Modeling Oregon and Washington Windmill Wind Speeds

Dusty Turner

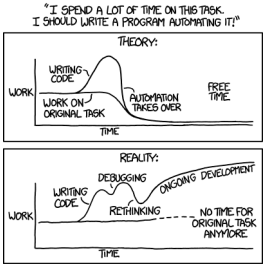
2 December 2022

Goal of Analysis

The major goals of this presentation are...

- By location, present an hourly "best" **ARIMA** model that considers **exogenous** variables
- Compare the results to **VAR** and **ETS** models
- **Forecast** using the best model
- Available online[a] [b]

[a] https://dustysturner.com/ts_presentation_fall_2022
[b] https://dustysturner.com/ts_presentation_fall_2022_dashboard



Model Building Overview

Modeling Overview

- Model every location
- Aggregate data by hour
- Training Set: 1 March - 31 May 2012
- Testing Set: 1 March - 31 May 2013

Model Building Strategy

- Address missing data
- Build ARIMA models
- Compare to VAR[a] and ETS[b] models
- Forecast with best models

[a] Vector Autoregressive Model
[b] Error-Trend-Seasonality

Cleaning the Data

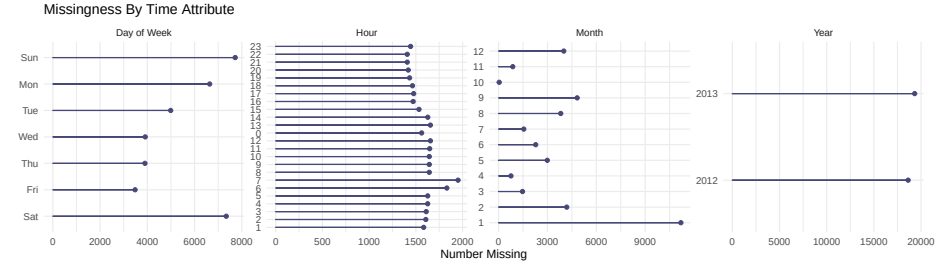
- Address “duplicated” data
- Address outliers
- Address “lowliers”
- Explicit missing data
- Implicit missing data

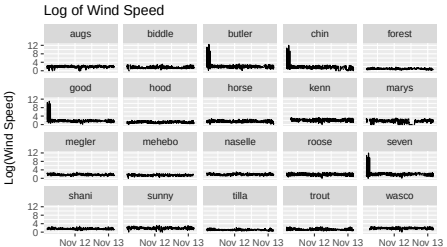
Windspeed Missingness Summary

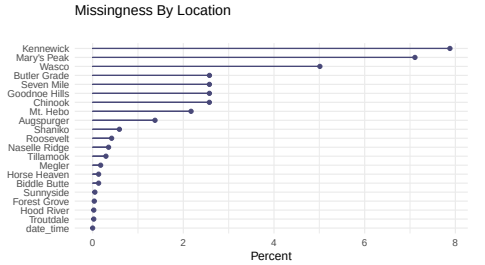
These locations will not be modeled

Location	Explicit	Implicit	Total	Percent
Kennewick	3819	4476	8,295	7.88%
Mary's Peak	7471	12	7,483	7.11%
Wasco	800	4476	5,276	5.01%
Butler Grade	2700	12	2,712	2.58%
Goodnoe Hills	2699	12	2,711	2.58%

Addressing Data Quirks

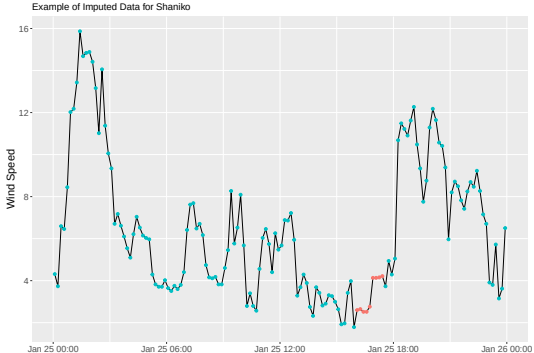




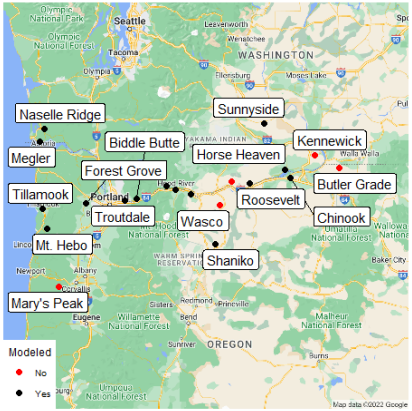


Cleaning Data Bottom Line

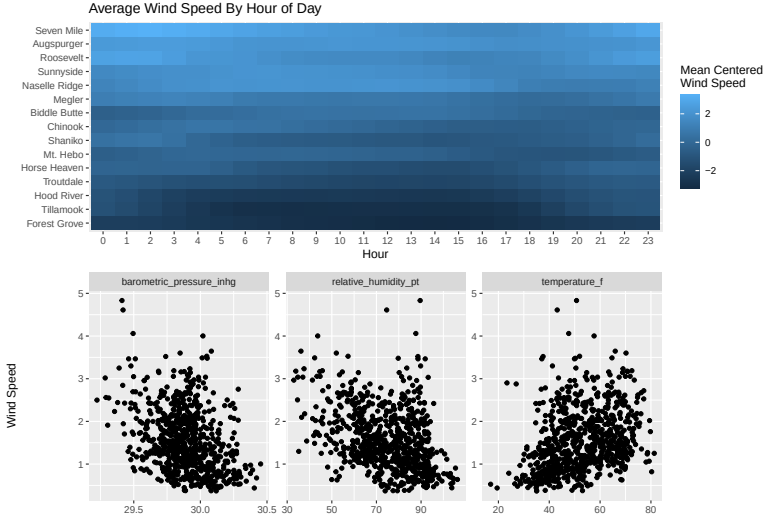
The following places will not be considered due to data issues:



Goodnoe Hills, Kennewick, Mary's Peak, Wasco, Butler Grade



Summarized Exploratory Data Analysis



ARIMA Model Building Strategy

- Model each location (March-May 2012)
- Consider exogenous variables (Temperature, Pressure, Humidity)
- Use `fable::model()` (similar to `astsa::sarima()`)[1]
- Finds best P, D, Q, p, d, and q values for each ARIMA formula[2]
- Selects the model with the lowest BIC for each set of exogenous variables (returns 225 models, 8 minutes)

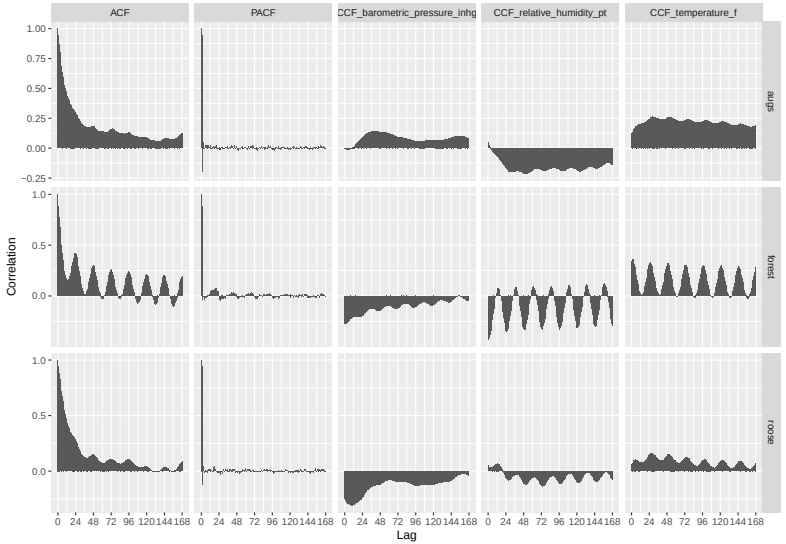
$$ws_t = \alpha + \sum_{n=1}^{p} \phi_n ws_{t-n} + \sum_{n=1}^{q} \theta_n ws_{t-n} + \sum_{n=1}^{P} \Phi_n ws_{t-sn} + \sum_{n=1}^{Q} \Theta_n ws_{t-sn} +$$

$$\beta_a P_{t-a} + \gamma_c T_{t-c} + \mu_e H_{t-e} + \epsilon_t$$

[1] Hyndman-Khandakar algorithm: https://otexts.com/fpp3/arima-r.html

[2] This is fitting about 3.2 million models

Look at ACF, PACF, CCFs



Fit one Seasonal ARIMA Model

Augspurger: ARIMA(3,0,0)(1,0,0)[24]

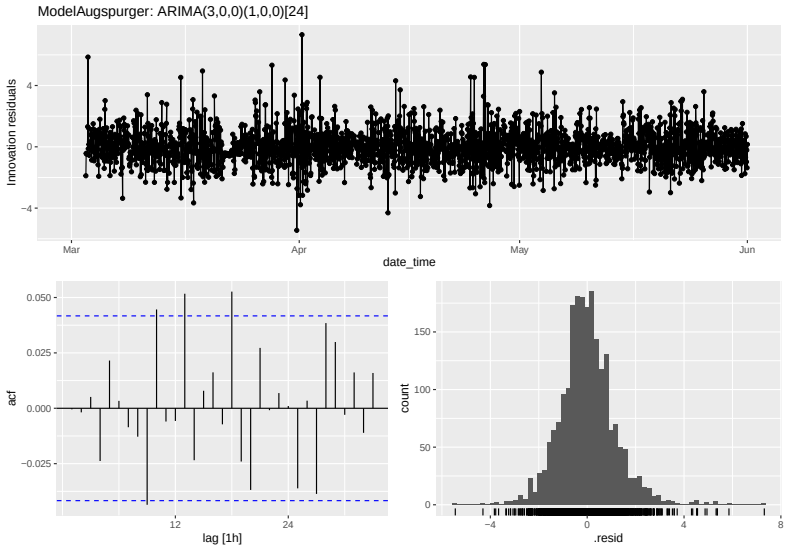
$$ws_t = 33.9 + 1.1ws_{t-1} - .3ws_{t-2} + .07ws_{t-3} - .04ws_{t-24} - 1.09P_{t-9} + .06H_{t-47} + .06T_{t-27}$$

Another Look

Best ARMIA Model for Forest According to BIC
ARIMA(1,0,0)(1,0,0)[24]

Term	Estimate	P Value
ar1	0.84	0.00
sar1	0.08	0.00
lag(barometric_pressure_inhg, 1)	−1.92	0.00
lag(relative_humidity_pt, 1)	−0.01	0.00
lag(temperature_f, 37)	−0.01	0.01
intercept	60.94	0.00

Model Diagnostics



Compare ARIMA to Other Model Types

Vector Autoregressive Models: $VAR(1)$

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

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

Error-Trend-Seasonality: ETS

ETS models estimate the following:
Error/Trend/Seasonality: Additive, Multiplicative, or None
ETS does not support Exogenous regressors

Best VAR and ETS models by BIC

VAR

Location	BIC	Model
Augspurger	8,279	VAR(16)
Biddle Butte	8,315	VAR(16)
Chinook	13,321	VAR(27)
Forest Grove	8,094	VAR(26)
Hood River	9,515	VAR(17)
Horse Heaven	14,502	VAR(26)
Megler	9,161	VAR(6)
Mt. Hebo	9,115	VAR(7)
Naselle Ridge	10,356	VAR(18)
Roosevelt	12,607	VAR(26)
Seven Mile	12,655	VAR(27)

ETS

Location	BIC	Model
Augspurger	17,783	<ETS(A,N,N)>
Biddle Butte	16,972	<ETS(A,N,N)>
Chinook	18,268	<ETS(A,N,N)>
Forest Grove	14,646	<ETS(A,N,A)>
Hood River	16,215	<ETS(A,N,A)>
Horse Heaven	18,327	<ETS(A,N,N)>
Megler	17,789	<ETS(A,N,N)>
Mt. Hebo	15,843	<ETS(A,N,N)>
Naselle Ridge	17,533	<ETS(A,N,N)>
Roosevelt	18,684	<ETS(A,A,A)>
Seven Mile	18,447	<ETS(A,N,A)>

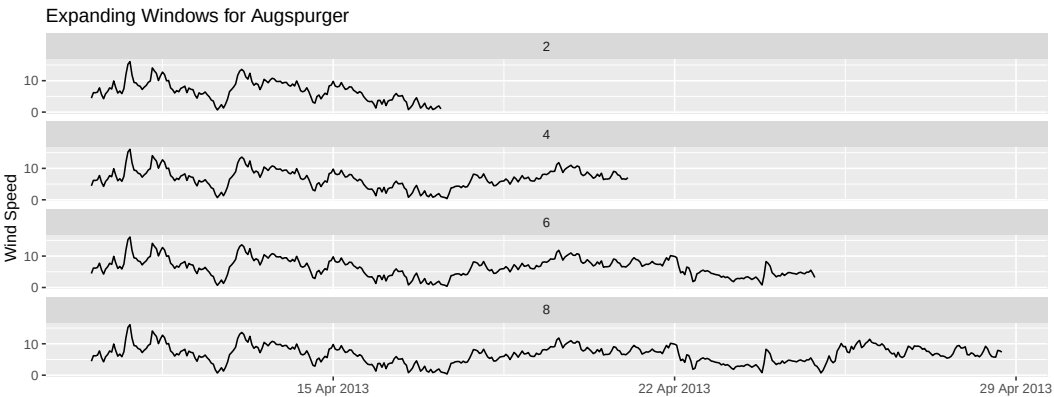
Choose a Final Model Based on BIC

Location	ARIMA	ETS	VAR
Forest Grove	**3,536**	14,646	8,094
Tillamook	**4,632**	15,715	10,482
Mt. Hebo	**4,949**	15,843	9,115
Hood River	**5,156**	16,215	9,515
Troutdale	**5,714**	16,833	10,209
Biddle Butte	**6,019**	16,972	8,315
Naselle Ridge	**6,623**	17,533	10,356
Shaniko	**6,733**	17,739	14,995
Augspurger	**6,776**	17,783	8,279
Megler	**6,852**	17,789	9,161
Horse Heaven	**7,296**	18,327	14,502
Chinook	**7,323**	18,268	13,321

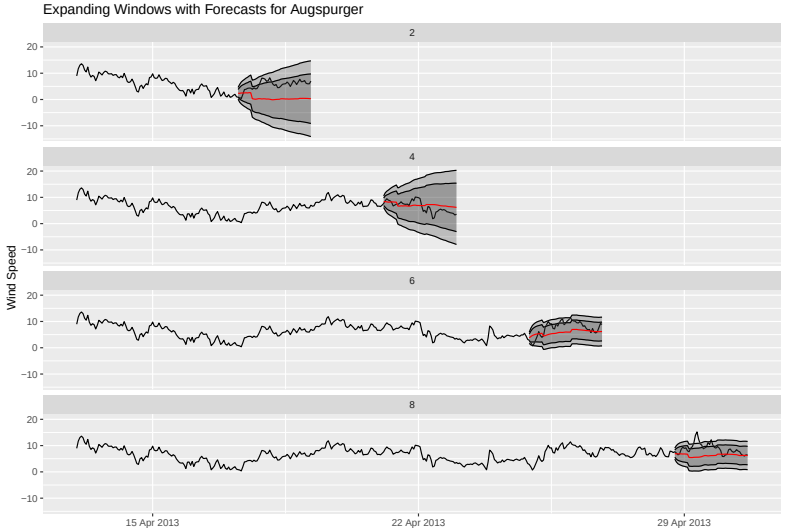
Forecast

Reminder
- Training: 1 March - 31 May 2012
- Testing: 1 March - 31 May 2013

Principals
- 48 hour expanding window
- Offset by 25 hours
- 15 locations and 25 models: $>$ 20 minutes



Make Predictions for Each Location and Each Window

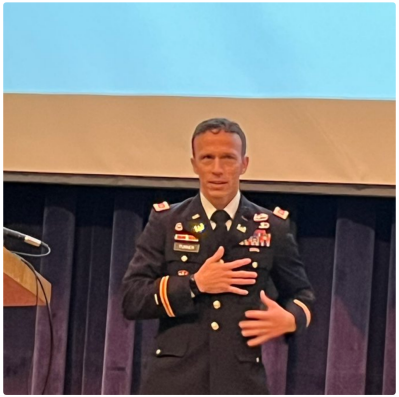


Evaluate Forecasts

Questions?

R Conference
@rstatsai

Want to win your next fantasy football auction draft? Use internet programming and r as described by @DTDusty at #rstatsgov #RStats

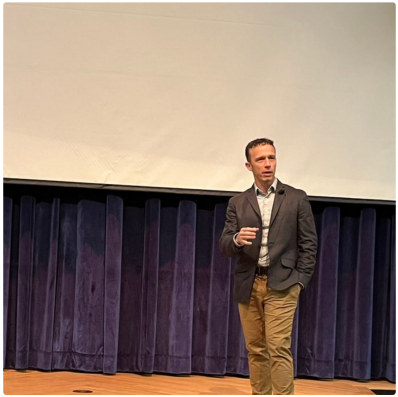


♡ 2 3:10 PM - Dec 1, 2022

See R Conference's other Tweets

R Conference
@rstatsai

Introducing the esteemed @jaredlander, we have @DTDusty #rstats #rstatsgov



♡ 2 11:14 AM - Dec 1, 2022

See R Conference's other Tweets