Predicting Cherry Blossom Bloom Dates

Joshua Wright and Taehoon Song

2/26/2022

Introduction

Cherry tree blossoms are celebrated across multiple cultures. Festivals and sight-seeing tourism often revolve around this natural event. Planning events around this event is difficult seeing as the date is not the same from year to year. As such, it is of interest to try and predict the date of the blossoming. In this research, we demonstrate our proposed model for predicting the peak bloom dates in the coming decade for Kyoto, Liestal, Washington D.C., and Vancouver.

For this report, we will go through each location individually as a different model is used for each. In each location's portion, we describe the logistic growth model's specification, the data used for fitting the model, and the predicted day of year for full blossoming in the years 2022 through 2031.

Washington DC

In the field of phenology, plant growth is often modeled by using growing degree-day (GDD) as a predictor. Since the National Park Service(NPS) publishes dates of various growth stages for 2004 through 2021, we use this to build a logistic growth model (also known as the Verhulst model) that helps us estimate the threshold GDD for which we can expect the stage for the flowers to bloom. The authors use 4 degrees Celsius as the base for calculating GDD and fit a logistic function to the phenological stages from side green as a function of accumulated GDD.

$$y = \frac{k}{1 + \left[\frac{k - n_0}{n_0} \cdot \exp(-r \cdot \mathrm{GDD})\right]}$$

where y is the phenological stage and k and r are empirical factors related to growth rate and limited growth factors. y is an integer value which ranges from 3 (green budding stage) and 8 (peak blossoming stage). Now we use this model to fit the growth stages published by the NPS as a function of cumulative GDD. In our data from the NPS, we only have data starting at green budding (stage 3), as such our n_0 will be 3.

Obtaining temperature and phenological data for DC

In order to determine at what day of year we start calculating GDD, we need to determine when our n_0 stage is estimated to begin. To do this, we will look at the range of dates that green budding is reported and take the minimum, which is found to be at 50 days into the year. Thus, we only use the average temperatures after the 49th day of the year to ensure that we start calculating GDD around when Stage 3 occurs.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.00   59.00   64.00   63.94   69.50   77.00
```

Building the growth model

As mentioned above, we calculate the GDD using a base of 4 degrees Celsius and starting from 49 days from the beginning of the year. Then, we fit the known phenological stages for the cherry blossoms in DC to the calculated GDDs.

```
thresh <- 4
GDD_final <- dc_temp %>%
  mutate(NewYear=year(date %m-% period("49 day"))) %>%
  mutate(DD = ifelse(tavg - thresh > 0, tavg - thresh, 0)) %>%
  group_by(NewYear) %>%
  mutate(GDD = cumsum(DD),doy = yday(date)) %>%
  ungroup()

# obtain GDD for each phenological stage for DC
stages <- tibble(stage=names(phenostage)[2:7],stagenum=seq(3,8))
pheno_GDD <- phenostage %>%
  select(-GreenBuds) %>% #This stage is the starting point for our model.
  pivot_longer(cols=FloretsVisible:FullBloom,names_to = "stage",values_to="doy") %>%
  left_join(GDD_final,by=c("year","doy")) %>%
  summarize(year,stage,doy,GDD) %>%
  left_join(stages,by='stage')

# Fit growth model
n0 <- 3 #Starting phenological stage for our model
growth_model <- nls(stagenum ~ k/(1+(k-n0)/n0*exp(-r*GDD)),
                    data=pheno_GDD,start=list(k=1,r=0.01))
summary(growth_model)
```

```
##
## Formula: stagenum ~ k/(1 + (k - n0)/n0 * exp(-r * GDD))
##
## Parameters:
##    Estimate Std. Error t value Pr(>|t|)
## k 10.223870   1.362473   7.504 4.72e-11 ***
## r  0.011973   0.001598   7.495 4.93e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9374 on 88 degrees of freedom
##
## Number of iterations to convergence: 8
## Achieved convergence tolerance: 7.789e-06
```

The model results show that both k and r are significant being estimated to be 10.22 and 0.012 respectively. What we are interested in is at what day of year do we expect the cherry tree to reach stage 8 (peak blossoming). with this model, we calculate the expected stage from a given GDD. Once we reach the first day that is estimated to be stage 8, that is our prediction for bloom date of the year. We can test the accuracy of this model with a few diagnostics.

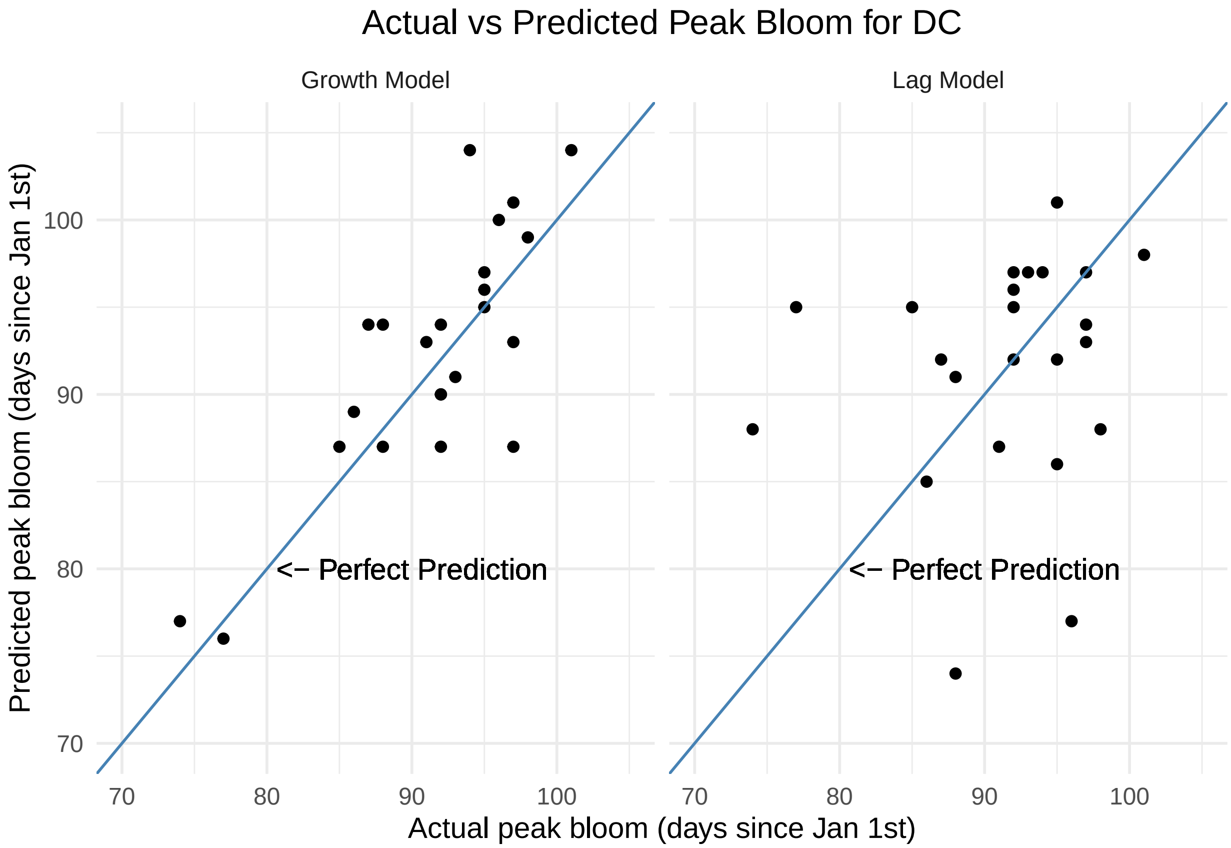
Checking performance on historical data

Now that we have the growth model, we can test its prediction performance out on historical data that we did not use. Using temperature data from 1980 to 2003, we will compare the logistic growth model to the lag model to see which is a better fit. We define the lag model to be where the predicted bloom date is just the previous year's date.

Logistic Growth Model MAE	Lag Model MAE
3.347826	6.304348

Our growth model has a much lower mean absolute error (MAE) compared to simply using the previous year's bloom date as a predictor. This is good because this is the bare minimum a useful model should be able to do. This is more apparent when we plot the two models against the actual peak bloom days since January 1st.

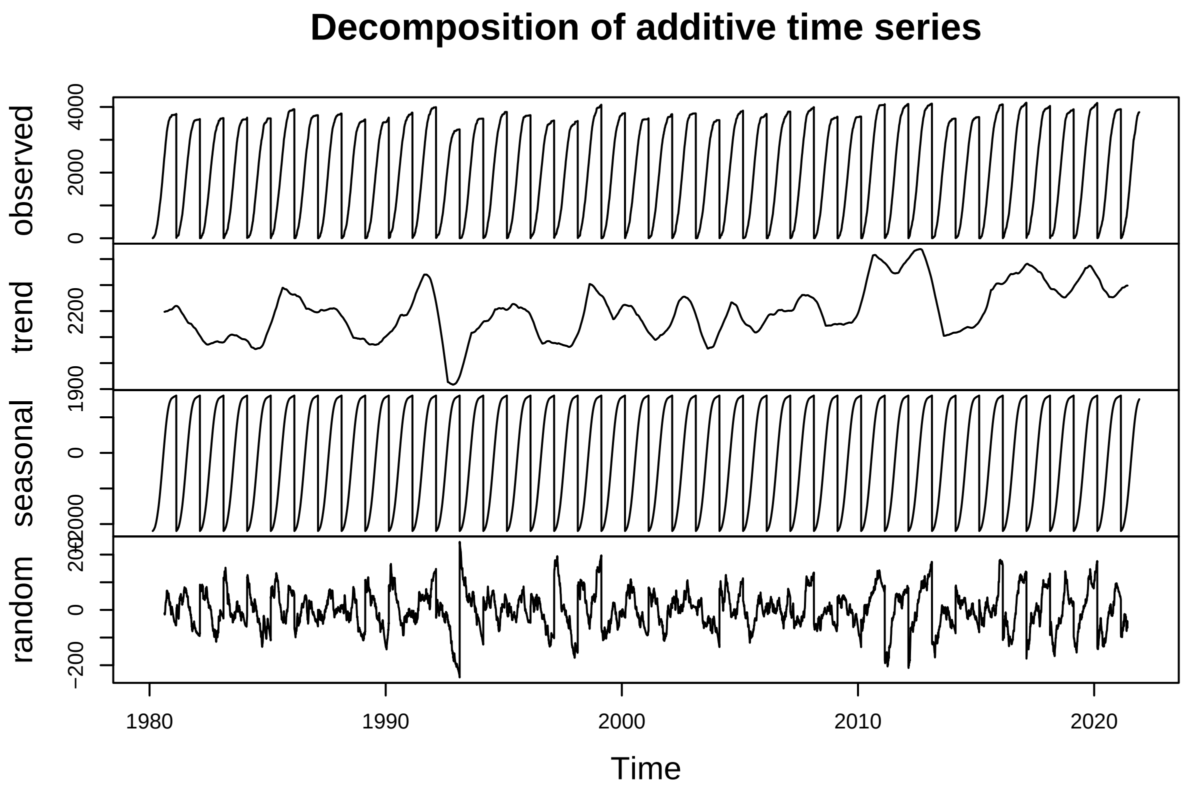
```
plt_data <- comp %>% rename(Actual=act.doy,GrowthModel=pred.doy,LagModel=lag_bloom) %>%
  pivot_longer(cols=c("GrowthModel","LagModel"),names_to="type",values_to="doy") %>%
  filter(!is.na(err_lag))
plt_data$type <- factor(plt_data$type,levels=c("GrowthModel","LagModel"),
                        labels=c("Growth Model","Lag Model"))

# Plot Actual vs Predicted for Growth Model
ggplot(plt_data,aes(x=Actual)) + geom_point(aes(y=doy)) +
  facet_grid(~type) +
  geom_abline(slope=1,col="steelblue") +
  geom_text(aes(x=90,y=80,label="<- Perfect Prediction")) +
  theme_minimal() +
  ggtitle("Actual vs Predicted Peak Bloom for DC") +
  xlab("Actual peak bloom (days since Jan 1st)") +
  ylab("Predicted peak bloom (days since Jan 1st)") +
  xlim(c(70,105)) + ylim(c(70,105)) +
  theme(plot.title=element_text(hjust=0.5))
```



Predicting future GDD and future peak bloom dates

For us to predict future peak bloom dates for Washington D.C., we must first predict future GDDs. We could predict future temperature and calculate GDD from that; however, that would be unnecessary because we always start calculating GDD at 50 days and use the base of 4 degrees Celsius. We can directly predict GDD to avoid that step. For GDD prediction, we will fit a model to the GDD using a time series seasonal ARIMA model.

```
# Create time series
ts_gdd <- ts(GDD_final$GDD[50:nrow(GDD_final)],frequency = 365,
             #Only start counting GDD from day 50 & 2021 DC data only goes to November
             start=c(1980,50),end=c(2021,334))
plot(decompose(ts_gdd))
```



```
if (file.exists('pred_dc.rds')){
  pred_dc <- readRDS(here("pred_dc.rds"))
} else{
  # Fit time series to GDD
  ts_fit <- auto.arima(ts_gdd, ic = 'aicc')
  ar <- ts_fit$arma[1:2]
  ma <- ts_fit$arma[3:4]
  dd <- ts_fit$arma[5:7]

  # Use result from fitted time series to predict GDD until May 2031
  preddays <- difftime("2031-05-31","2021-12-01",units="days")
  pred_dc <- sarima.for(ts_gdd,as.numeric(preddays),ar[1],dd[2],ar[2],ma[1],dd[3],ma[2],dd[1])
  saveRDS(pred_dc,here("pred_dc.rds"))
}
```

With a time series model fit, we can now predict the peak bloom day for DC using the predicted GDD as an input for the logistic growth model. Feeding our predicted GDD values into our model outputs the following predictions:

```
# Forecasted GDD and growth stage
newdata <- tibble(date=seq.Date(as.Date("2021-12-01"),length.out=length(pred_dc$pred),by="day"),
                  GDD=pred_dc$pred,doy=yday(date),year=year(date))
future_growthstage <- predict(growth_model,newdata=newdata)
forecasts_dc <- newdata %>% mutate(pred.gs=future_growthstage) %>%
  group_by(year) %>%
  filter(doy >= 50 & doy <= 200) %>%
  filter(round(pred.gs) >= 8) %>%
  filter(row_number()==1) %>%
  select(year,doy)

colnames(forecasts_dc)<-c('Year', 'Day of Year')
kable(forecasts_dc)
```

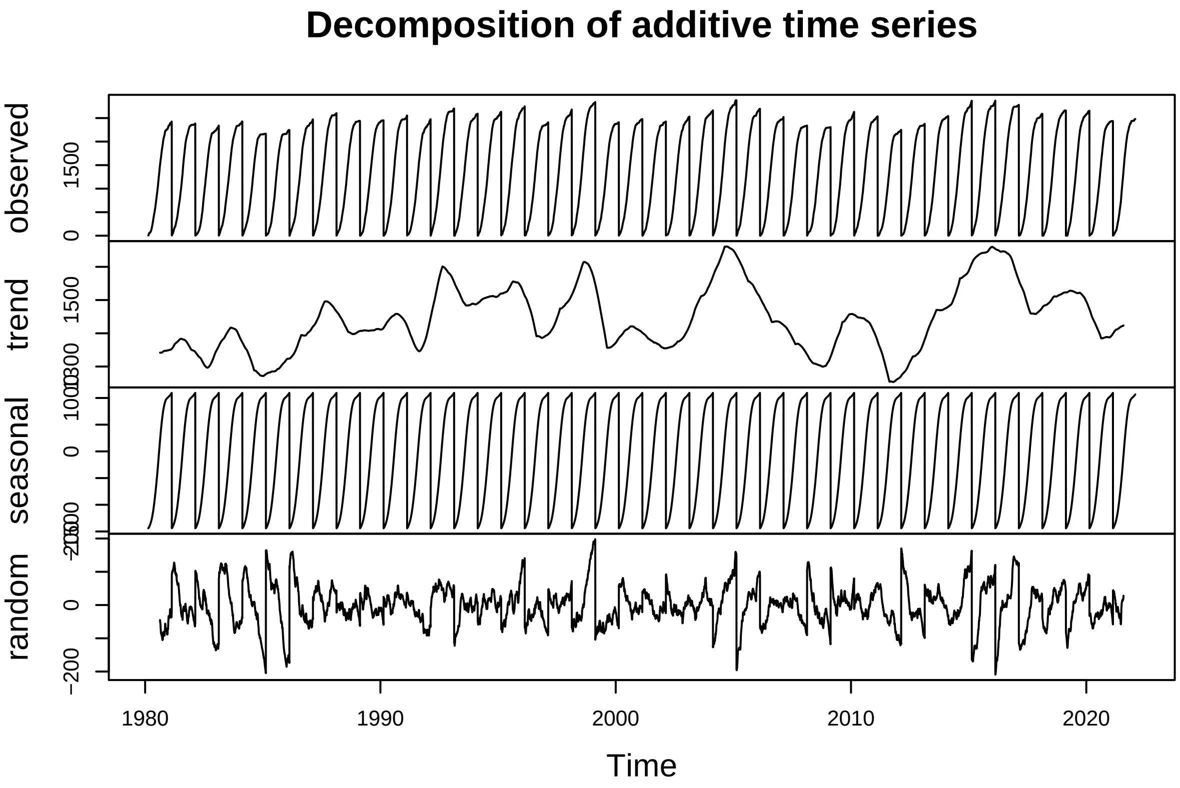
Year	Day of Year
2022	86
2023	86
2024	86
2025	85
2026	85
2027	85
2028	84
2029	83
2030	83
2031	83

Vancouver

Vancouver doesn't provide historical peak bloom dates. However, we use the growth model from DC to predict future peak bloom dates because the tree species are identical (or close to identical) and the peak bloom definition is the same. We will use temperature data to once again calculate the historical GDD values and fit another time series model to predict future bloom dates.

Obtaining historical weather data

Similar to DC, we obtained the historical weather data for Vancouver via DayMet. However, this dataset only goes from January 1980 to December 2019, so we use the `rnoaa` *R* package provided by NOAA to obtain the remaining data up to Jan 2022. We undergo the same process as we did for Washington D.C. in order to fit the time series model and use the exact same logistic growth model for bloom day predictions.



Year	Day of Year
2022	109
2023	108
2024	108
2025	107
2026	106
2027	106
2028	106
2029	105
2030	104
2031	104

What we have are later predictions later than what we saw in Washington D.C., but that is to be expected given the lower average temperatures in Vancouver.

Liestal

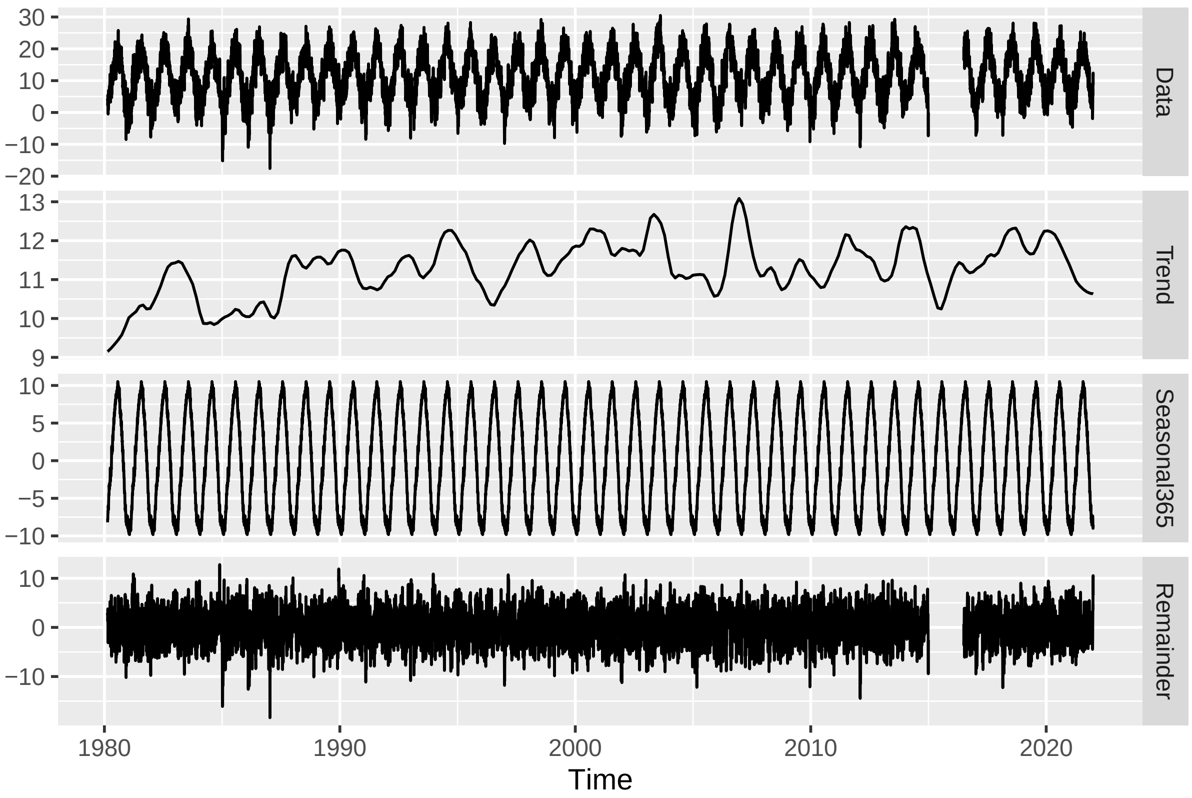
The cherry tree species in Liestal is a different species from the ones in Washington D.C. and Vancouver. Additionally, the percentage of blooming for the tree to be declared as at the stage of peak bloom is less as well. Ideally, we would want to fit a new growth model using historical phenological stages. Unfortunately, such a dataset is not readily available. Thusly, we use the same growth model parameters k and r from original to calculate the mean growth stage for historical peak bloom dates. This is to say that we will use the cumulative GDD from each of the historical bloom dates to see what y value our model would predict at that GDD. We will once again start calculating GDD at the 50th day of the year. We use the average growth stage for the past 5 years as the cutoff for our prediction.

```
# Check historical GDD and corresponding predicted growth stages for actual bloom dates in Liestal
testdata <- cherry %>%
  left_join(GDD_final %>% select(date,GDD), by=c('bloom_date'='date')) %>%
  filter(year >= 2016 & location=='liestal') %>%
  select(GDD)

summary(predict(growth_model,newdata=testdata))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   5.995   6.532   6.773   6.799   7.170   7.524       1
```

```
li_thresh <- mean(predict(growth_model,newdata=testdata),na.rm=TRUE)
```

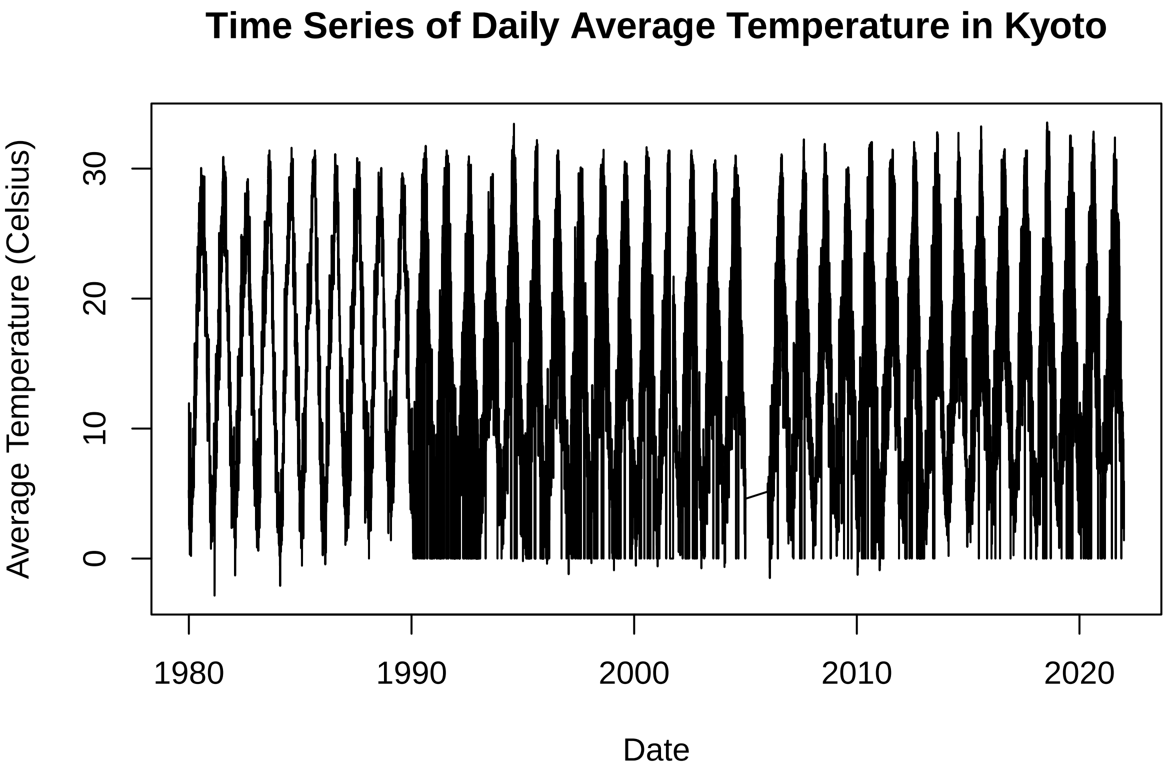
We can approximate the peak bloom for the cherry trees in Liestal (as defined by MeteoSwiss) using the growth model for the trees in Washington D.C. if we use 6.8 as the predicted growth stage for full blossoming. Another change we use here is directly predicting the temperature to then find the GDD as opposed to just fitting a time series model to the GDD. This is done due to the new circumstances making GDD time series model fits less reliable than just predicting temperature. After predicting the temperature, we then calculate the GDD and find which day of year is the first to reach stage 6.8. Our predictions are listed below.



Year	Day of Year
2022	88
2023	88
2024	88
2025	87
2026	87
2027	87
2028	87
2029	86
2030	86
2031	86

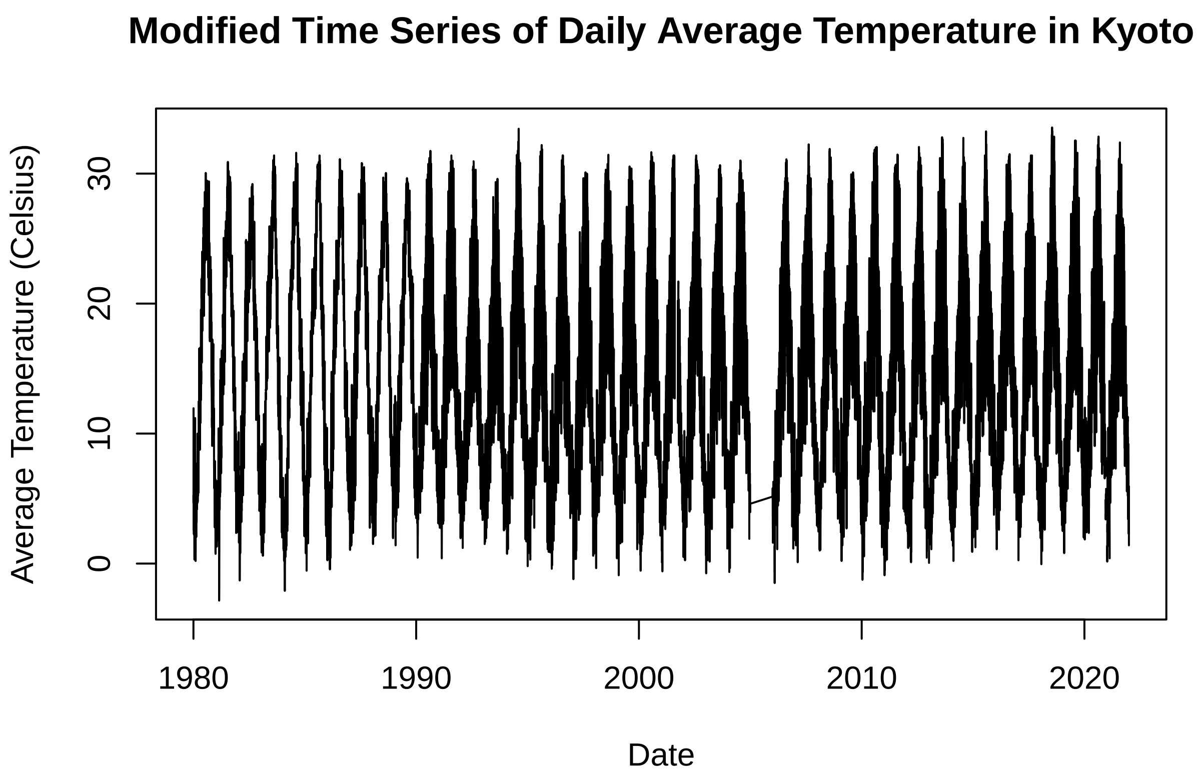
Kyoto

Similar to predicting peak bloom dates for Liestal, the peak bloom dates for Kyoto (as defined by a local newspaper in Arashiyama) need to be approximated using the growth model for the Washington D.C. cherry trees. Alongside the issues of having a different species of cherry tree without phenological stage data and a differing definition of peak bloom, we also have the issue of incorrect historical temperature data provided by NOAA.



Notice that the data abruptly cuts off at 0 commonly starting at the year 1990. This is likely due to missing values due to these freezing temperatures being abruptly different than previous temperatures and sometimes even in the middle of summer. We change 0's to NA's so that the algorithm knows that the value is actually missing. Kyoto is a very warm city and rarely dips below 0 degrees Celsius, so we should not lose too much information with this substitution. We're missing a month of data for September 2009, but that shouldn't be too much of a problem for predicting 2022 and onward given the amount of data we have. Modifying the data as we mentioned, we have this new time series as seen below.

```
# Substitute 0s for NA
jp_temp2 <- jp_temp %>%
  mutate(year=year(date), month=month(date)) %>%
  group_by(year,month) %>%
  mutate(tavg=ifelse(tavg==0,NA,tavg))

plot(jp_temp2$date,jp_temp2$tavg,type="l", ylab = "Average Temperature (Celsius)", xlab = "Date", main = "Modi
```



With the substitution, the odd behavior is gone mostly accounted for, so we can proceed to calculate the historical GDD for Kyoto. Just like with Liestal, we will examine what stage the previous bloom dates in 2016 to 2021 were according to our model.

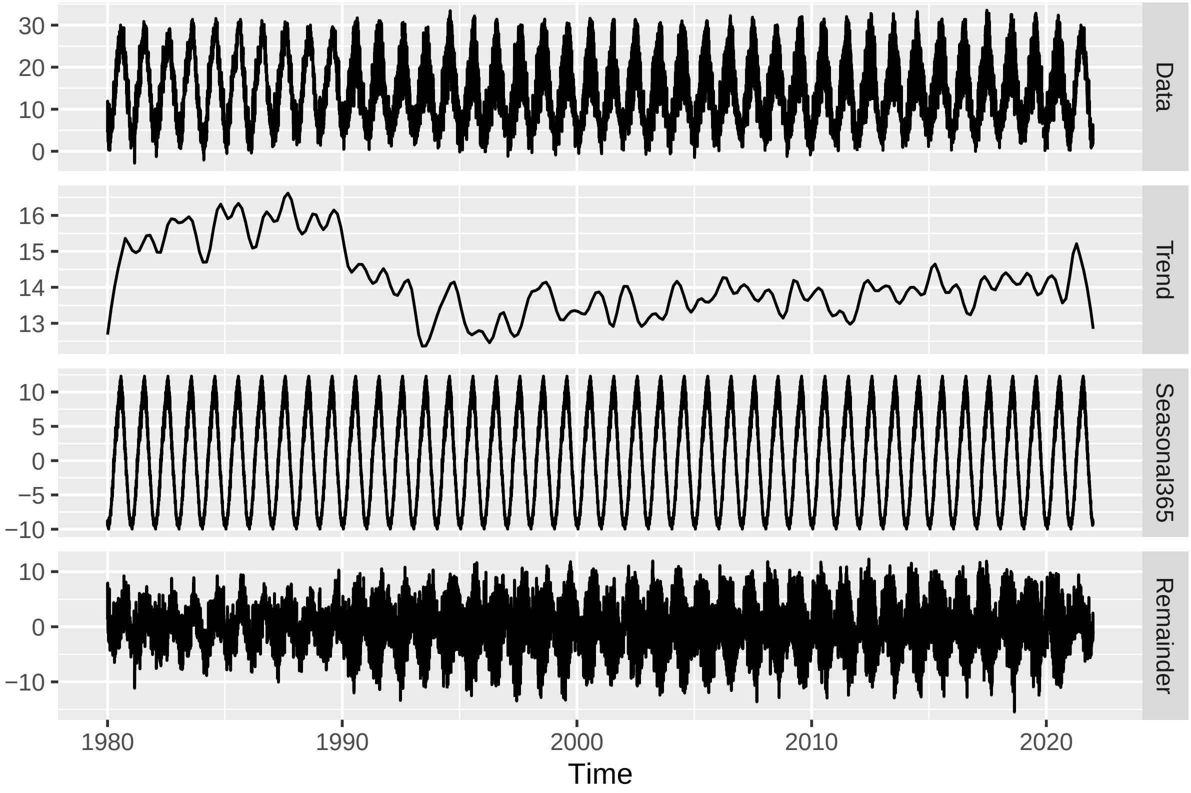
```
# GDD calculations
GDD_final <- jp_temp2 %>%
  mutate(NewYear=year(date %m-% period("49 day"))) %>%
  mutate(DD = ifelse(tavg - thresh > 0, tavg - thresh, 0)) %>%
  group_by(NewYear) %>%
  mutate(GDD = cumsum(DD),doy = yday(date)) %>%
  ungroup()

# Check historical GDD and corresponding predicted growth stages for actual bloom dates in Liestal
testdata <- cherry %>%
  left_join(GDD_final %>% select(date,GDD), by=c('bloom_date'='date')) %>%
  filter(year >= 2016 & location=='kyoto') %>%
  select(GDD)

summary(predict(growth_model,newdata=testdata))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   8.234   8.268   8.303   8.303   8.337   8.372       4
```

```
jp_thresh <- mean(predict(growth_model,newdata=testdata),na.rm=TRUE)
```

What we find is that we want our model to predict when the stage is 8.3. This makes sense with the peak bloom definition being greater than that of Washington D.C.. With this stage number, we fit a time series model to the daily average temperature of Kyoto to predict future GDD and thusly the bloom day of year as seen below.

```r
# Create time series
ts_temp <- ts(GDD_final$tavg[1:nrow(GDD_final)],frequency = 365,
             start=c(1980,1),end=c(2021,365))

ts_temp %>%
  mstl(s.window="periodic") %>%
  autoplot()
```



Year	Day of Year
2022	79
2023	79
2024	79
2025	79
2026	79
2027	79
2028	79
2029	78
2030	78
2031	78

Summary of Predictions

We have described the methodology used for bloom day predictions for each of the locations. The summary table below shows the final predictions for the four locations from 2022 to 2031.

Year	Kyoto	Liestal	Washington D.C.	Vancouver
2022	79	88	86	109
2023	79	88	86	108
2024	79	88	86	108
2025	79	87	85	107
2026	79	87	85	106
2027	79	87	85	106
2028	79	87	84	106
2029	78	86	83	105
2030	78	86	83	104
2031	78	86	83	104

Closing thoughts

If historical phenological stages were recorded for all locations in this competition, we would have been able to fit a growth model for each species and location. This would have allowed us to create more accurate models specific to that cherry tree species and environmental factors. Additionally, weather data was extremely limited or unreliable for locations outside the United States. If historical weather data was freely and readily available to the public, we would have been able to create a better time series model for GDD.

Details can be found in Phenological Models of Flower Bud Stages and Fruit Growth of 'Montmorency' Sour Cherry Based on Growing Degree-day Accumulation by C. Zavalloni, J. Adresen, and J. Flore.