The Conference for Users of R
July 10-13 , 2018
Brisbane, Australia

MODELLING FIELD OPERATION CAPACITY USING GENERALISED ADDITIVE MODEL AND RANDOM FOREST

Scan QR Code for Slides

Timothy Wong

Senior Data Scientist, Centrica plc

timothy.wong@centrica.com

@timothywong731

timothywong731.github.io

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

CENTRICA - WHO ARE WE?

- 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

Driven by many factors

Customer Contact → Creates → Job Demand

Job Demand → Booking → Initial Appointment

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

Initial Appointment → Done → Closed → Time taken? → Duration

2nd Appointment → Done → Closed → Time taken? → Duration

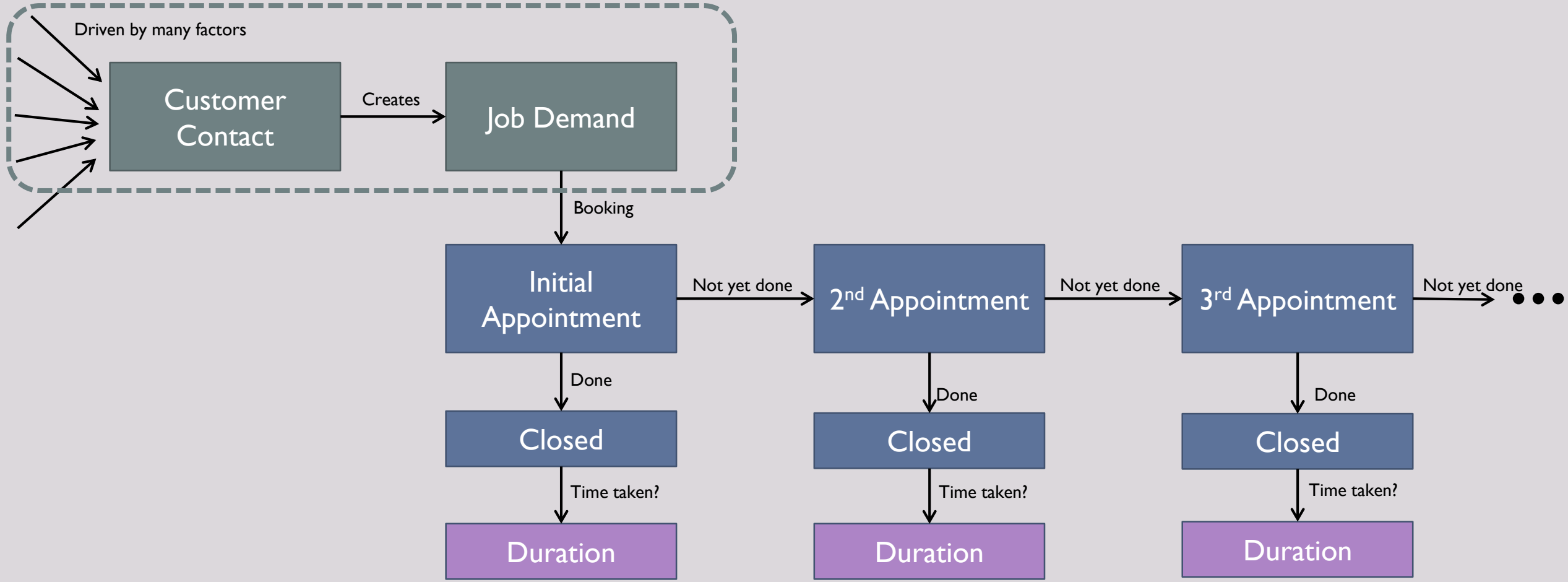
3rd Appointment → Done → Closed → Time taken? → Duration

OBJECTIVES

- Forecast boiler breakdown demand
  1. Number of expected jobs (#)
  2. Number of expected appointments (#)
  3. Total amount of time (hours)

- Service patch level model at daily granularity

STEP 1: JOB MODELLING



JOB MODELLING

- 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
```

- Poisson GLM

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

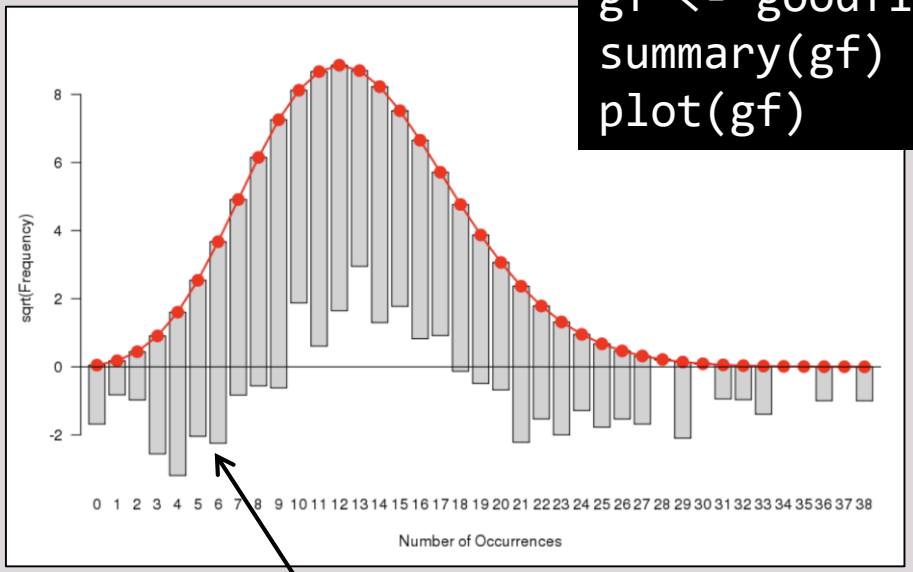
Assumption:

$$y_i \sim Poisson(\lambda)$$

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

- Response variable y_i is contact count

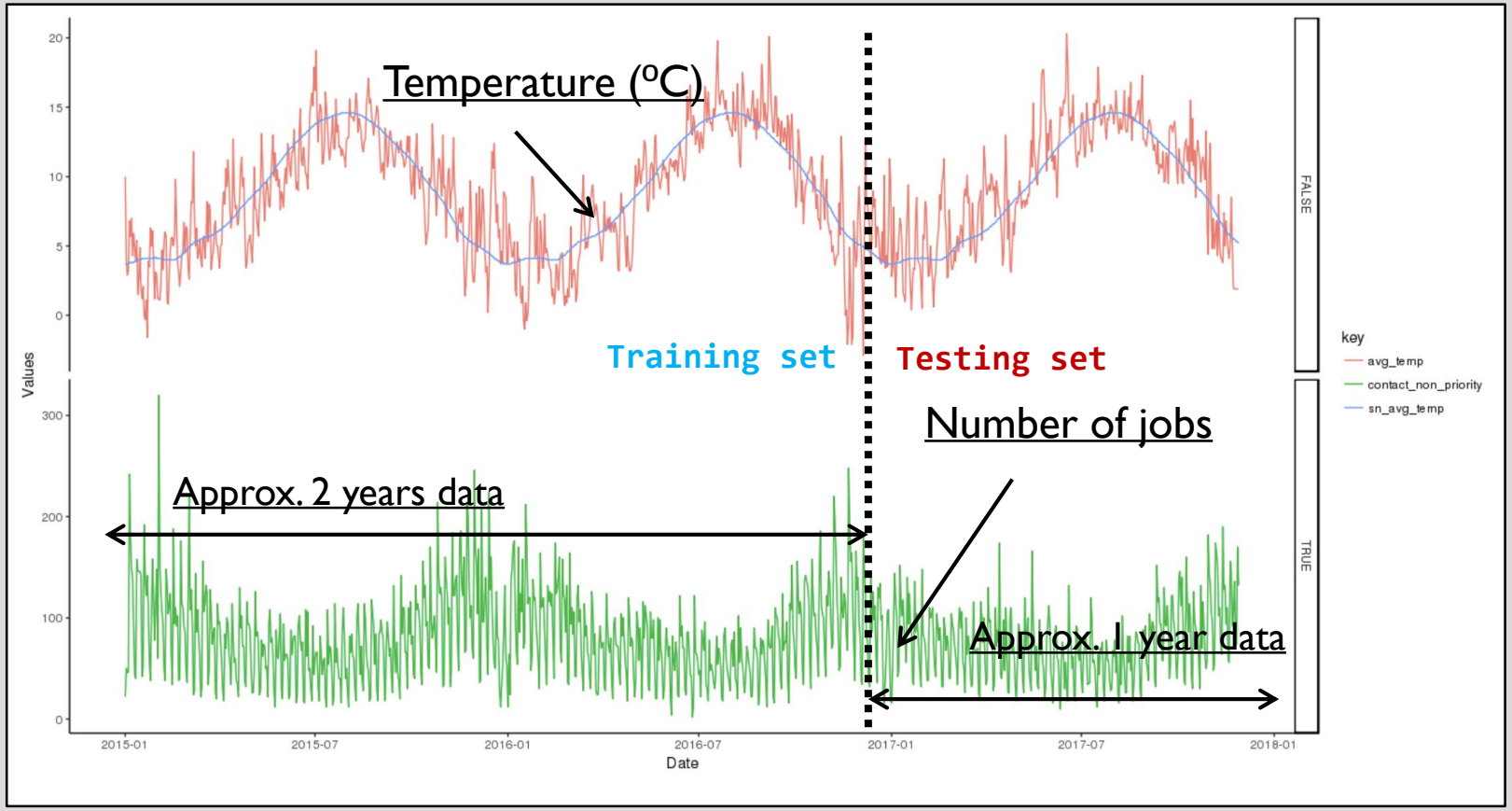
```
library(vcd)
gf <- goodfit(x)
summary(gf)
plot(gf)
```



Number of jobs

INPUT VARIABLES

- For each service patch (173 in total):
  - Weather measurements
  - Date effects
- Strong correlation / causality
  - Cold weather → More jobs
  - Weekend → Fewer jobs



GENERALISED ADDITIVE MODEL (GAM)

- Variables may have non-linear relationship

  e.g. warm weather → low demand,

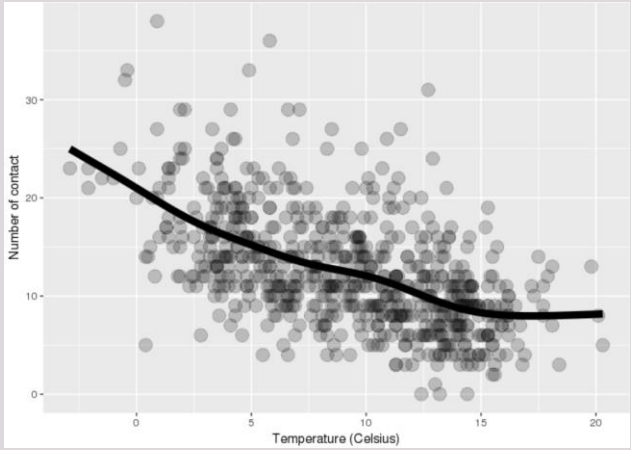
  but we don't expect zero demand on extremely hot day

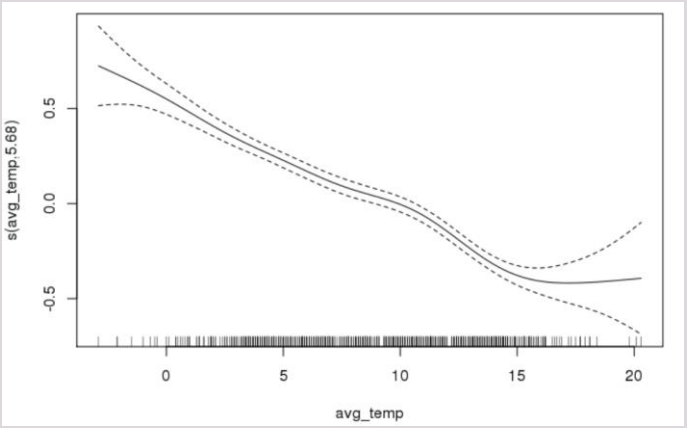
- GAM deals with smoothing splines (basis functions)

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

```
Family: poisson
Link function: log

Formula:
job ~ 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





GAM VS GLM

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

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

AIC = 4263

AIC = 4260

AVOVA:
Check reduction of sum of squared

Statistically significant

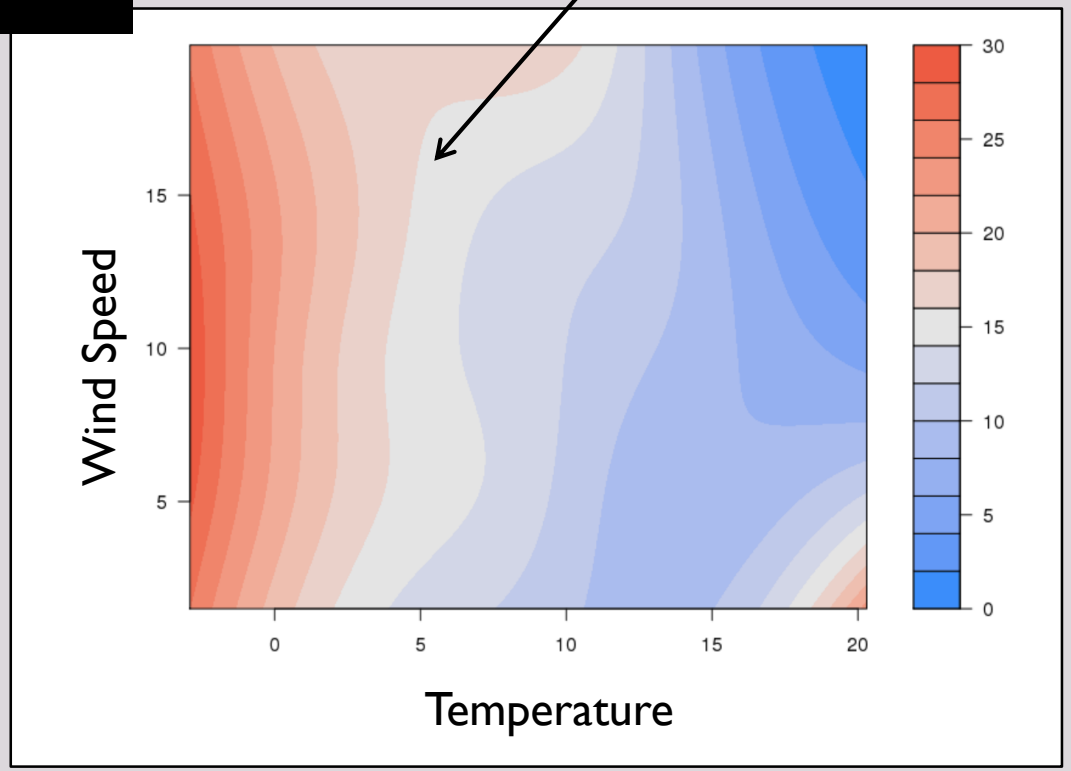
```
> anova(myGLM, myGAM, test="Chisq")
Analysis of Deviance Table

Model 1: job ~ avg_temp
Model 2: job ~ 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 = job ~ te(avg_temp, avg_wind),
              data = myData,
              family = poisson())
```

```
Family: poisson

Link function: log
Formula: job ~ 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
```

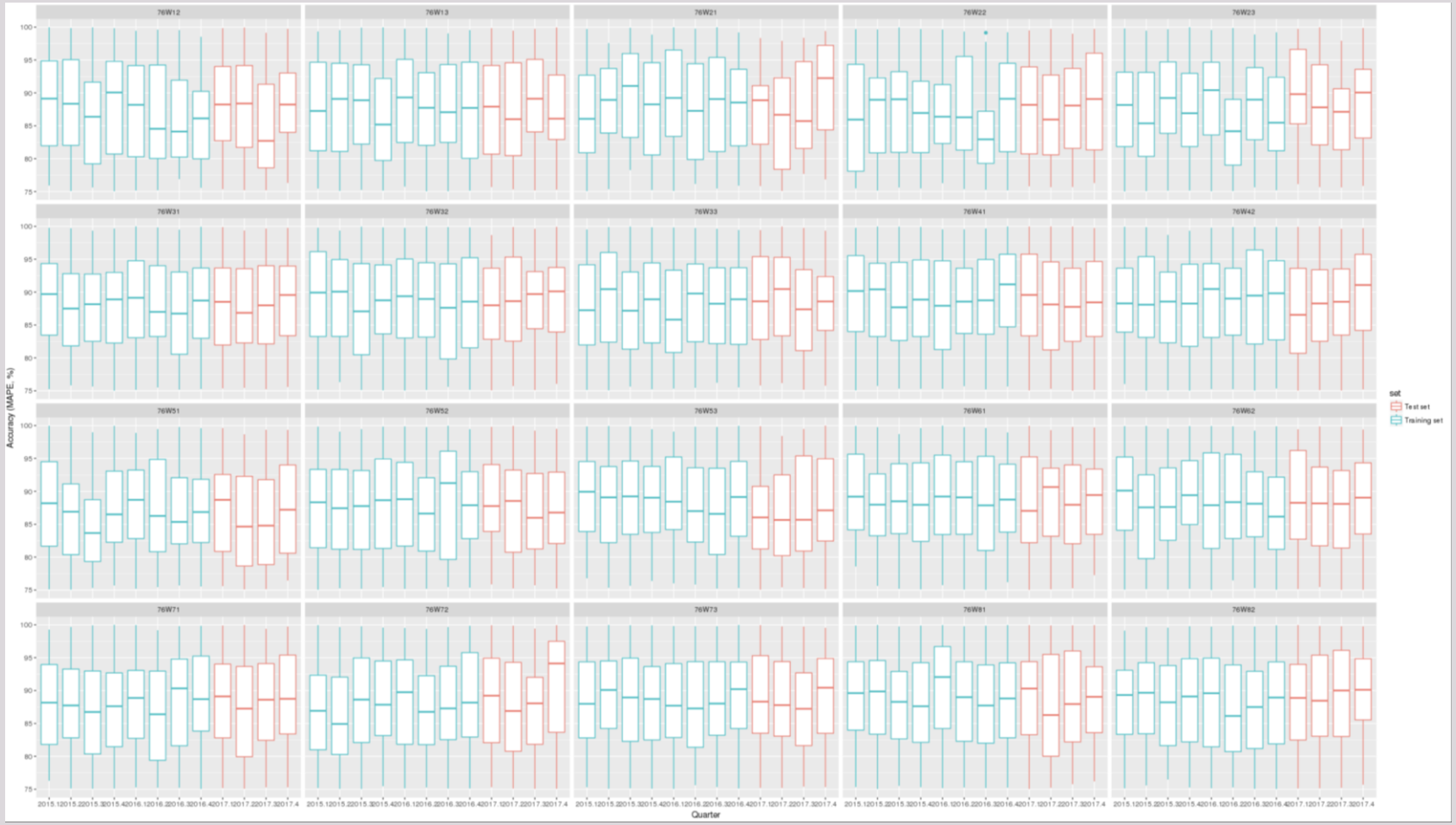
Colour showing density



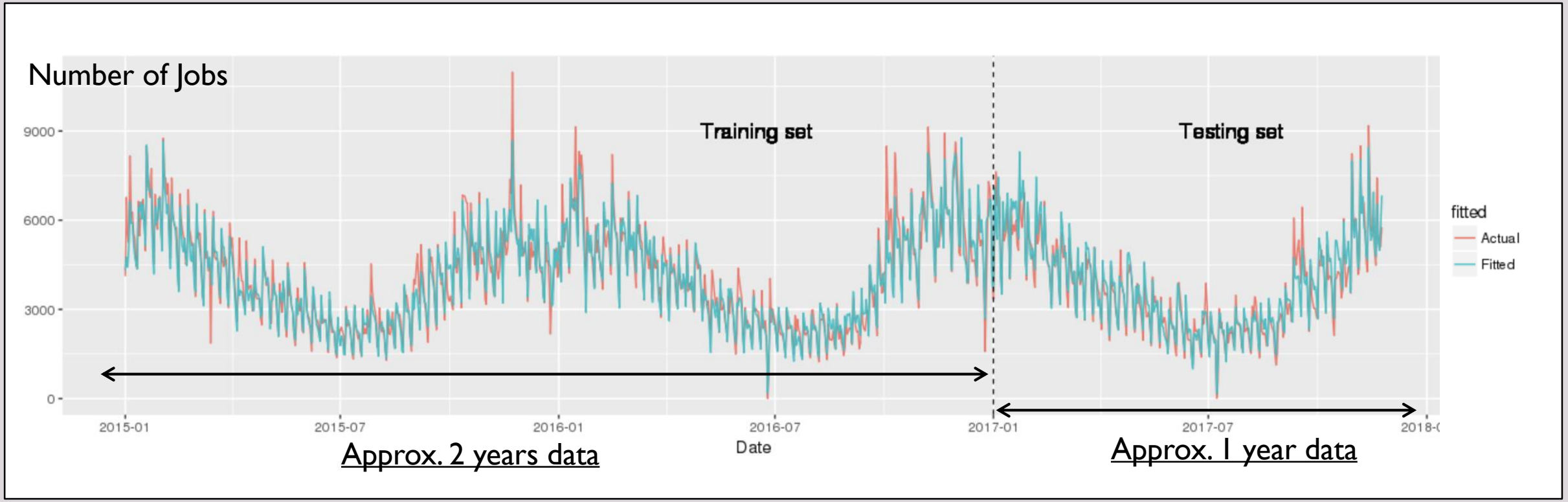
GAM RESULTS: PATCH LEVEL VIEW

Consistent results across patches

Example: London area



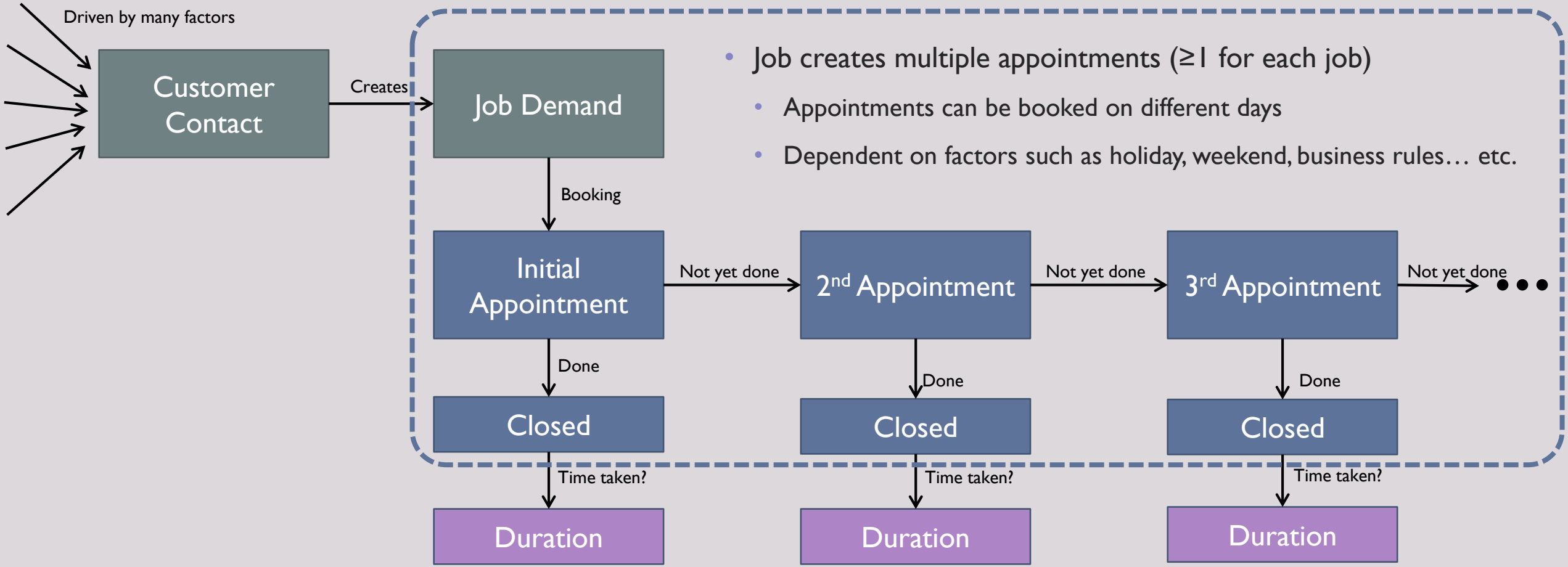
GAM RESULTS: AGGREGATED VIEW



ALTERNATIVE WAYS

- Feature transformation
  - Combine and transform existing variables
  - Easier statistical interpretation
- Multilevel regression ("Mixed-effect model")
  - Service patches as groups
  - Single model for all patches
- Time Series Approach (ARIMA)
- Trees / Forests
  - Works nicely with ordinal / categorical variables
- Recurrent Nets (RNN)

STEP 2: APPOINTMENT BOOING MODELLING

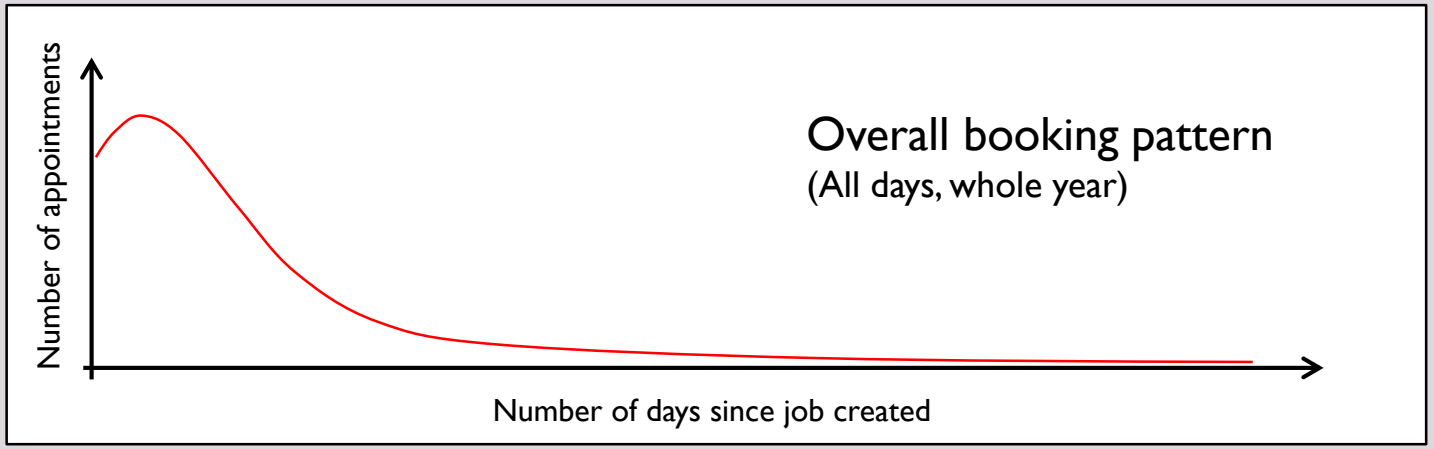


- Job creates multiple appointments (≥1 for each job)
  - Appointments can be booked on different days
  - Dependent on factors such as holiday, weekend, business rules… etc.

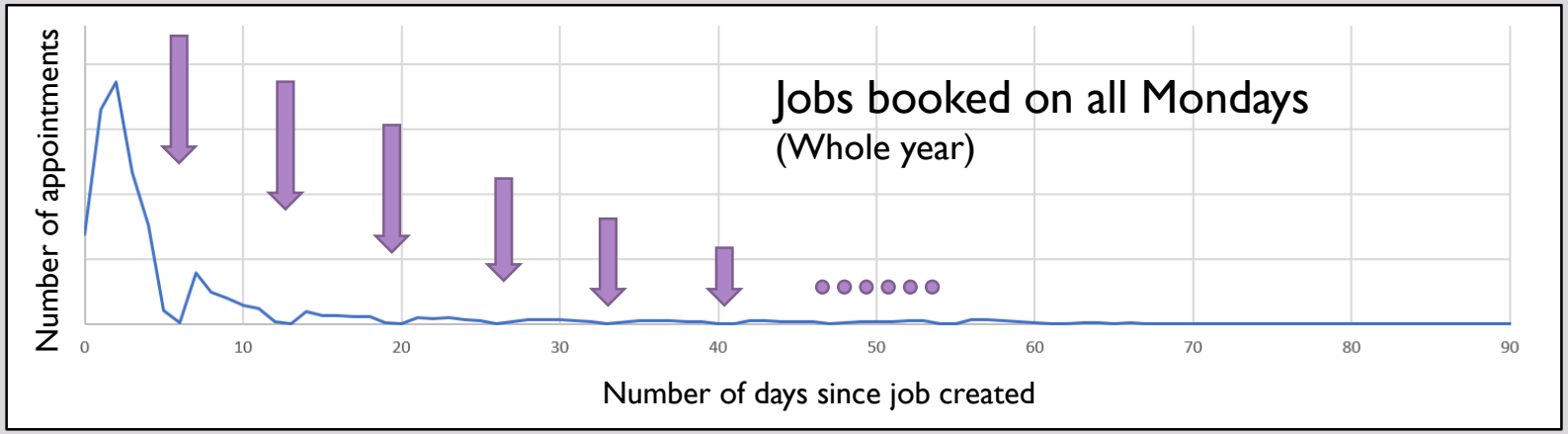
APPOINTMENT BOOING PATTERN

Visualise

JobID	ApptID	JobDate	ApptDate
J0001	A0001	2018-01-01	2018-01-02
J0002	A0002	2018-01-03	2018-01-03
	A0003		2018-01-04
	A0004		2018-01-05
J0003	A0005	2018-01-04	2018-01-06
	A0006		2018-01-10
J0004	A0007	2018-01-04	2018-01-05
	A0008		2018-01-06
…	…	…	…



Filter



APPOINTMENT BOOING PATTERN

Original data

JobID	ApptID	JobDate	ApptDate
J0001	A0001	2018-01-01	2018-01-02
J0002	A0002	2018-01-03	2018-01-03
	A0003		2018-01-04
	A0004		2018-01-05
J0003	A0005	2018-01-04	2018-01-06
	A0006		2018-01-10
J0004	A0007	2018-01-04	2018-01-05
	A0008		2018-01-06
…	…	…	…

Transformed data

JobID	JobDate	D0	D1	D2	D3	D4	D5	D6	…
J0001	2018-01-01	0	1	0	0	0	0	0	…
J0002	2018-01-03	1	1	1	0	0	0	0	…
J0003	2018-01-04	0	0	1	0	0	0	1	…
J0004	2018-01-04	0	1	1	0	0	0	0	…
…	…	…	…	…	…	…	…	…	…

N is quite large
(>300 million rows)

JobID	JobDate	ApptDate	IsBooked
J0001	2018-01-01	2018-01-01	0
J0001	2018-01-01	2018-01-02	1
J0001	2018-01-01	2018-01-03	0
J0001	2018-01-01	2018-01-04	0
…	…	…	…

RANDOM FOREST MODEL

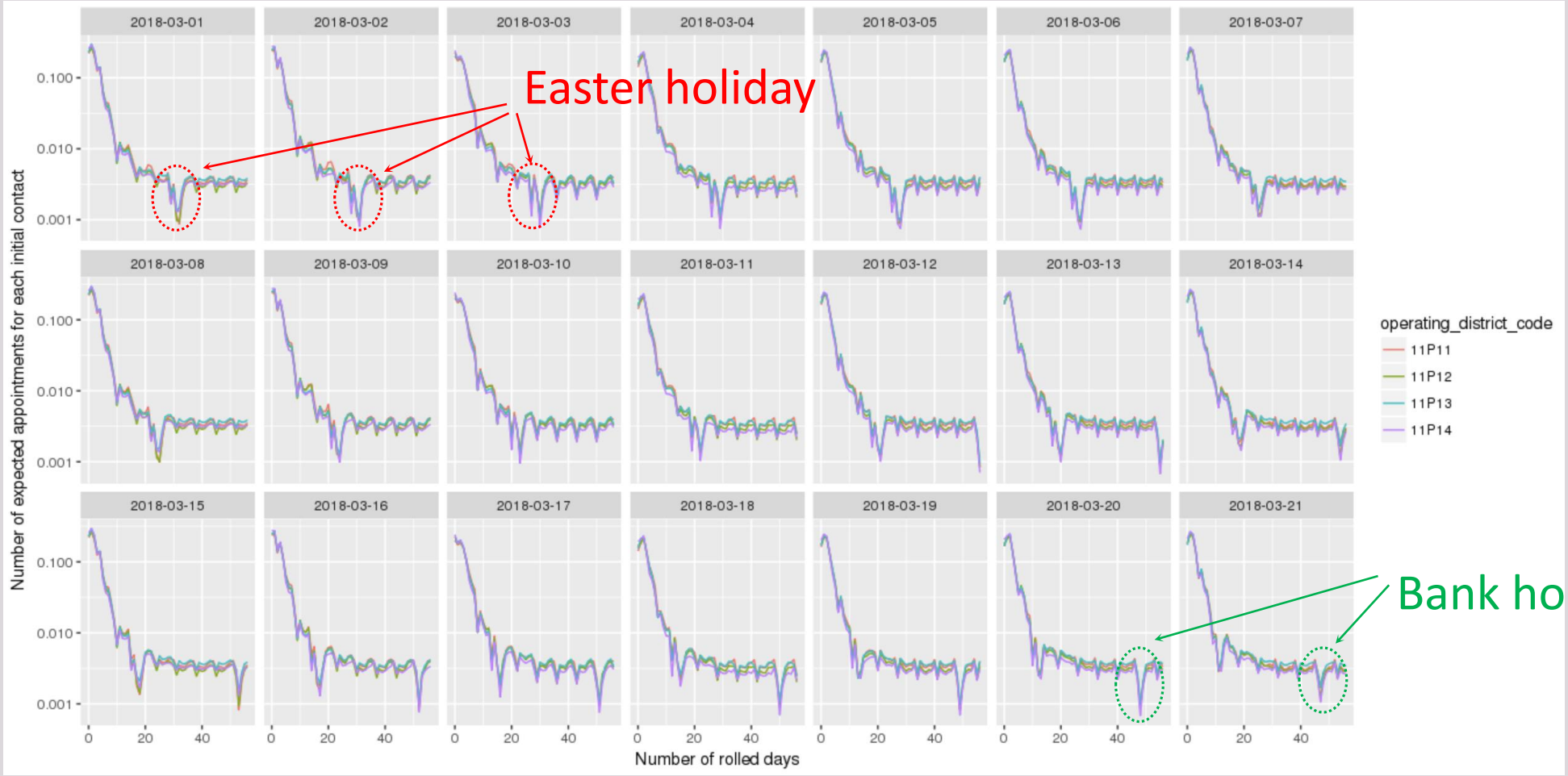
- Compute new variables ('mutate') as features for random forest model
  - Most of them are categorical / ordinal variables
  - Large dataset – used '`rxDForest()`' in the '`RevoScaleR`' package for distributed model training

M variables (M is small)

JobID	JobDate	ApptDate	IsBooked	Feature1	Feature2	Feature3	Feature4	…
J0001	2018-01-01	2018-01-01	0	…	…	…	…	…
J0001	2018-01-01	2018-01-02	1	…	…	…	…	…
J0001	2018-01-01	2018-01-03	0	…	…	…	…	…
J0001	2018-01-01	2018-01-04	0	…	…	…	…	…
…	…	…	…	…	…	…	…	…

N is quite large
(>300 million rows)

APPOINTMENT BOOING CURVES



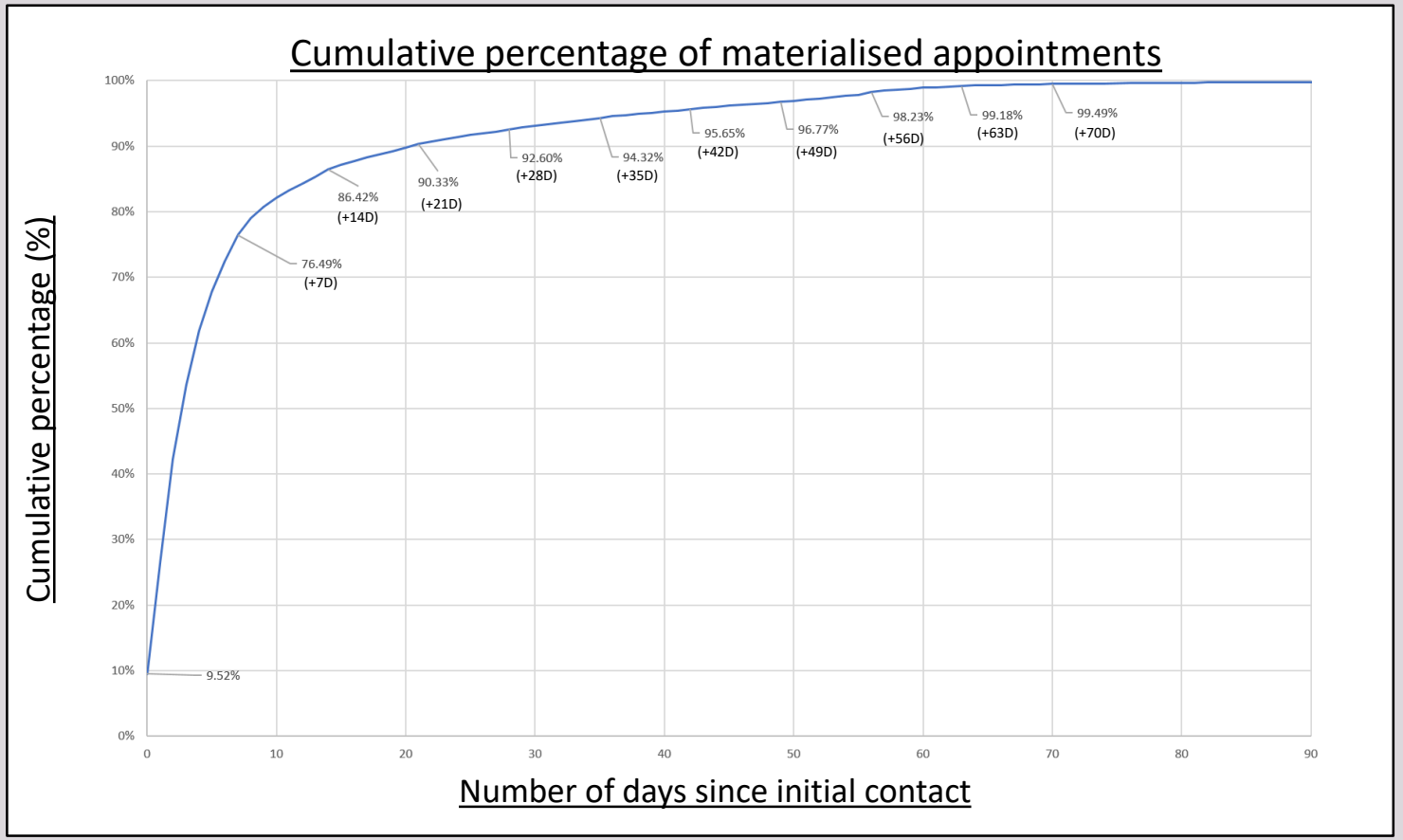
CALCULATING EXPECTED NUMBER OF APPOINTMENTS

- Expected number of appointment $\mathbb{E}(A)$ is determined by jobs created in the past T days.
  - Each day in the past T days has an appointment booking curve $\{B_0, B_1, B_2, B_3 \dots B_T\}$, derived from the random forest model
  - The expected number of jobs $\mathbb{E}(J)_t$ on each day $t = 1,2,3, \dots, T$ is derived from the GAM model
  - The expected number of appointment $\mathbb{E}(A)$ is calculated as:

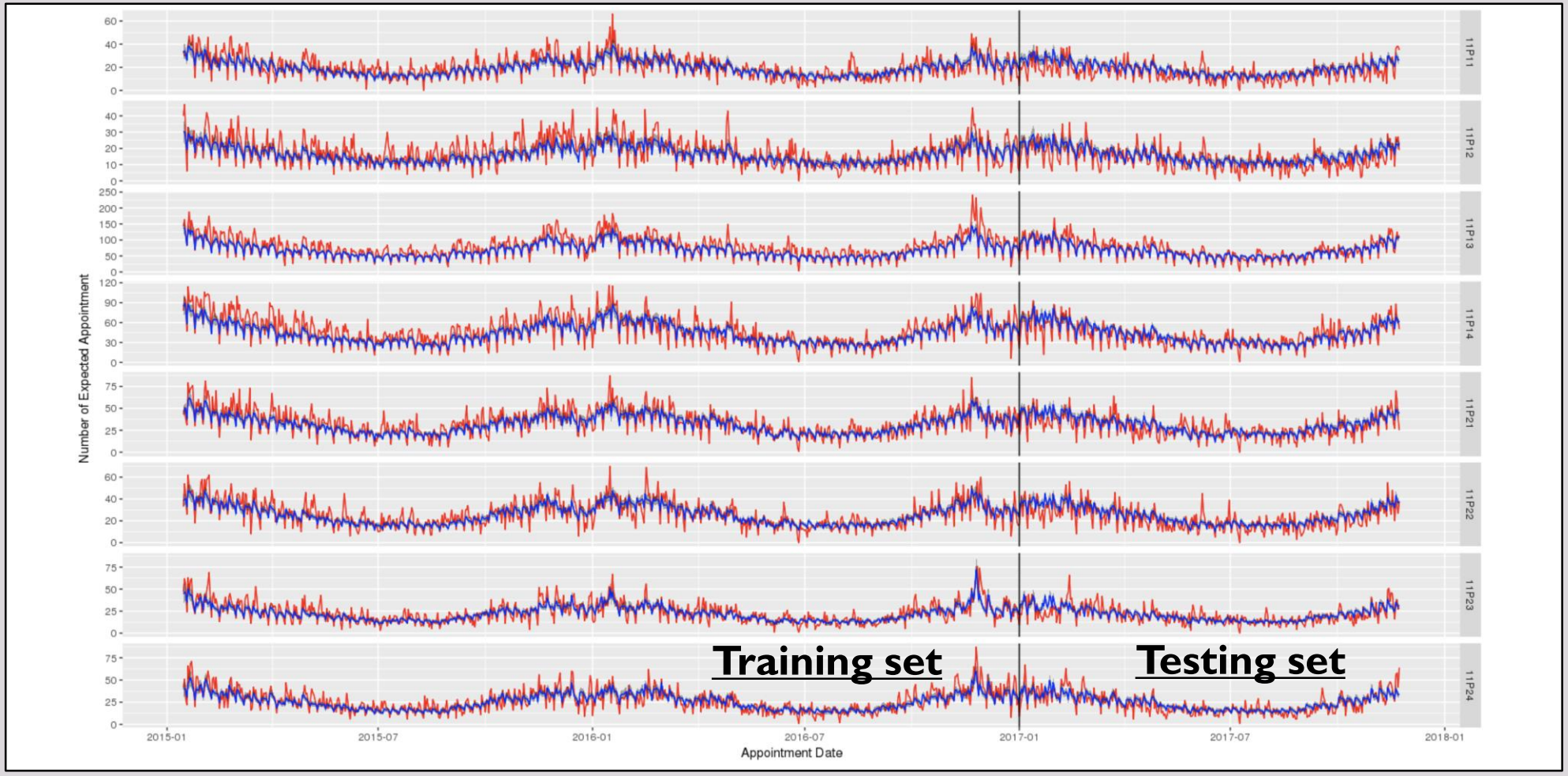
$$\mathbb{E}(A) = \sum_{t=0}^{T} \mathbb{E}(J)_t B_{(T,t)}$$

DETERMINING HOW FAR IN THE PAST TO INCLUDE

- Determining T
  - Find out how many days it took to empirically materialise 99% of all appointments.
  - In our sample = approx. 56 days (8 weeks)

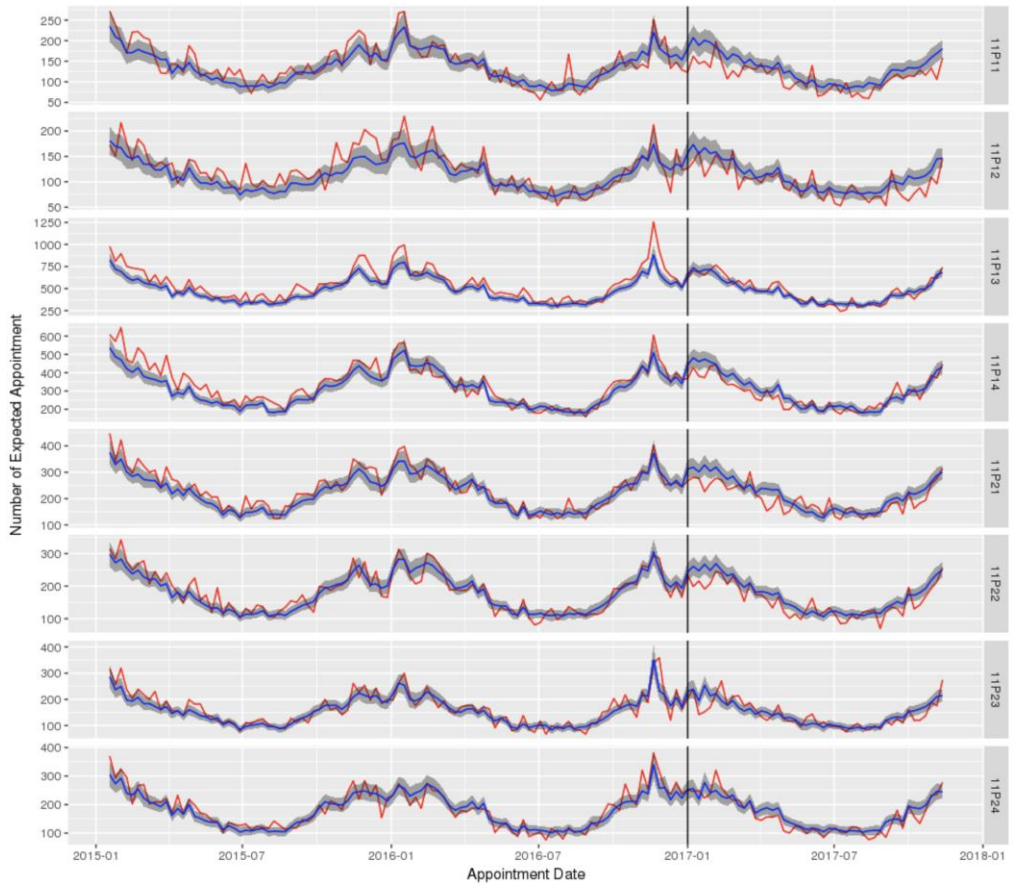


EXPECTED NUMBER OF APPOINTMENTS (DAILY)

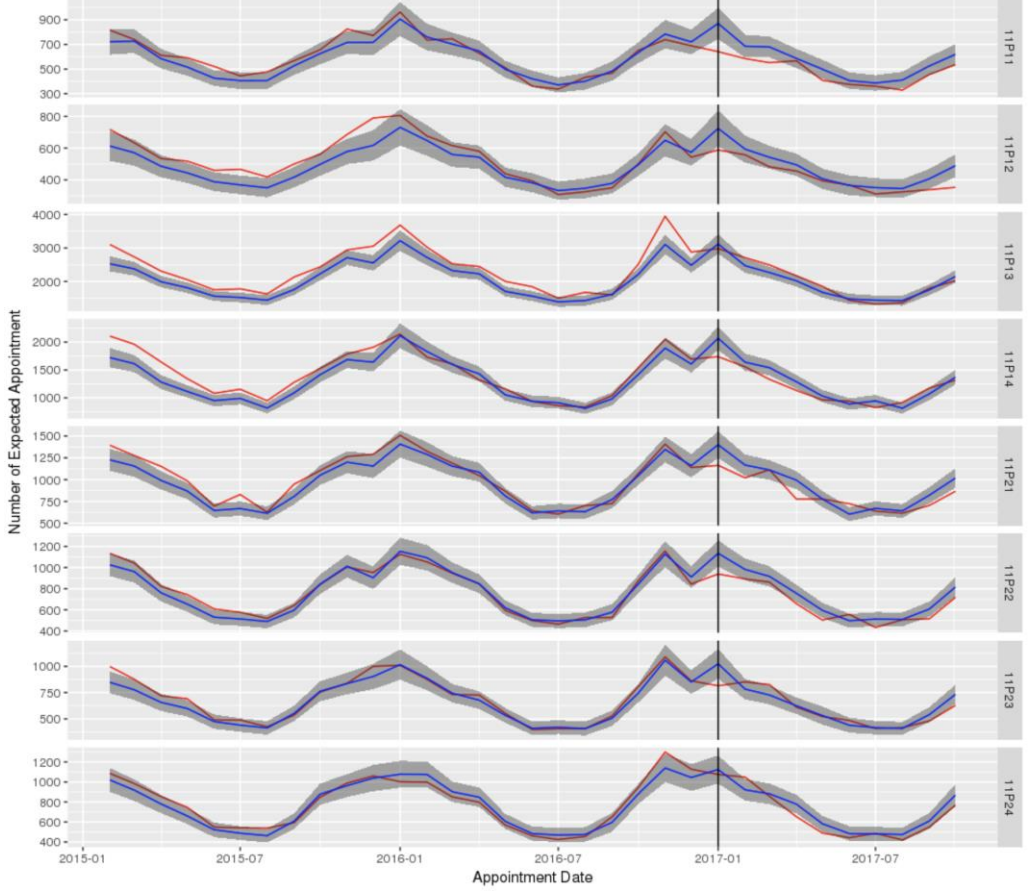


EXPECTED NUMBER OF APPOINTMENTS (WEEKLY/MONTHLY)

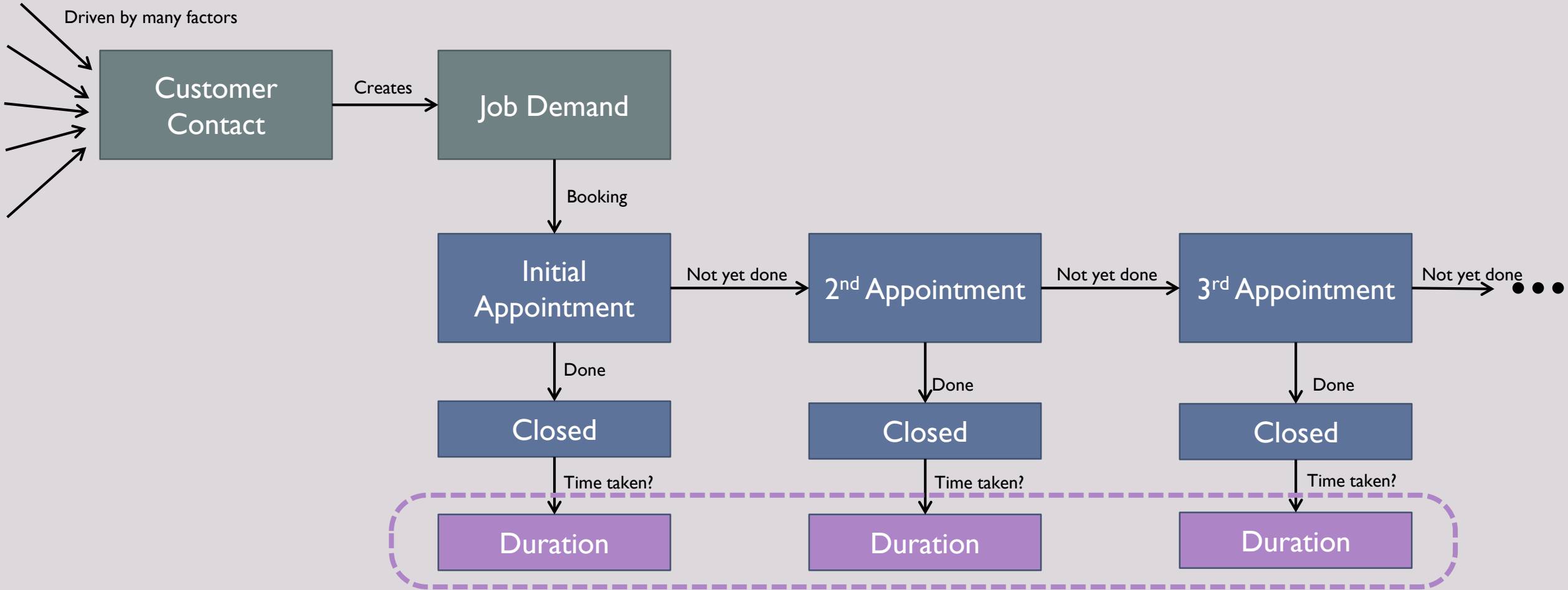
Weekly



Monthly



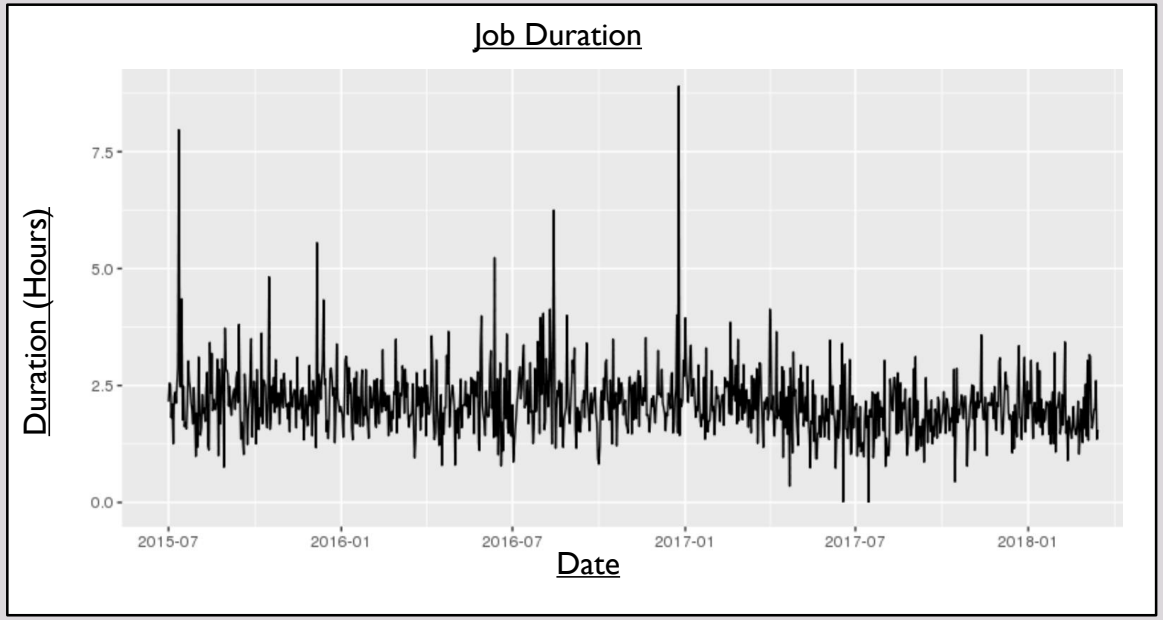
STEP 3: DURATION ANALYSIS

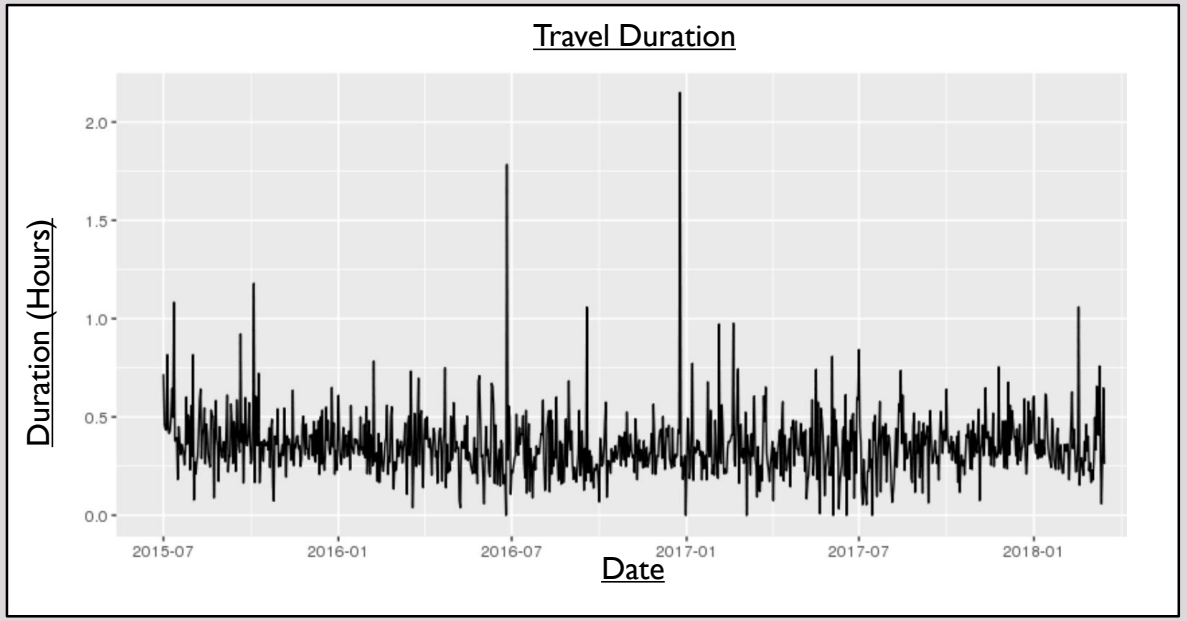


STEP 3: DURATION ANALYSIS

Appointment Duration = Travel + Job time

- No clear weekly / annual seasonality
- No clear trend

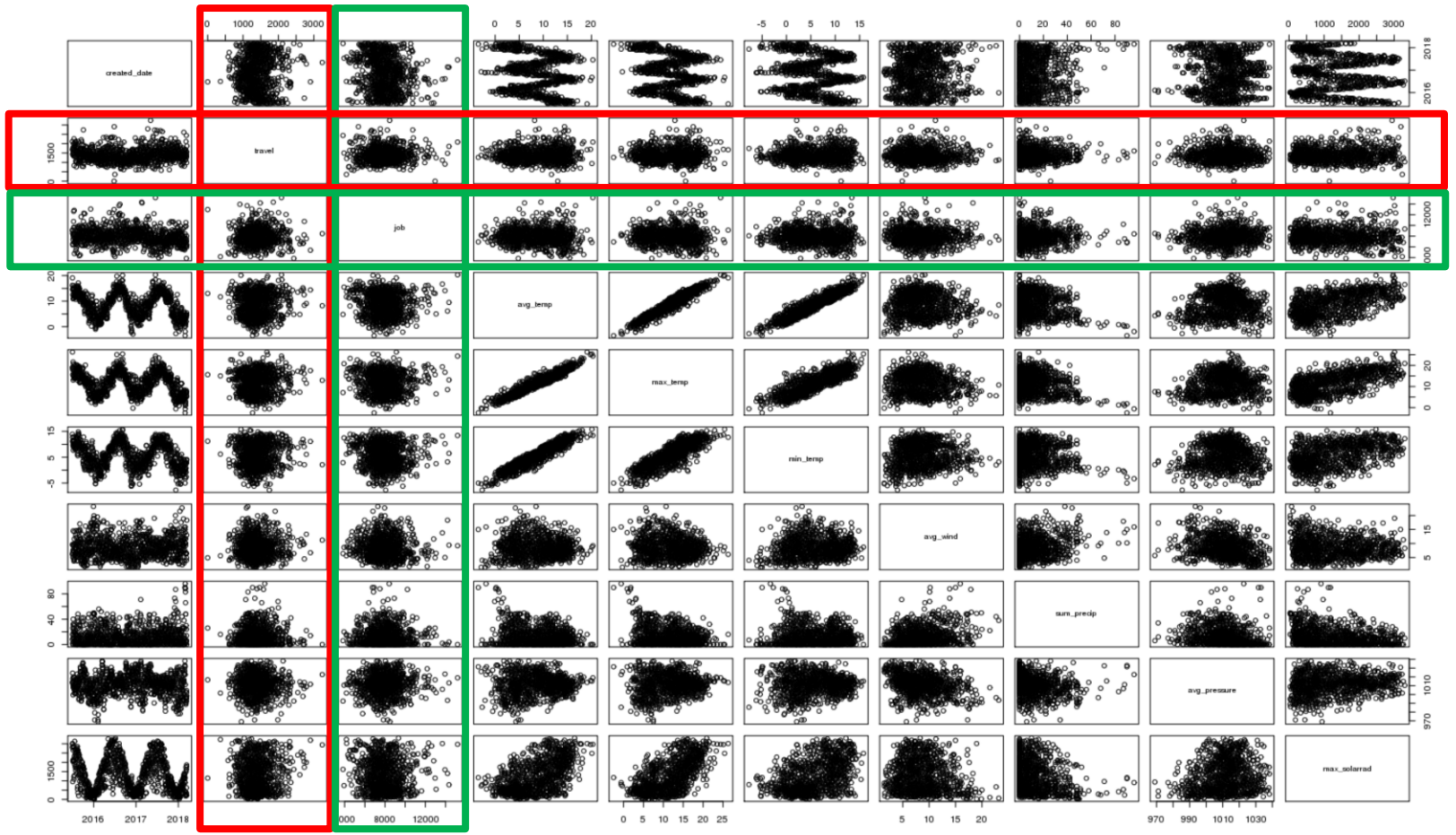




CORRELATION WITH VARIABLES

Travel

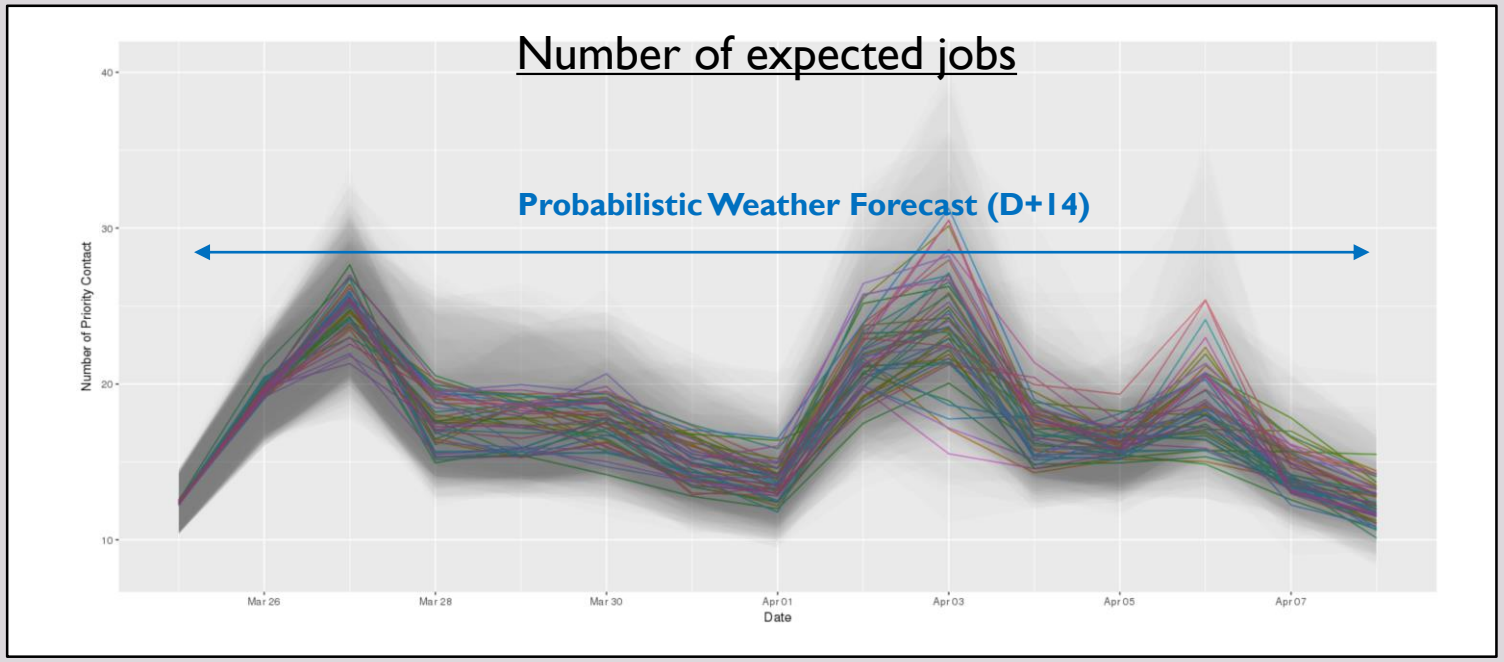
Job



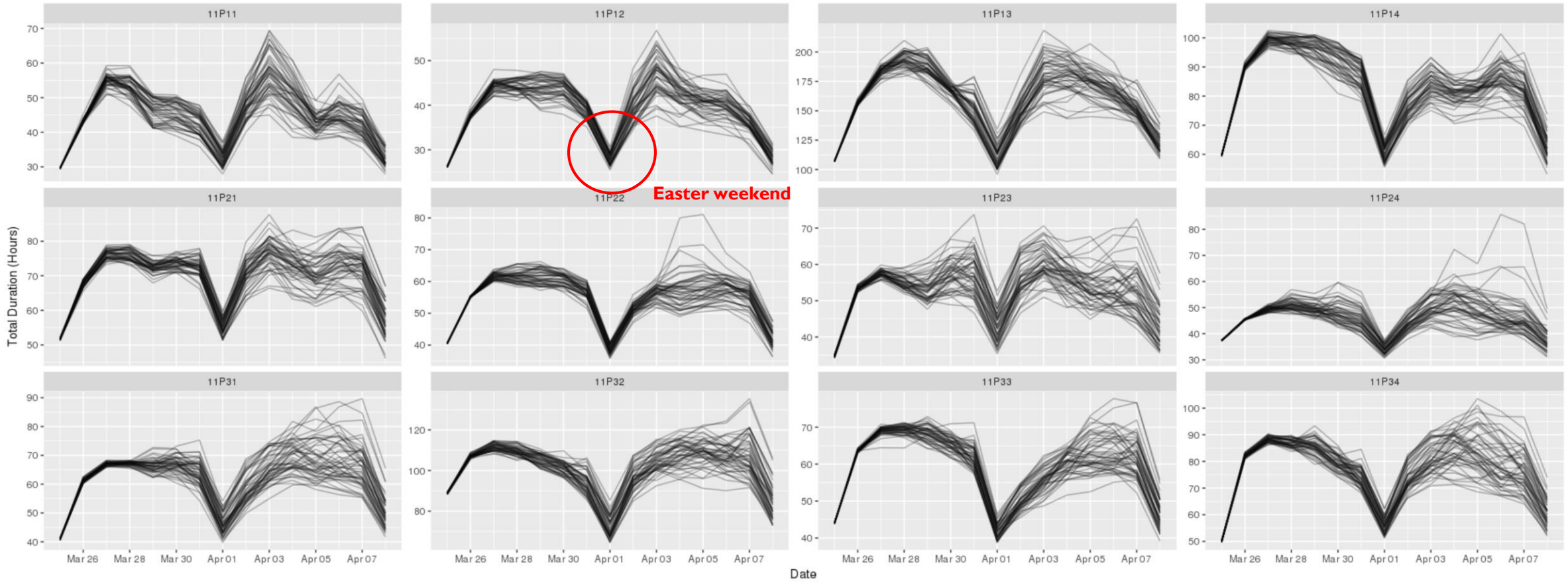
- Rolling mean was used to estimate travel / job duration

DEPLOYING MODEL

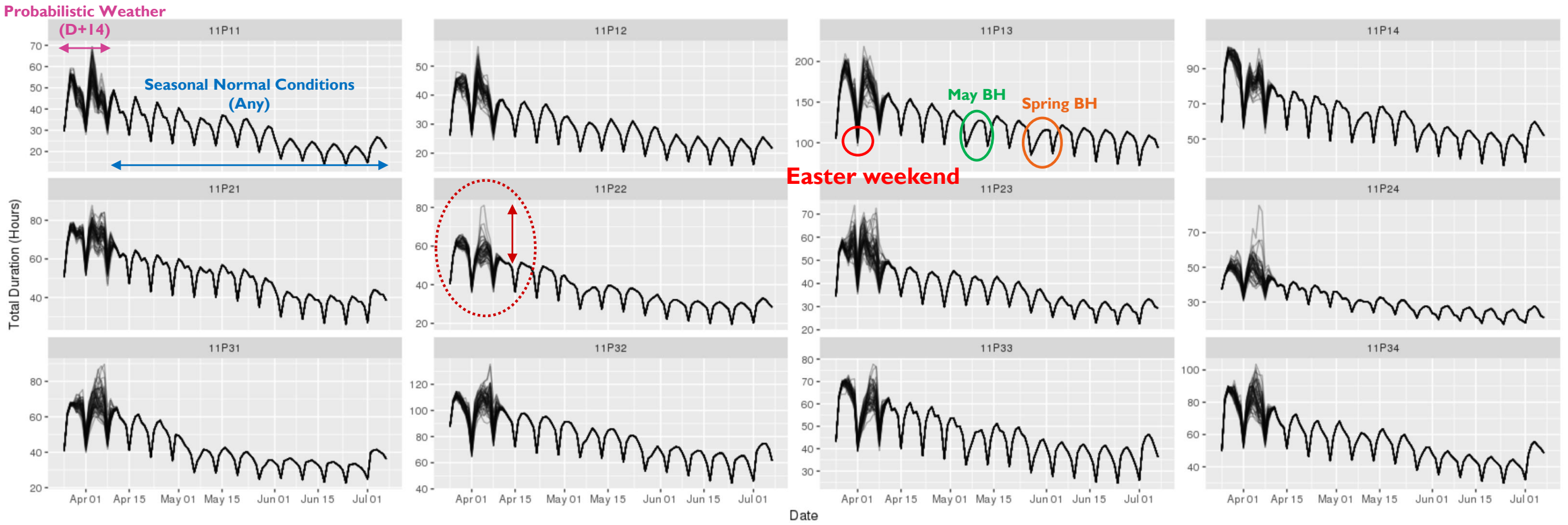
- Using proprietary ensemble weather forecast
  - 50 ensemble members
  - Distribution can be used to assess risks (e.g. worse scenario)



EXAMPLE OUTPUT: PROBABILISTIC FORECAST (14 DAYS)



EXAMPLE OUTPUT: EXTENDED PROBABILISTIC FORECAST (90 DAYS)



SUMMARY

- Patch level forecast (Long term, beyond 14 days)
- Risk-based planning using probabilistic model
- Data transformation is important
- Handling large dataset is important

THANKS – Q&A

Scan QR Code for Slides

timothywong731.github.io/talks/user2018

Timothy Wong

Senior Data Scientist, Centrica plc

timothy.wong@centrica.com

@timothywong731

timothywong731.github.io

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