Heterogeneous methods for heterogeneous data?

Anastasia Ushakova

University of Edinburgh

@apavluhina

March 7th, PyData Edinburgh

Big Data

- Big Data as a promising resource for academic and industry research
- Are we clear on how to approach?

Big Data?

- **Volumous**
- **Variable**
- **High Velocity**
- **Exhaustive**
- **Relational**
- **Flexible and Scalable**
- **Ans also: rich, exciting, insightful...**

Well...

This is the same as saying that fruit salad is made of homogeneous fruits which are **juicy, ripe, sweet, colourful!**
So here is the scenario: your friend said 'Get me some fruits for fruit salad!'...

You said fruit?

You said big data?

- Do we all assume big data means the same thing across various settings?
- Or that all big data can be analysed in the same way?

Just for the sake of example

I decided to study the energy use...

Complex and granular data

Smart meter data

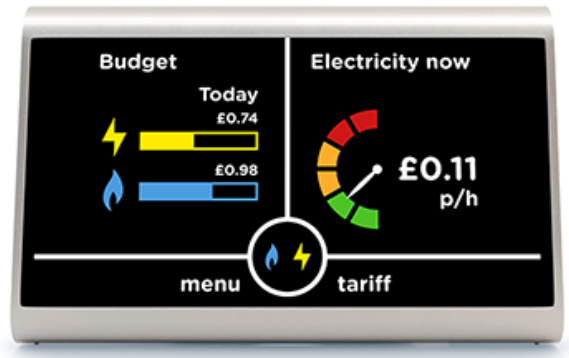


Figure: Smart Meter Display

Data Visualisations

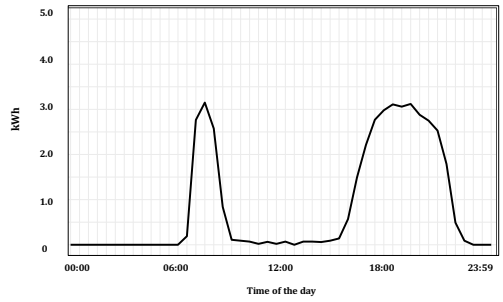


Figure: Example 1: 48 half hourly profile of energy use

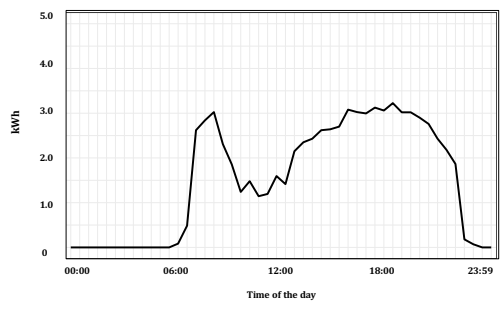
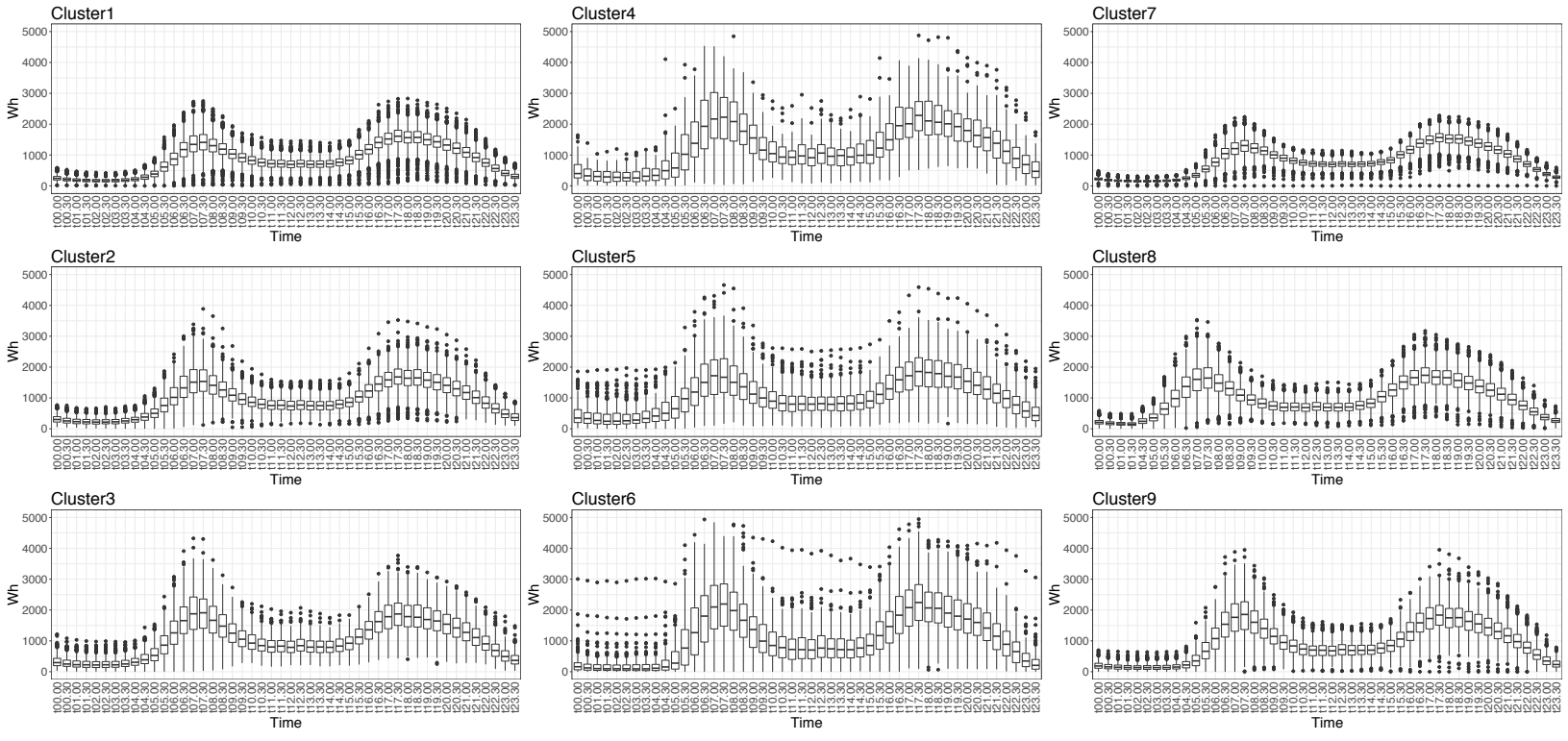


Figure: Example 2: 48 half hourly profile of energy use

I have about 400 k of energy customers with their annual patterns (365 days * 48 readings a day)
So could do with...clustering of course!

Clustering: Gaussian Mixture Models

Aggregated

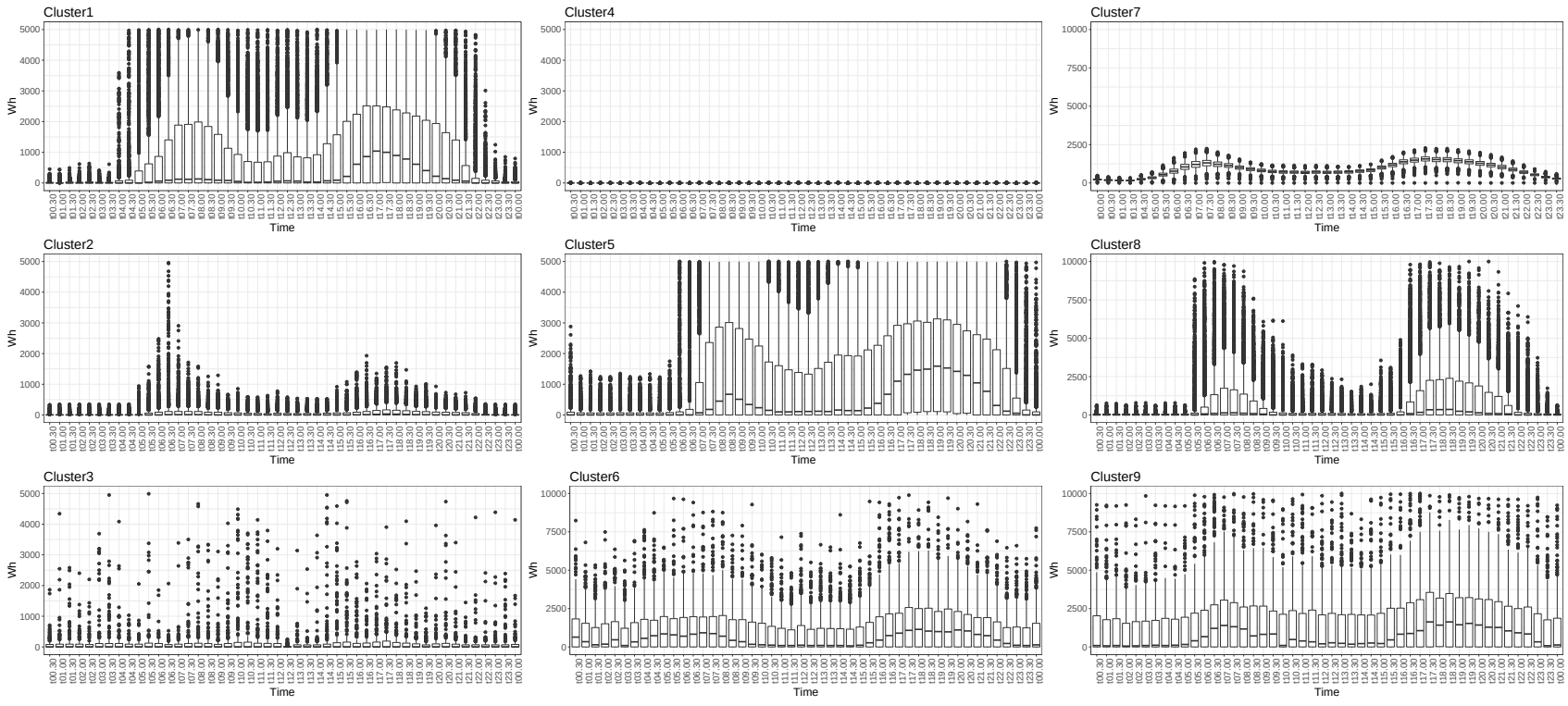


Clustering: Gaussian Mixture Models

Everyone looks the same! What?

Clustering: Gaussian Mixture Models

Disaggregated



What if we want to do something else?

- I want to predict!
- Think Generalised Adittive Models may be a good idea? (for model overview please see Wood (2006))
- The secret I am hiding is that it took me good three years to find out this was appropriate!!!!

Prediction and Regression Analysis

Lets consider some examples...

Prediction and Regression Analysis

GAMs: Experimental Results for Rural Resident

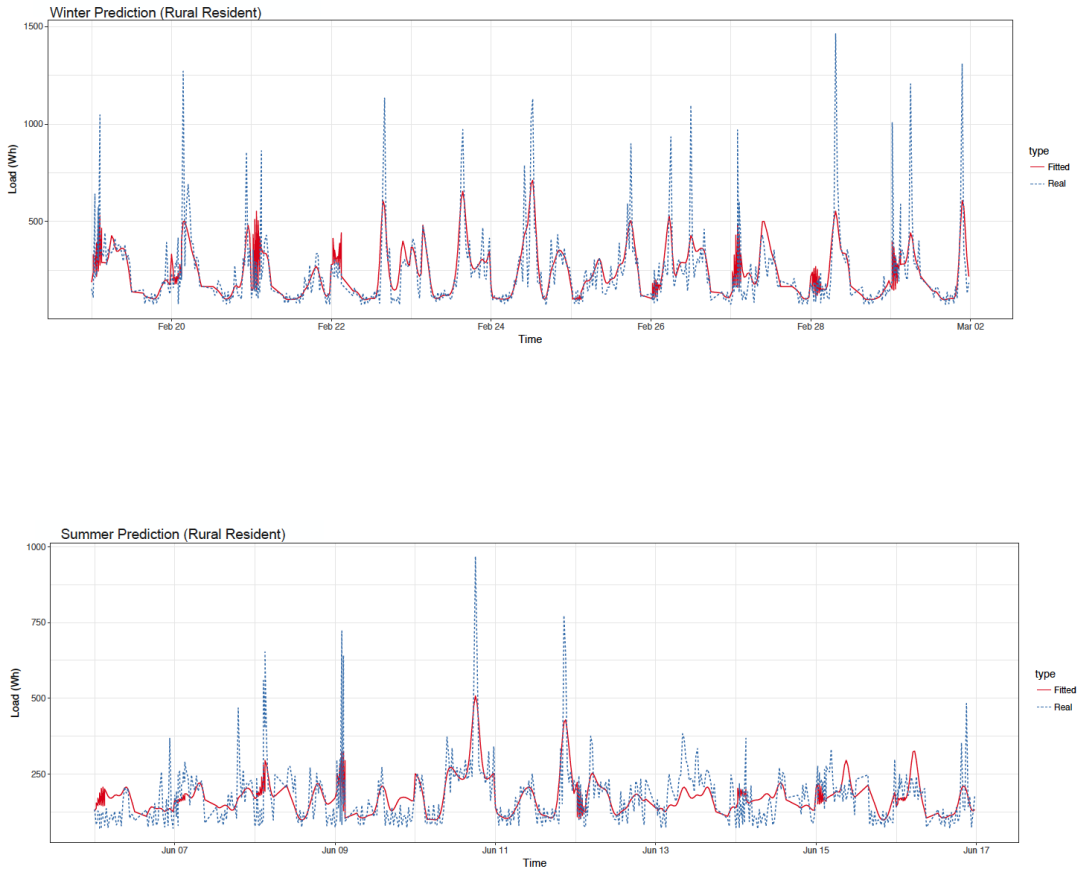


Figure: GAM fit for a customer that belongs to OA characterised as 'Rural Residents'

Prediction and Regression Analysis

GAMs: Experimental Results for Rural Resident

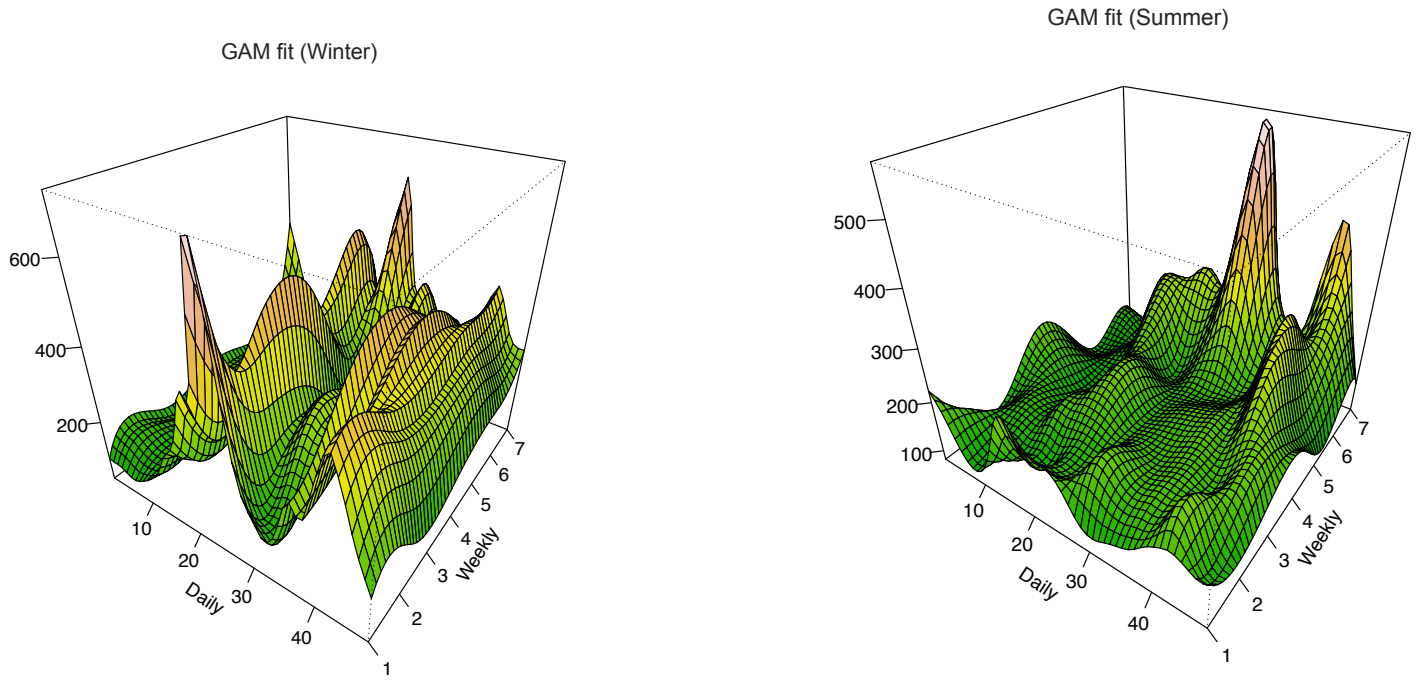


Figure: GAM fit for a customer that belongs to OA characterised as 'Rural Resident' in 3D.

Prediction and Regression Analysis

GAMs: Experimental Results for Urban Resident

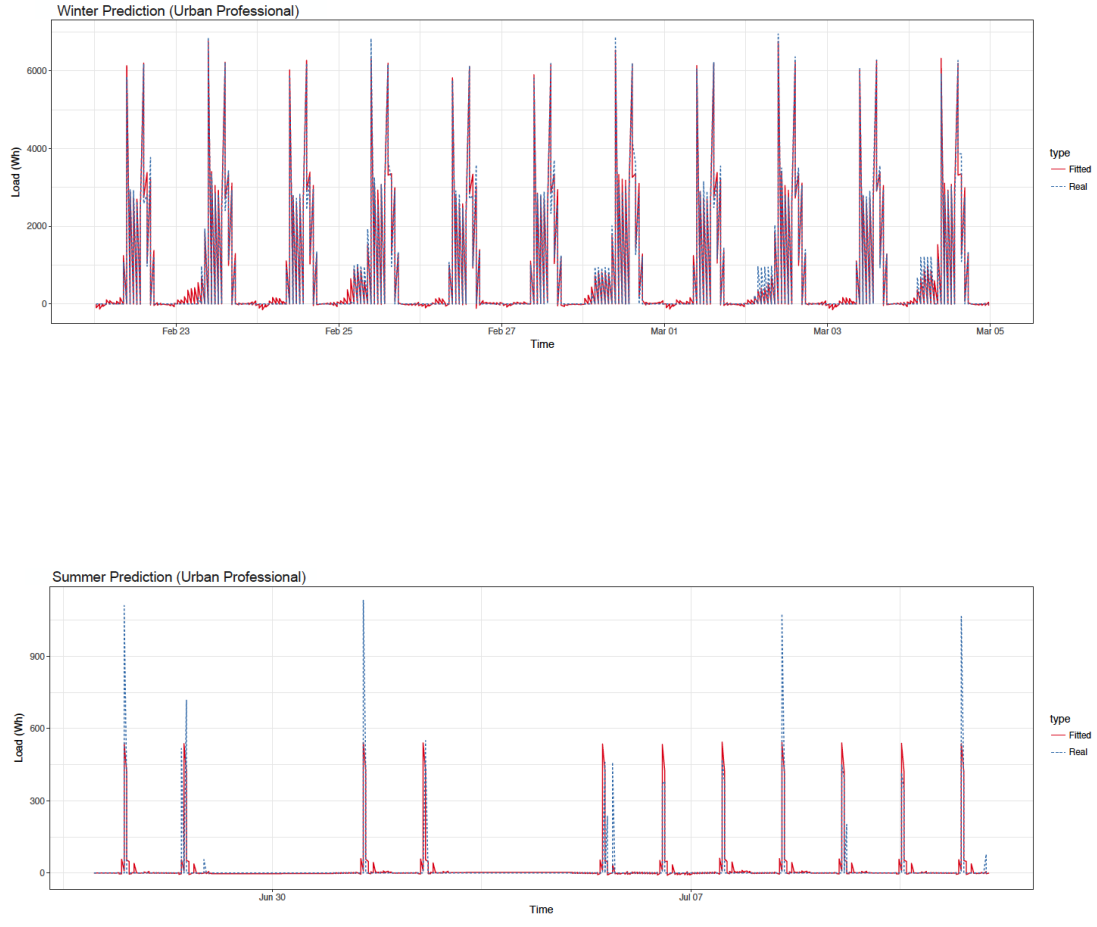


Figure: GAM fit for a customer that belongs to OA characterised as 'Urban Professionals'

Prediction and Regression Analysis

GAMs: Experimental Results for Urban Resident

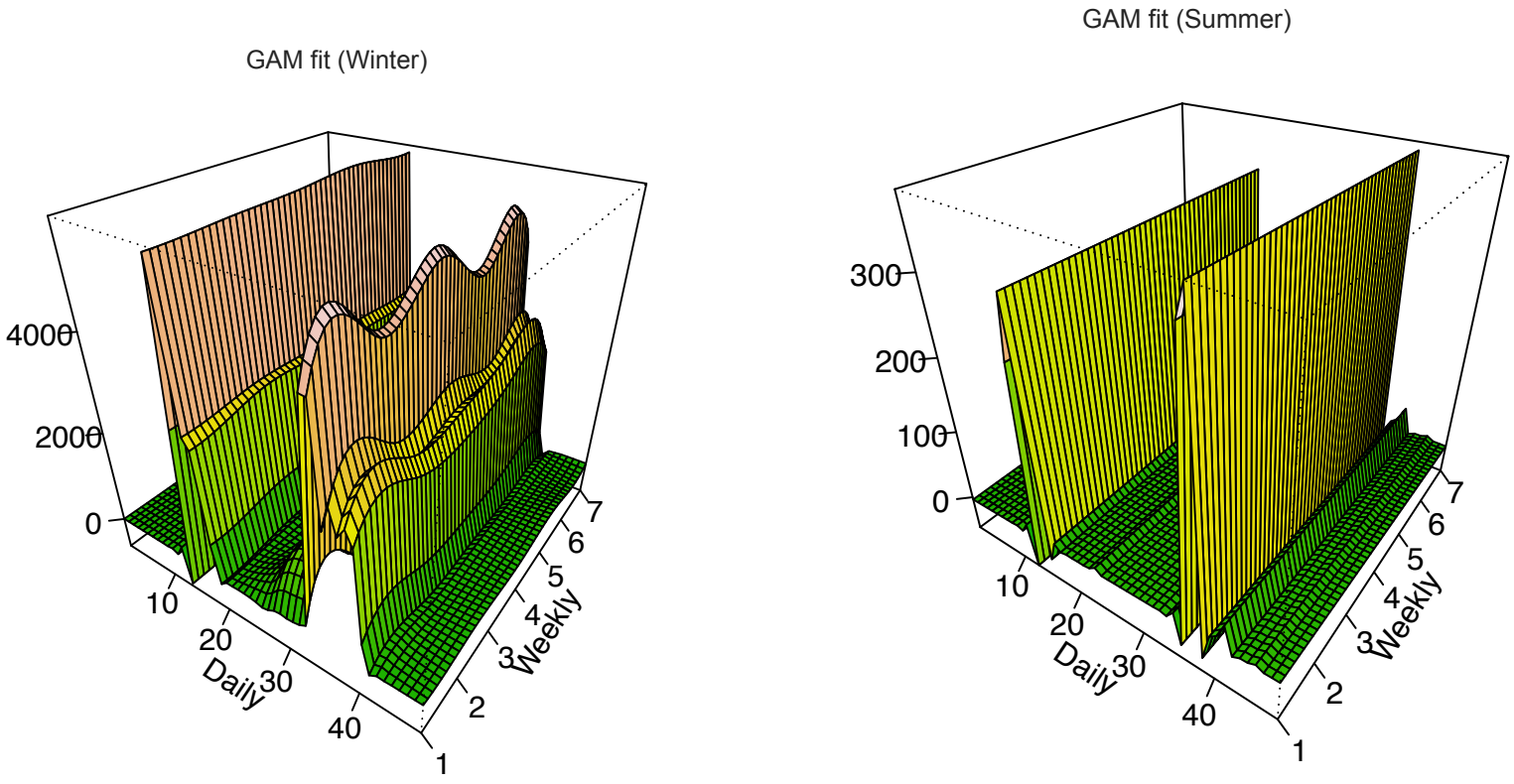


Figure: GAM fit for a customer that belongs to OA characterised as 'Urban Professionals' in 3D.

Some Immediate Findings and Conclusions

- So, what happened?
- This was just an example
- We all deal with different things, different data...
- And we all have different fruit in minds when we hear 'fruit'
- Could extra data save us?
- Extra methods? extra training? extra diagnostics?
- Extra? This is mad :)

Treat your data and problem as a unique one...

'The goals in statistics are to use data to predict and to get information about the underlying data mechanism. Nowhere is it written on a stone tablet what kind of model should be used to solve problems involving data. To make my position clear, I am not against data models per se. In some situations, they are the most appropriate way to solve the problem. But the emphasis needs to be on the problem and on the data' - Breiman (2001)

Thank you

Thank you for listening!

Twitter: @apavluhina
email: anastasia.ushakova@ed.ac.uk