Applied Data Science Assignment 2

1 Part I

1.1 Question 1

We are measuring heat intensity in a continuous manner, which suggests a normal distribution (Occam's razor).

$$t_i \sim \mathrm{N}(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + f(x_i, y_i)$$

Where f is a thin plate regression spline (isotropic).

1.2 Question 2

We are effectively measuring performance of adverts by intensity of clicks. This suggests a normal distribution (Occam's razor). We might expect the advert size to follow some non-linear relationship. This is because regardless of size, adverts at the top of the webpage will get more views than adverts at the bottom of the webpage.

$$C_i \sim \mathrm{N}(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + f_1(x_i, y_i) + f_2(A_i)$$

Where f_1 is a thin plate regression spline (isotropic) and f_2 is a cubic regression spline.

1.3 Question 3

We have 100 plants segregated into groups of 20, each group sprayed with a different concentration. This suggests a binomial distribution with 20 trials and a probability of infection for the i-th group. We choose the logit link transformation since we require probabilities. We might expect the relationship to be non-linear since when increasing concentration gradually, the infection will reduce faster up until a certain point, where it will slow down (i.e. increasing the concentration of the spray after a certain concentration is not significantly more effective than the previous concentration).

$$I_i \sim \mathrm{Binomial}(20, p_i), \quad \mathrm{logit}(p_i) = \beta_0 + f(c_i)$$

Where p_i is the probability of infection for group i, and f is a cubic regression spline.

1.4 Question 4

We try the default distribution (normal) since there are no obvious choices from the context of the problem (Occam's razor). We expect some seasonal non-linear pattern with temperature. It is difficult to say whether there will be a seasonal trend with monthly costs of wool since if the Scottish manufacturer buys wool in bulk from a global market, then this trend could be very non-linear and not necessary cyclic (economy is becoming increasingly unstable). However, the number of hats they need to produce is strongly related to how cold it is which suggests a non-linear interaction term between temperature and wool.

$$c_i \sim \mathrm{N}(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + f_1(t_i) + f_2(w_i) + f_3(t_i, w_i)$$

Where f_1 is a cyclic cubic regression spline, f_2 is a cyclic cubic regression spline, and f_3 is a tensor product spline (wool and temperature are not on the same scale)

1.5 Question 5

From context, two events happen: elephant in the $i-$th grid square, elephant not in $i-$th grid square. This implies binomial with 1 trial i.e. bernoulli. We choose the logit link transformation since we require probabilities.

$$E_i \sim \text{Bernoulli}(p_i), \quad \text{logit}(p_i) = \beta_0 + f(x_i, y_i)$$

where p_i is the probability of finding an elephant in the i-th grid cell, and f is a thin plate regression spline (isotropic).

2 Part II

2.1 Question 1: Briefly examine and comment on the data using some simple summaries/plots.

We initially conducted some exploratory data analysis

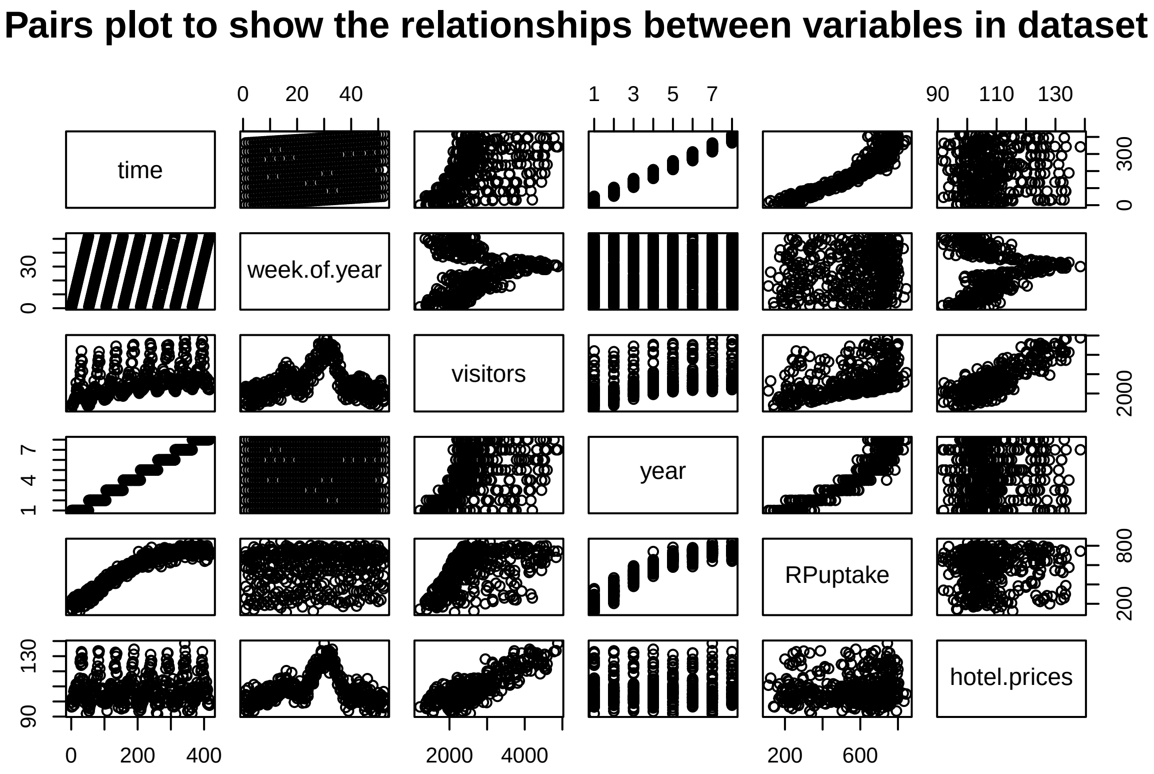

Figure 1: Pairs plot to show the relationships between variables in dataset

We observe some relationships we expect to see:

- `visitors` is positively correlated with `RPuptake`, `hotel.prices`, `time` and `year`. Reduced MoMO tickets for residents increase accessibility, and increase in hotel prices is an indicator of increasing number of tourists in the area.

- There is no significant relationship between `hotel.prices` and `RPuptake`. Intuitively, residents are unlikely to book hotels near where they live.
- There non-linear relationship between `week.of.year` and `hotel.prices` which indicates seasonal increases and decreases in hotel prices. This relationship is very similar to the relationship between `visitors` and `week.of.year`. We can also visualize this relationship further in the following time series plots.

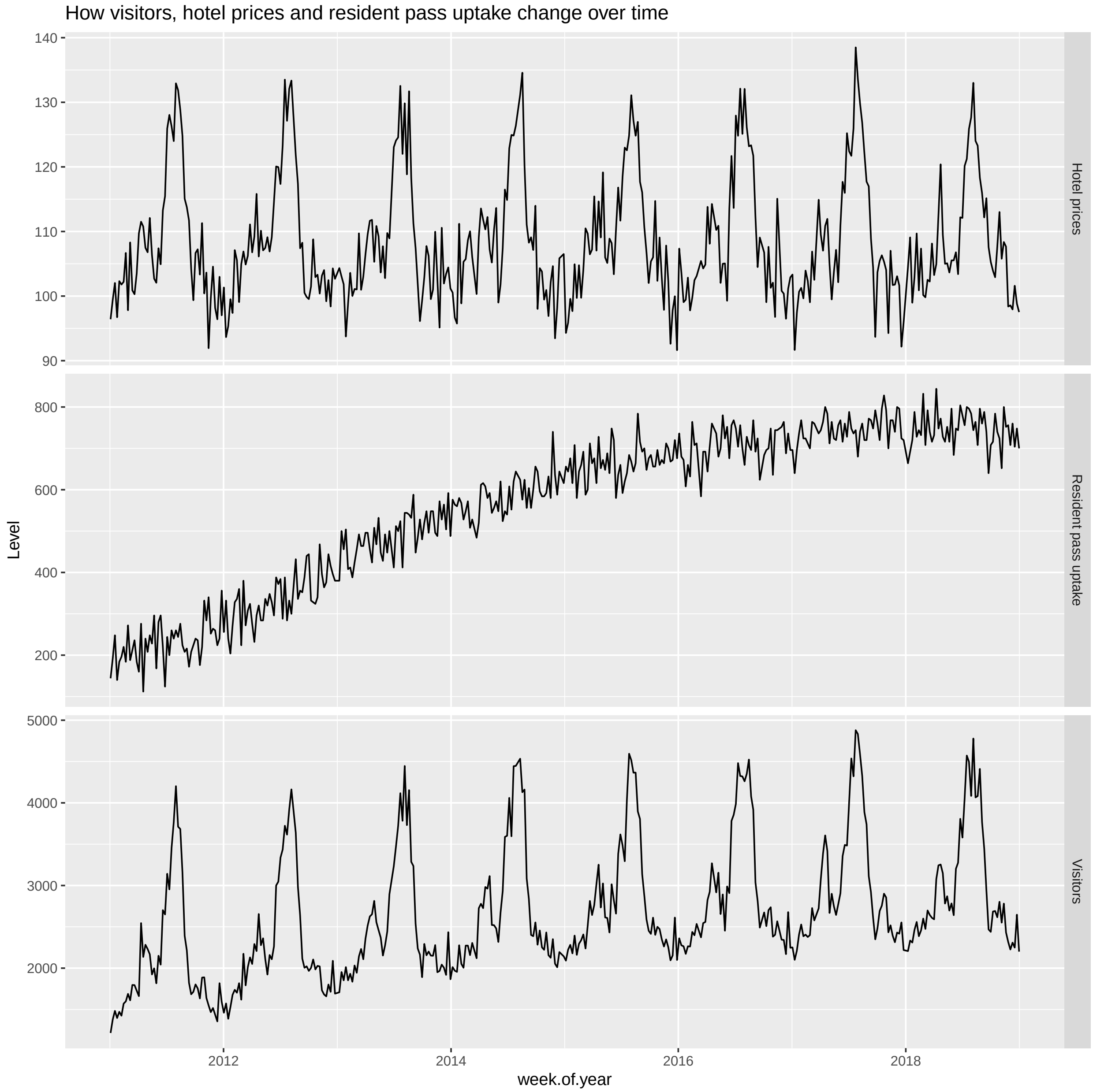

Figure 2: How visitors, hotel prices and resident pass uptake change over time

We observe that the `hotel.prices` time series and `visitors` time series seem to contain a fair amount of short-term noisy auto-correlation, and a strong seasonal trend in the form of a yearly cycle. It is also clear that there seems to be an increase in average number of visitors every year. The uptake of resident passes seems to increase and then stagnate.

We note that our `time` data is not equally spaced (Appendix question 1). This is important to note since normally when applying the relevant time series methodologies we theoretically should have regularly spaced data with no missing values.

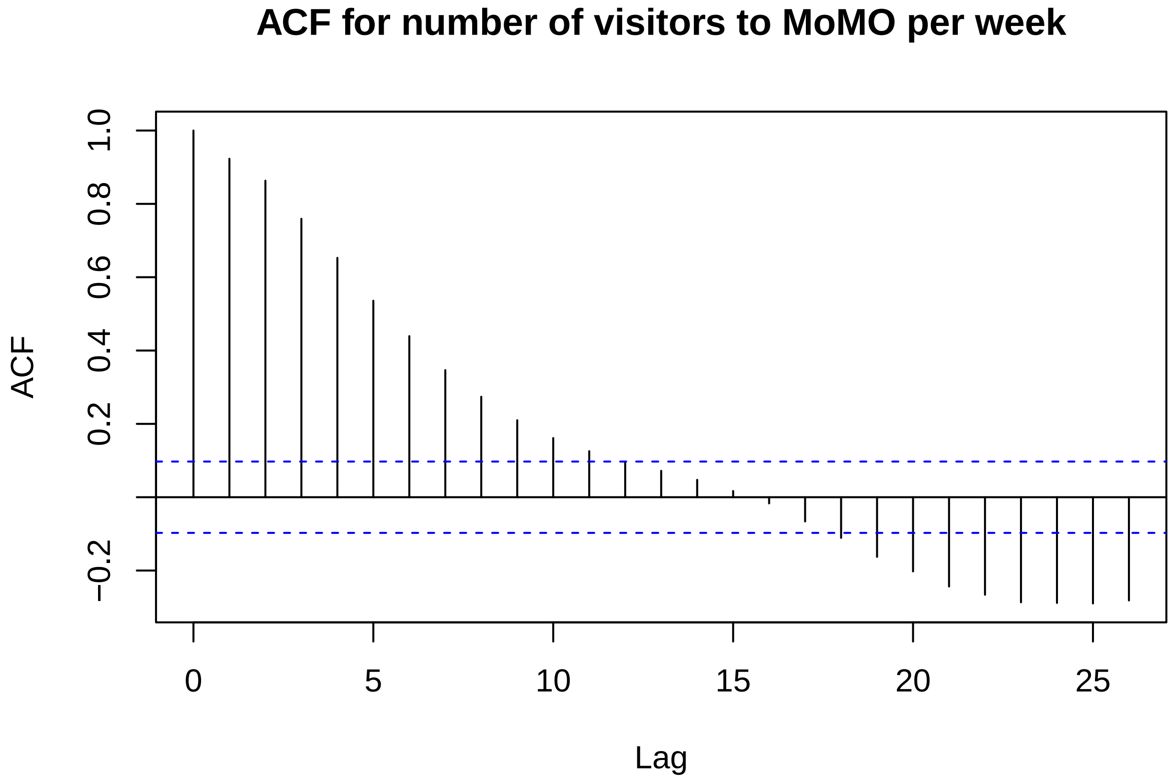

Figure 3: ACF for number of visitors to MoMO per week (MA30085: Time Series)

We observe strong auto correlations at a large range of lags. It will be critical to take of this auto-correlation when trying to fit a model if we would like to understand any predictor relationships.

2.2 Question 2: Fit a model with a normal response distribution and response variable ‘square root of `visitors`’ that estimates how visitor numbers are changing over time

2.2.1 Bullet Point 1: Explain your model specification choices and summarise what you conclude from the output.

When choosing a good fitting model, it is often a good idea to look at similar problems for inspiration. Our model specifications are therefore inspired by Emiko Dupont’s ‘Temperature of Ciaro’ model. From question 1, we observed that there was seasonal cyclic behaviour within seasons for the number of visitors visiting MoMO, with the average number of visitors increasing year by year. This is very similar to the ‘Temperature of Ciaro’ model, where seasonal increases and decreases in temperature, gradually increasing year by year where observed, perhaps due to global warming. Therefore, we consider models in the form,

$$\sqrt{\text{visitors}_i} = \beta_0 + f_1(\text{time}_i) + f_2(\text{week.of.year}_i)$$

We investigate the cases where f_1 is a cyclic cubic regression spline or p-spline respectively. We also consolidate our findings with a comparison to a polynomial approach (Appendix question 2)

```
mod1 = gam(sqrt(visitors)~s(time ,bs = 'cr', k=10) + s(week.of.year, bs="cp",k=10),
           data = dat, gamma = 1.5)
mod2 = gam(sqrt(visitors)~s(time ,bs = 'cr', k=20) + s(week.of.year, bs="cp",k=20),
           data = dat, gamma = 1.5)
mod3 = gam(sqrt(visitors)~s(time ,bs = 'cr', k=30) + s(week.of.year, bs="cp",k=30),
           data = dat, gamma = 1.5)
```

Table 1: Model summary statistics table

K	EDF time	EDF week of the year	Deviance explained	GCV	AIC
10	2.981847	8.878018	94.41%	3.258674	1619.988
20	3.067564	14.759928	95.23%	2.916083	1567.808
30	3.082177	16.436424	95.31%	2.904673	1564.128

One could argue that when K was either 10 or 20, the EDF of f_2 was getting close to the maximum complexity of f_2. Therefore, incorporating the aim of keeping the basis size as small as possible, taking K equal to 30 certainly seemed like a reasonable basis size ($16 \ll 29$). Indeed, we observe that a basis size of 30 for both smooth terms is able to capture the behaviour of the response variable well with a deviance explained of 95.31%. We observe that the AIC and GCV lowest for basis size of 30, indicating that the increased basis size is significant, and that we are not overfitting. There is no significant difference between the use of a cyclic p-spline or cubic regression spline (table 8). However, the cyclic p-spline specification for f_2 had very marginally lower GCV and AIC.

GCV with $\gamma = 1.5$ was chosen over REML due to superior prediction performance measured using 10-fold cross validation (table 9). We believe that a low prediction error is very important in this application since we would like to predict missing values of visitors in question 3.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## sqrt(visitors) ~ s(time, bs = "cr", k = 30) + s(week.of.year,
##     bs = "cp", k = 30)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.01272    0.08022   635.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df     F p-value
## s(time)         3.082   3.85 429.4  <2e-16 ***
## s(week.of.year) 16.436  29.00 206.7  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## R-sq.(adj) =  0.951   Deviance explained = 95.3%
## GCV = 2.9047  Scale est. = 2.613     n = 406
```

The approximate p-values from the `summary` command show that both smooths f_1 and f_2 are very significant. The significant p-value associated with f_1 suggests that there is a significant trend of increasing numbers of visitors to the museum over time.

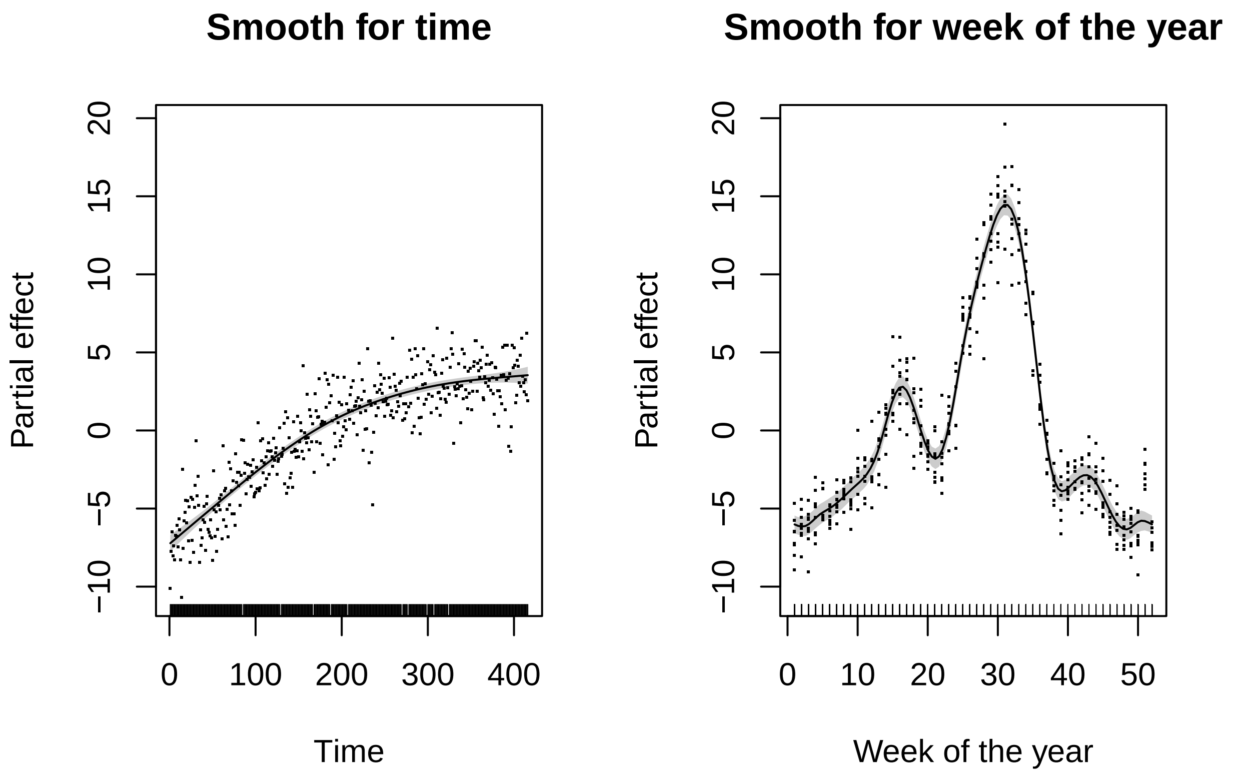

Figure 4: Partial effect plots for mod3

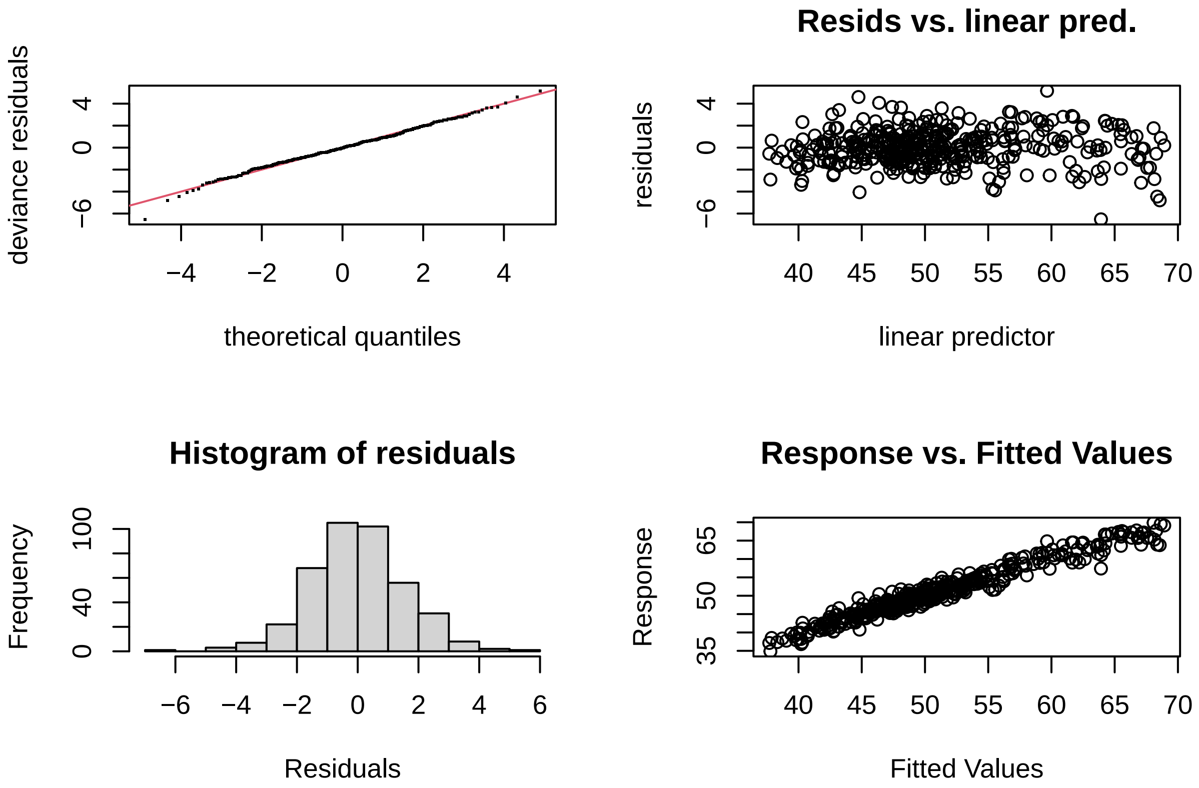

Figure 5: Model fit diagnostics for mod3

Table 1 and figure 4 show the estimated smooth terms and corresponding partial effect plots for `mod3`. We are not surprised to see a wiggly seasonal pattern for f_2. This is because at MoMO there are special exhibitions and events through the year, in particular, during school holidays. We observe three peaks: Easter break (peak approximately week 16), summer break (peak approximately week 30) and Autumn half term (peak approximately week 42). The f_1 smooth confidence interval excludes 0, and therefore provides more evidence that there is an significant increase in visitors over time. The residual plots also confirms that it is a reasonable fit.

2.2.2 Bullet point 2: What might be the reason for modelling the square root of `visitors` rather than `visitors`.

We choose to perform a variance stabilization transform on the response variable, `visitors`; precisely, we choose to model the response variable as $\sqrt{\text{visitors}}$. This is because we observe, from the model diagnostic plots, that the variance seems to be non-constant when `visitors` is modelled as the response variable. In the linear predictor vs residuals plot, the variation around zero changes for increasing values of the linear predictor, which is inconsistent with our white noise assumption.

```
mod_var_unstab = gam(visitors~s(time ,k=30) + s(week.of.year, bs="cp",k=30),
                     data = dat, gamma=1.5)
```

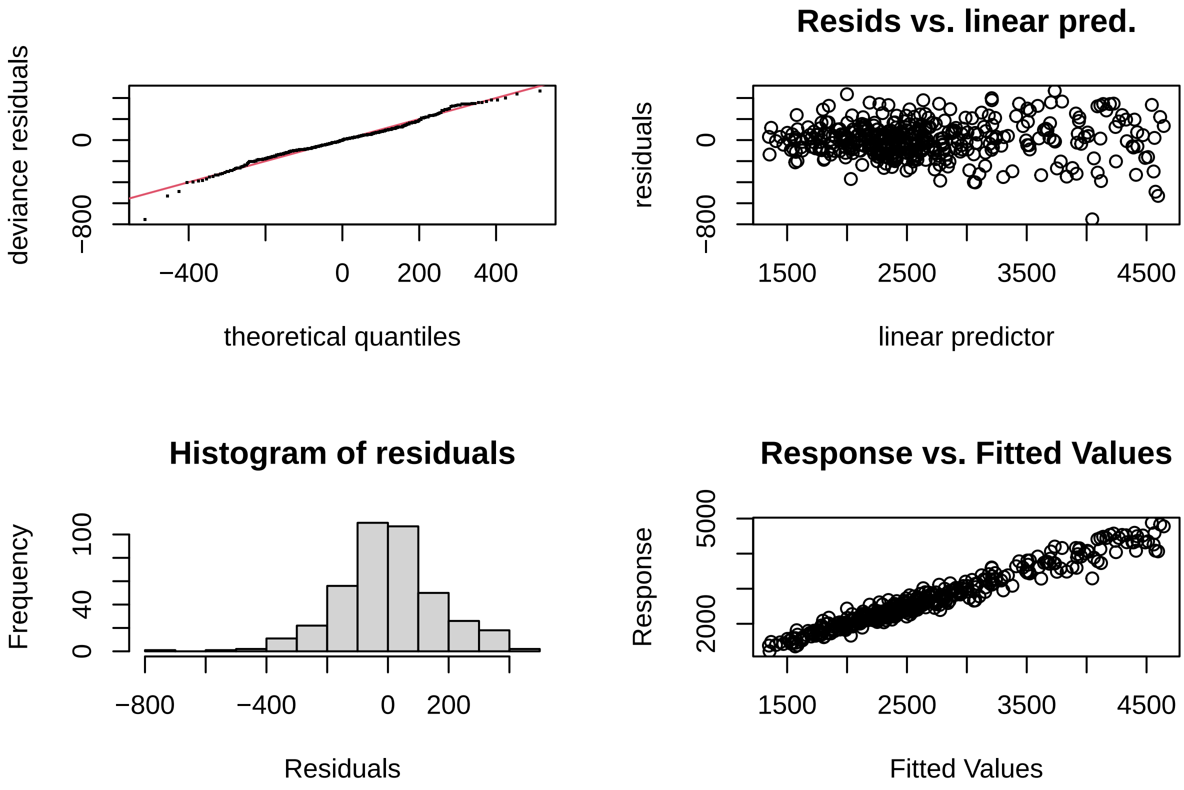

Figure 6: Model fit diagnostics for model fit without variance stabilisation

2.2.3 Bullet point 3: Create a line plot of the data overlaid by the predictions from your model

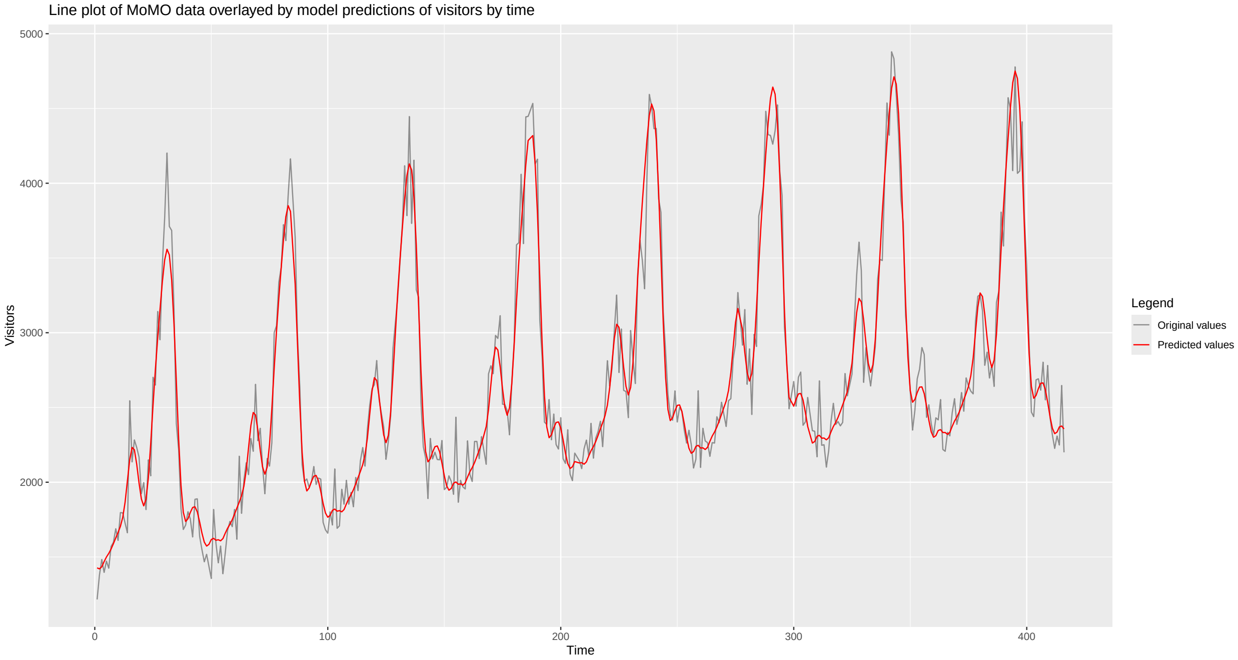

Figure 7: Line plot of MoMO data overlayed by model predictions of visitors by time

2.3 Question 3: There are some weeks during the period where the visitor numbers were not recorded

2.3.1 Bullet point 1: Use your model to estimate these missing values.

Table 2: Model (mod3) estimates of missing visitor values

Time	Week of the year	Predicted number of visitors
85	33	3636
129	25	3002
167	11	2303
187	31	4363
207	51	2146
270	10	2497
277	17	3140
299	39	2490
307	47	2297
324	12	2677

2.3.2 Bullet point 2: Create a line plot of the data showing the predicted points

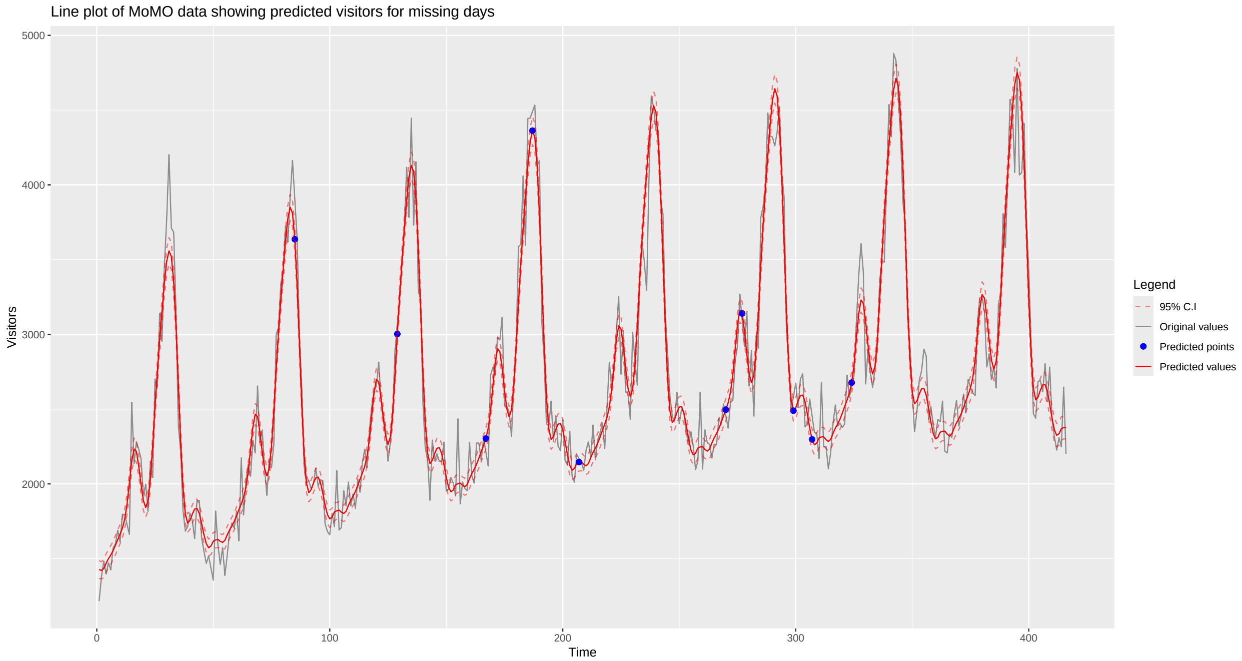

Figure 8: Line plot to show the predicted visitor values by the model over the time, including predictions for missing values

2.3.3 Bullet point 3: Comment on the appropriateness of using the model for this purpose

We observe from figure 8 that there seems to be some slight underfitting for earlier times in peak summer periods. The model response variable was squared to undo the variance stabilizing transform, which slightly increases the uncertainty. Therefore, the number of visitors at `time` 187 seems to be slightly underestimated. However, the rest of the predicted points (blue points on figure 8) seem to be visually sensible considering the original data. We somewhat expect this. We use a linear combination of different basis functions associated with smooths f_1 and f_2 in a fixed time interval (1 to 416). We have a high deviance explained which means that the basis functions will be good at explaining the data in their respective local neighbourhoods. We have also taken care to reduce the chance of the model being overfitted. The model has a low AIC and has been fitted by minimizing the GCV which captures in-sample prediction error. In conclusion, we feel that the model is in general appropriate for this purpose.

2.4 Question 4: in January 2011 the city council introduced a residents' pass that gives local residents heavily discounted entry to museums, including MoMO. The variable RPuptake records the weekly uptake of this pass. You have also been given the data hotel.prices which records the average price per night of a hotel room in the city.

2.4.1 Bullet point 1: Can these variables be used to improve your model from Q2 above?

Initially we try models of the following form,

$$\sqrt{\text{visitors}_i} = \beta_0 + f_1(\text{time}_i) + f_2(\text{week.of.year}_i) + f_3(\text{hotel.prices}_i) + f_4(\text{RPuptake}_i)$$

```
mod4 = gam(sqrt(visitors)~s(time ,bs = 'cr', k=30) + s(week.of.year, bs="cp",k=30) +
             s(RPuptake, k=30) + s(hotel.prices, k=30), data = dat, gamma = 1.5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## sqrt(visitors) ~ s(time, bs = "cr", k = 30) + s(week.of.year,
##     bs = "cp", k = 30) + s(RPuptake, k = 30) + s(hotel.prices,
##     k = 30)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.01272    0.06409     796   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df       F p-value
## s(time)         13.65  16.43   1.374   0.184
## s(week.of.year) 17.64  29.00  37.445  <2e-16 ***
## s(RPuptake)      1.00   1.00 155.805  <2e-16 ***
## s(hotel.prices)  1.00   1.00  55.009  <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.969   Deviance explained = 97.1%
## GCV = 2.0019  Scale est. = 1.6676    n = 406
```

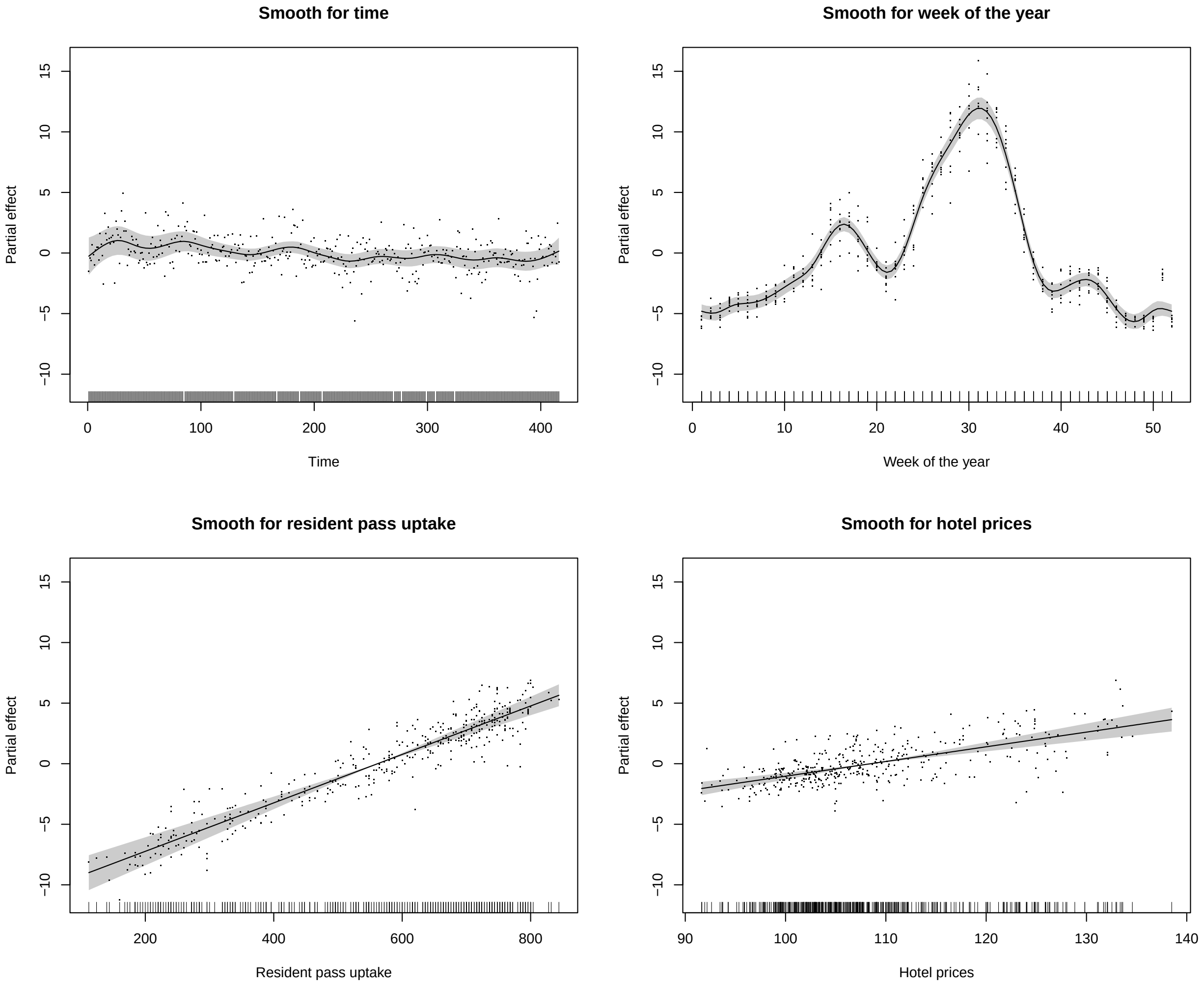

Figure 9: Partial effect plots for mod4

We notice that the p-value for the `time` smooth, f_1, is no longer significant, and indeed, the confidence interval includes 0. The p-values associated with the smooths f_3 and f_4 are significant. The smooths f_3 and f_4 clearly steal the explanatory power away from f_1, indicating that the variables `RPuptake` and `hotel.prices` are more important in explaining the trend in increasing visitors. The EDF value for smooths f_3 and f_4 is 1, and remains equal to 1 when increasing the number of basis functions significantly higher than 30. This indicates that a linear relationship is perhaps preferred (with two very similar competing models it is best to use the simpler one (Occam's razor)). Therefore, we use the variables `hotel.prices` and `RPuptake` additively in the model without being wrapped by a smooth function. This prompts the following model.

$$\sqrt{\text{visitors}_i} = \beta_0 + f_2(\text{week.of.year}_i) + \text{hotel.prices}_i + \text{RPuptake}_i$$

```
k= 30
mod5 = gam(sqrt(visitors)~ s(week.of.year, bs="cp",k=k) +
             RPuptake + hotel.prices, data = dat, gamma = 1.5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## sqrt(visitors) ~ s(week.of.year, bs = "cp", k = k) + RPuptake +
##     hotel.prices
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.754e+01  1.806e+00  15.249  < 2e-16 ***
## RPuptake     1.776e-02  3.494e-04  50.821  < 2e-16 ***
## hotel.prices 1.245e-01  1.659e-02   7.499 4.46e-13 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df     F p-value
## s(week.of.year) 17.61     29 38.81  <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.967   Deviance explained = 96.8%
## GCV =  1.955  Scale est. = 1.7578    n = 406
```

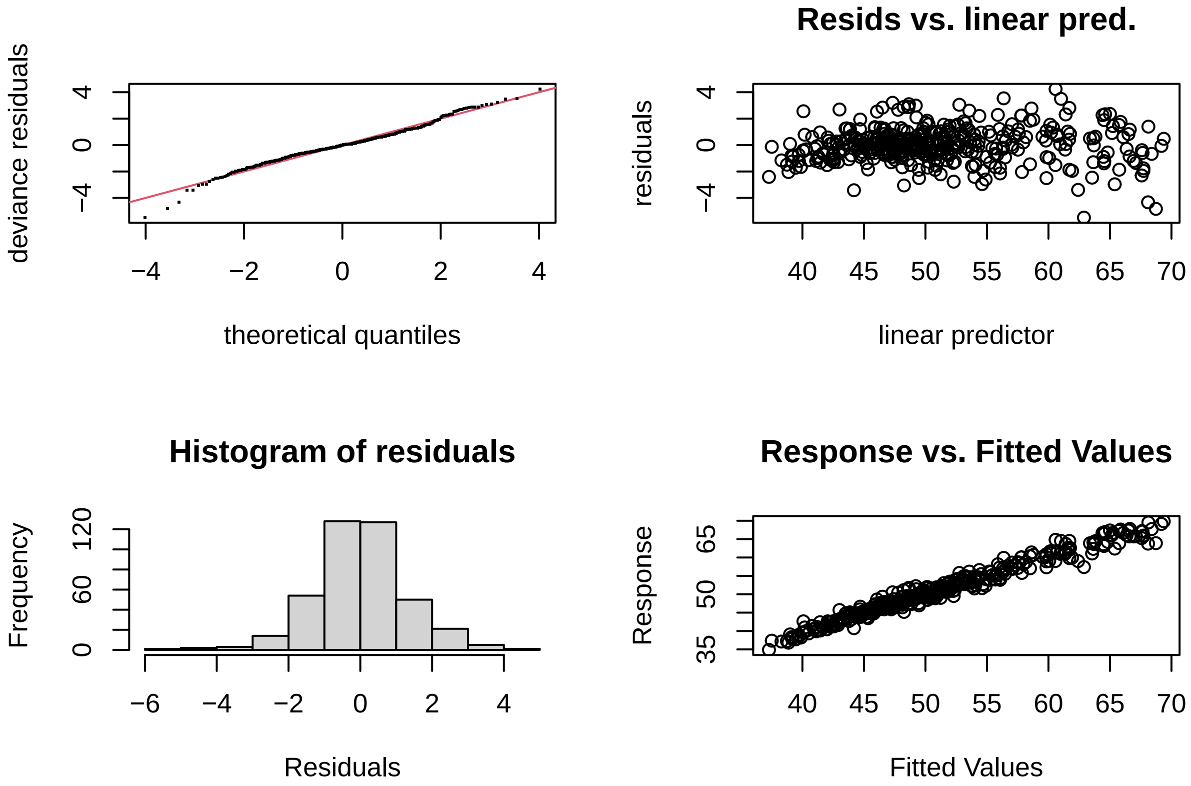

Figure 10: Model diagonstic plots for mod5

Table 3: Model summary statistics table

Model	Deviance explained	GCV	AIC
mod3	95.31%	2.904673	1564.128
mod5	96.85%	1.954953	1403.251

We observe that with the new specification (`mod5`) all terms are significant. The deviance explained of `mod5` is larger than `mod3` which implies that the inclusion of the terms `hotel.prices` and `RPuptake` has given the model more explanatory power. It seems like when separating the seasonal yearly cycle component, the variables `RPuptake` and `hote.prices` are more important at explaining the increase in visitors to MoMO than the variable `time`. From table 3 we also observe that the GCV and AIC for `mod5` are significantly smaller than `mod3`. We also observe that the residual plots confirm that `mod5` is a reasonable fit. Therefore, we can be confident that `mod5`, which includes the variables `RPuptake` and `hotel.prices`, is more adequate at explaining the number of visitors over the years 2011 to 2018 than `mod3`.

2.4.2 Bullet point 2: Describe and discuss the differences between different modelling approaches.

We investigate three alternative modelling approaches for `mod5`

- Polynomial approaches (A similar investigation is in Appendix Question 2)
- No smoothing
- Interaction terms

2.4.2.1 Investigating polynomial approaches

We proceed to investigate a different modelling approach by approximating $f_2(\text{week.of.year}_i)$ by a low degree polynomial (degree of 2) and a higher degree polynomial (degree of 10). The aim of this illustration is to show that GAMs can be better at modelling non-linear relationships.

```r
degree=2
formula1<-as.formula(paste("sqrt(visitors)~ hotel.prices + RPuptake+ ",
                           paste0("I(week.of.year^",1:degree,")",collapse="+"),sep=""))
mod_poly = lm(formula=formula1,data=dat)

degree=10
formula2<-as.formula(paste("sqrt(visitors)~ hotel.prices + RPuptake+ ",
                           paste0("I(week.of.year^",1:degree,")",collapse="+"),sep=""))
mod_poly1 = lm(formula=formula2,data=dat)
```

```
##
## Call:
## lm(formula = formula1, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7865 -1.5526 -0.0356  1.5312  7.5025
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.740e+01  1.582e+00 -10.996   <2e-16 ***
## hotel.prices       5.063e-01  1.635e-02  30.967   <2e-16 ***
## RPuptake           1.753e-02  6.191e-04  28.320   <2e-16 ***
## I(week.of.year^1)  3.913e-01  4.190e-02   9.338   <2e-16 ***
## I(week.of.year^2) -7.264e-03  7.710e-04  -9.422   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.35 on 401 degrees of freedom
## Multiple R-squared:  0.8969, Adjusted R-squared:  0.8959
## F-statistic: 872.2 on 4 and 401 DF,  p-value: < 2.2e-16
```

```
## [1] "AIC: 1853.0827194893"
```

```
##
## Call:
## lm(formula = formula2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3569 -1.1654 -0.1035  1.2089  5.4772
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.095e+00  2.696e+00  -0.777 0.437663
## hotel.prices        2.747e-01  1.986e-02  13.833  < 2e-16 ***
## RPuptake            1.766e-02  4.837e-04  36.507  < 2e-16 ***
## I(week.of.year^1)   1.169e+01  2.226e+00   5.253 2.45e-07 ***
## I(week.of.year^2)  -4.530e+00  8.575e-01  -5.283 2.11e-07 ***
## I(week.of.year^3)   8.111e-01  1.598e-01   5.075 5.99e-07 ***
## I(week.of.year^4)  -7.778e-02  1.677e-02  -4.638 4.80e-06 ***
## I(week.of.year^5)   4.353e-03  1.071e-03   4.065 5.80e-05 ***
## I(week.of.year^6)  -1.477e-04  4.315e-05  -3.423 0.000686 ***
## I(week.of.year^7)   3.044e-06  1.103e-06   2.760 0.006042 **
## I(week.of.year^8)  -3.664e-08  1.733e-08  -2.115 0.035051 *
## I(week.of.year^9)   2.303e-10  1.526e-10   1.509 0.132014
## I(week.of.year^10) -5.506e-13  5.764e-13  -0.955 0.340002
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.835 on 393 degrees of freedom
## Multiple R-squared:  0.9384, Adjusted R-squared:  0.9365
## F-statistic: 498.8 on 12 and 393 DF,  p-value: < 2.2e-16
```

```
## [1] "AIC: 1660.10509935189"
```

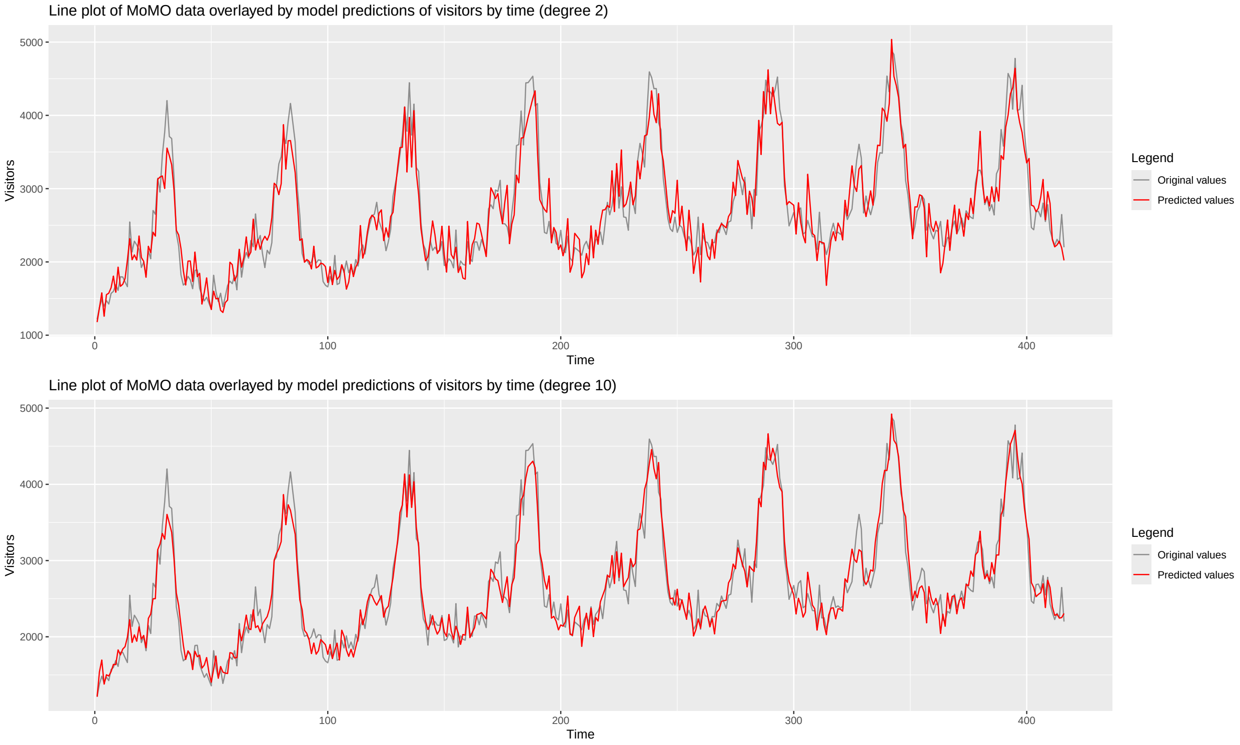

Figure 11: Line plots of MoMO data overlayed by model predictions of visitors by time for polynomials of different degrees

Although all terms for the degree two model are significant, it is clear from figure 11 that the it fit is not ideal. Naturally, increasing the polynomial order the predicted values become closer to the data. We observe this in the summary statistics with increasing R^2 values. However, not all the terms for the degree 10 model are significant; the interpretation being that the model has a few terms which are quite similar so divides out all the explanatory power. With the GAM approach, we have far more flexibility and interpretability. Interpreting one single significant smooth is far easier and more meaningful than interpreting high order polynomial coefficients.

2.4.2.2 Investigation into the use of no smoothing

We now investigate how the use of smoothing parameters impacts model fit and performance.

```
k=30
mod6 = gam(sqrt(visitors)~ s(week.of.year, bs="cp",k=k, fx=TRUE) +
             RPuptake + hotel.prices, data = dat, gamma = 1.5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## sqrt(visitors) ~ s(week.of.year, bs = "cp", k = k, fx = TRUE) +
##     RPuptake + hotel.prices
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.860e+01  1.841e+00   15.54  < 2e-16 ***
## RPuptake     1.775e-02  3.486e-04   50.91  < 2e-16 ***
## hotel.prices 1.147e-01  1.692e-02    6.78  4.7e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                 edf Ref.df     F p-value
## s(week.of.year)  29     29 40.51  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.967   Deviance explained =   97%
## GCV = 2.0705  Scale est. = 1.7476    n = 406
```

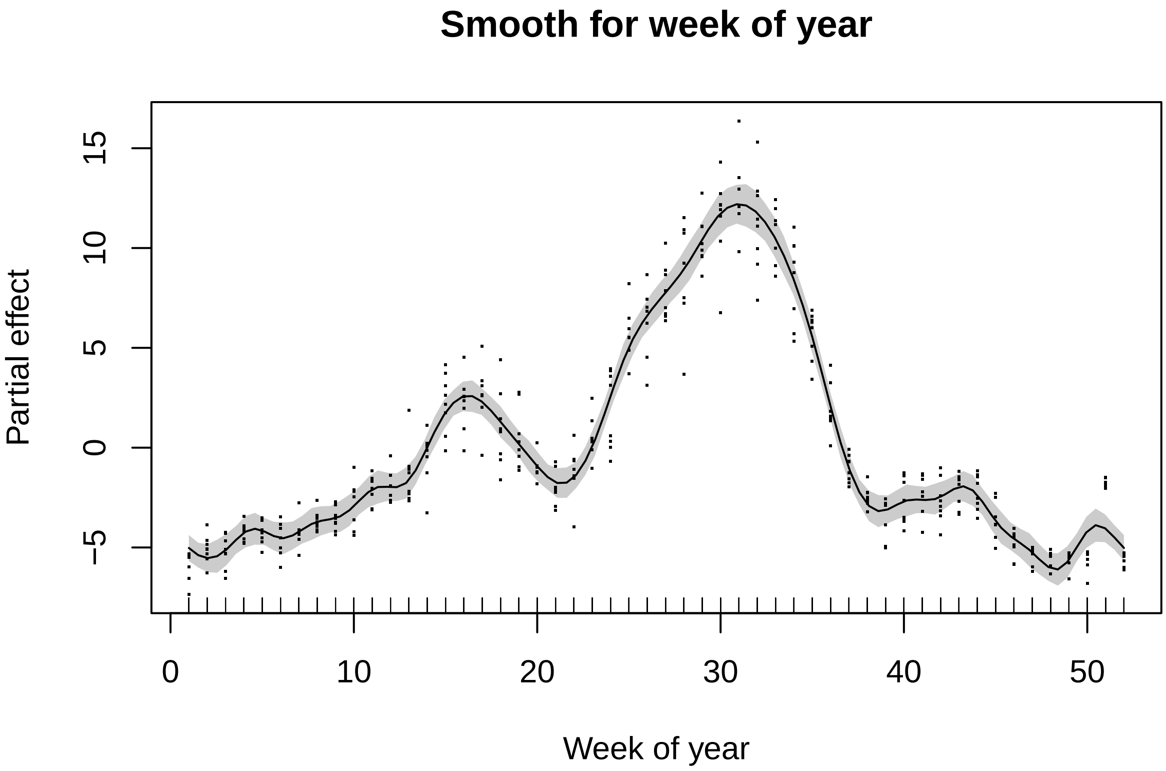

Figure 12: Partial effects plot for mod6 (mod5 without smoothing)

We observe that although the deviance explained for `mod6` is higher than `mod5`, the smooth is now very wiggly. This might imply that the smooth is modelling the noise, rather than any underlying pattern in the data (or overfitting the data). We therefore expect that the model with no smoothing parameter (`mod6`) will have a higher cross validation score than the model with smoothing parameter (`mod5`).

Table 4: Model statistics summary table

Model	10-fold cross validation score	Deviance explained	GCV	AIC
mod6	1.942308	96.96%	2.070507	1411.496
mod5	1.887205	96.85%	1.954953	1403.251

We observe that the model with the smoothing parameter (`mod5`) has a lower 10-fold cross validation score than the model with no smoothing parameter (`mod6`). The use of no smoothing has also increased the AIC and GCV which is undesirable. We can therefore confirm that smoothing is beneficial for our fitted GAM.

2.4.2.3 Investigation into modelling with interactions

We might suspect that if there is a particularly large uptake on the resident pass, and there is a spike in the hotel prices, then this may indicate a larger effect on the number of visitors to MoMO. We investigate this hypothesis by investigating the use of a tensor product spline in `mod5` to account for an interaction effect between `hotel.prices` and `RPuptake`.

```
mod7 = gam(sqrt(visitors)~ s(week.of.year, bs="cp", k=k) +
             te(hotel.prices, RPuptake, bs = c('cr','cr'), k=c(10,10)),
           data = dat, gamma = 1.5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## sqrt(visitors) ~ s(week.of.year, bs = "cp", k = k) + te(hotel.prices,
##     RPuptake, bs = c("cr", "cr"), k = c(10, 10))
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.01272    0.06334   805.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                              edf Ref.df     F p-value
## s(week.of.year)            18.04     29  42.1  <2e-16 ***
## te(hotel.prices,RPuptake)  3.00      3 968.7  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.969   Deviance explained = 97.1%
## GCV = 1.8258  Scale est. = 1.629     n = 406
```

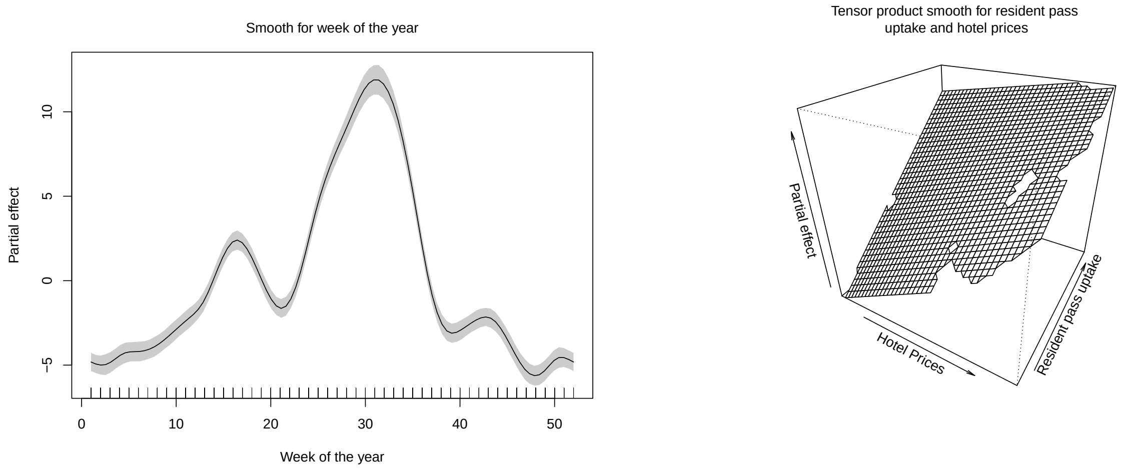

Figure 13: Partial effect plots for mod7

Table 5: Model summary statistics table

Model	Deviance explained	GCV	AIC
mod5	96.85%	1.954953	1403.251
mod7	97.09%	1.825823	1373.719

We notice that with the inclusion of the interaction term, both smooths are significant. Large values of `hotel.prices` and `RPuptake` result in a larger effect than small values of `RPuptake` and `hotel.prices`. Although, since the EDF of the tensor product spline is three, we may investigate the following model for comparison.

```
mod8 = gam(sqrt(visitors)~ s(week.of.year, bs="cp",k=30) +
             RPuptake + hotel.prices + RPuptake:hotel.prices, data = dat, gamma = 1.5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## sqrt(visitors) ~ s(week.of.year, bs = "cp", k = 30) + RPuptake +
##     hotel.prices + RPuptake:hotel.prices
##
## Parametric coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.551e+01  2.805e+00   5.529 5.96e-08 ***
## RPuptake               3.911e-02  3.881e-03  10.078  < 2e-16 ***
## hotel.prices           2.353e-01  2.583e-02   9.111  < 2e-16 ***
## RPuptake:hotel.prices -1.967e-04  3.561e-05  -5.524 6.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df     F p-value
## s(week.of.year) 18.02     29 41.91  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.969   Deviance explained = 97.1%
## GCV = 1.8259  Scale est. = 1.6292    n = 406
```

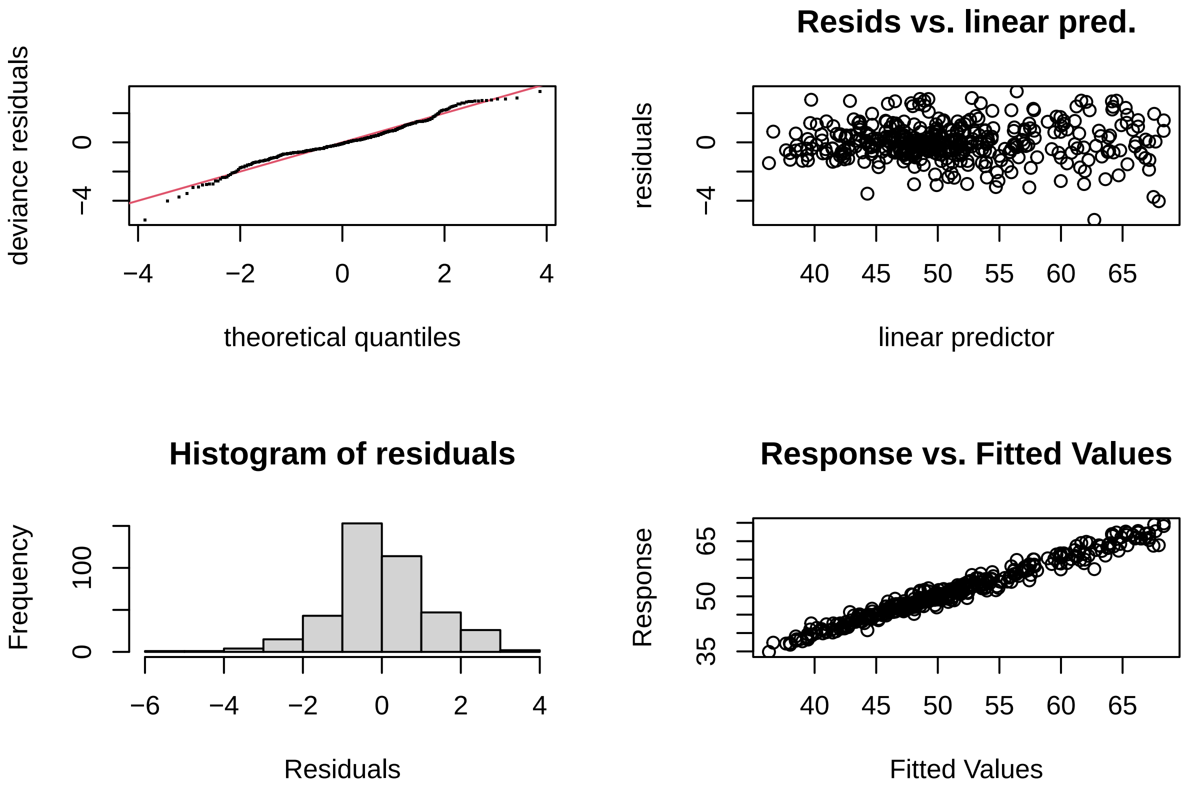

Figure 14: Model diagnostic plots for mod8

Table 6: Model summary statistics table

Model	Deviance explained	GCV	AIC
mod5	96.85%	1.954953	1403.251
mod7	97.09%	1.825823	1373.719
mod8	97.09%	1.825883	1373.751

The GCV and AIC for `mod5` are higher than for `mod7` and `mod8`, of which the respective scores are approximately the same. We also verify that the residual plots are sensible for both `mod7` (figure 19) and `mod8` (figure 14). We can confirm that the interaction terms in `mod7` and `mod8` are better at modelling the weekly visitors at MoMO over 2011-2018.

2.5 Question 5a: The museum director would like to be able to forecast how many visitors to expect next week, next month or even next year. Discuss whether your models can be used for this purpose.

2.5.1 Fitted model from question 2 (`mod3`):

We observed strong auto-correlation in figure 3. Therefore, we should verify that the model treats this auto-correlation appropriately.

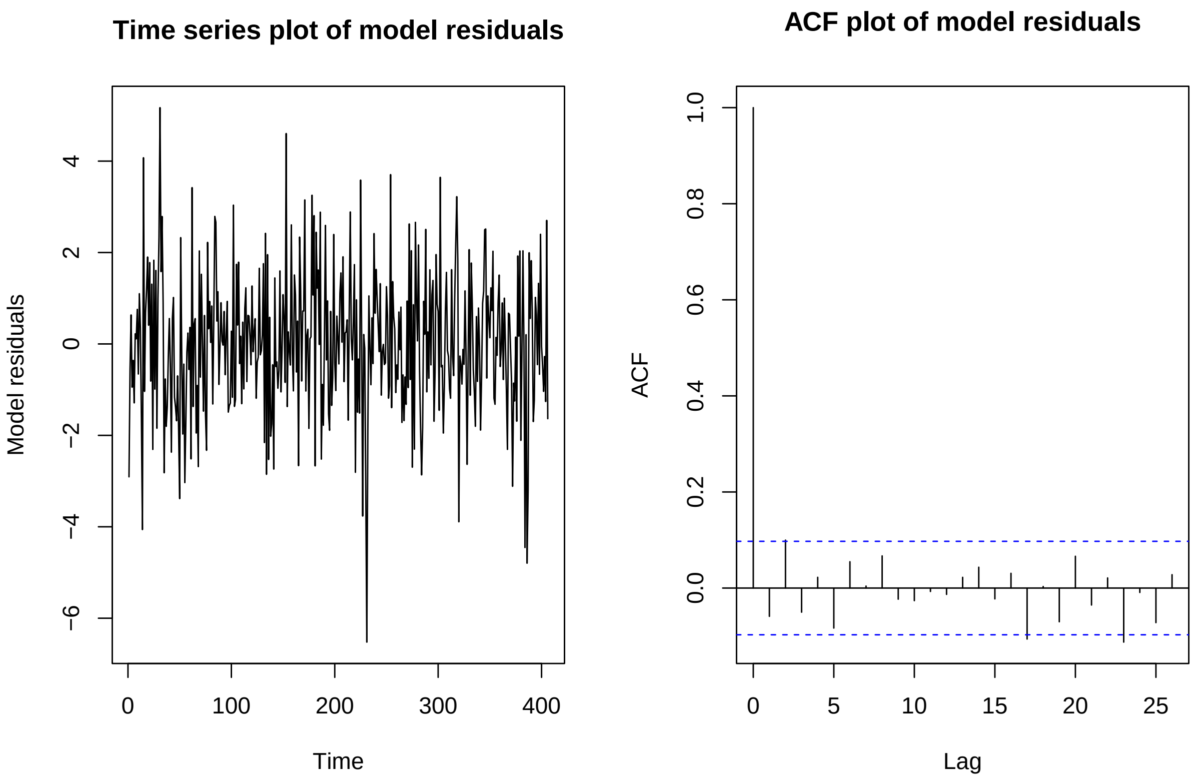

Figure 15: Time series model diagnostic plots for mod3 (MA30085 Time Series)

We observe that the mean of the residuals is close to zero, the spread of residuals is constant over time, and the sample autocorrelation function between the residuals is negligible. Therefore, we can confirm that we have controlled for that non-linear trend in the auto-correlation certainly for the time period we have data for.

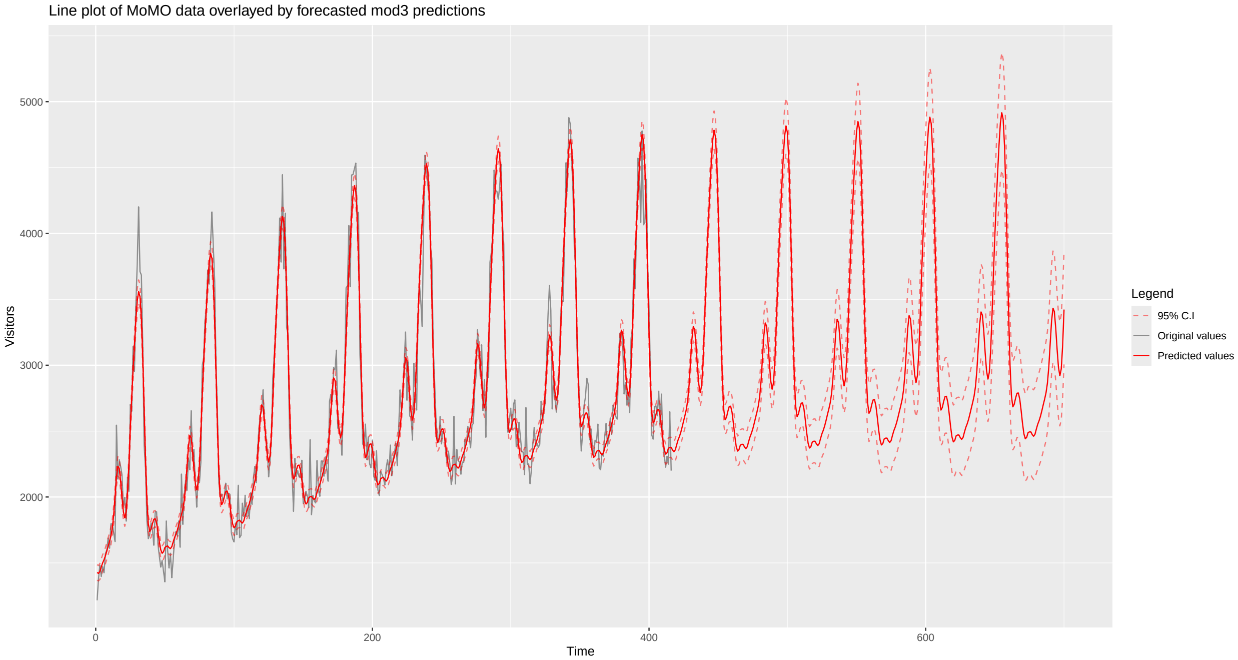

Figure 16: Extrapolation of mod3 predictions for weekly visitors outside historical time interval

Using figures 15 and 16, we can conclude that very short term forecasting, i.e. forecasting the number of visitors next week, seems okay.

However, we should be concerned about using `mod3` for forecasting months or years into the future. We observe that the forecast seems to follow the same behaviour that was observed at the end of the observed time interval. Despite the predicted continued yearly cycle, the trend is not that similar to the trend observed historically (i.e. the data the model learnt on). Although we expect the uncertainty to increase as we forecast years into the future, it seems to be growing in a less constrained way. This is because the smooth basis functions corresponding to `time` only have local knowledge. To elaborate, the splines are segmented across a fixed time interval (1 to 416), so they must only have knowledge about what is happening in their unique local neighbourhood on the fixed time interval. Therefore, unreasonable forecasting behaviour is observed when predicting years into the future for this particular model.

2.5.2 Fitted model from question 4 (`mod5`, `mod8`):

Since there is only one smooth present, that cannot be extrapolated, these models do not have the same forecasting issue as `mod3`. However, the inclusion of the terms `hotel.prices` and `RPuptake` make forecasting, even short term, difficult due to the need to forecast these variables too. If these values could be predicted accurately, then `mod5` and `mod8` are certainly good for forecasting due to the models ability to control the auto-correlation (figures 20 and 21) and being well fitted (discussed in question 4). Therefore, one added practical bonus of `mod5` and `mod8` is the ability to produce accurate forecasts for different economic scenarios based on hypothetical values for `hotel.prices` and `RPuptake`. We know, by the benefit of hind sight, that COVID-19 forced UK lockdowns in 2020. Therefore, modelling hypothetical economic scenarios would help MoMO mitigate future damages.

3 Appendix

3.1 Part II

3.1.1 Question 1

Summary of MoMO data using `summary` command.

```
##       time        week.of.year      visitors          year        RPuptake    
##  Min.   :  1.0   Min.   : 1.00   Min.   :1216   2011   :52   Min.   :112.0  
##  1st Qu.:103.2   1st Qu.:14.00   1st Qu.:2152   2015   :52   1st Qu.:412.0  
##  Median :208.5   Median :26.50   Median :2458   2018   :52   Median :620.0  
##  Mean   :208.1   Mean   :26.47   Mean   :2655   2012   :51   Mean   :561.7  
##  3rd Qu.:313.8   3rd Qu.:39.75   3rd Qu.:2997   2013   :51   3rd Qu.:720.0  
##  Max.   :416.0   Max.   :52.00   Max.   :4879   2017   :51   Max.   :844.0  
##                                                 (Other):97                  
##   hotel.prices   
##  Min.   : 91.63  
##  1st Qu.:101.97  
##  Median :106.43  
##  Mean   :108.50  
##  3rd Qu.:112.49  
##  Max.   :138.50  
## 
```

3.1.2 Question 2

3.1.2.1 Polynomial example case study

We illustrate an example of modelling the change in weekly visitors over time using the polynomial approach. We try increasingly high order polynomials for `week.of.year`, and add the `time` variable to account for the increasing trend in weekly visitors.

```r
degree=2
formula1<-as.formula(paste("sqrt(visitors)~ time + ",
                           paste0("I(week.of.year^",1:degree,")",collapse="+"),sep=""))
mod_poly = lm(formula=formula1,data=dat)

degree=10
formula2<-as.formula(paste("sqrt(visitors)~ time + ",
                           paste0("I(week.of.year^",1:degree,")",collapse="+"),sep=""))
mod_poly1 = lm(formula=formula2,data=dat)

degree=20
formula3<-as.formula(paste("sqrt(visitors)~ time + ",
                           paste0("I(week.of.year^",1:degree,")",collapse="+"),sep=""))
mod_poly2 = lm(formula=formula3,data=dat)
```

```
## 
## Call:
## lm(formula = formula1, data = dat)
## 
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -9.0045 -3.1621 -0.6066  2.5386 13.5951 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       33.938423   0.772296   43.95   <2e-16 ***
## time               0.026302   0.001870   14.06   <2e-16 ***
## I(week.of.year^1)  1.253650   0.060678   20.66   <2e-16 ***
## I(week.of.year^2) -0.023298   0.001111  -20.96   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 4.52 on 402 degrees of freedom
## Multiple R-squared:  0.6179, Adjusted R-squared:  0.615 
## F-statistic: 216.7 on 3 and 402 DF,  p-value: < 2.2e-16

## [1] "AIC: 2383.00411679407"

## 
## Call:
## lm(formula = formula2, data = dat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.9853 -1.9655  0.0911  1.7234  7.4118 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         2.768e+01  2.781e+00   9.952  < 2e-16 ***
## time                2.601e-02  1.069e-03  24.320  < 2e-16 ***
## I(week.of.year^1)   1.467e+01  3.117e+00   4.706 3.50e-06 ***
## I(week.of.year^2)  -5.741e+00  1.199e+00  -4.788 2.39e-06 ***
## I(week.of.year^3)   1.026e+00  2.235e-01   4.593 5.89e-06 ***
## I(week.of.year^4)  -9.635e-02  2.348e-02  -4.103 4.96e-05 ***
## I(week.of.year^5)   5.152e-03  1.502e-03   3.429 0.000669 ***
## I(week.of.year^6)  -1.614e-04  6.065e-05  -2.661 0.008101 ** 
## I(week.of.year^7)   2.896e-06  1.552e-06   1.866 0.062773 .  
## I(week.of.year^8)  -2.657e-08  2.438e-08  -1.090 0.276489    
## I(week.of.year^9)   7.735e-11  2.146e-10   0.360 0.718693    
## I(week.of.year^10)  2.487e-13  8.094e-13   0.307 0.758757    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 2.584 on 394 degrees of freedom
## Multiple R-squared:  0.8776, Adjusted R-squared:  0.8742 
## F-statistic: 256.9 on 11 and 394 DF,  p-value: < 2.2e-16

## [1] "AIC: 1936.71430074575"

## 
## Call:
## lm(formula = formula3, data = dat)
## 
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6934 -1.3091 -0.0182  1.3637  5.5341
##
## Coefficients: (5 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.660e+01  7.528e+00   3.533 0.000459 ***
## time                2.590e-02  8.285e-04  31.266  < 2e-16 ***
## I(week.of.year^1)   2.450e+01  1.416e+01   1.730 0.084402 .
## I(week.of.year^2)  -1.773e+01  9.808e+00  -1.808 0.071399 .
## I(week.of.year^3)   6.565e+00  3.495e+00   1.878 0.061130 .
## I(week.of.year^4)  -1.411e+00  7.426e-01  -1.901 0.058091 .
## I(week.of.year^5)   1.913e-01  1.021e-01   1.874 0.061655 .
## I(week.of.year^6)  -1.723e-02  9.529e-03  -1.808 0.071414 .
## I(week.of.year^7)   1.067e-03  6.220e-04   1.716 0.086985 .
## I(week.of.year^8)  -4.636e-05  2.875e-05  -1.613 0.107605
## I(week.of.year^9)   1.415e-06  9.375e-07   1.510 0.131901
## I(week.of.year^10) -2.978e-08  2.105e-08  -1.414 0.158043
## I(week.of.year^11)  4.050e-10  3.044e-10   1.330 0.184148
## I(week.of.year^12) -2.913e-12  2.313e-12  -1.259 0.208679
## I(week.of.year^13)         NA         NA      NA       NA
## I(week.of.year^14)  1.366e-16  1.185e-16   1.153 0.249689
## I(week.of.year^15)         NA         NA      NA       NA
## I(week.of.year^16) -8.791e-21  8.116e-21  -1.083 0.279383
## I(week.of.year^17)         NA         NA      NA       NA
## I(week.of.year^18)  3.386e-25  3.269e-25   1.036 0.300890
## I(week.of.year^19)         NA         NA      NA       NA
## I(week.of.year^20)         NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.002 on 389 degrees of freedom
## Multiple R-squared:  0.9275, Adjusted R-squared:  0.9245
## F-statistic: 310.9 on 16 and 389 DF,  p-value: < 2.2e-16
```

```
## [1] "AIC: 1734.27007680628"
```

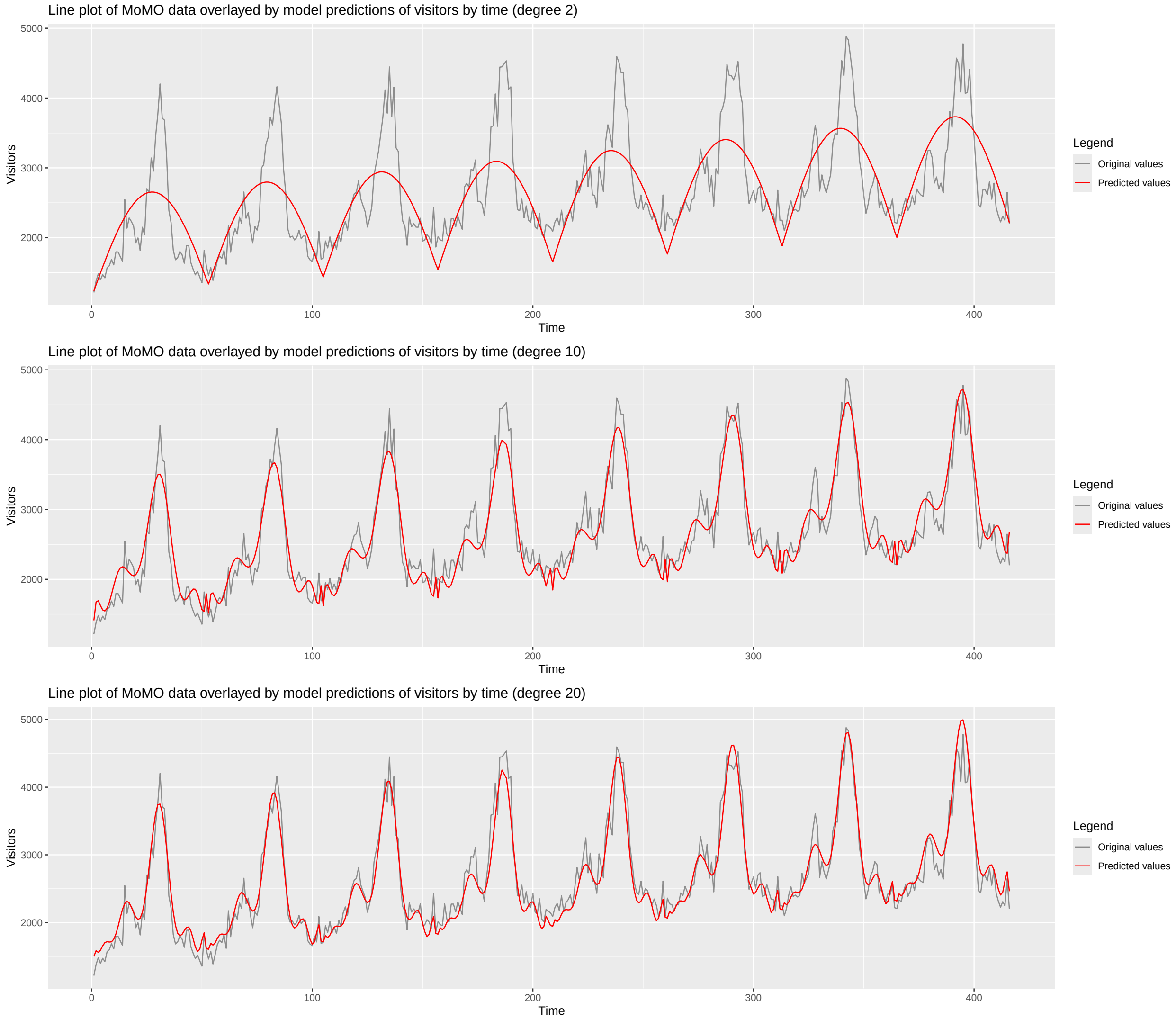

Figure 17: Predictions for increasingly high polynomial degree fits

We observe that when modelling `week.of.year` with a degree polynomial, despite all the parameters being significant, underfitting occurs, which results in a low R^2 and adjusted R^2 score. We observe that increasing the polynomial order results in an increasingly high R^2 and adjusted R^2 value, but fewer of the model terms are significant. This is because the model now has a lot of terms that are quite similar, so divides the explanatory power, so less of the coefficients are important in explaining significant parts of the data individually. For a fitted polynomial degree of 20, we observe some `NA` values for coefficients simply because some terms in the model matrix are so similar that they are numerically linearly dependent. This behaviour is observed because the required fit is very non-linear. This motivates the need to use a GAM to get a good smooth fit.

3.1.2.2 Figures and tables

Table 7: mod3 with cyclic cubic regression spline summary statistics table

K	EDF time	EDF week of the year	Deviance explained	GCV	AIC
10	2.874910	7.935843	92.68%	4.236844	1727.761
20	3.064853	14.947582	95.22%	2.927631	1569.186
30	3.082855	17.063303	95.32%	2.917061	1565.071

Table 8: Cross validation comparison table (MA20278 Machine Learning 1)

	10-fold cross validation score
REML	2.82053
GCV	2.81269

3.1.3 Question 4

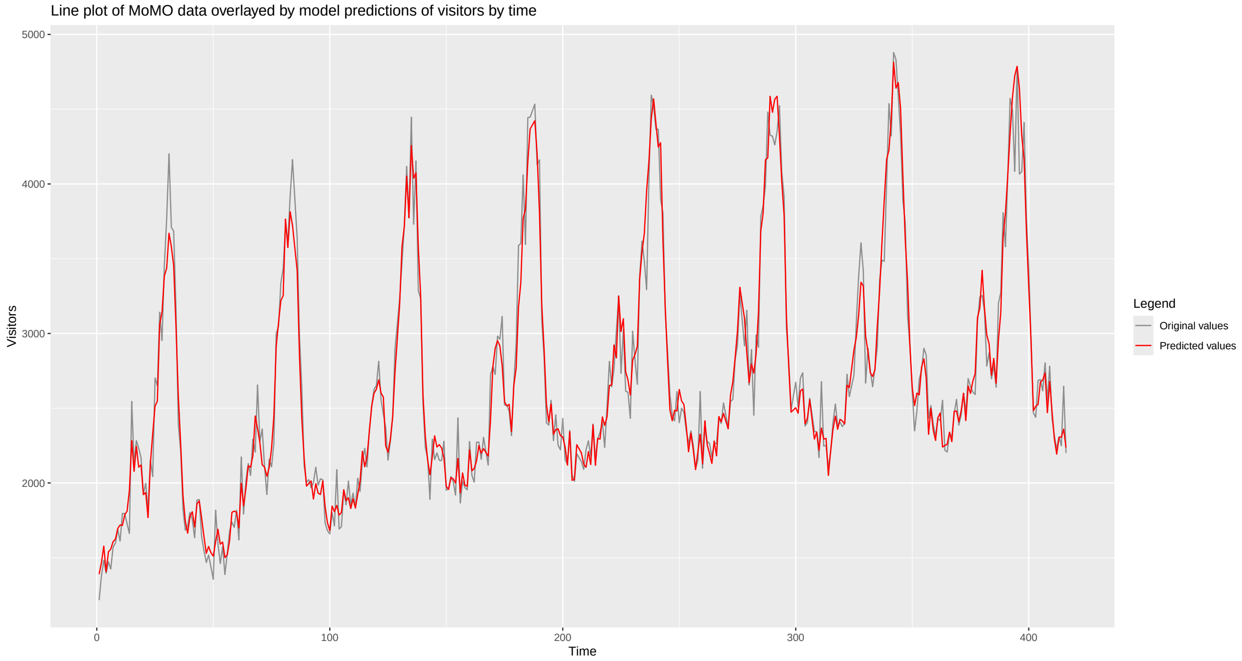

Figure 18: Line plot for mod5 predictions of visitors over time

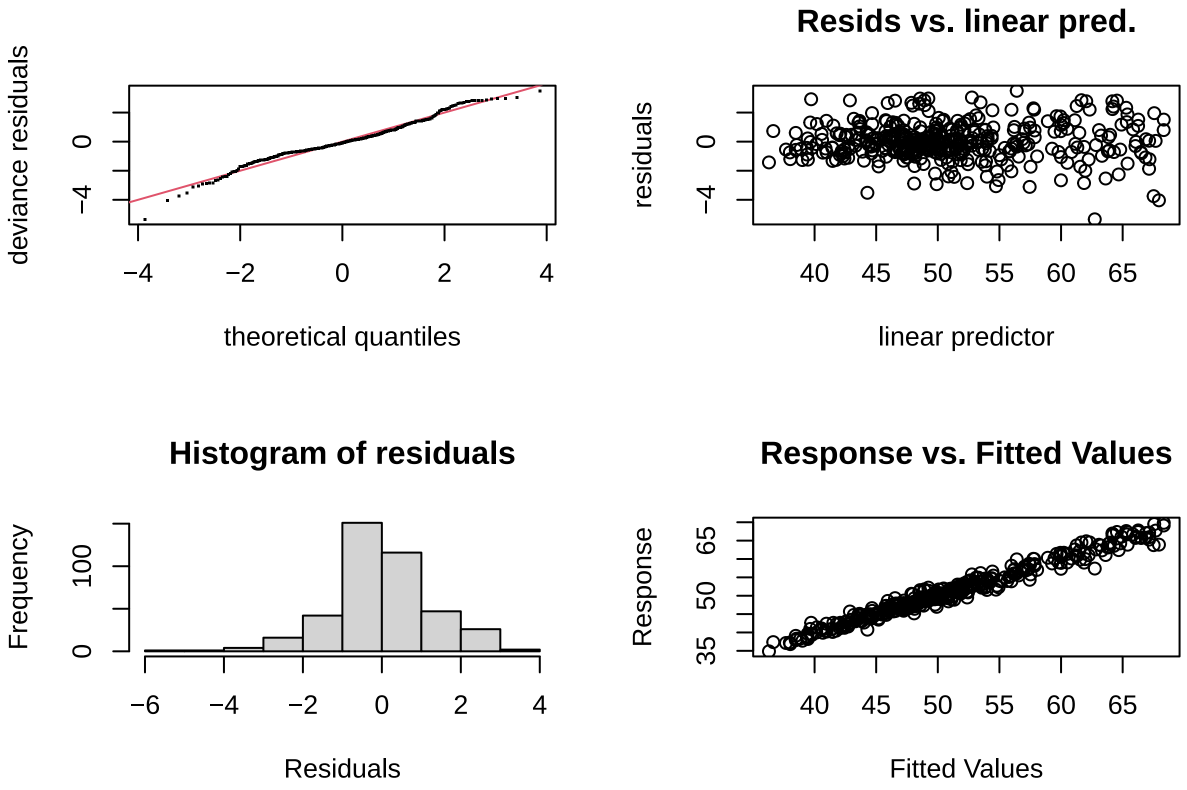

Figure 19: Model diagnostics for mod7

3.1.4 Question 5

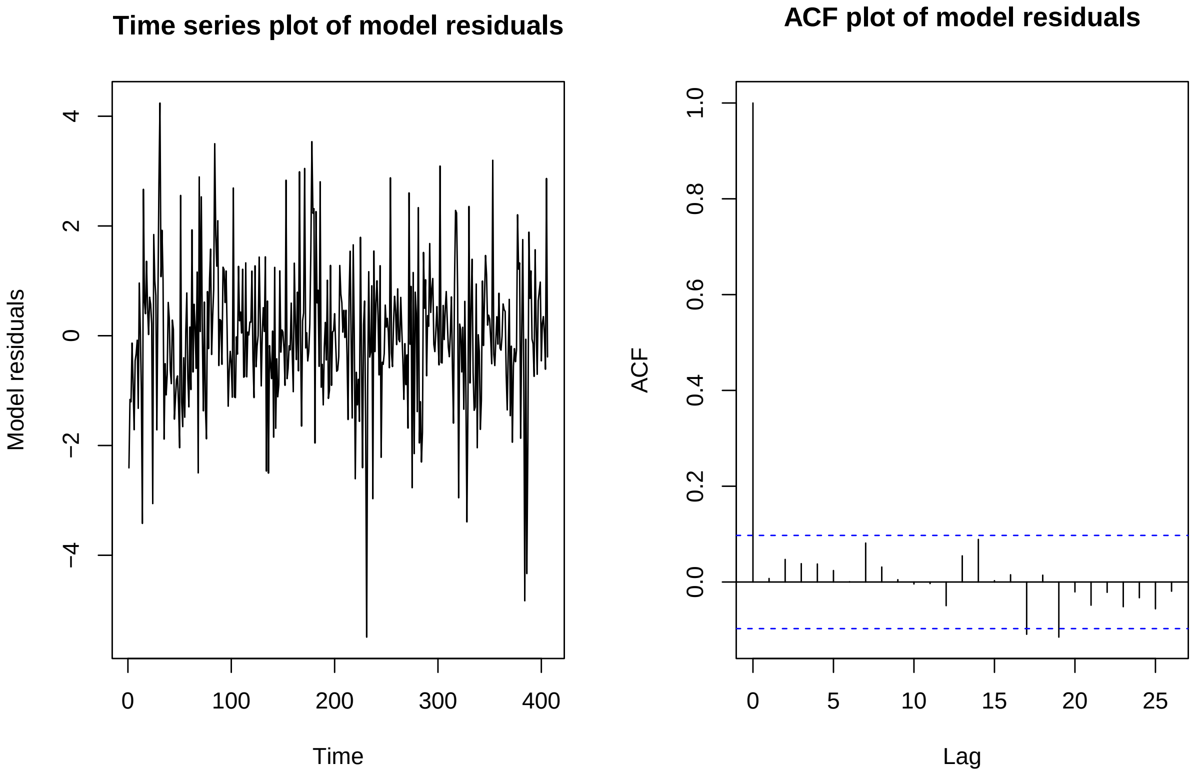

Figure 20: Time series model diagnostic plots for mod5 (MA30085 Time Series)

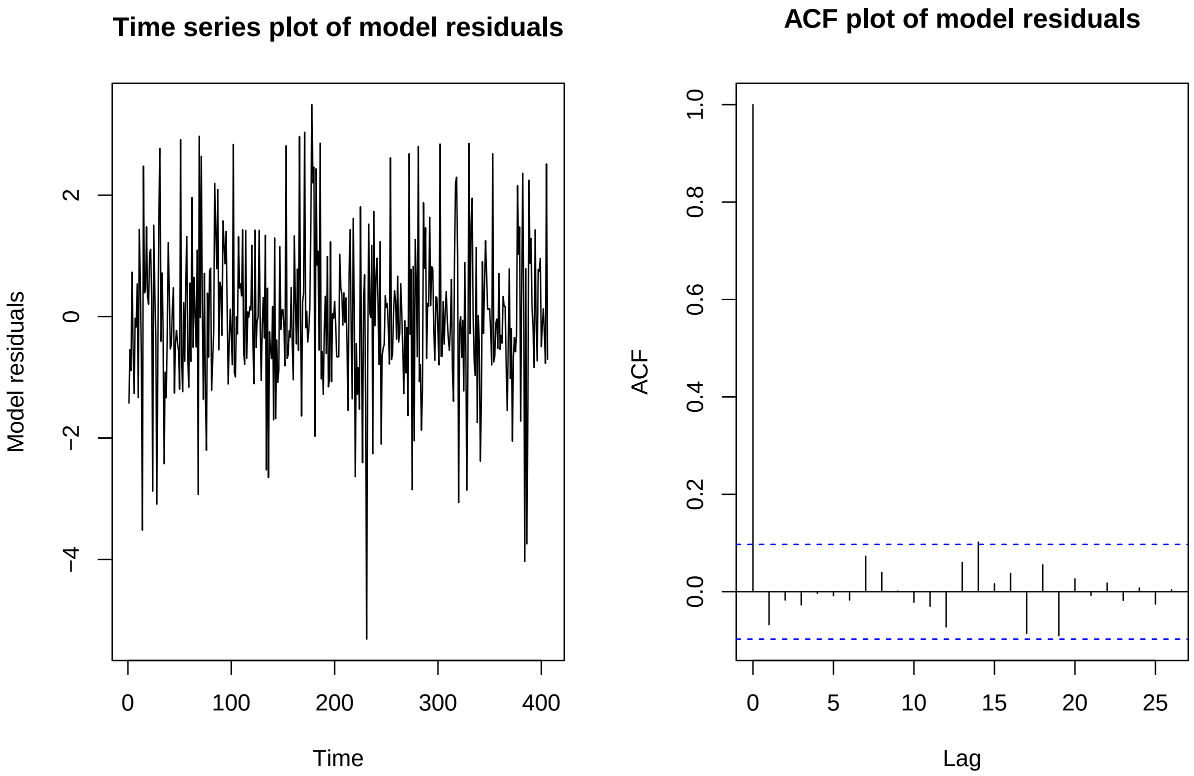

Figure 21: Time series model diagnostic plots for mod8 (MA30085 Time Series)

4 Generative AI statement

No generative AI was used in the creation of this report