Predict "peak bloom date" for cherry trees in 4 different locations

Summary

Because cherry blossoms don't last long, forecasting the number of cherry blossom days is very important for a country to attract tourists. When we read some literature, most of them propose that temperature is the main effect of cherry blossom bloom. Avoiding the model overfit the data, we split it into 90-10 training-testing dataset for weather and bloom day of the year. After forecasting the weather by Seasonal Autoregressive Integrated Moving Average Models, we can forecast the bloom day of the year for Washington DC, Kyoto, and Liestal by Autoregressive Integrated Moving Average Models or Time series Linear Models. For Vancouver, without any historic cherry blossom days of the year, we use an Autoregressive Integrated Moving Average Model and a Naïve Model by extrapolating from the other three locations to forecast the bloom day of the year in Vancouver.

Introduction

The impact of climate change on the wild has recently attracted the attention of international academics. Climate-related changes in the sequence of environmental events, such as the date of snowmelt at high altitudes, which starts off the planting conditions, have had major implications for some typical annual herb flower species. Another example, due to global warming, 78 percent of all studies in 21 European countries indicated earlier flowering, with an overall phenological activity movement of 2.5 days per decade. For flowering cherry trees, the peak date in Kyoto has been gradually shifting from mid-April to the beginning of the month. Washington's cherry trees are blooming six days earlier than they were a century ago. Simultaneously, weather station measurements in the US capital reveal that the temperature has risen by 1.6 degrees Celsius (2.88F). In Liestal, flowering has been fairly consistent in April, but there has been a trend towards earlier flowering dates since the 1990s. Because their flowering timing is highly sensitive to temperatures, especially throughout the winter and early spring, flowering cherry trees are a useful indication of the effects of climate change on phenology. From a cultural standpoint, precise phenology prediction is critical, as many festivals across the world revolve around specific phenology timing, and flower watching tours are in high demand. In Japan, there is a saying "seven days of cherry blossoms". It means that a cherry blossom takes about 7 days from the time it opens to the time it fades. Accurate prediction of the flowering period of ornamental plants can provide guidance for public travel arrangements.

The peak bloom dates of previous years in 3 locations have been provided. In Washington DC, the historical bloom dates data range from 1921 to 2021; in Kyoto, they range from 1888 to 2021; in Liestal, they range from 1894 to 2021; and in Vancouver, there is no historic data. We used the bloom day of the year as the dependent variable. For different locations, we used the website NOAA (https://www.noaa.gov/) to collect covariates data based on similar latitude, longitude, and altitude. The data set we found has 26 variables, such as extreme temperature (EMNT and EMXT), temperature (TAVG, TMAX, and TMIN), and precipitation (EMXP and PRCP).

The main goal of this study was to forecast the peak bloom days of the year from 2022 to 2033 at four locations, including Washington DC, Kyoto Japan, Liestal Switzerland, and Vancouver Canada.

Covariate's collection

In brief, because the cherry blossom bloom date is highly sensitive to temperature, we chose temperature-related covariates. Due to the high number of missing values for mean temperatures (TVAG), minimum temperatures (TMIN) and extreme minimum temperatures (EMNT) in some locations, we chose extreme maximum temperatures (EMXT) and maximum temperatures (TMAX) as covariates.

Based on historical peak bloom dates, we roughly split the data into three groups (i.e., early bloom date, regular bloom date & late bloom date). The average TMAX of each month for these three groups is calculated. Fig. 1 shows information about the monthly maximum temperature over time. Overall, in Washington DC, Liestal, and Kyoto, these three groups have a similar trend. The lowest value occurs in January, followed by a slight increase in February and there is a jump of TMAX between February and March. When the three groups are compared in February and March, the early bloom date group has the highest maximum temperature. As a result, we believe that the TMAX has a correlation with peak bloom date. We conducted the same analysis on extreme temperatures and discovered that, in Fig. 2, while the overall difference between the three groups was not as significant as the average temperature, extreme temperatures have an impact on blooming time.

Then we visualize the correlation. From Fig. 3-5, we discover that the covariate March maximum temperature is negatively correlated with peak bloom date In Washington DC (r = -0.74), Kyoto (r = -0.77), and Liestal (r = -0.7), while the covariate December extreme temperature is positively correlated with peak bloom date in Washington DC (r = 0.2). In accordance with the above findings, we collect these two covariates maximum temperature from March and extreme temperature from December for our study.

Preprocessing

We obtain covariates for Washington DC from 1945 to 2021, covariates for Kyoto from 1951 to 2021, and covariates for Liestal from 1901 to 2021; therefore, only historical data of peak bloom date for these years are used. Regarding missing values of TMAX and EMXT, in the Washington covariate data, January and February 2022 have missing values, so those two rows are deleted; in the Liestal covariate data, the row of January 2022 with missing values is deleted, and the missing values of January 2017 are handled with "tsclean()"; in the Kyoto covariate data, the rows of 1978-06, 2004-07, 2004-08, 2011-01, 2012-03, 2012-11, and 2012-12 have the missing values which are similarly handled using "tsclean()." tsclean() is an useful function that identifies and replaces outliers as well as missing values in a time series. Any missing values are replaced with linearly interpolated replacements. It does allow us to utilize forecasting models that are sensitive to outliers or do not handle missing values.

We split the entire datasets into 90% training and 10% testing, and we only used TMAX from March and EMXT from December as covariates in this analysis.

Methods

(a) Autoregressive integrated moving average models (ARIMA (p, d, q)). ARIMA is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. ARIMA with p, d, and q is a standard notation for ARIMA models, where integer values replace the parameters to indicate the kind of ARIMA model utilized. Here p is number of autoregressive terms (AR order), d is number of non-seasonal differences, and q is number of moving-average terms (MA order). The auto.arima() function in R uses a variation of the Hyndman-Khandakar algorithm (Hyndman &

Khandakar, 2008), which combines unit root tests, minimisation of the AICc and MLE to obtain an ARIMA model. The arguments to auto.arima() provide for many variations on the algorithm.

(b) Seasonal Autoregressive Integrated Moving Average (SARIMA) is an ARIMA extension that supports univariate time series data with a seasonal component explicitly. It adds three new hyperparameters for the seasonal component of the series: autoregression (AR), differencing (I), and moving average (MA), as well as an additional parameter for the seasonality period. There are four seasonal elements that must be setup that are not part of ARIMA: P: Seasonal autoregressive order, D: Seasonal difference order, Q: Seasonal moving average order, m: The number of time steps for a single seasonal period. An SARIMA model is denoted by the following notation: SARIMA(p,d,q)(P,D,Q).

(c) Time series linear model (TSLM). TSLM is used to fit linear models to time series including trend and seasonality components. It's similar to the commonly used lm() function for linear models, but tslm() has more features for dealing with time series.

(d) Ljung-Box test. It is used to check if exists autocorrelation in a time series. The statistic is

$$q = n(n+2) \cdot \sum_{j=1}^{h} \hat{\rho}(j)^2/(n-j)$$

with n the number of observations and $\hat{\rho}(j)$ the autocorrelation coefficient in the sample when the lag is j. LSTS_lbtp computes q and returns the p-values graph with lag j.

Utilizing the SARIMA to predict and forecast the weather, TMAX and EMXT for whole years. There are two models. One is ARIMA_auto which we use the auto.arima() function on R, and another is ARIMA_manual which we choose by checking the acf2(). Hence, the ARIMA_manual is better, so we use the outcomes of ARIMA_manual to be our predicted weather data. Having the predicted and forecast weather data, we apply 6 main models to predict and forecast days of the year of cherry blossom in Washington DC, Kyoto, and Liestal, which have the historic bloom day data. First, ARIMA_without_covariate is without any covariates. Second, the model with TMAX in March and EMXT in December is ARIMA_with_covariates. ARIMA_with_TMAX_covariate only includes the TMAX 3. ARIMA_with_EMXT_covariate includes EMXT in December. And we also have TSLM_with_covariates, which also contain TMAX in March, EMXT in December, and trend. Last but not least, TSLM_with_TMAX_covariate is similar to TSLM_with_covariates but excludes EMXT in December.

For Vancouver, in order to predict cherry blossom's peak bloom date; first, we make a large data frame of all locations' data (Washington D.C., Kyoto, and Liestel) alongside their monthly covariates. Once we have our combined dataset, we considered a linear model predicting "blooms_doy" using the covariates. We see that both predictors "TMAX_3" and "EMXT_12" are significant. Then, we use time series to predict the value of these predictors for Vancouver.

Results
ARIMA_auto has wide confidence intervals for future times when we predict and forecast TMAX and EMXT for DC, in Fig. 6 and 7, respectively. ARIMA_manual will not have the issue, but some real-data temperature (black line) doesn't in the predicted confidence interval, such as TMAX in March 2017 in Fig. 6. However, we use the outcomes of ARIMA_manual to be our predicted weather data, because it doesn't have wide confidence intervals in the future. We also use the same methods for Kyoto and Liestal to obtain the predicted weather data.

In Table 1, we can see that TSLMs (AIC: 206.83 and 209.44) outperform the ARIMA models on the Washington DC dataset (AIC: 447.74, 388.29, 395.84, and 443.54). However, the residuals of the TSLM reject the null hypothesis of the Ljung-Box test in Table 2, implying that the residuals are serially correlated. We can only use the ARIMA models for the Washington DC dataset. For the Kyoto and Liestal datasets, the ARIMA models (Kyoto AIC: 359.90, 314.39, 312.95 and 361.53; Liestal AIC: 813.41, 737.08, 737.06 and 812.66) are also poor than TSLMs (Kyoto AIC: 133.5 and 134.89; Liestal AIC: 440.93 and 442.20) in Tables 1. We check the Ljung-Box test, the residuals of all models in Kyoto fail to reject the null hypothesis of significant 0.1 in Table 2, so none of the models show a significant lack of fit, which means we can use all models for the Kyoto dataset. However, the TSLM with the TMAX covariate in Liestal doesn't pass the Ljung-Box test at a significance level of 0.1 in Table 2, so we remove the model. In Table 1, the performance of the ARIMA model with EMXT covariates (BIC: 450.15) for the Washington DC datasets is only marginally better than the ARIMA model without any covariates (BIC: 452.15), so we remove the ARIMA models with EMXT covariates. On the contrary, the ARIMA model with EMXT covariates (BIC: 367.91; 820.68) for the other two location datasets are only slightly worse than the ARIMA models without any covariates (BIC: 364.16; 818.76), we also remove the models for these two locations.

Using the predicted weather data to be the covariates, we predict the bloom day of the year and compare the results. In Table 3, ARIMA with TMAX and EMXT covariates has the lowest RMSE (5.02) and MAPE (4.27) for predicting DC cherry bloom day. For Kyoto, TSLM with the TMAX covariate has the smallest RMSE (5.02) and MAPE (4.27) when we predict blooming days during the year from 2014 to 2021 in Table 3. Last, ARIMA with TMAX covariate fits the Liestal testing dataset well in Table 3, in which RMSE (8.40) and MAPE (8.09) are the lowest. In Fig. 8-10, the ARIMA models with no covariates for three locations have wider confidence intervals than others. We find that our model does not capture some real data for early flowering years within the confidence interval; for example, the real data of bloom days of the year in 2018 and 2020 in Kyoto are not within the 95% confidence interval for most models in Fig. 9. Using the best model for each location, we forecast the bloom day from 2022 to 2033 in Table 4.

For Vancouver, by looking at Fig. 14, there's a positive trend between TMAX_3 and time (meaning as time passes by TMAX_3 increases). Then, we look at the time plot of change in TMAX_3 (Fig. 15) where we don't see a trend that could help us predict the value of TMAX_3. We build two models, one Naive model, and one Arima model. We see that the Arima model had both a lower standard deviation and a better residual plot compared to the Naive model (Fig. 16 and 17). So, using the Arima model, we predicted the TMAX_3 to be 50. Next, we try and predict the EMXT_12.

Similar to TMAX_3, by looking at Fig. 18, we see that there's a positive trend between EMXT_12 and time (as time passes, EMXT_12 increases). Then, we look at the time plot of change in EMXT_12 (Fig. 19) where we see no trend. We compared two-time series models to predict EMXT_12: one Naive model and one Arima model. We see that the Arima model had both a lower standard deviation and a better residual plot (Fig. 20 and 21). So, using the Arima model, we predicted the EMXT_12 to be 51.

Conclusion
In Table 4, the cherry bloom which opens around 91 days of the year in DC is earlier than in other locations (around 95 days). Kyoto has the narrowest confidence interval (about 11 days for 95% CI), and Liestal has the widest confidence interval (about 32 days for 95% CI). And the peak bloom date will be for Vancouver which we get to be 104 days.

The bloom days of the year are hard to predict. This is because we cannot predict the weather very well. That is, when the temperature is high, the temperature we predict is not high enough, which makes the prediction of flowering days inaccurate.

In Table 5, when we utilize the same models with real weather datasets, all models give the extremely lower RMSE and MAPE for prediction than the models with predicted weather covariates. We can find that ARIMA with TMAX and EMXT covariates is the best model for those three locations, because MAPEs are the lowest across all models in Table 5, and 80% confidence intervals including most real test data are shown in Fig. 11-13. If we can predict the weather well in the future, then ARIMA models with TMAX and EMXT covariates will be proper for predicting days of the years of cherry blossom.

We try a lot of methods for forecasting the weather dataset, such as Holt-Winters, Double Seasonal Holt-Winters and Anomaly Detection; however, the results are not good enough. We also use vector autoregressive model to weather datasets, but not all the covariates are stationary, so we cannot apply the method. Even though we find the temperature dataset from the Copernicus website (https://cds.climate.copernicus.eu/). The temperature dataset from 1950 to 2100 by using the climate models for Washington DC. However, there is a weak correlation between temperature and the bloom day of year.

Hence, only if we can find a good forecast climate method to forecast temperature, we can easily forecast the cherry blossom days of the year by ARIMA model with monthly maximum temperature from March and extreme maximum temperature from December.

Reference:

David W. Inouye. Effects of climate change on phenology, frost damage, and floral abundance of montane wildflowers. 01 February 2008, doi: 10.1890/06-2128.1.

P. Q. Craufurd., T. R. Wheeler. Climate change and the flowering time of annual crops. Journal of Experimental Botany. Volume 60, Issue 9, July 2009, Pages 2529–2539.

NBC Washington Staff (March 28, 2021). Cherry Blossom Trees Reach Peak Bloom Early Amid Warm Temps.

Impact of global warming on a group of related species and their hybrids: cherry tree (Rosaceae) flowering at Mt. Takao, Japan. American Journal of Botany 94: 1470–1478.

Autoregressive Integrated Moving Average (ARIMA)., By ADAM HAYES., Updated October 12, 2021., https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp.

Fit a linear model with time series components — tslm • forecast., https://pkg.robjhyndman.com/forecast/reference/tslm.html.

Ljung-Box Test Plot in LSTS: Locally Stationary Time Series., https://rdrr.io/cran/LSTS/man/Box.Ljung.Test.html.

Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

Martynas Rakickis, Time Series Forecasting with R Using Belgium Flights Data https://rpubs.com/marakis/belgium_flights_forecasting

Ubiqum Code Academy (2019), Forecasting non-seasonal time series with ARIMA models, https://rpubs.com/Mentors_Ubiqum/Tutorial_ARIMA_models
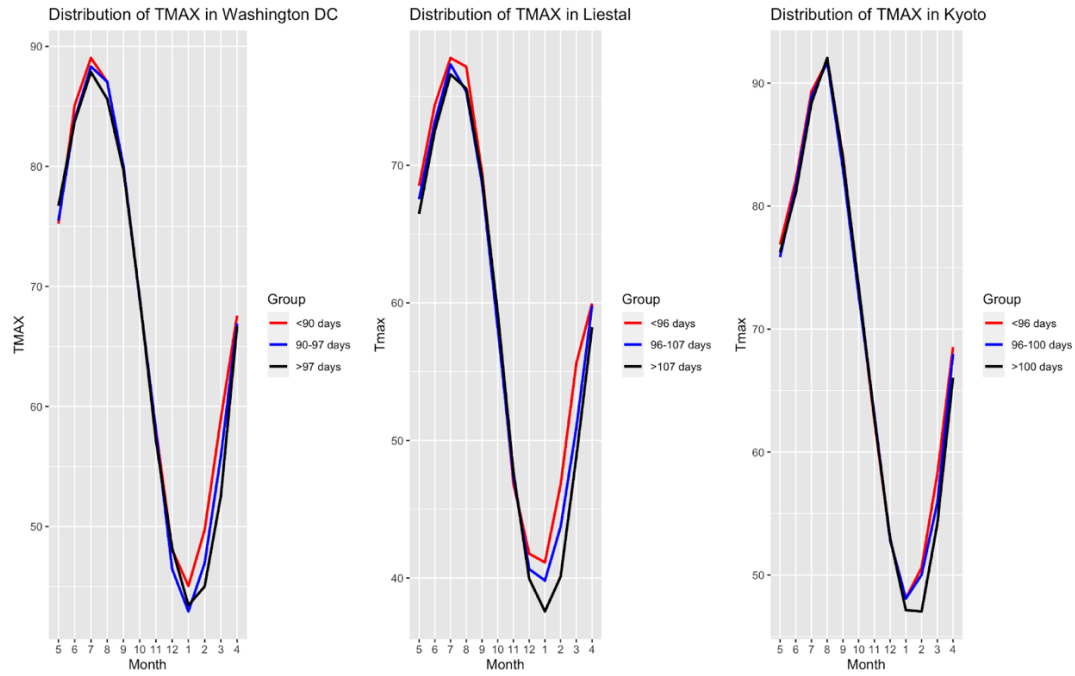
Fig. 1: Monthly Maximum Temperature in Washington DC, Liestal, and Kyoto.
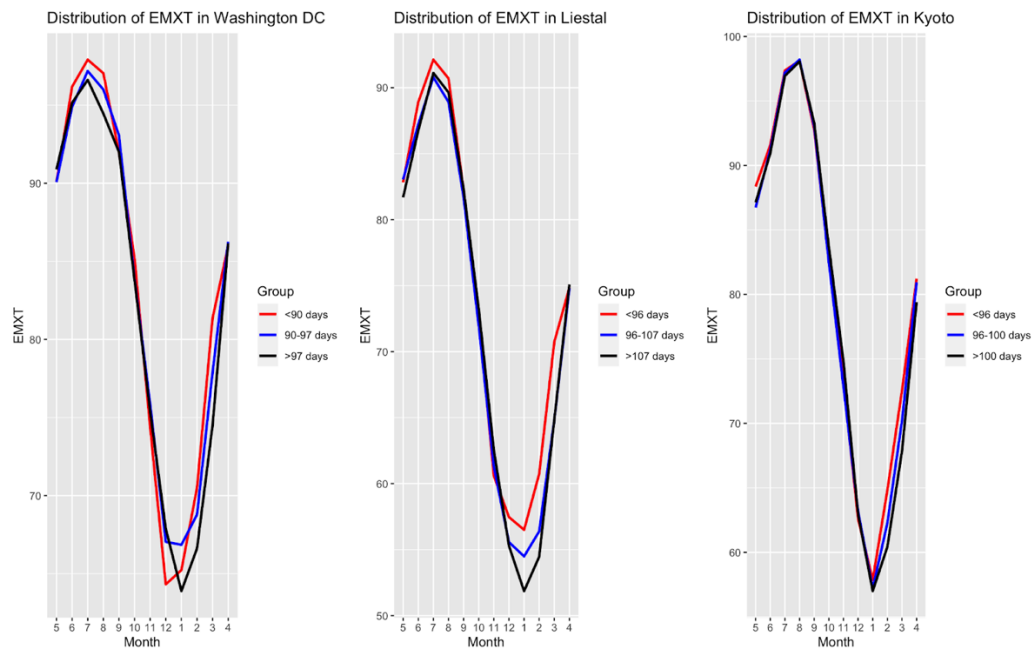
Fig. 2: Monthly Extreme Temperature in Washington DC, Liestal, and Kyoto.
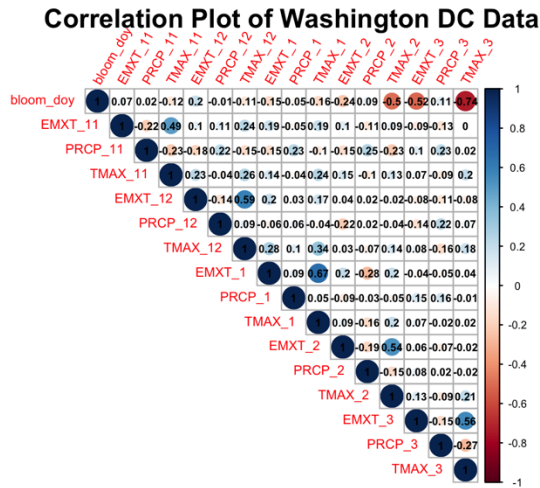
Fig. 3: Correlation between bloom date and covariates in Washington DC, USA.
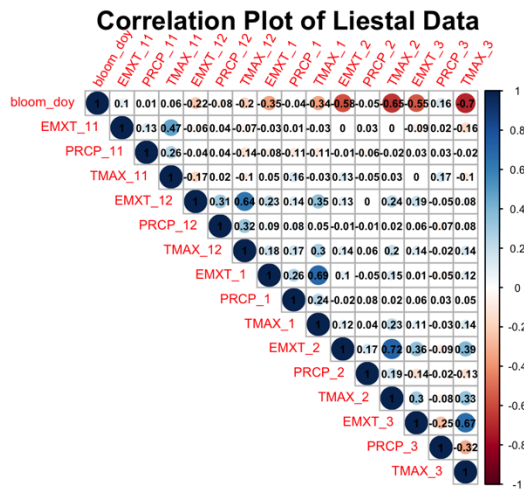
Fig. 4: Correlation between bloom date and covariates in Liestal, Switzerland.
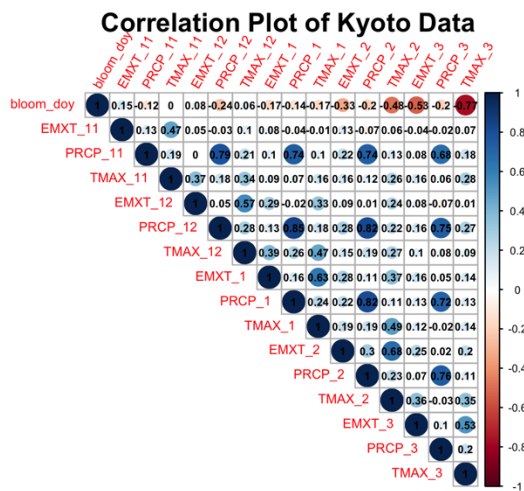
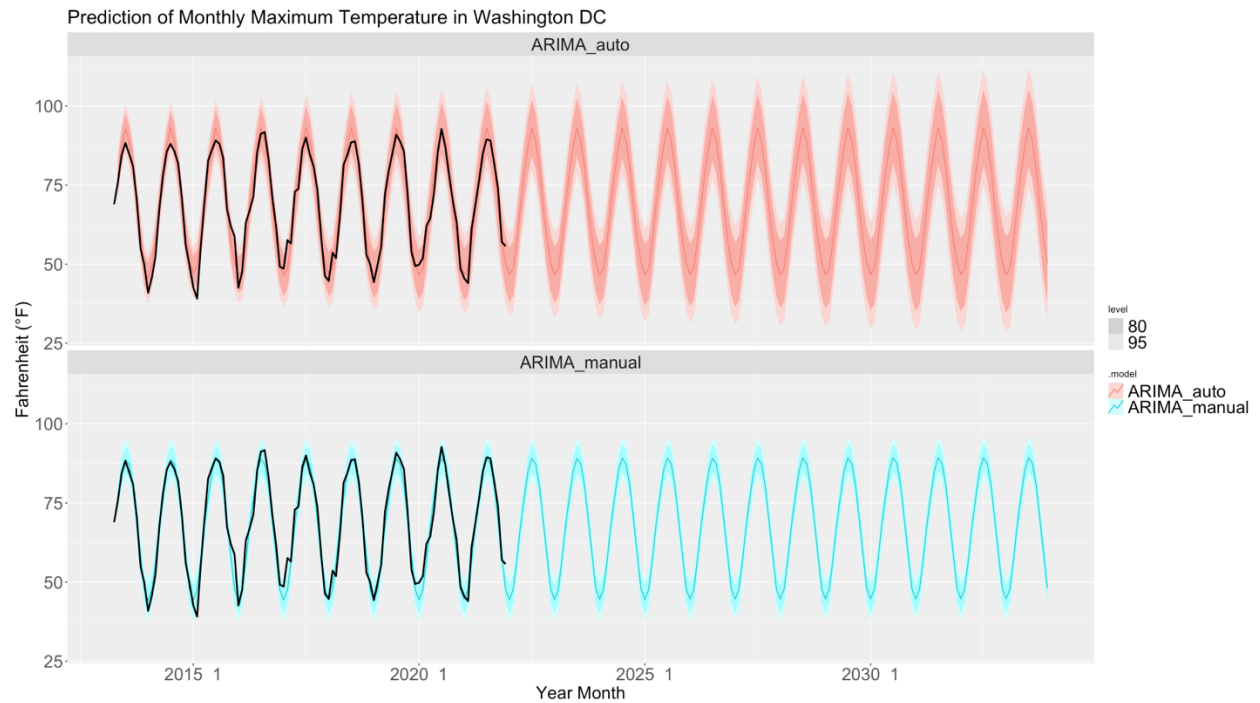Fig. 5: Correlation between bloom date and covariates in Kyoto, Japan.

Fig. 6: Prediction and Forecast monthly maximum temperature (TMAX) in Washington DC, from May 2013 to December 2033. (Black bold line is testing data.)
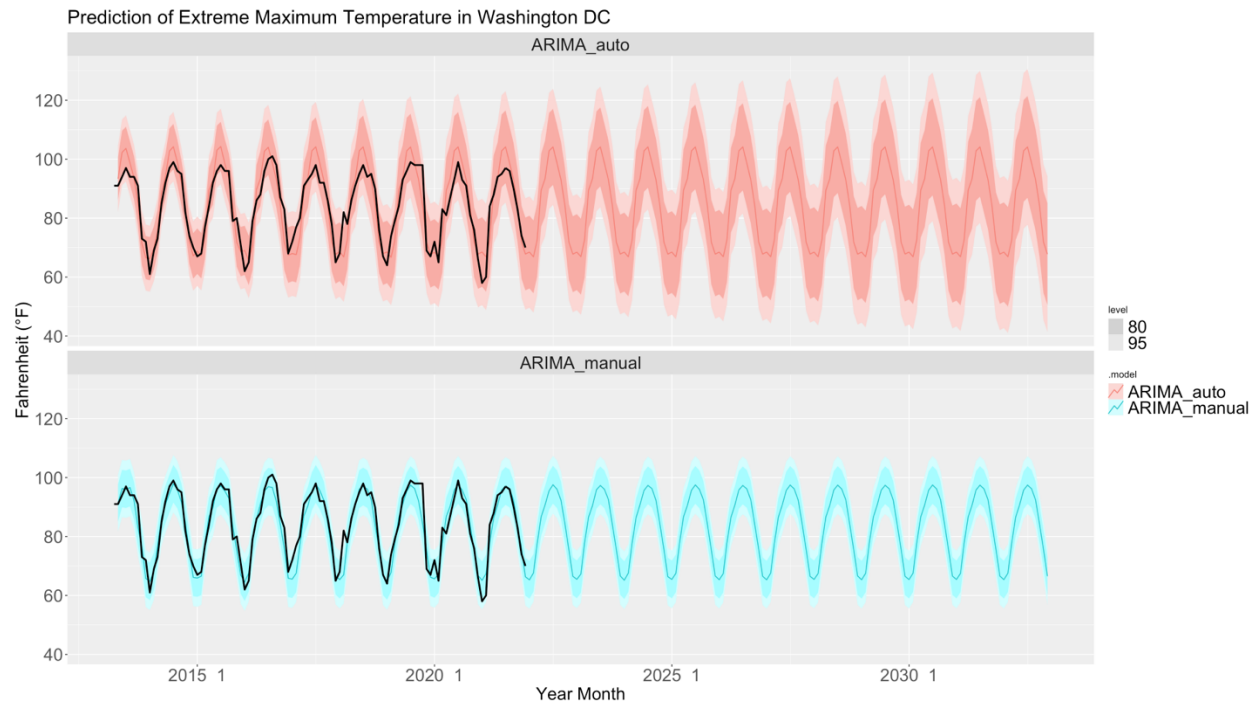
Fig. 7: Prediction and Forecast monthly extreme maximum temperature (EMXT) in Washington DC, from May 2013 to December 2032. (Black bold line is testing data.)

Table 1: Compare Models for Training data (left: Washington DC, middle: Kyoto, right: Listal).

DC Models Compare				Kyoto Models Compare				Liestal Models Compare			
model	AIC	AICc	BIC	model	AIC	AICc	BIC	model	AIC	AICc	BIC
ARIMA_without_covariate	447.74	447.93	452.15	ARIMA_without_covariate	359.90	360.10	364.16	ARIMA_without_covariate	813.41	813.53	818.76
ARIMA_with_covariates	388.29	389.27	399.31	ARIMA_with_covariates	314.39	315.09	322.90	ARIMA_with_covariates	737.08	737.47	747.77
ARIMA_with_TMAX_covariate	395.84	396.48	404.72	ARIMA_with_TMAX_covariate	312.95	313.36	319.33	ARIMA_with_TMAX_covariate	737.06	737.29	745.08
ARIMA_with_EMXT_covariate	443.54	443.92	450.15	ARIMA_with_EMXT_covariate	361.53	361.95	367.91	ARIMA_with_EMXT_covariate	812.66	812.89	820.68
TSLM_with_covariates	206.83	207.80	217.93	TSLM_with_covariates	134.89	135.94	145.60	TSLM_with_covariates	440.93	441.52	454.34
TSLM_with_TMAX_covariate	209.44	210.07	218.32	TSLM_with_TMAX_covariate	133.50	134.19	142.07	TSLM_with_TMAX_covariate	442.20	442.59	452.93

Table 2: Ljung–Box Test of the Models' Residuals (left: Washington DC, middle: Kyoto, right: Listal).

Ljung–Box test for DC			Ljung–Box test for Kyoto			Ljung–Box test for Liestal		
model	lb_stat	lb_pvalue	model	lb_stat	lb_pvalue	model	lb_stat	lb_pvalue
ARIMA_without_covariate	0	0.99	ARIMA_without_covariate	2.72	0.10	ARIMA_without_covariate	0.02	0.88
ARIMA_with_covariates	0.11	0.74	ARIMA_with_covariates	0	0.98	ARIMA_with_covariates	0.37	0.54
ARIMA_with_TMAX_covariate	0.02	0.89	ARIMA_with_TMAX_covariate	2.71	0.10	ARIMA_with_TMAX_covariate	0.03	0.85
ARIMA_with_EMXT_covariate	0.24	0.62	ARIMA_with_EMXT_covariate	0	0.96	ARIMA_with_EMXT_covariate	0.77	0.38
TSLM_with_covariates	9.63	0	TSLM_with_covariates	2.09	0.15	TSLM_with_covariates	2.73	0.10
TSLM_with_TMAX_covariate	8.16	0	TSLM_with_TMAX_covariate	1.99	0.16	TSLM_with_TMAX_covariate	3.55	0.06

Table 3: Compare Models for Testing data with predicted weather data (left: Washington DC, middle: Kyoto, right: Listal).

| DC models	ME	RMSE	MAE	MPE	MAPE	Kyoto model	ME	RMSE	MAE	MPE	MAPE	Liestal model	ME	RMSE	MAE	MPE	MAPE
ARIMA_with_covariates	-0.61	6.87	6.12	-1.27	6.85	ARIMA_with_covariates	-3.95	5.72	4.57	-4.48	5.11	ARIMA_with_covariates	-2.08	8.47	7.26	-3.09	8.16
ARIMA_with_TMAX_covariate	-2.40	7.28	6.51	-3.26	7.41	ARIMA_with_TMAX_covariate	-3.85	5.65	4.49	-4.37	5.03	ARIMA_with_TMAX_covariate	-1.89	8.40	7.21	-2.87	8.09
ARIMA_without_covariate	-0.12	7.03	6.25	-0.74	6.95	ARIMA_without_covariate	-3.89	5.73	4.62	-4.42	5.16	ARIMA_without_covariate	-3.24	8.82	7.44	-4.36	8.45
						TSLM_with_TMAX_covariate	-2.91	5.02	3.81	-3.35	4.27	TSLM_with_covariates	-4.22	9.14	7.56	-5.42	8.66
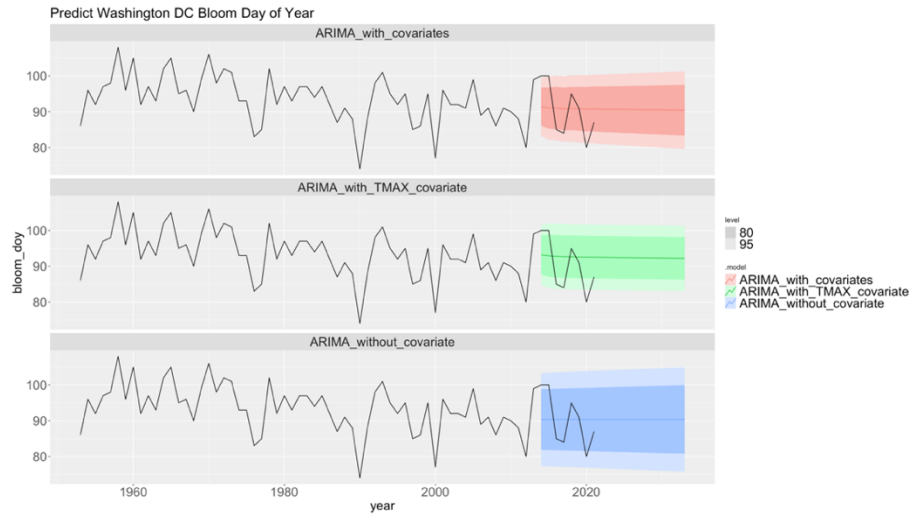
Fig. 8: Prediction and Forecast Bloom data of year in Washington DC with predicted weather data. (Black bold line is real data.)
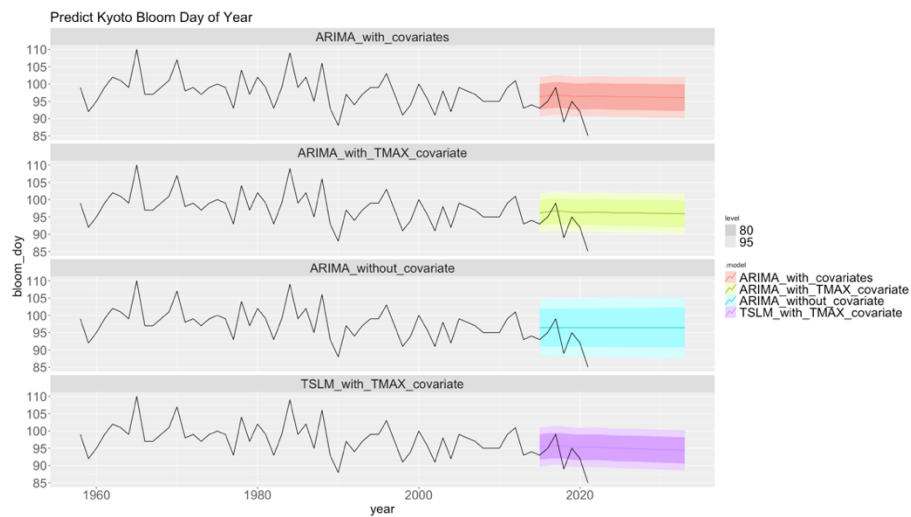

Fig. 9: Prediction and Forecast Bloom data of year in Kyoto with predicted weather data. (Black bold line is real data.)
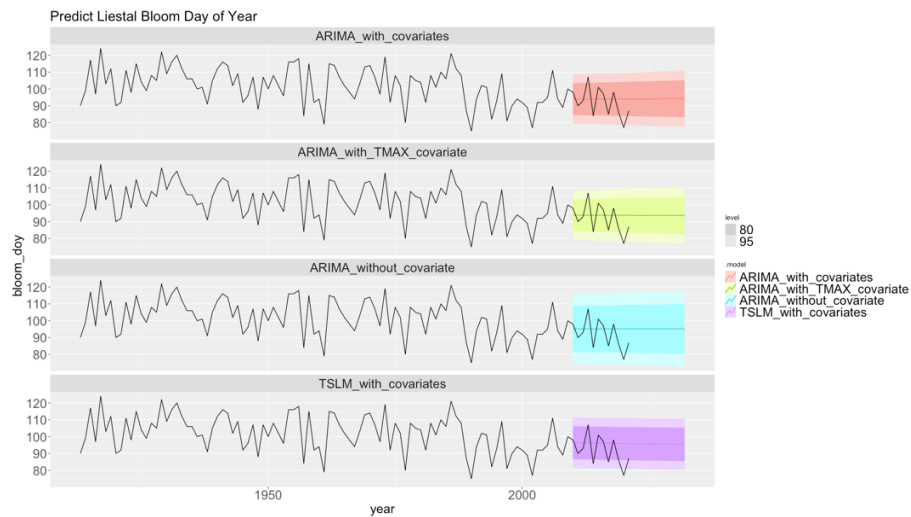

Fig. 10: Prediction and Forecast Bloom data of year in Liestal with predicted weather data. (Black bold line is real data.)

Table 4. Forecast bloom day of the year (left: Washington DC, middle: Kyoto, right: Listal).

DC	mean	80% CI		95% CI		Kyoto	mean	80% CI		95% CI		Liestal	mean	80% CI		95% CI	
2022	91	84	97	81	100	2022	95	92	99	90	101	2022	94	83	104	78	110
2023	91	84	97	81	100	2023	95	91	99	90	101	2023	94	83	104	78	110
2024	91	84	97	81	101	2024	95	91	99	89	101	2024	94	83	104	78	110
2025	91	84	97	81	101	2025	95	91	99	89	101	2025	94	83	104	78	110
2026	91	84	97	81	101	2026	95	91	99	89	101	2026	94	83	104	77	110
2027	91	84	97	80	101	2027	95	91	99	89	101	2027	94	83	104	77	110
2028	91	84	97	80	101	2028	95	91	99	89	101	2028	94	83	104	77	110
2029	91	84	97	80	101	2029	95	91	98	89	100	2029	94	83	105	77	110
2030	91	84	97	80	101	2030	95	91	98	89	100	2030	94	83	105	77	110
2031	90	84	97	80	101	2031	95	91	98	89	100	2031	94	83	105	77	110
2032	90	83	97	80	101	2032	94	91	98	89	100	2032	94	83	105	77	111
2033	90	83	98	80	101	2033	94	91	98	89	100	2033	94	83	105	77	111

Table 5: Compare Models for Testing data with real weather data (left: Washington DC, middle: Kyoto, right: Listal).

DC models	ME	RMSE	MAE	MPE	MAPE	Kyoto model	ME	RMSE	MAE	MPE	MAPE	Liestal model	ME	RMSE	MAE	MPE	MAPE
ARIMA_with_covariates	-0.19	3.93	3.05	-0.47	3.42	ARIMA_with_covariates	0.29	1.49	1.23	0.26	1.35	ARIMA_with_covariates	0.85	6.15	4.11	0.64	4.61
ARIMA_with_TMAX_covariate	-1.62	4.42	3.49	-2.08	3.95	ARIMA_with_TMAX_covariate	0.43	1.52	1.29	0.41	1.39	ARIMA_with_TMAX_covariate	0.5	6.46	4.39	0.26	4.91
ARIMA_without_covariate	-0.12	7.03	6.25	-0.74	6.95	ARIMA_without_covariate	-3.89	5.73	4.62	-4.42	5.16	ARIMA_without_covariate	-3.24	8.82	7.44	-4.36	8.45
						TSLM_with_TMAX_covariate	1.2	1.92	1.66	1.24	1.77	TSLM_with_covariates	-1.19	5.88	4.17	-1.61	4.69

Fig. 11: Prediction and Forecast Bloom data of year in Washington DC with real weather data. (Black bold line is real data.)
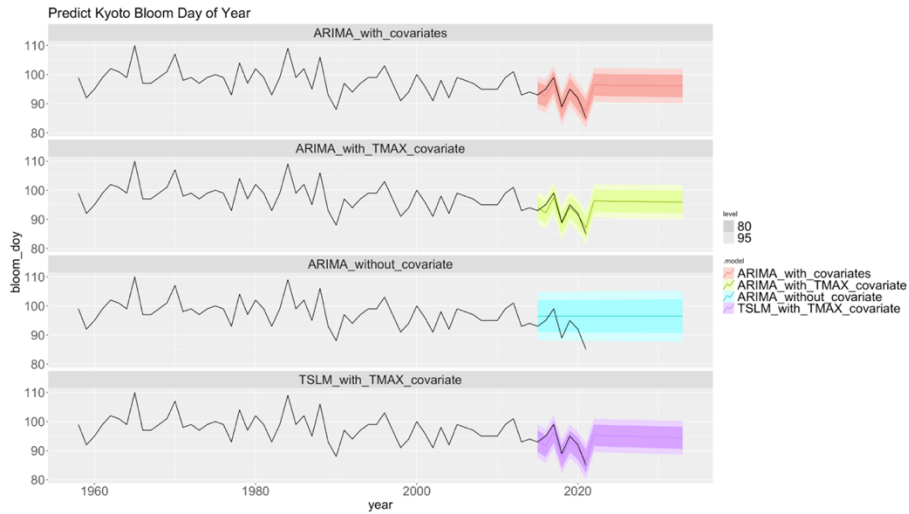

Fig. 12: Prediction and Forecast Bloom data of year in Kyoto with predicted weather data. (Black bold line is real data.)
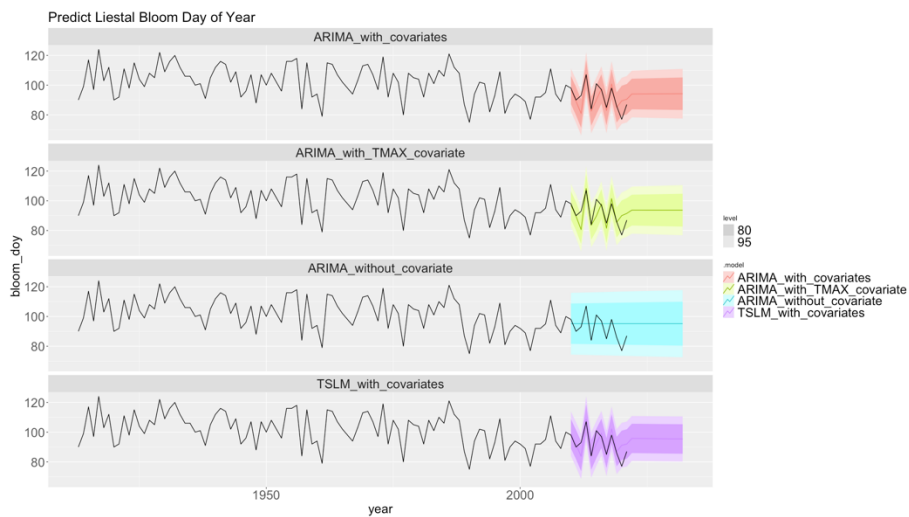

Fig. 13: Prediction and Forecast Bloom data of year in Liestal with predicted weather data. (Black bold line is real data.)
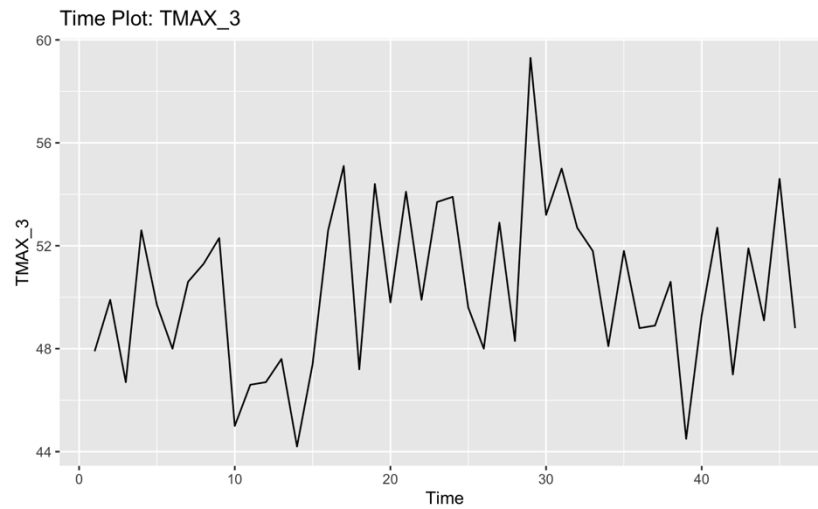
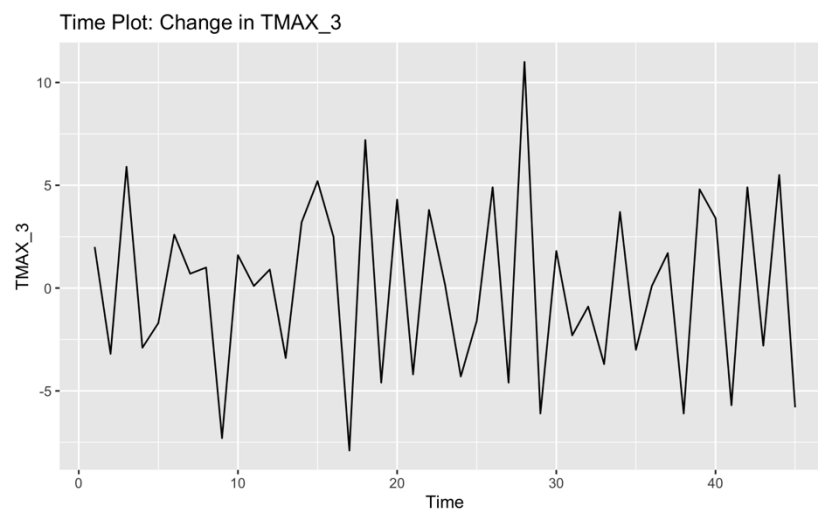Fig. 14: Time plot of TMAX_3 for Vancouver, Canada.
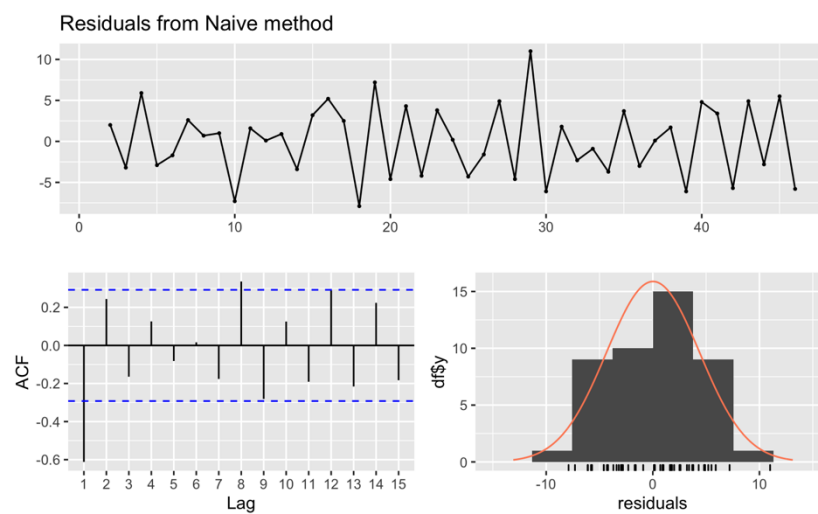

Fig. 15: Time plot of change in TMAX_3 for Vancouver, Canada.

Fig. 16: Residual plot of Naïve model for TMAX_3 for Vancouver, Canada.
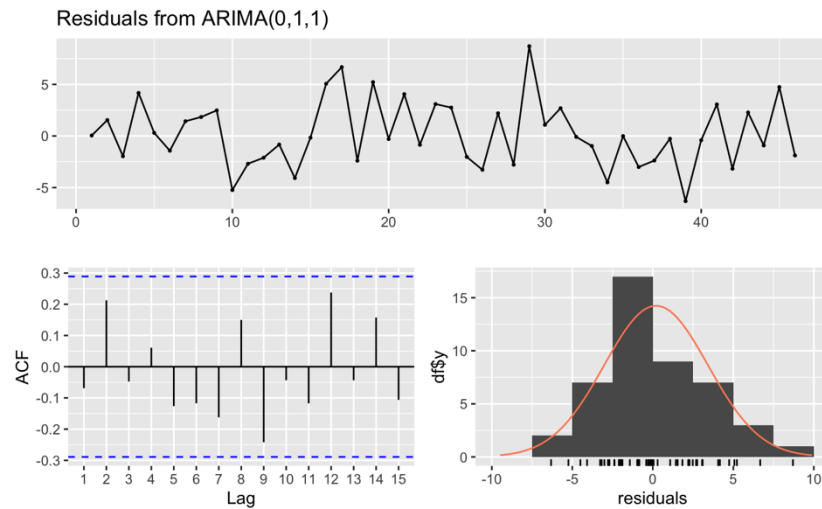
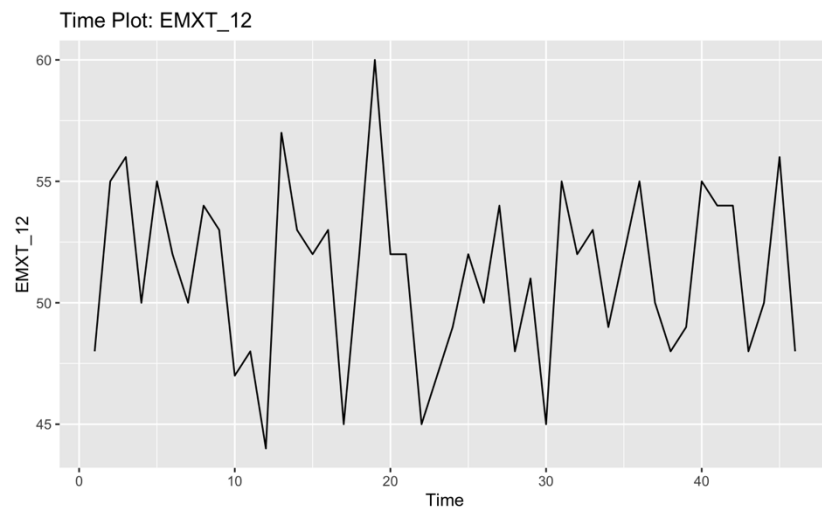Fig. 17: Residual plot of Arima model for TMAX_3 for Vancouver, Canada.

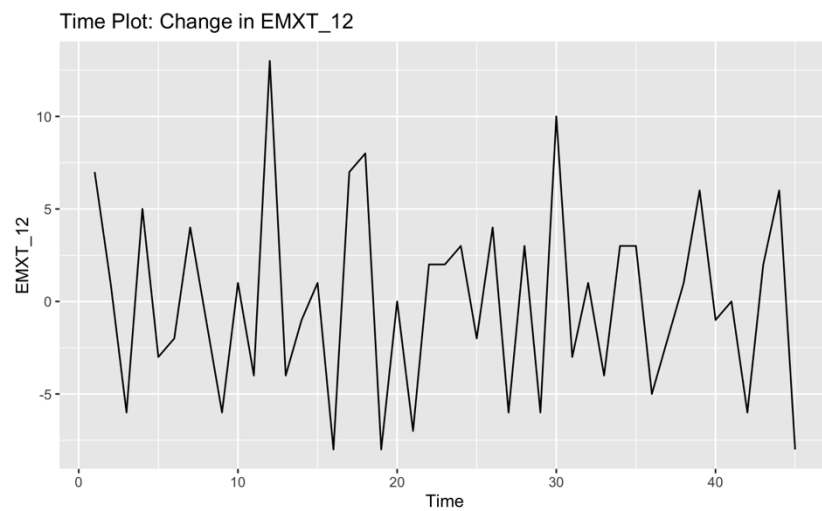

Fig. 18: Time plot EMXT_12 for Vancouver, Canada.

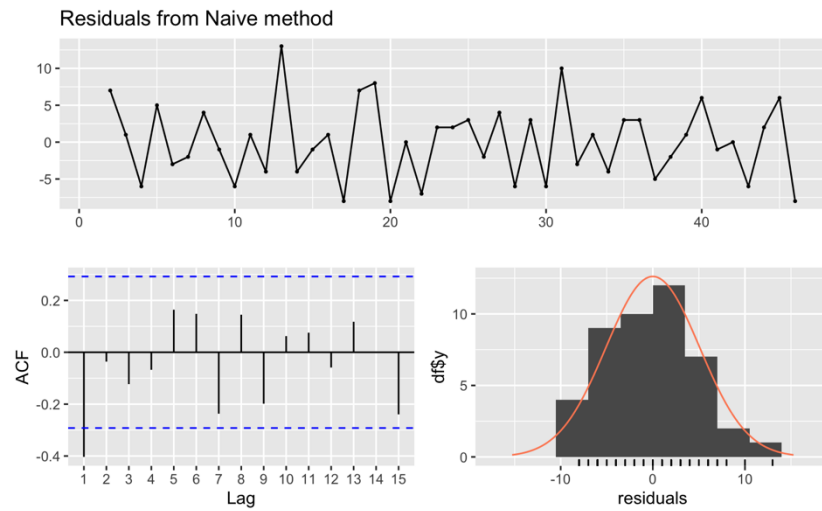Fig. 19: The time plot of change in EMXT_12 for Vancouver, Canada.

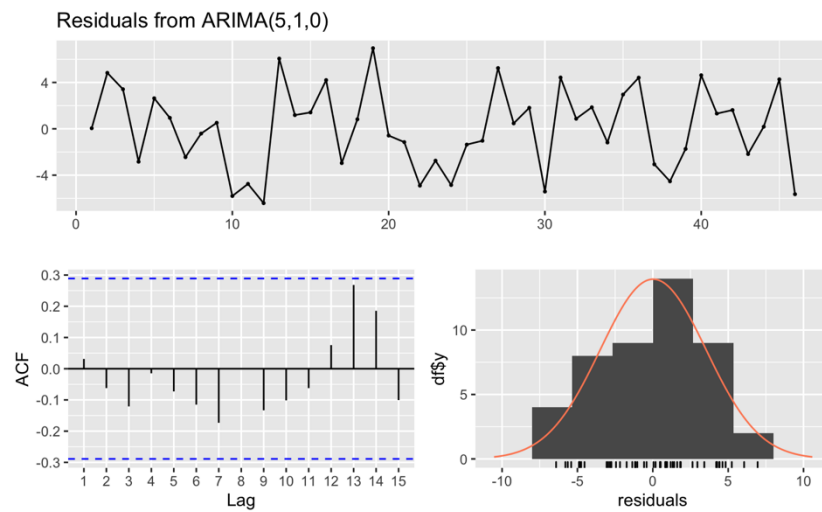Fig. 20: The Residual plot of Naïve model for EMXT_12 for Vancouver, Canada.

Fig. 21: The Residual plot of Arima model for EMXT_12 for Vancouver, Canada.