 |
| **3-day forecast** | | | | | | |
| | Prophet | TBATS | ARIMA | STLF | tvReg | tvRegOptim |
| RMSE | 19.9227 | 13.6121 | 25.9987 | 12.5956 | 9.1497 | 8.7471 |
| MAPE | 0.1705 | 0.1178 | 0.2208 | 0.1063 | 0.0752 | 0.0701 |
| sMAPE | 0.1884 | 0.1194 | 0.2656 | 0.1105 | 0.0748 | 0.0697 |

# Forecast results

Lancaster University

sMAPE scores for time-series cross validation

| Window | Prophet | TBATS | ARIMA | STLF | tvReg | tvRegOptim |
|---|---|---|---|---|---|---|
| 1 | 0.117 | 0.133 | 0.150 | 0.149 | 0.094 | 0.101 |
| 2 | 0.195 | 0.116 | 0.178 | 0.119 | 0.080 | 0.077 |
| 3 | 0.089 | 0.066 | 0.126 | 0.077 | 0.077 | 0.074 |
| 4 | 0.064 | 0.058 | 0.114 | 0.065 | 0.066 | 0.060 |
| Overall mean | 0.116 | 0.093 | 0.142 | 0.103 | 0.079 | 0.078 |

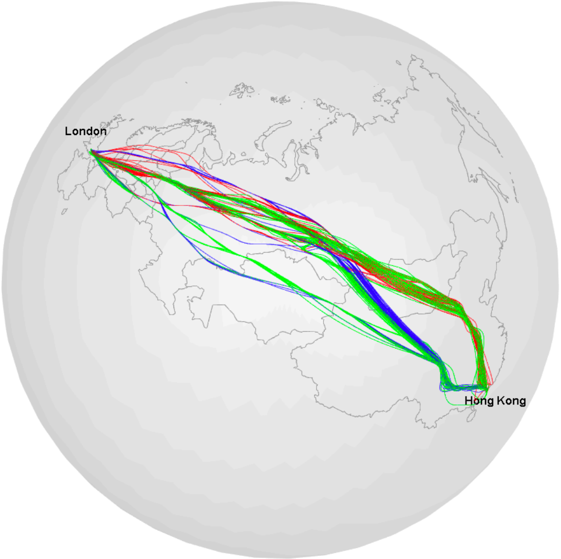
# Challenges and other topics

# Challenges

- Functional data analysis is relatively "new" in statistics
- Adding more covariates
- Hierarchical modeling (university campus → building → floor)
- Dynamic model (with differential equation)
- Functional sparsity (coefficient values are close to zero at certain period)

# Topics in functional data analysis

- Functional time series model
- Bayesian Nonparametric Functional Data Analysis
- Functional data over multidimensional domains (time, space, etc)
- Functional data analysis on non-Euclidean space
- Functional data analysis on Riemannian manifolds and spheres
- etc




https://anson.ucdavis.edu/~mueller

# References


Chiou, J.-M., and Müller, H.-G. Diagnostics for functional regression via residual processes. Computational Statistics & Data Analysis 51, 10 (2007), 4849–4863.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4), 757-779.

Ramsay, J.O., & Silverman, B.W. (2005). *Functional Data Analysis*. Springer, New York.

Hall, P., Müller, H. G., & Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, 34(3), 1493-1517.

Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis*. Chapman and Hall/CRC, Florida.

Wu, H., & Zhang, J. T. (2006). Nonparametric regression methods for longitudinal data analysis: mixed-effects modeling approaches. John Wiley & Sons.

# End of slide