Modeling Oregon and Washington Windmill Wind Speeds

Dusty Turner

2 December 2022

Baylor University

Goal of analysis

The major goal of this presentation is...

To present a "Best" model for forecasting hourly wind speeds two weeks into the future (ambitious).

Adapted Model Fitting Lifecycle

- Explore the data
- Transform the data
- Identify model type(s)
- Estimate parameters
- Perform model diagnostics
- Choose a final model based on BIC
- Forecast

Explore the Data

Data Quirks

- Duplicated data
- Outliers
- “Lowliers”
- Missing Data

Duplicated Data

Every location has duplicated data from 0705-0800 1 March 2012

Location	Duplicated Rows
augs	24
biddle	24
butler	24
chin	24
forest	24
good	24
hood	24
horse	24
kenn	24
marys	24

Location	Duplicated Rows
megler	24
mehebo	24
naselle	24
roose	24
seven	24
shani	24
sunny	24
tilla	24
trout	24
wasco	24

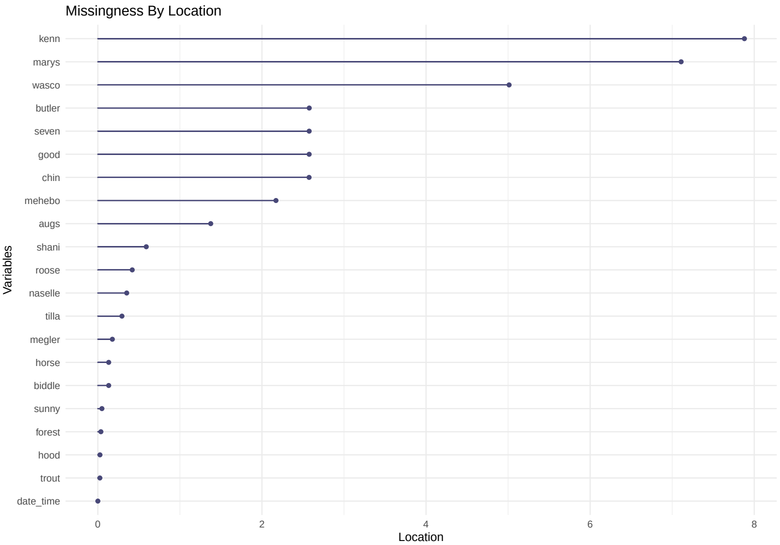
Windspeed Missingness

Windspeed Missingness Summary

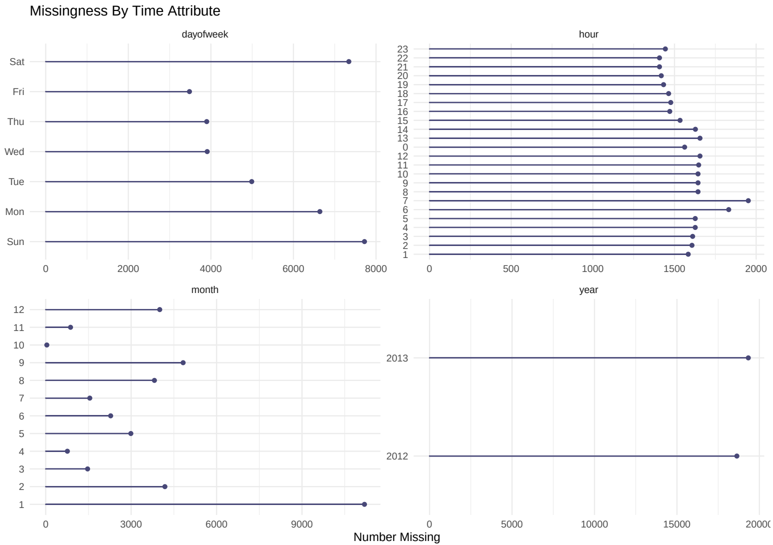
These locations will not be modeled

Location	Explicit_Missing	Implicit_Missing	Total_Missing	Percent
kenn	3,819	4,476	8,295	7.88%
marys	7,471	12	7,483	7.11%
wasco	800	4,476	5,276	5.01%
butler	2,700	12	2,712	2.58%
good	2,699	12	2,711	2.58%

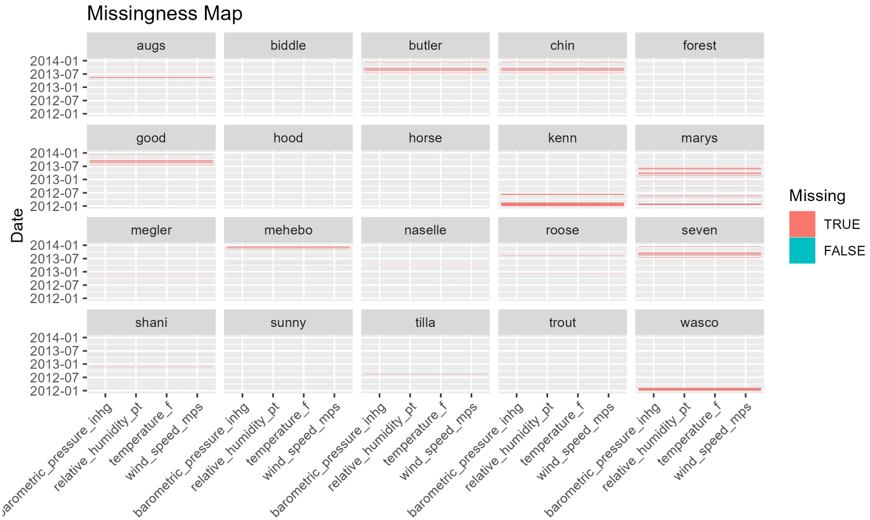
Windspeed Missingness



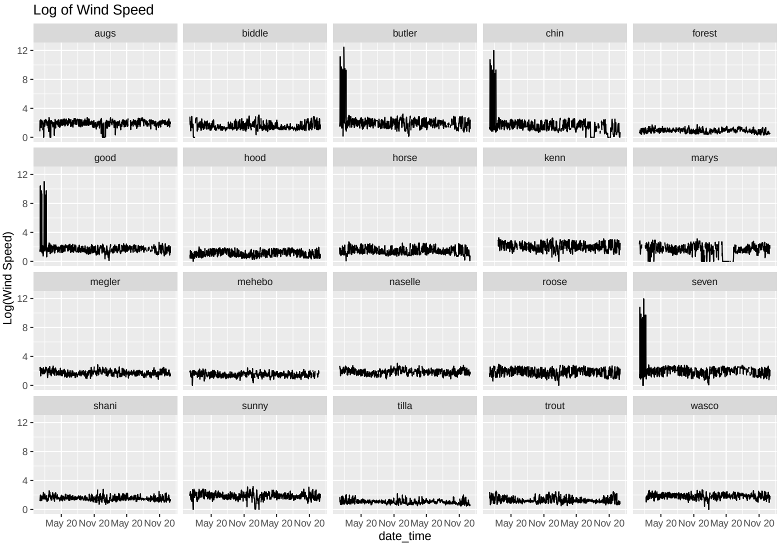
Windspeed Missingness



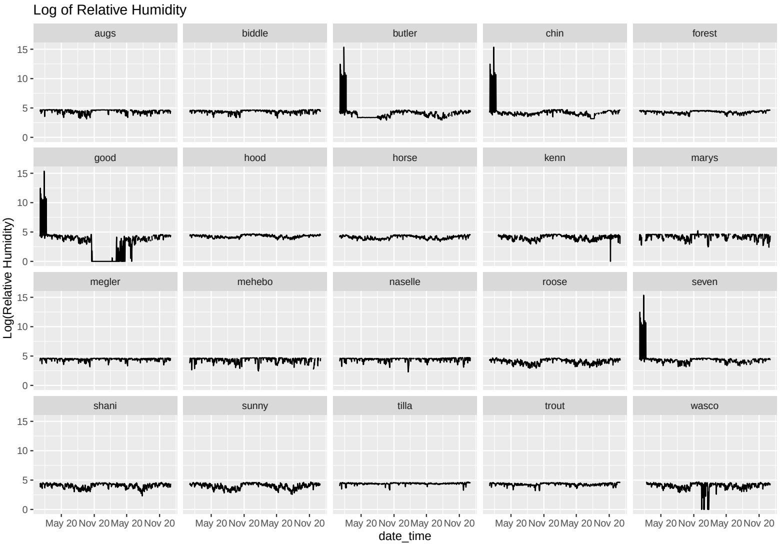
Overall Missingness



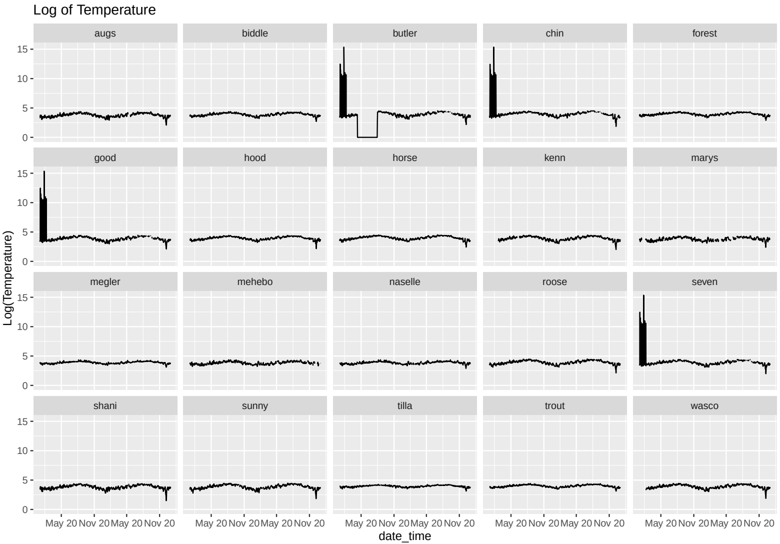
High Outliers



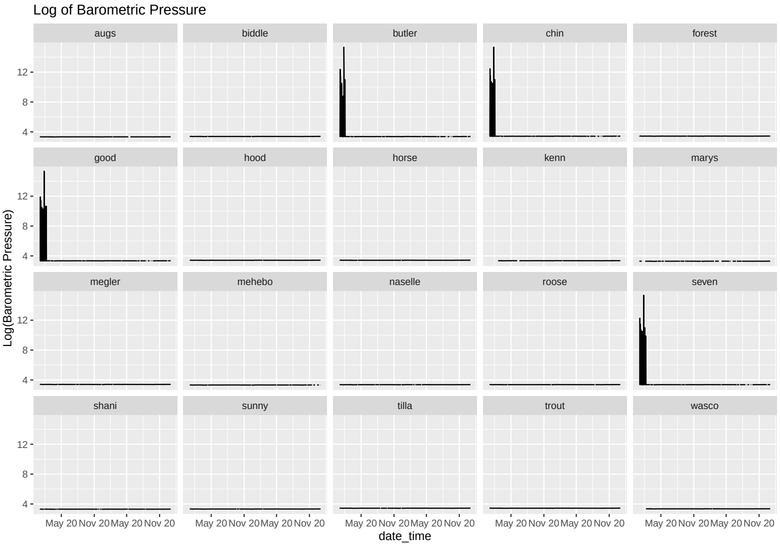
High Outliers



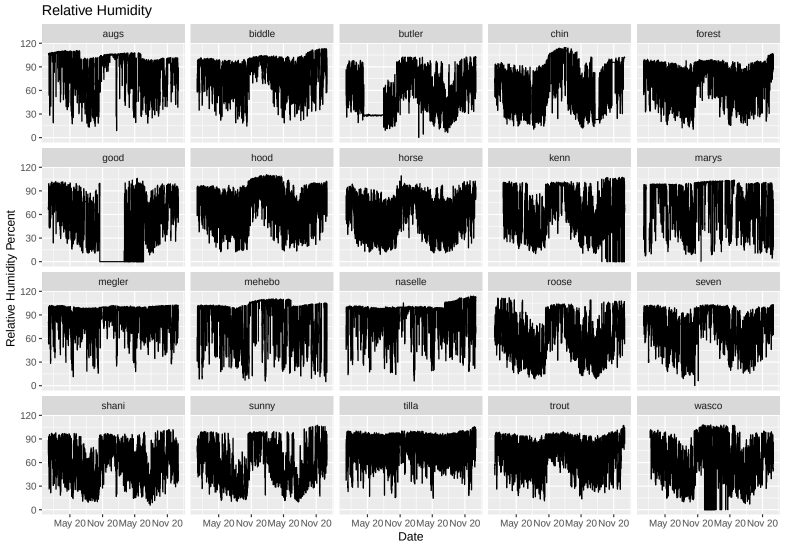
High Outliers



High Outliers

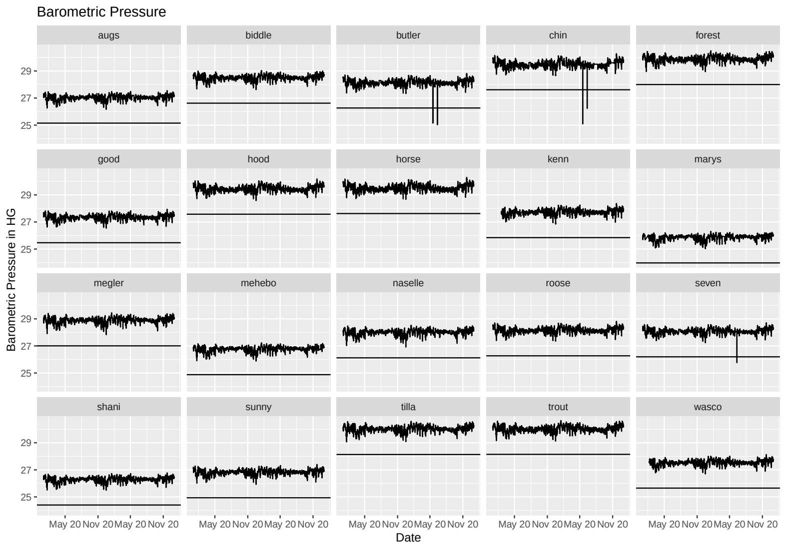


Low Outliers



0 percent humidity is not possible

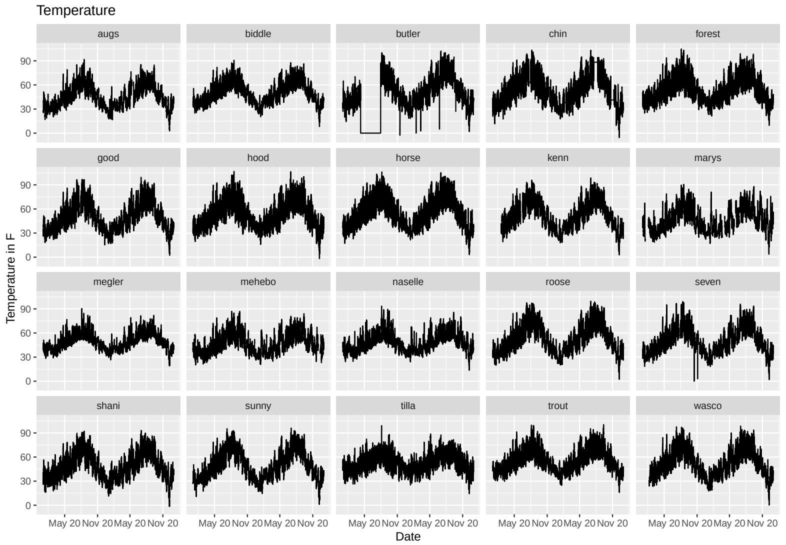
Low Outliers



Below 32 is not feasible

Remove below 2 standard deviations

Low Outliers



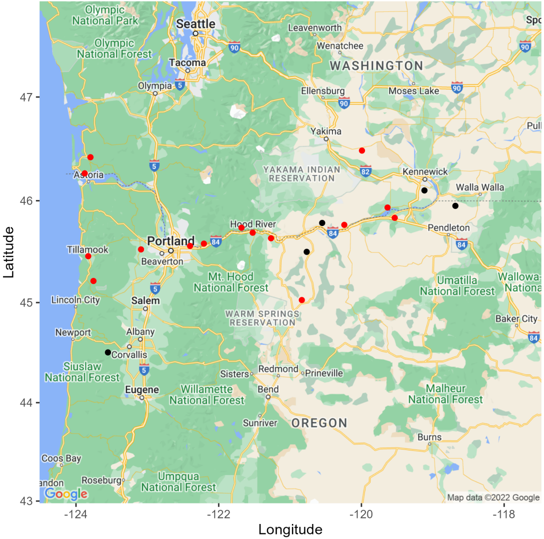
Will remove Butler because of missing data.

Data Quirks In Summary

- Good, Kenn, Marys, Wasco, or Butler will not be considered.
  - Too much missing data
  - Not missing at random
- Treat outliers as NAs
- Impute all NAs with Exponential Weighted Moving Average (EWMA) with window of 4.

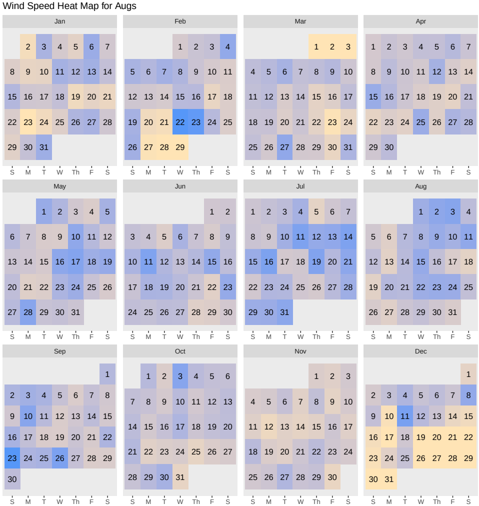
Location	Not Missing	Missing	Percent Missing
good	380,064	40,992	9.74%
kenn	386,397	34,659	8.23%
marys	390,534	30,522	7.25%
wasco	396,104	24,952	5.93%
butler	410,129	10,927	2.60%
seven	410,134	10,922	2.59%
chin	410,140	10,916	2.59%
mehebo	411,958	9,098	2.16%
augs	415,306	5,750	1.37%
shani	418,611	2,445	0.58%
roose	419,334	1,722	0.41%

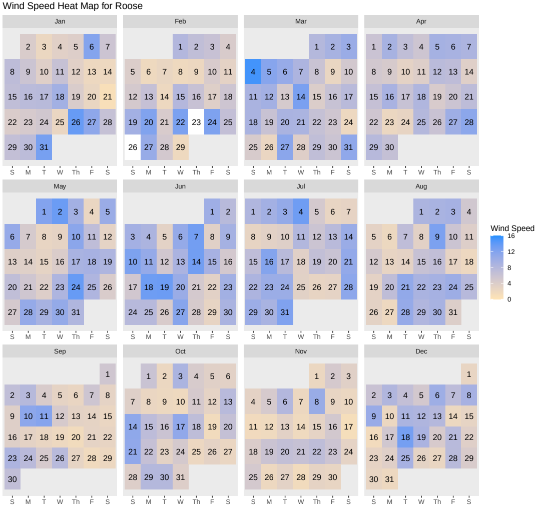
View Locations



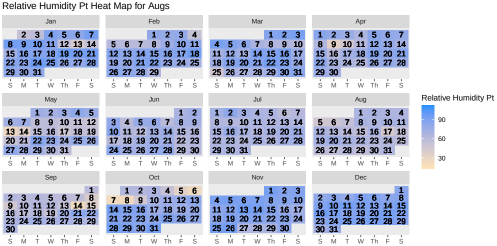
Explore Trends

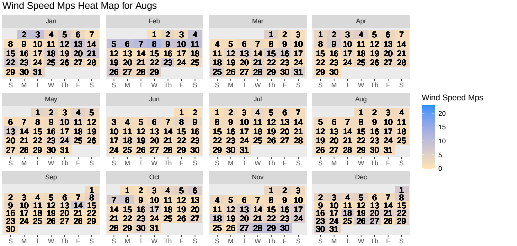
General Trends in Temperature

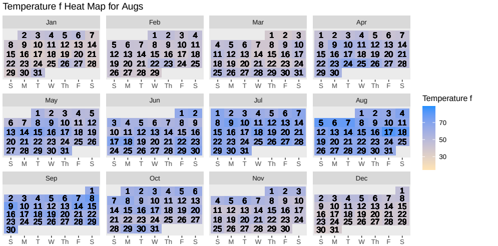


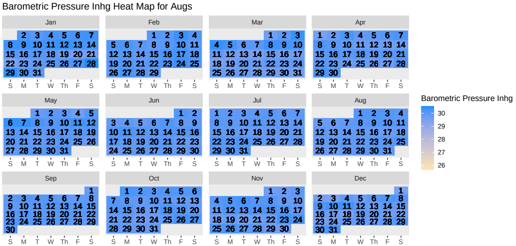


Trends in Wind, Pressure, Temperature, and Humidity in Augs

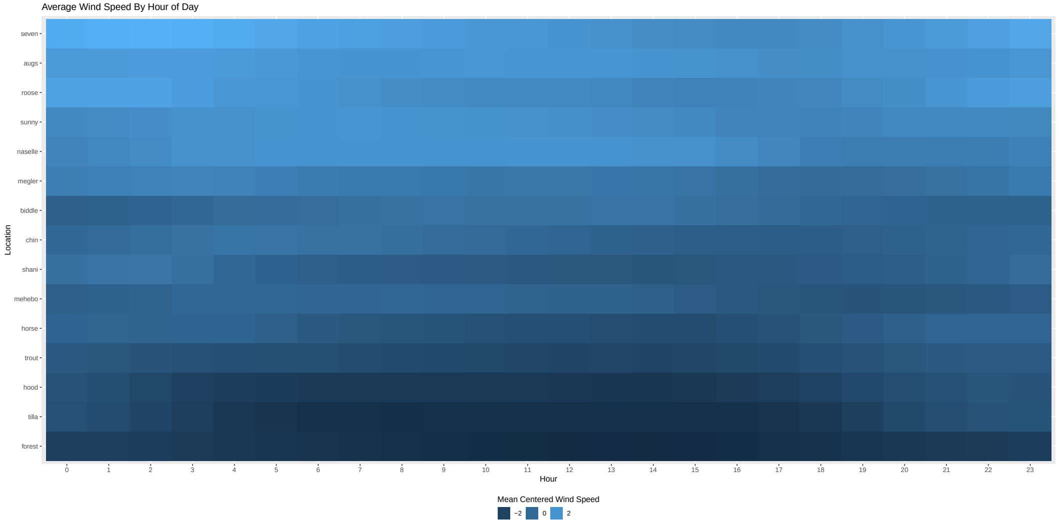




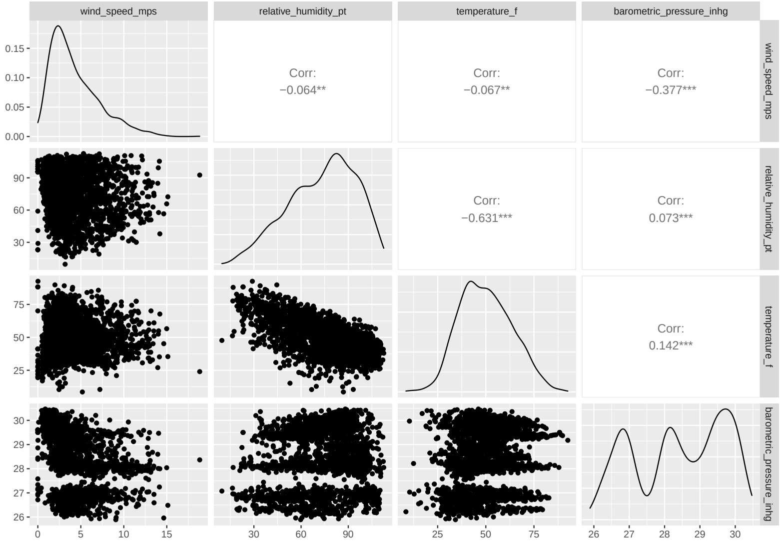




Hourly Wind Speed Fluxuations



Correlations

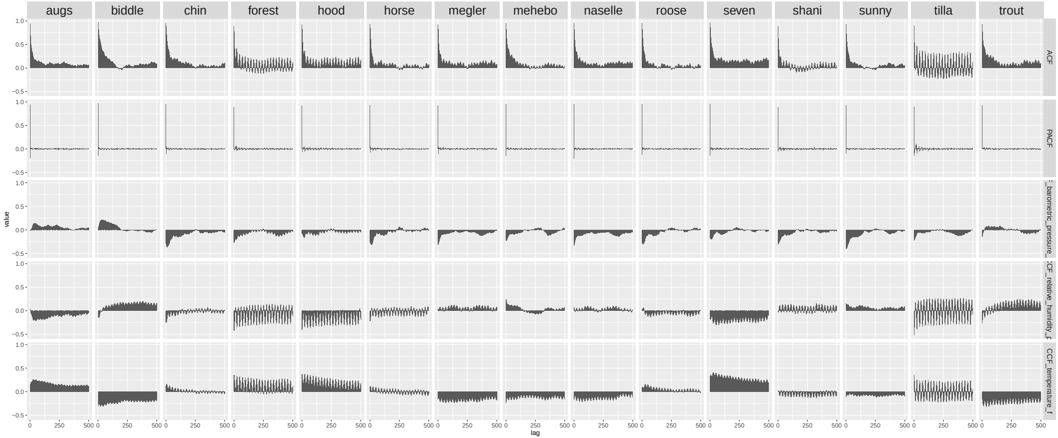


Check Back in on Adapted Model Fitting Lifecycle

- Explore the data
- Transform the data
- Identify model type(s)
- Estimate parameters
- Perform model diagnostics
- Choose a final model based on BIC
- Forecast

Identify Model types(s)

Cross Correlation Between Wind Speed and Other Factors



Modeling Strategy

- Model by location
- Find CCF lags that may by leading indicators of wind speed for each location
- Create multiple 'ARIMA formulas' consisting of each set of lags to build 15 models for each location (total of 15 x 15 $=$ 115 models)
- Use the `{fable}` package to determine best P, D, Q, p, d, and q values for each ARIMA formula.

Example formulas on next slide

Example of several formulas

Example 1

Wind Speed $\sim$ lag(Barometric Pressure, 4) + lag(Barometric_Pressure, 41) + lag(Relative Humidity, 1) + lag(Relative Humidity, 51) + lag(Temperature, 29) + lag(Temperature, 329) + Relative Humidity + Temperature + Barometric Pressure

Example 2

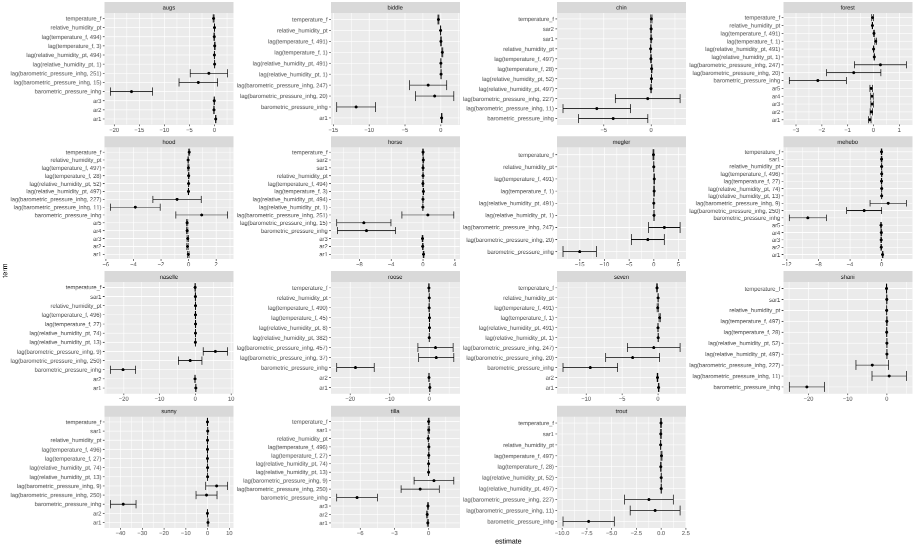
Wind Speed $\sim$ lag(Barometric Pressure, 37) + lag(Barometric_Pressure, 457) + lag(Relative Humidity, 8) + lag(Relative Humidity, 382) + lag(Temperature, 45) + lag(Temperature, 490) + Relative Humidity + Temperature + Barometric Pressure

Fable

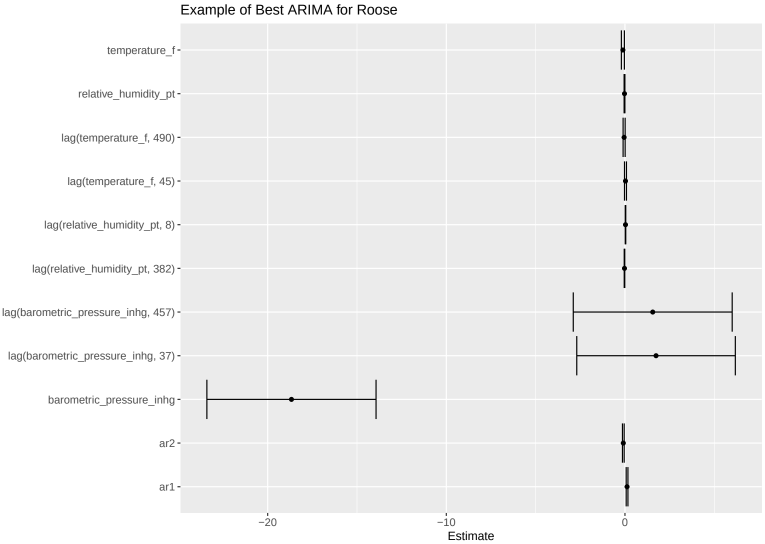
`purrr::map()` over the function below to execute each ARIMA model formula for each of the 15 locations.

```
my_maple <- function(data, formula) {
  model(.data = data,
        arima = ARIMA(formula = as.formula(formula),
                      order_constraint = p + q + P + Q <= 200 &
                        (constant + d + D <= 32)
                      )
        )
}
```

Model Results



Single Model

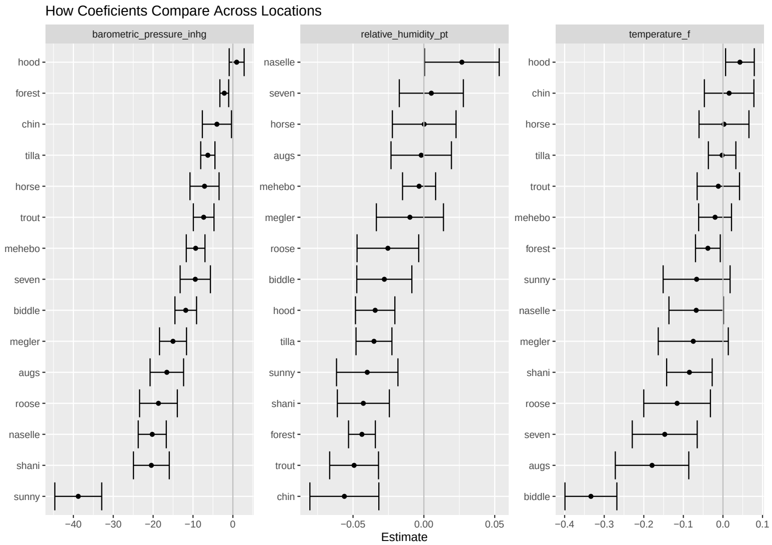


Single Model

Output for 'best' ARMIA Model for Roose: <LM w/ ARIMA(2,1,0)>

term	estimate	std.error	statistic	p.value
barometric_pressure_inhg	−18.67	2.42	−7.73	0.00
ar1	0.12	0.03	4.70	0.00
lag(relative_humidity_pt, 8)	0.03	0.01	4.17	0.00
ar2	−0.09	0.02	−3.68	0.00
lag(relative_humidity_pt, 382)	−0.03	0.01	−3.24	0.00
temperature_f	−0.12	0.04	−2.70	0.01
relative_humidity_pt	−0.03	0.01	−2.29	0.02
lag(temperature_f, 490)	−0.05	0.03	−1.58	0.11
lag(temperature_f, 45)	0.03	0.03	1.06	0.29
lag(barometric_pressure_inhg, 37)	1.74	2.26	0.77	0.44
lag(barometric_pressure_inhg, 457)	1.55	2.27	0.69	0.49

How Exogenous Variables Impact Each Location

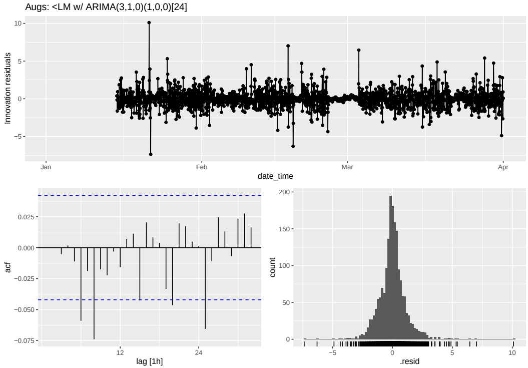


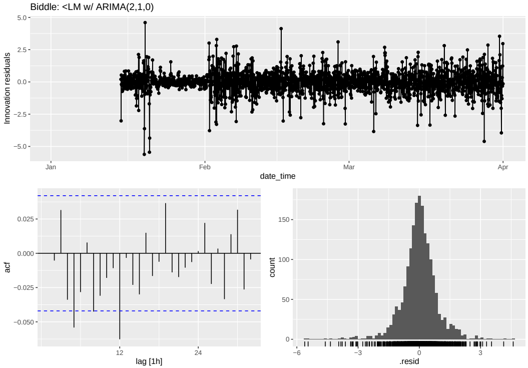
Check Back in on Adapted Model Fitting Lifecycle

- Explore the data
- Transform the data
- Identify model type(s)
- Estimate parameters
- Perform model diagnostics
- Choose a final model based on BIC
- Forecast

Perform model diagnostics

Model Diagnostics with {feasts}





Compare ARIMA to Other Model Types

Other model options: Vector Autoregressive Models (VAR)

Vector Autoregressive Models: "The structure is that each variable is a linear function of past lags of itself and past lags of the other variables."

$VAR(1)$ model can be specified as follows:

$$
\begin{aligned}
x_{t,1} &= \alpha_1 + \phi_{11}x_{t-1,1} + \phi_{12}x_{t-1,2} + \phi_{13}x_{t-1,3} + \phi_{14}x_{t-1,4} + w_{t,1} \\
x_{t,2} &= \alpha_2 + \phi_{21}x_{t-1,1} + \phi_{22}x_{t-1,2} + \phi_{23}x_{t-1,3} + \phi_{24}x_{t-1,4} + w_{t,2} \\
x_{t,3} &= \alpha_3 + \phi_{31}x_{t-1,1} + \phi_{32}x_{t-1,2} + \phi_{33}x_{t-1,3} + \phi_{34}x_{t-1,4} + w_{t,3} \\
x_{t,4} &= \alpha_4 + \phi_{41}x_{t-1,1} + \phi_{42}x_{t-1,2} + \phi_{43}x_{t-1,3} + \phi_{44}x_{t-1,4} + w_{t,4}
\end{aligned}
$$

where

$x_{t,i}$ is the model prediction at time t for $i \in$ (wind speed, relative humidity, temperature, barometric pressure)

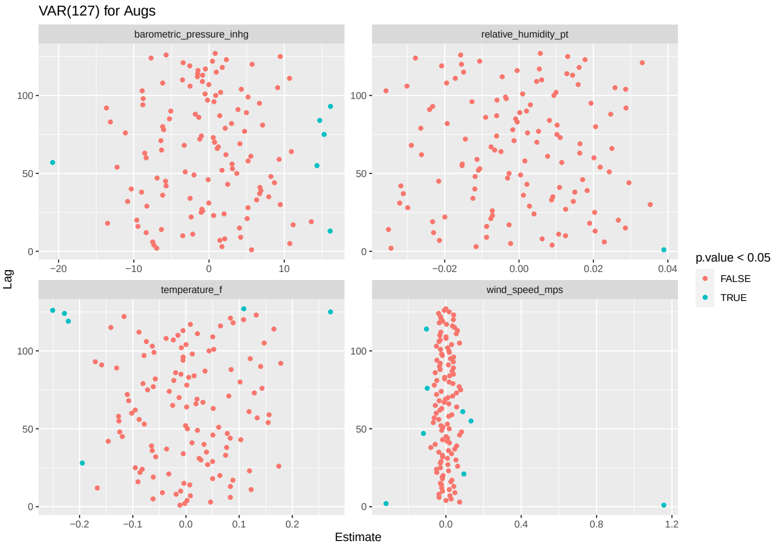
VAR in R {fable}

For each location, this function finds the best model VAR(P) model (24 x 7 x 2 x 15) models.

Best VAR(P) Model for Each Location

Location	BIC	VAR(P)
augs	$19,197$	127
biddle	$18,685$	127
chin	$23,532$	127
forest	$17,510$	127
hood	$20,646$	127
horse	$23,317$	127
megler	$20,345$	127
mehebo	$20,539$	127
naselle	$20,673$	127
roose	$23,182$	127
seven	$23,371$	127
shani	$23,863$	127
sunny	$22,786$	127

VAR(P) Model for Augs



Other Model Options: Error-Trend-Seasonality (ETS)

Error-Trend-Seasonality Models estimate the following:

Error: Additive or Multiplicative
Trend: Additive, Multiplicative, or None
Seasonality: Additive, Multiplicative, or None

From the docs: "If more than one method is specified, `{fable}` will consider all combinations of the specified models and select the model which best fits the data (minimizing) ic"

ETS does not support Exogenous regressors

ETS in R {fable}

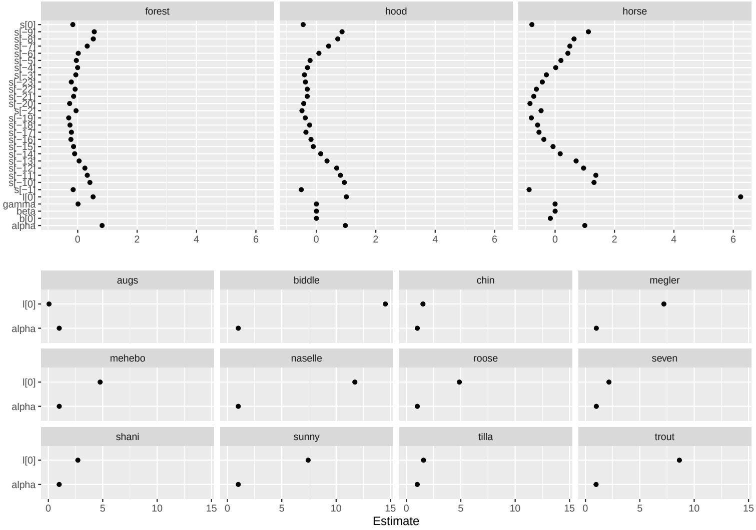
```r
tictoc::tic()
fit_ets <-
  progressr::with_progress(
    model(.data = data_imputed_limited_hourly_small,
          ets = ETS(
            wind_speed_mps ~ trend(method = c("N", "A", "M")) +
              error(method = c("A", "M")) +
              season(method = c("N", "A", "M")
                 ) # None ("N")
            )          # Additive("A")
          )            # Multiplicative("M")
    )
tictoc::toc()
```

```
32.94 sec elapsed
```

Best ETS Model for Each Location

Location	BIC
augs	$17,549$
biddle	$16,441$
chin	$18,162$
forest	$14,047$
hood	$15,923$
horse	$17,992$
megler	$17,716$
mehebo	$15,940$
naselle	$17,877$
roose	$18,422$
seven	$18,002$
shani	$18,161$
sunny	$18,713$

ETS Models


forest
hood
horse
s[0]
l[0]
gamma
beta
b[0]
alpha
0
2
4
6
augs
biddle
chin
megler
mehebo
naselle
roose
seven
shani
sunny
tilla
trout
l[0]
alpha
0
5
10
15
Estimate

Adapted Model Fitting Lifecycle

- Explore the data
- Transform the data
- Identify model type(s)
- Estimate parameters
- Perform model diagnostics
- Choose a final model based on BIC
- Forecast

Choose a Final Model

Choose a Final Model Based on BIC

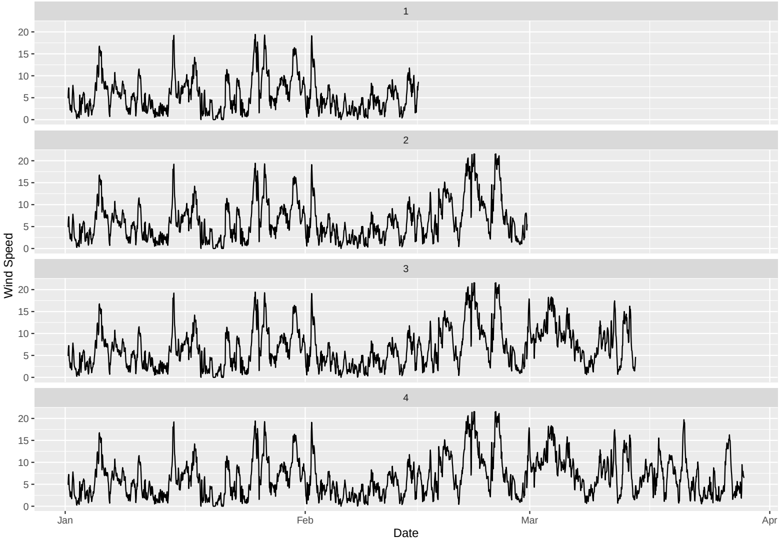
Location	Arima	ETS	VAR
forest	2,517	14,047	17,510
mehebo	4,010	15,940	20,539
hood	4,015	15,923	20,646
tilla	4,019	15,921	21,451
biddle	4,278	16,441	18,685
trout	4,684	16,814	21,233
augs	5,186	17,549	19,197
megler	5,471	17,716	20,345
naselle	5,500	17,877	20,673
horse	5,663	17,992	23,317
shani	5,851	18,161	23,863
seven	5,864	18,002	23,371

Adapted Model Fitting Lifecycle

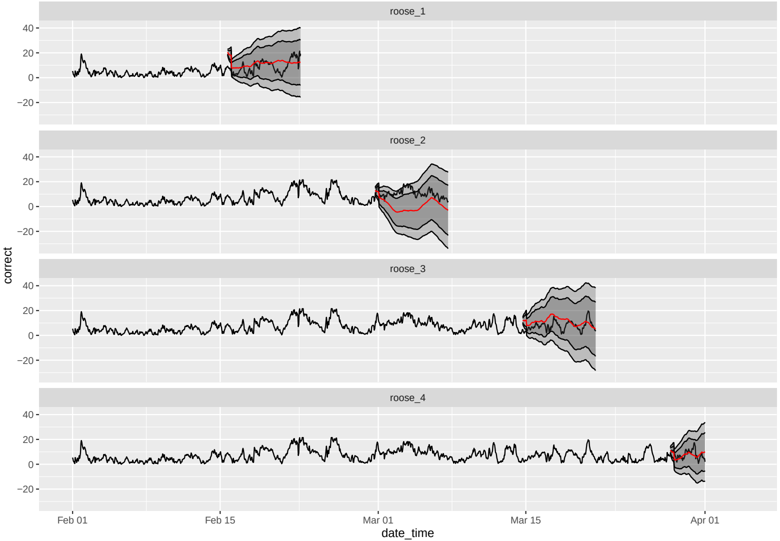
- Explore the data
- Transform the data
- Identify model type(s)
- Estimate parameters
- Perform model diagnostics
- Choose a final model based on BIC
- Forecast

Forecast

Create expanding windows for each loction



Make Predictions for Each Location and Each Window



Observe Accuracy for Each Location

Location	RMSE
augs	7.99
biddle	3.29
chin	5.51
forest	2.38
hood	2.32
horse	8.20
megler	3.58
mehebo	2.62
naselle	4.69
roose	7.15
seven	7.20
shani	6.68
sunny	7.08

Questions?