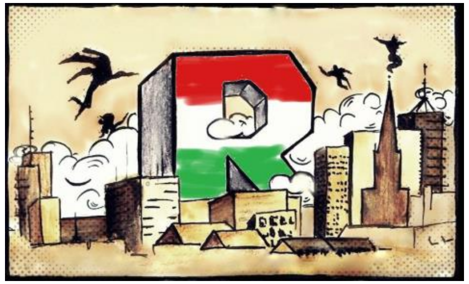
Service & Repair Demand Forecasting

Timothy Wong (Senior Data Scientist, Centrica plc)



European R Users Meeting
14th -16th May, 2018
Budapest, Hungary

We supply energy and services to over 27 million customer accounts

Supported by around 12,000 engineers and technicians

Our areas of focus are Energy Supply & Services, Connected Home, Distributed Energy & Power, Energy Marketing & Trading

Bord Gáis Energy British Gas centrica

Direct Energy DYNO HIVE Local Heroes

Overview

British Gas

Driven by many factors

Customer Contact → Creates → Job Demand

Booking

Initial Appointment → Not yet done → 2nd Appointment → Not yet done → 3rd Appointment → Not yet done …

Done → Closed

Done → Closed

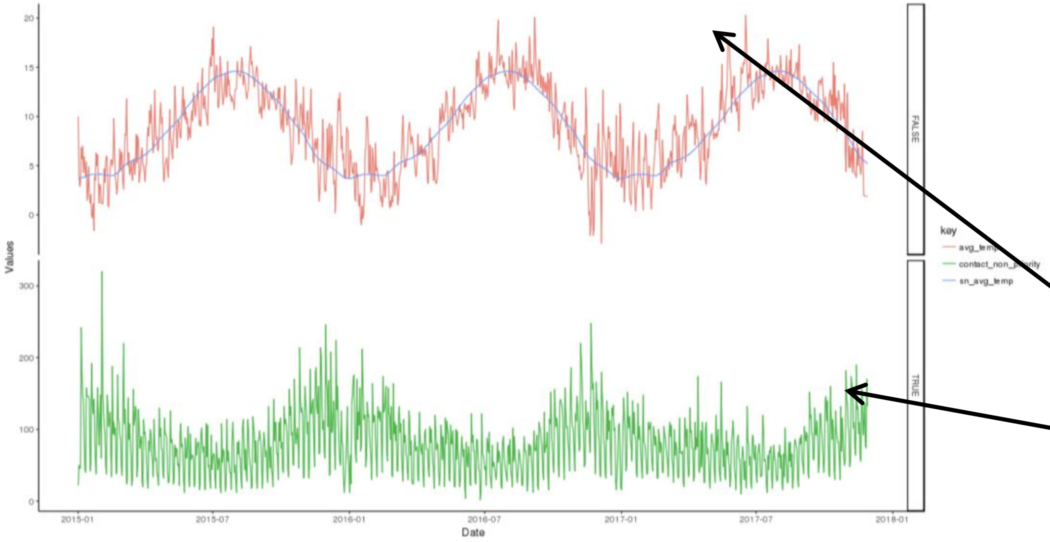
Done → Closed

My gas boiler is not working.

We can help. Would you like to book an appointment?

Gas boiler service & repair demand

- Strong causality, e.g.:
  - Cold weather → use more gas → high repair demand
  - Holiday → away from home → less repair demand
- 173 service patches in the UK
  - Each has dependent variables, e.g. weather observations.



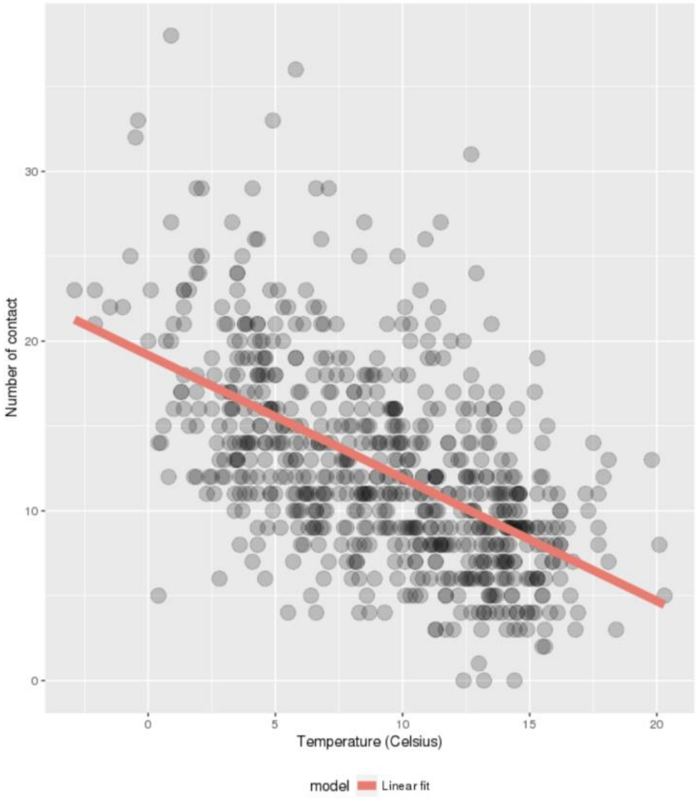
Temperature : **Independent variable**

Number of contact : **Dependent variable**

Linear Models

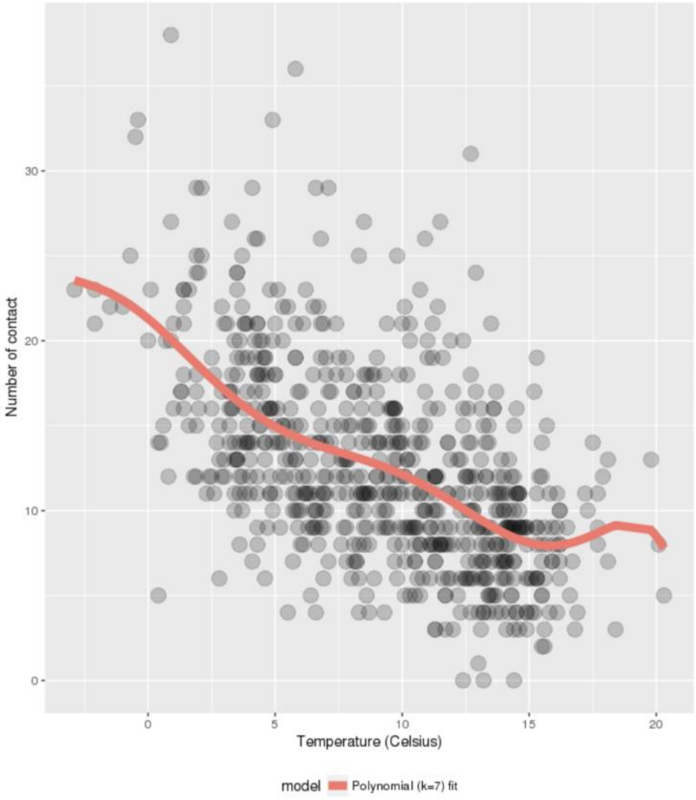
Linear fit

$$\hat{y} = \beta_0 + \beta_1 x$$



Polynomial fit

$$\hat{y} = \beta_0 + \sum_{k=1}^{K} \beta_k x^k$$



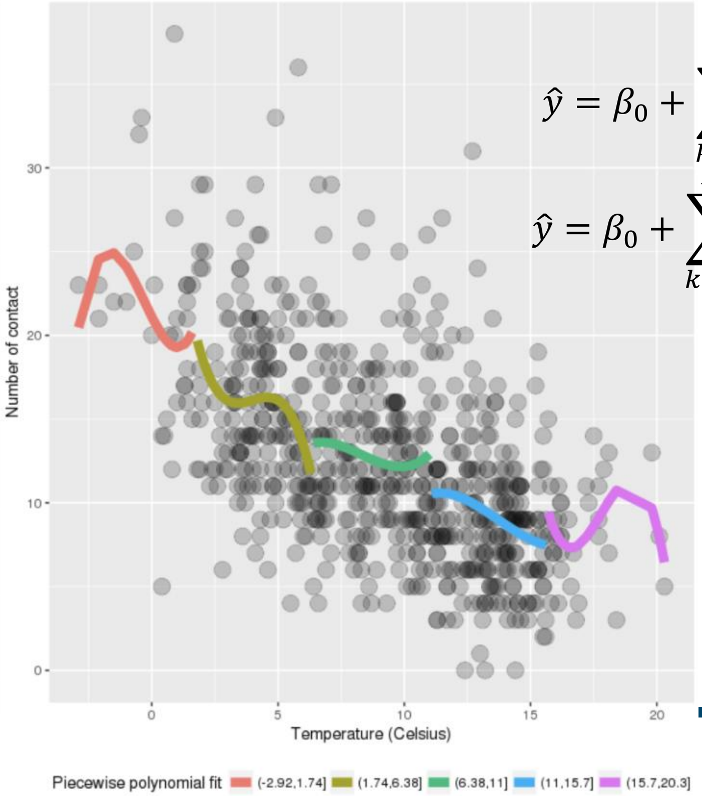
Piecewise polynomial fit

$$\hat{y} = \beta_0 + \sum_{k=1}^{K} \beta_k x^k \mid x \in (0,5]$$

$$\hat{y} = \beta_0 + \sum_{k=1}^{K} \beta_k x^k \mid x \in (5,10]$$

$$\hat{y} = \beta_0 + \sum_{k=1}^{K} \beta_k x^k \mid x \in (10,15]$$

$$\dots$$

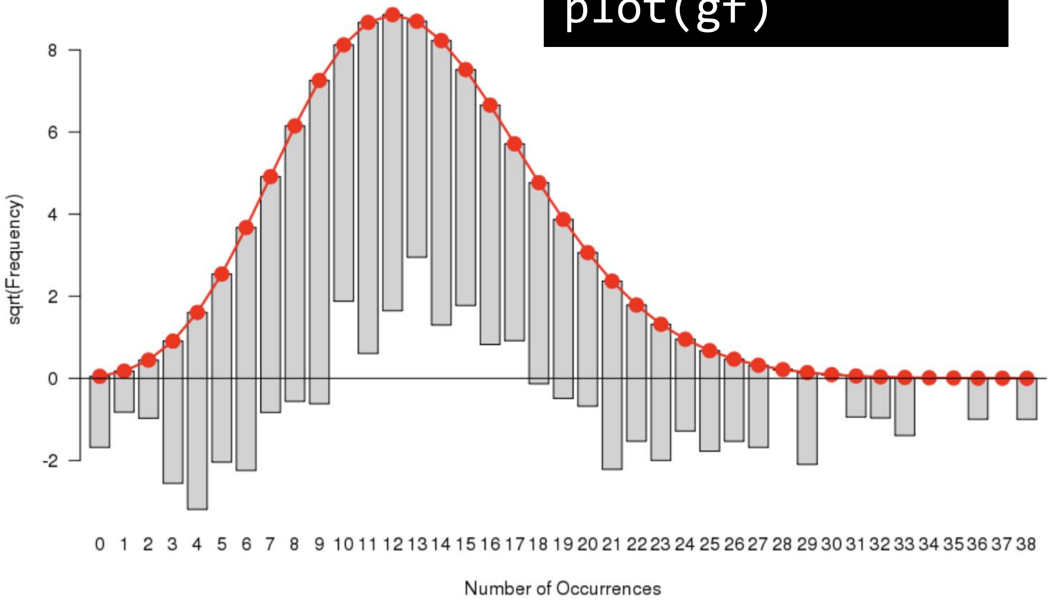


Poisson Distribution

- Goodness-of-fit test for Poisson distribution

```
> summary(gf)
Goodness-of-fit test for poisson distribution
                          X^2          df            P(> X^2)
Likelihood Ratio          543.702      32            2.288901e-94
```

```
library(vcd)
gf <- goodfit(x)
summary(gf)
plot(gf)
```



- Poisson GLM

$$\boldsymbol{y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + \epsilon_i}$$

Assumption:

$$y_i \sim Poisson(\lambda)$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- Response variable y_i is contact count.

Generalised Additive Model (GAM)

- Variables may have non-linear relationship

  e.g. warm weather → low demand,

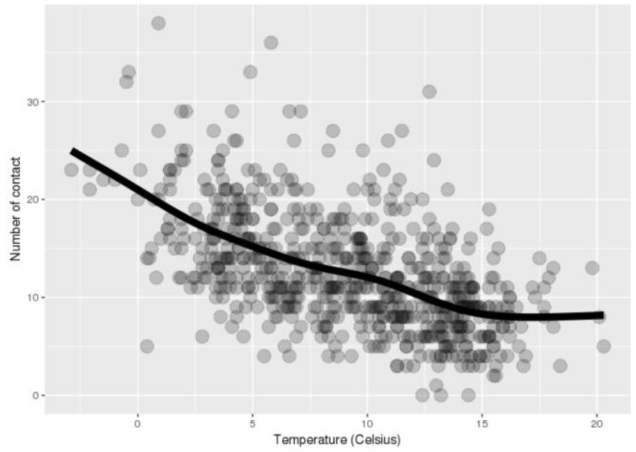
  but we don't expect zero demand on extremely hot day

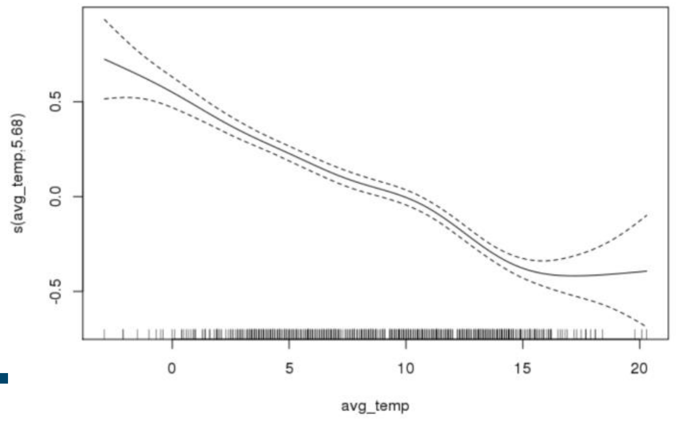
- GAM deals with smoothing splines (basis functions)

$$s(x) = \sum_{k=1}^{K} \beta_k b_k(x)$$

```
Family: poisson
Link function: log

Formula:
contact_priority ~ s(avg_temp)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.49418    0.01109   224.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(avg_temp) 5.681  6.858  588.6  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.315   Deviance explained = 31.5%
UBRE = 0.88378  Scale est. = 1         n = 694
```

GAM: Spline function





GLM vs GAM

AIC = 4263

AIC = 4260

```
myGLM <- glm(formula = contact_priority ~ avg_temp,
                   data = myData,
                   family = poisson())

myGAM <- gam(formula = contact_priority ~ s(avg_temp),
                   data = myData,
                   family = poisson())
```

Statistically significant

AVOVA:
Check reduction of sum of squared

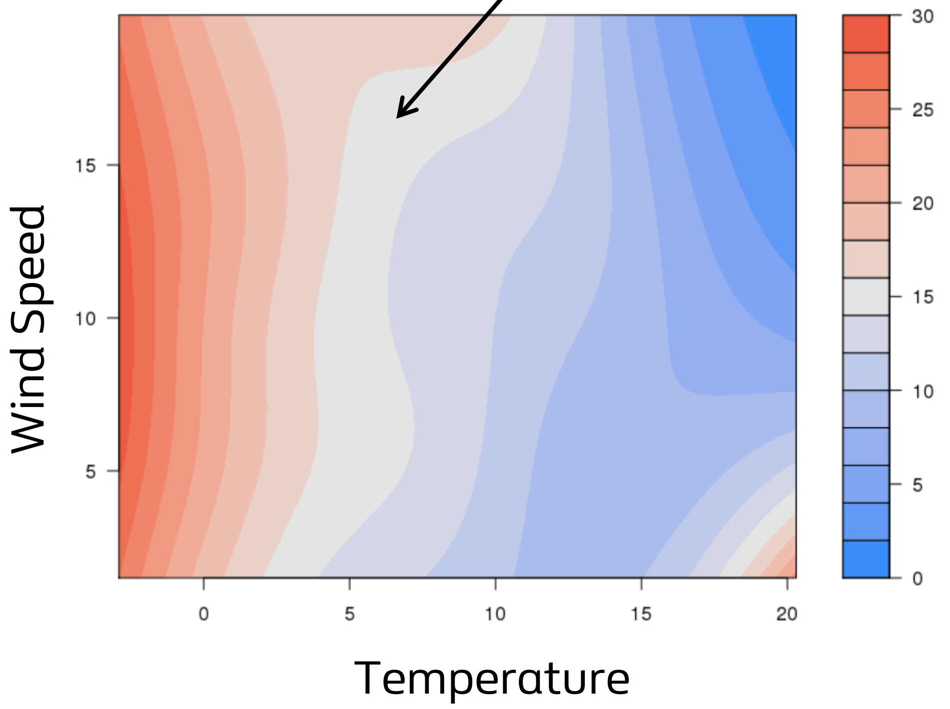
```
anova(myGLM, myGAM, test="Chisq")
Analysis of Deviance Table

Model 1: contact_priority ~ avg_temp
Model 2: contact_priority ~ s(avg_temp)

Resid. Df  Resid. Dev Df         Deviance   Pr(>Chi)
1 692.00   1307.1
2 687.32   1294.0     4.6808     13.087     0.01813 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

More Variables

```
myGAM2 <- gam(formula = contact_priority ~ te(avg_temp, avg_wind),
              data = myData,
              family = poisson())
```

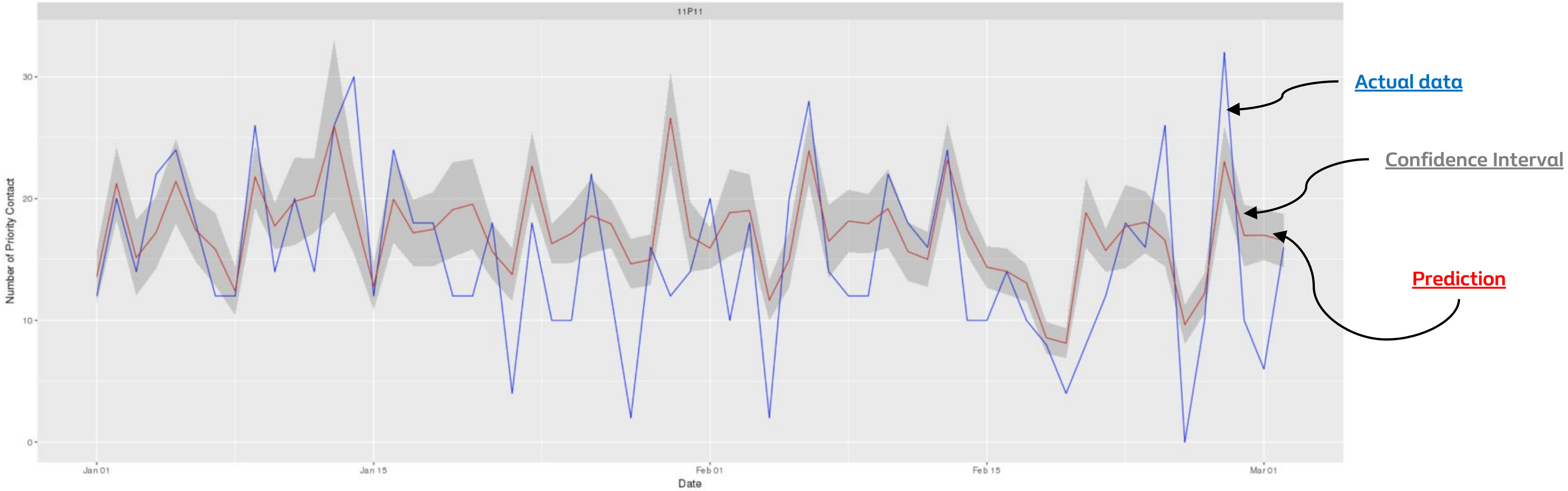
```
Family: poisson

Link function: log
Formula: contact_priority ~ te(avg_temp, avg_wind)

Parametric coefficients:
            Estimate   Std. Error  z value    Pr(>|z|)
(Intercept) 2.4927     0.0111      224.5      <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                       edf         Ref.df      Chi.sq      p-value
te(avg_temp,avg_wind)  14.12       16.52       613.6       <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.321 Deviance explained = 33.1%
UBRE = 0.86457 Scale est. = 1 n = 694
```

Colour showing output density



Results

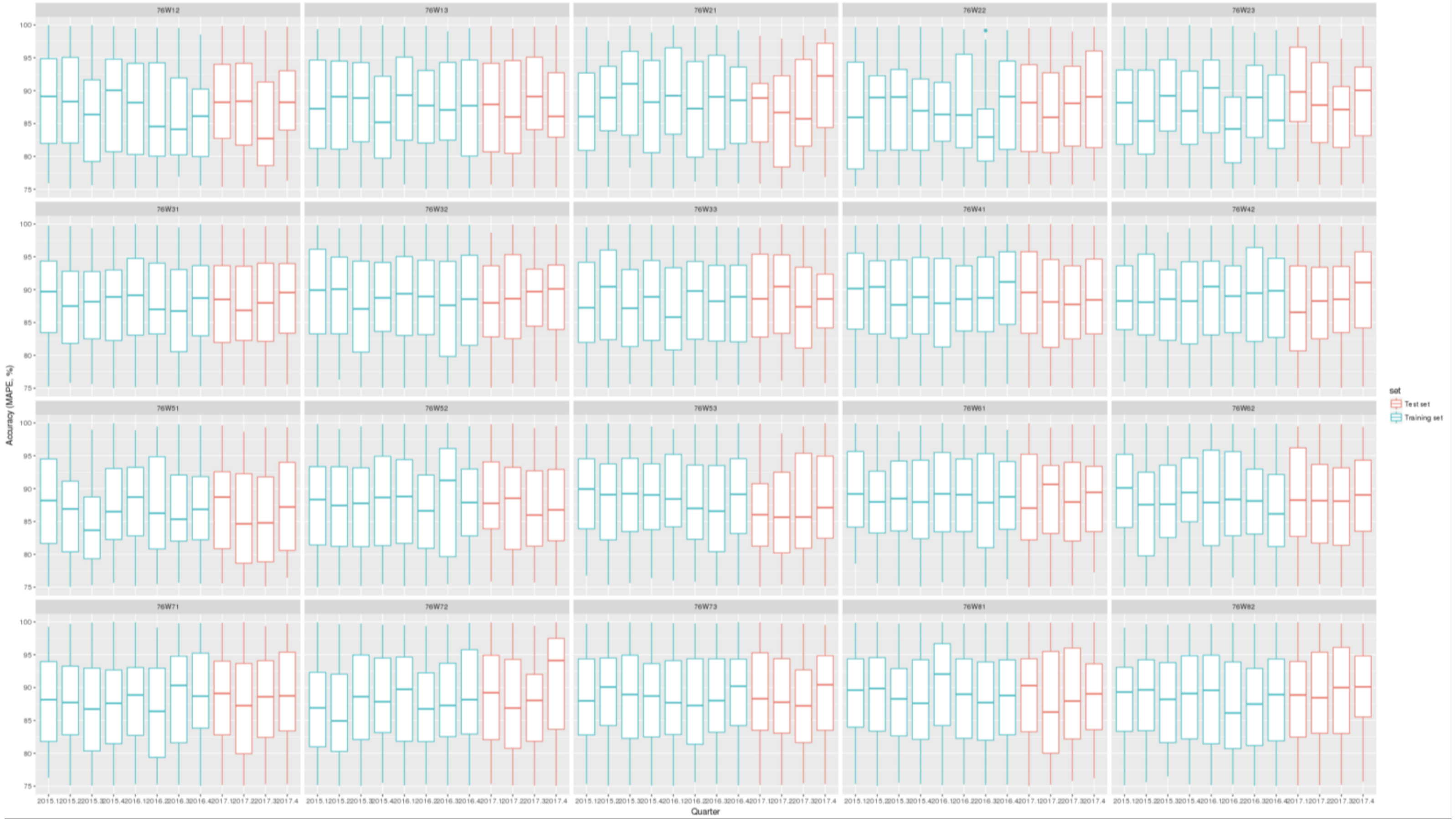
- For each response variable y we also know the standard error
  - Establish confidence interval



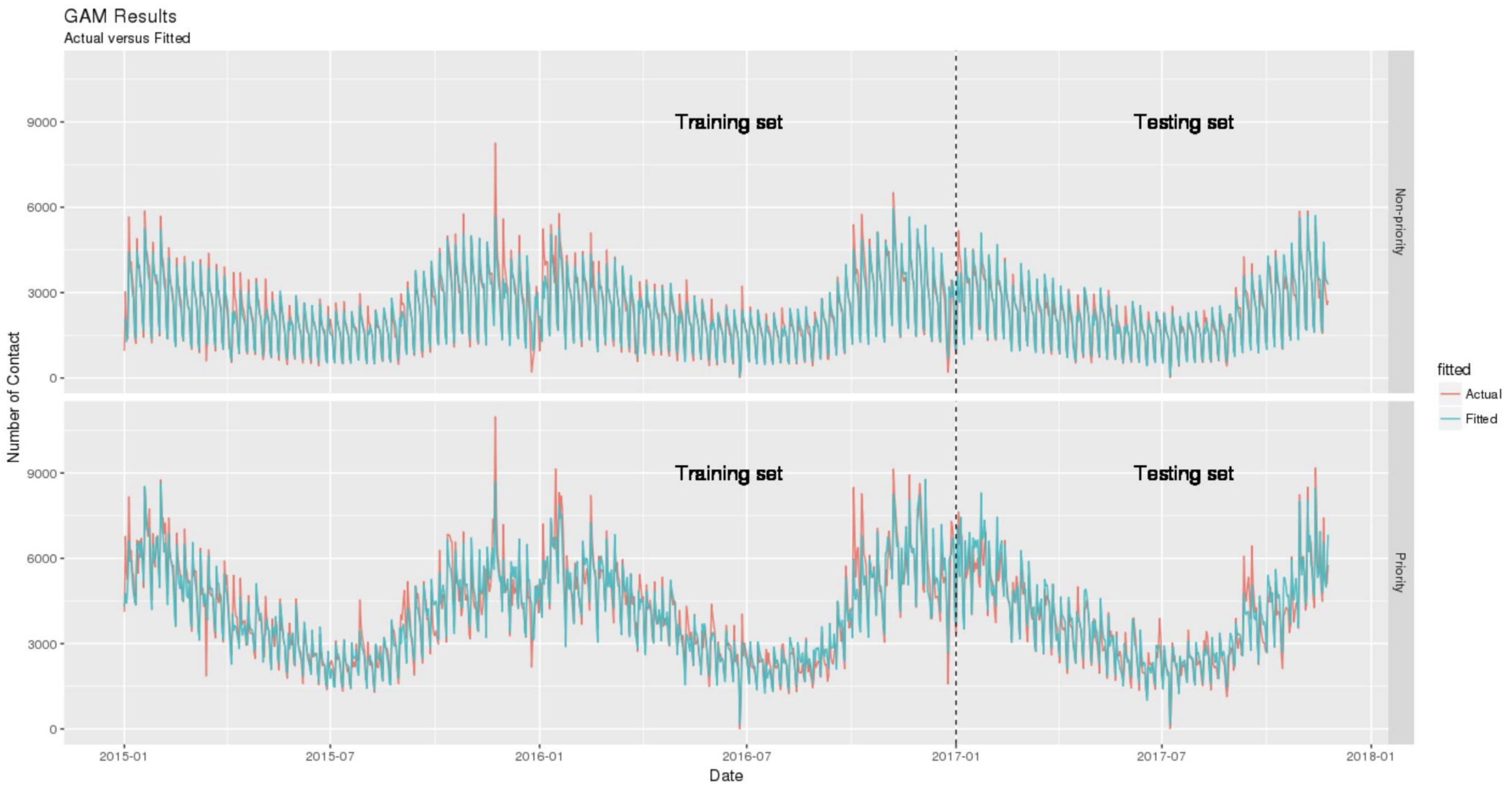
Accuracy measurement

Consistent results across patches

<u>London area:</u>



GAM Results: Aggregated View



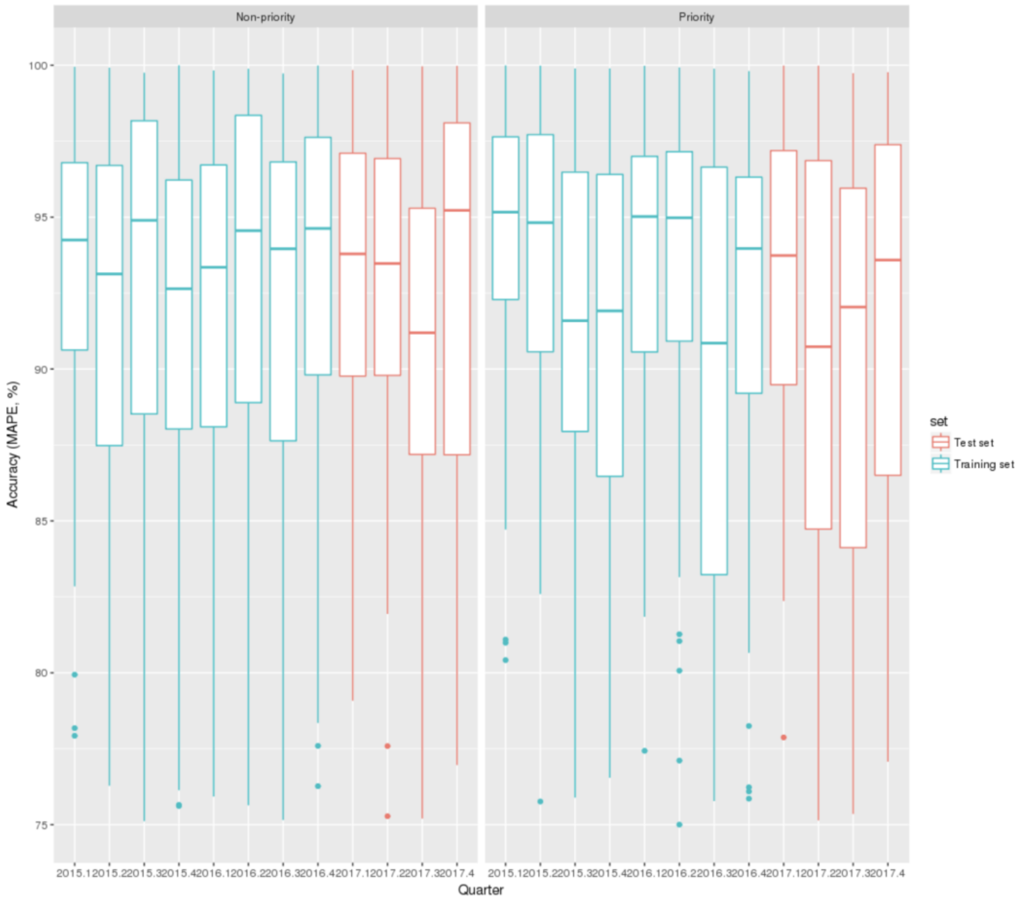
Accuracy measurement

- Defined as 1-MAPE (%)

  MAX(0, 1 - ABS(Forecast – Actual)/Actual)

  <u>Average accuracy of each quarter:</u>

	year_quarter	set	`Non-priority`	Priority
*	<fctr>	<chr>	<dbl>	<dbl>
1	2015.1	Training set	90.92	92.94
2	2015.2	Training set	86.77	92.42
3	2015.3	Training set	90.48	89.41
4	2015.4	Training set	87.40	89.47
5	2016.1	Training set	87.34	92.85
6	2016.2	Training set	87.28	90.79
7	2016.3	Training set	90.06	87.99
8	2016.4	Training set	89.50	89.84
9	2017.1	Test set	90.92	92.69
10	2017.2	Test set	88.68	89.55
11	2017.3	Test set	87.90	86.42
12	2017.4	Test set	91.44	90.32



Potential Improvements

- Feature transformation
  - Manually hand-craft *linear* features
  - Combine and transform existing variables
  - Use linear methods
  - Easier to interpret

- GAM + Bagging

- Multilevel linear regression ("Mixed-effect model")
  - Service patches as groups
  - Single model for all patches

Potential Improvements

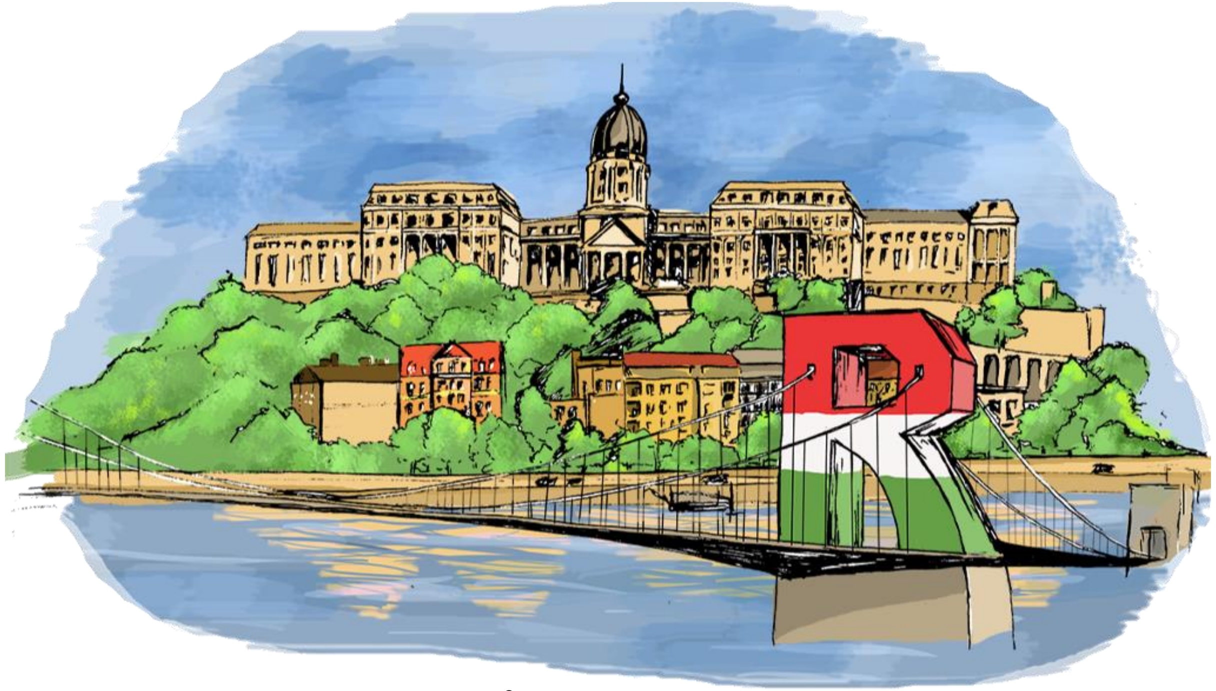
- Time Series Approach
  - ARMA (Auto-Regressive Moving Average) / ARIMA
  - Analyse seasonality

- Other machine learning techniques
  - Boosted trees
  - Random Forest
    - Works nicely with ordinal/categorical variables
  - Neural net (RNNs)
    - Substantially longer model training time

Less interpretable,
No confidence interval

Thanks

Project Team
(Names in alphabetical order)
Angus Montgomery
Hari Ramkumar
Harriet Carmo
Kerry Wilson Morgan
Martin Thornalley
Matthew Pearce
Philip Szakowski
Terry Phipps
Timothy Wong
Tonia Ryan



European R Users Meeting
14th -16th May, 2018
Budapest, Hungary

Timothy Wong

Senior Data Scientist

Centrica plc

timothy.wong@centrica.com

@timothywong731

github.com/timothywong731

linkedin.com/in/timothy-wong-7824ba30