

➔ Road map!

- Module 1- Introduction to Deep Forecasting
- Module 2- Setting up Deep Forecasting Environment
- **Module 3- Exponential Smoothing**
- Module 4- ARIMA models
- Module 5- Machine Learning for Time series Forecasting
- Module 6- Deep Neural Networks
- Module 7- Deep Sequence Modeling (RNN, LSTM)
- Module 8- Transformers (Attention is all you need!)
- Module 9- Prophet and Neural Prophet





Simple Methods

- Recall:
- Naïve forecasts

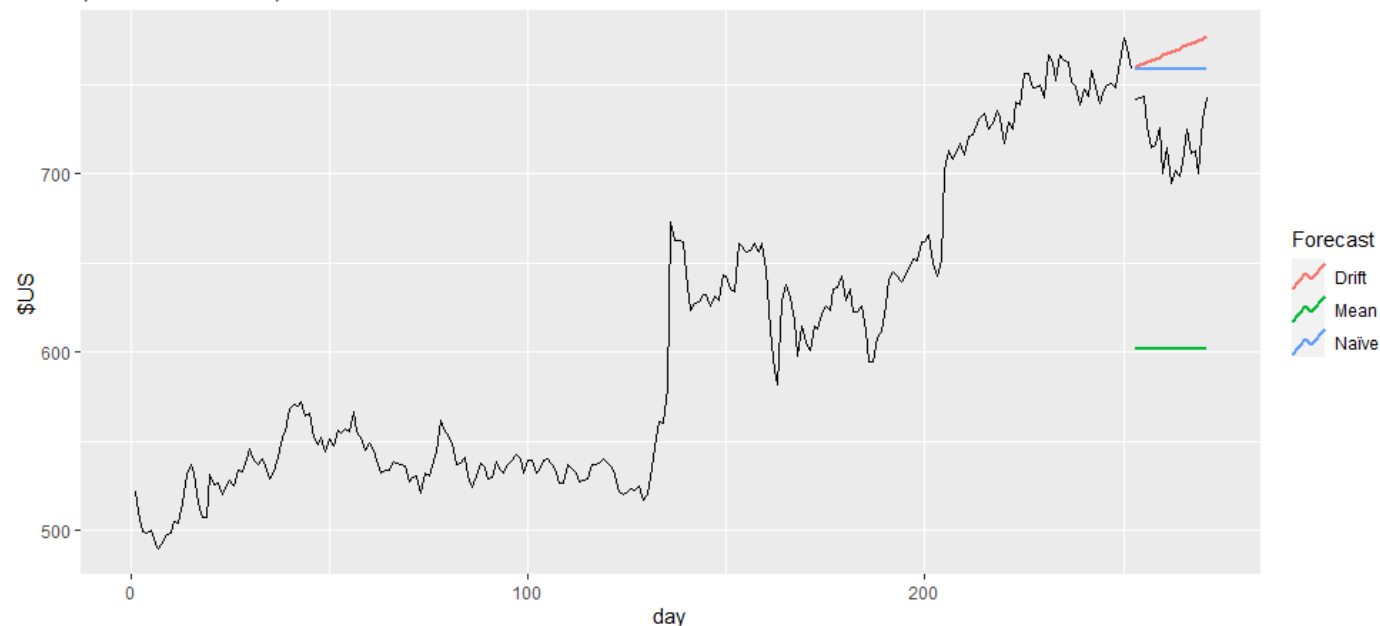
$$\hat{y}_{T+h|T} = Y_T$$

- Mean forecasts

$$\hat{y}_{T+h|T} = \frac{1}{T} \sum(Y_t)$$

- We want something in between
- Giving more weight to more recent data

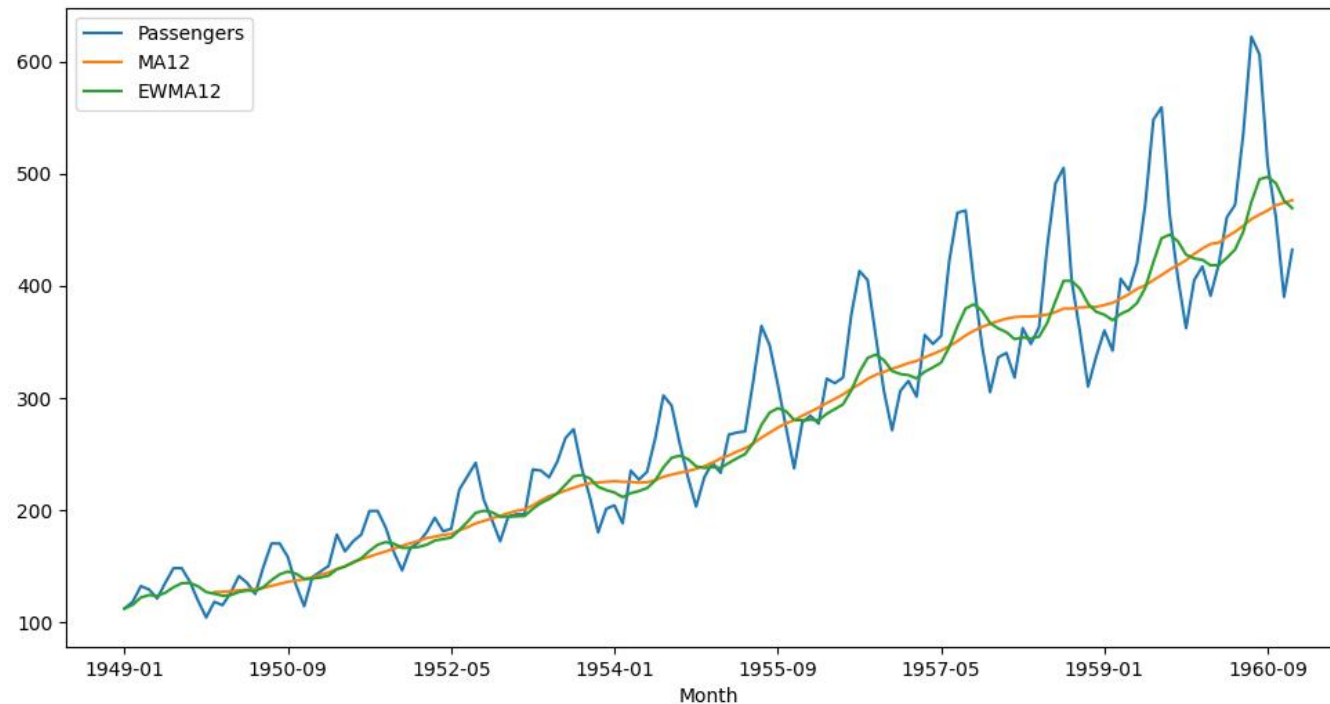
Google daily closing stock prices
(Jan 2015 - Jan 2016)





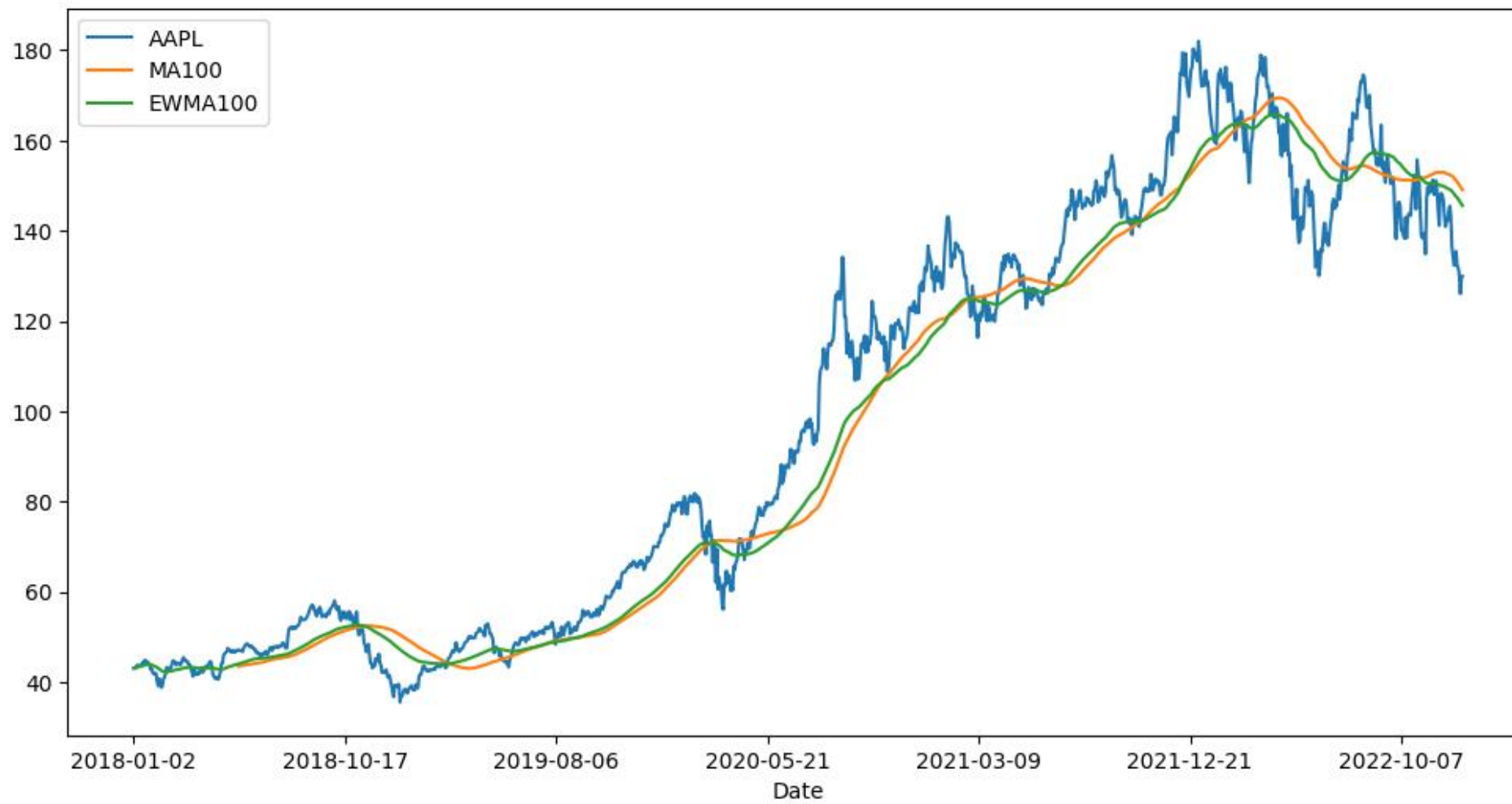
Simple and Exponential Weighted Moving Average

- Module 3 is all about **moving averages**!
- Simple Moving Average: **SMA**
- Exponentially Weighted Moving Average: **EWMA**
- EWMA will allow us to **reduce the lag effect from SMA** and it will put more weight on values that occurred more recently





SMA vs EWMA



→ Exponential Smoothing

- Exponential smoothing was proposed in the **late 1950s** (Brown, 1959; Holt, 1957; Winters, 1960), and has **motivated** some of the most successful forecasting methods
- A forecast generated by exponential smoothing uses weighted averages of past observations, with the **weights decaying** exponentially over time.
- In other words, the more recent the observation the higher the associated weight.
- In this module:
 - First, we present the **mechanics** of the most important exponential smoothing **methods**
 - Then, we present the **statistical models** that **underlie** exponential smoothing methods. These models generate **identical** point forecasts to the methods discussed in the first part of the chapter, but also generate prediction **intervals**.



Module 3- Part I

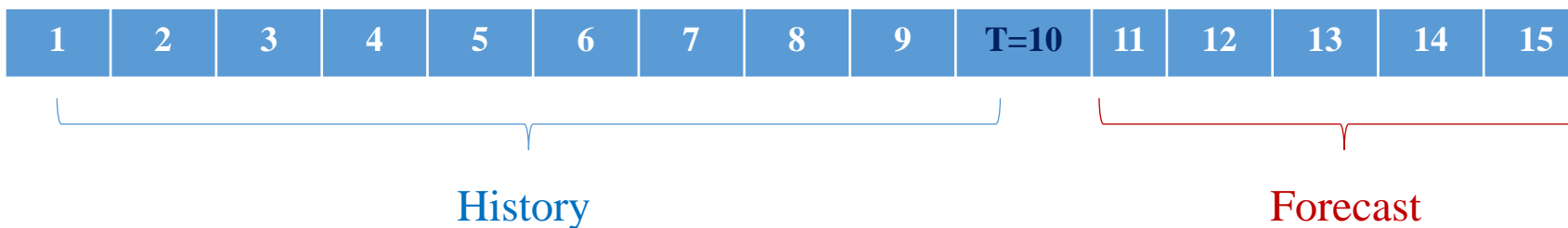
Exponential Smoothing (methods)

Method	Data Pattern	Forecast Equation
SES	No trend No seasonality	$\hat{y}_{t+h t} = l_t$
Holt's linear trend	Trend No seasonality	$\hat{y}_{t+h t} = l_t + hb_t$
Damped trend	Damped Trend No seasonality	$\hat{y}_{t+h t} = l_t + (\phi + \phi^2 + \dots + \phi^h)b_t$
Holt Winter	Trend Seasonality	$\hat{y}_{t+h t} = l_t + hb_t + s_{t+h-m(k+1)}$ $\hat{y}_{t+h t} = (l_t + hb_t) * s_{t+h-m(k+1)}$
Holt-Winter's Damped	Damped Trend Seasonality	$\hat{y}_{t+h t} = [l_t + (\phi + \phi^2 + \dots + \phi^h)] b_t * s_{t+h-m(k+1)}$

Forecasting notation

$$\hat{y}_{t+h|t} = f(y_t, h)$$

- y_t itself can be decomposed into different components (level, trend, seasonality)
- Fitted values at time $t = 1 \dots T$, are $\hat{y}_{t|t-1}$ ($h = 0$)
- One-step ahead forecast at time $T + 1$ (T last observation in train data) and $h = 1$.
- Multi-step ahead forecast: $h = 2, 3, 4, \dots$
 - One-output at a time
 - Multi-output at once





Simple Exponential Smoothing (SES)

- SES is suitable for forecasting data with **no clear trend or seasonal pattern**
- **Naïve** forecast can be thought of as a weighted average where all the weight is given to the last observation.

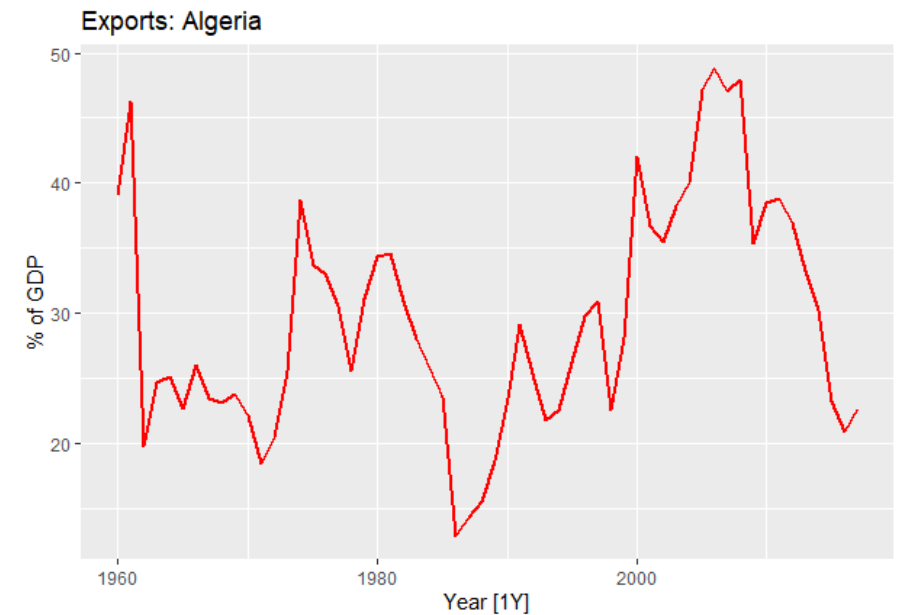
$$\hat{y}_{T+h|T} = y_T$$

- **Mean** forecast assumes all observations are of equal importance and give them same weights.

$$\hat{y}_{T+h|T} = \frac{1}{T} \sum_{t=1}^T y_t$$

- We want **something in between** these two extremes. ex, attach larger weights to more recent observations! This is **SES**.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1 - \alpha)y_{T-1} + \alpha(1 - \alpha)^2 y_{T-2} + \dots$$





SES weights

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1 - \alpha)y_{T-1} + \alpha(1 - \alpha)^2 y_{T-2} + \dots$$

- $0 < \alpha < 1$ is the **smoothing parameter**.
- For any α between 0 and 1, the weights attached to the observations **decrease exponentially** as we go back in time, hence the name “exponential smoothing”.
- Sum of the weights is approximately one.

	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
y_T	0.2000	0.4000	0.6000	0.8000
y_{T-1}	0.1600	0.2400	0.2400	0.1600
y_{T-2}	0.1280	0.1440	0.0960	0.0320
y_{T-3}	0.1024	0.0864	0.0384	0.0064
y_{T-4}	0.0819	0.0518	0.0154	0.0013
y_{T-5}	0.0655	0.0311	0.0061	0.0003

SES weights

→ Equivalent forms of SES

- There are two equivalent forms of SES:
 1. **Weighted Average Form**: the **forecast** is equal to weighted average between the most recent observation and the previous **forecast**.

$$\hat{y}_{T+1|T} = \alpha y_T + (1 - \alpha) \hat{y}_{T|T-1}$$

2. **Component Form**: For simple exponential smoothing, the only component included is the level, l_t

Forecast equation	$\hat{y}_{t+h t} = l_t$
Smoothing equation	$l_t = \alpha y_t + (1 - \alpha) l_{t-1}$



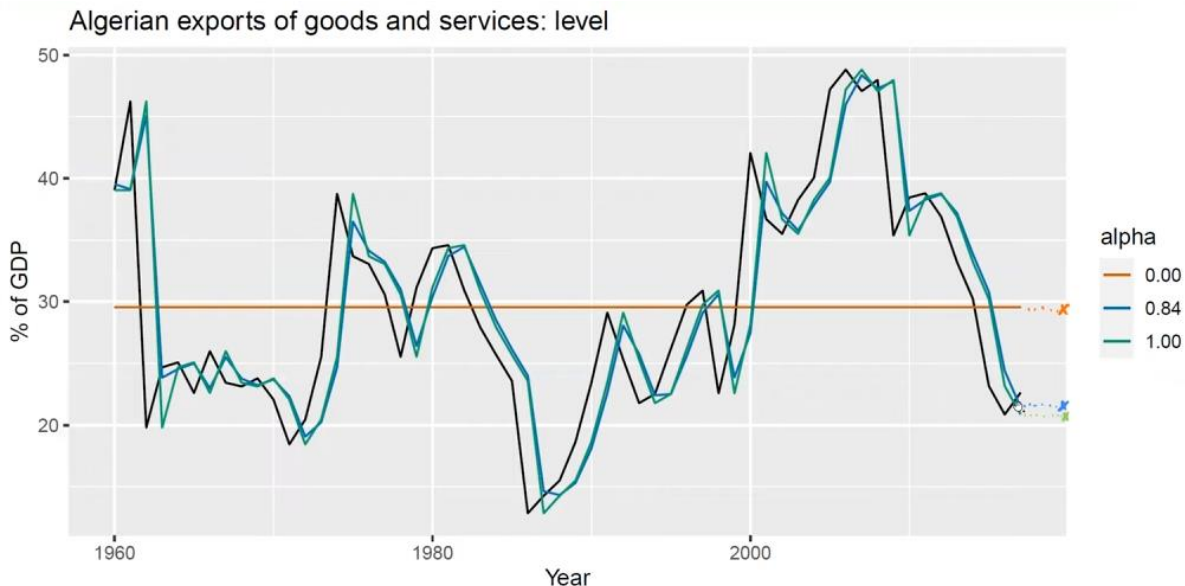
SES forecasts

- Simple exponential smoothing has a “**flat**” forecast function:

$$\hat{y}_{T+h|T} = \hat{y}_{T+1|T} = \ell_T, \quad h = 2, 3, \dots$$

- All forecasts take the same value, equal to the **last level component**.
- Remember that these forecasts will only be suitable if the time series has **no trend or seasonal component**.
- The parameters of SES model (alpha and level zero) can be optimized by minimizing SSE.

Year	Time	Observation	Level	Forecast
	t	y_t	ℓ_t	$\hat{y}_{t t-1}$
1959	0		39.54	
1960	1	39.04	39.12	39.54
1961	2	46.24	45.10	39.12
1962	3	19.79	23.84	45.10
2016	57	20.86	21.43	24.39
2017	58	22.64	22.44	21.43
	h			$\hat{y}_{T+h T}$
2018	1			22.44
2019	2			22.44
2020	3			22.44
2021	4			22.44
2022	5			22.44



Exponential Smoothing

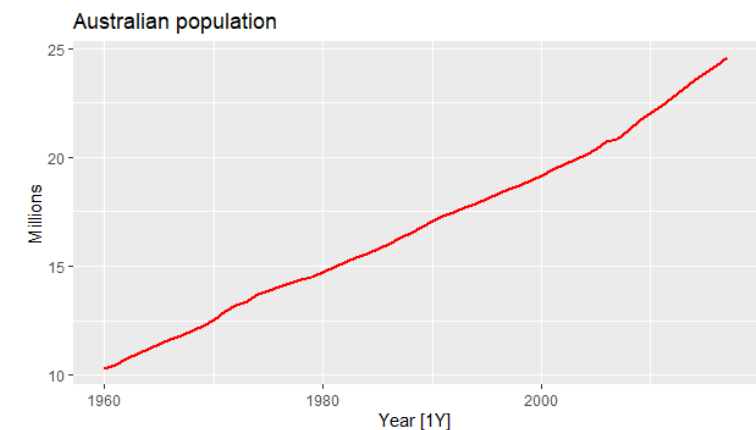
Methods with trend

→ Holt's linear trend method

- **Holt** (1957) extended simple exponential smoothing to allow the **forecasting of data with a trend**.
- This method is suitable for forecasting data **with clear trend but no seasonal pattern**
- This method involves a forecast equation and two smoothing equations:

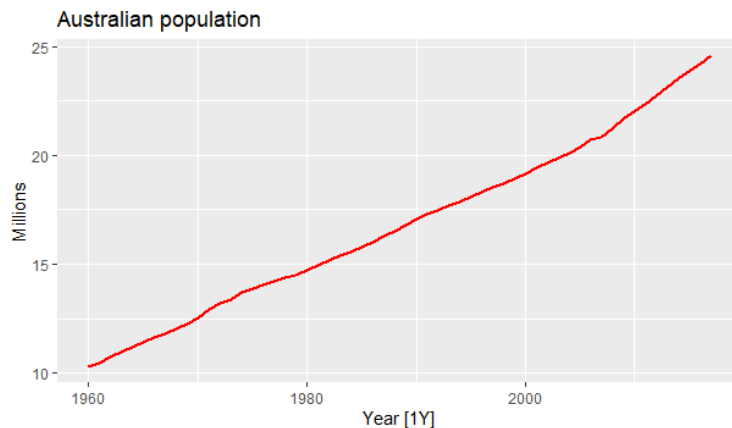
Forecast equation	$\hat{y}_{t+h t} = \ell_t + hb_t$
Level equation	$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$
Trend equation	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1},$

- ℓ_t denotes an estimate of the **level** of the series at time t
- b_t denotes an estimate of the **trend** (slope) of the series at time t
- α and β^* are the smoothing parameters for level and trend.



➔ Holt's linear trend forecasts

- The forecast function is **no longer flat but trending**. $\hat{y}_{t+h|t} = \ell_t + hb_t$
- The h -step-ahead forecast is equal to the **last estimated level** + h times **the last estimated trend value**. Hence the forecasts are a linear function of h
- The smoothing parameters, and the initial values are estimated by minimizing the SSE for the one-step training errors



Year	Time	Observation	Level	Slope	Forecast
	t	y_t	ℓ_t		$\hat{y}_{t+1 t}$
1959	0		10.05	0.22	
1960	1	10.28	10.28	0.22	10.28
1961	2	10.48	10.48	0.22	10.50
1962	3	10.74	10.74	0.23	10.70
2016	57	24.21	24.21	0.36	24.21
2017	58	24.60	24.60	0.37	24.57
	h				$\hat{y}_{T+h T}$
2018	1				24.97
2019	2		$\hat{y}_{t+h t} = \ell_t + hb_t$ $\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$		25.34
2020	3				25.71
2021	4				26.07
2022	5				26.44

➔ Damped trend methods

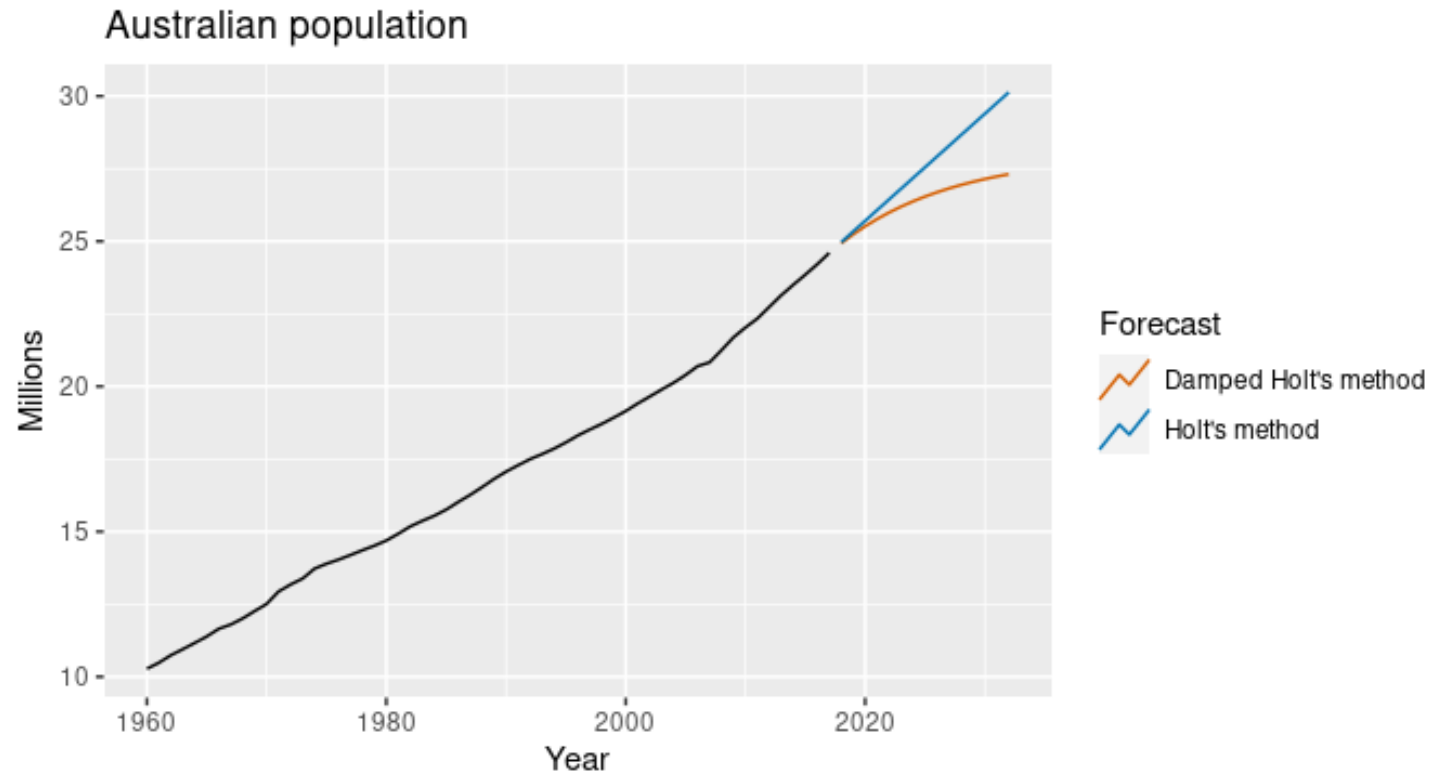
- The forecasts generated by Holt's linear method display a constant trend (increasing or decreasing) **indefinitely into the future**
- Empirical evidence indicates that these methods tend to **over-forecast**, especially for **longer forecast horizons**.
- Gardner & McKenzie (1985) introduced a parameter that “**dampens**” the trend to a flat line some time in the future (adding a damping parameter $0 < \phi < 1$)
- ϕ dampens the trend so that it approaches a constant some time in the future. This means that **short-run forecasts are trended, while long-run forecasts are constant**.

Forecast equation	$\hat{y}_{t+h t} = \ell_t + hb_t$
Level equation	$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$
Trend equation	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1},$

$\hat{y}_{t+h t} = \ell_t + (\phi + \phi^2 + \dots + \phi^h)b_t$
$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})$
$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)\phi b_{t-1}$

➔ Methods with trend, Example

- Forecasting annual Australian population (millions) over 2018-2032. For the damped trend method, $\phi = 0.90$



Exponential Smoothing

Methods with trend and seasonality

→ Holt-Winters method

- Holt (1957) and Winters (1960) extended Holt's method to capture seasonality.
- This method comprises the forecast equation and three smoothing equations:
 1. l_t for the level component with corresponding smoothing parameter α
 2. b_t for the trend component with corresponding smoothing parameter β^*
 3. s_t for the seasonal component with corresponding smoothing parameter γ
- There are two variations to this method that differ in the nature of the seasonal component.
 1. The additive method: when the seasonal variations are roughly constant through the series
 2. the multiplicative method: when the seasonal variations are changing proportional to the level of the series.

Holt-Winters' additive vs multiplicative methods

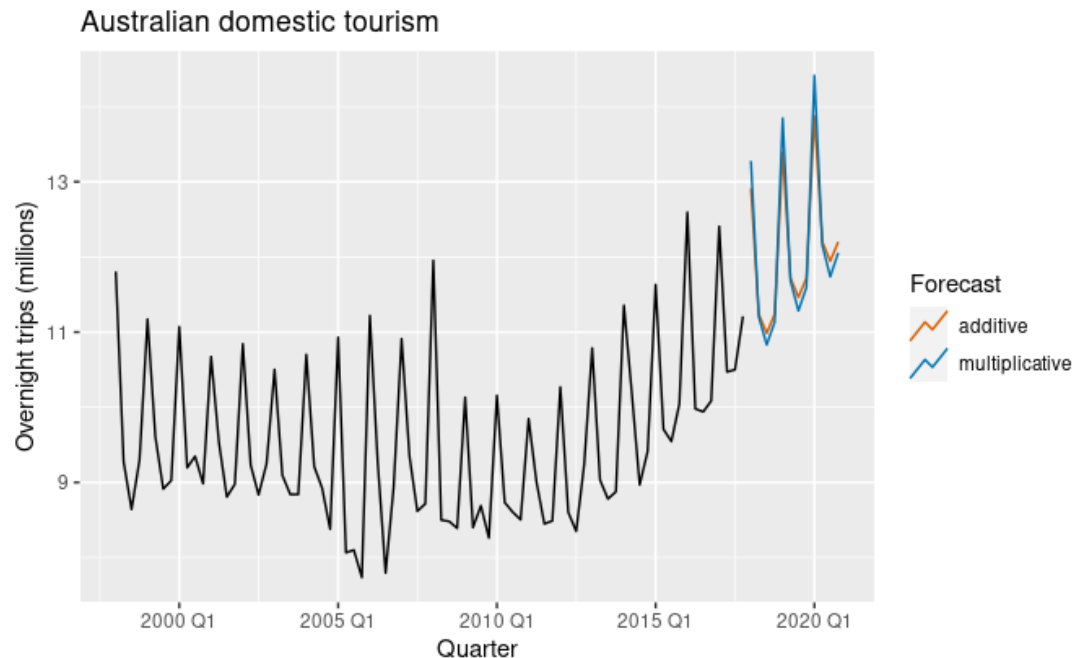
- m to denote the **period** of the seasonality. For quarterly data $m = 4$ and monthly, $m = 12$.

$$\begin{aligned}\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(k+1)} \\ \ell_t &= \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1} \\ s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m},\end{aligned}$$

$$\begin{aligned}\hat{y}_{t+h|t} &= (\ell_t + hb_t)s_{t+h-m(k+1)} \\ \ell_t &= \alpha \frac{y_t}{s_{t-m}} + (1 - \alpha)(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1} \\ s_t &= \gamma \frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1 - \gamma)s_{t-m}.\end{aligned}$$

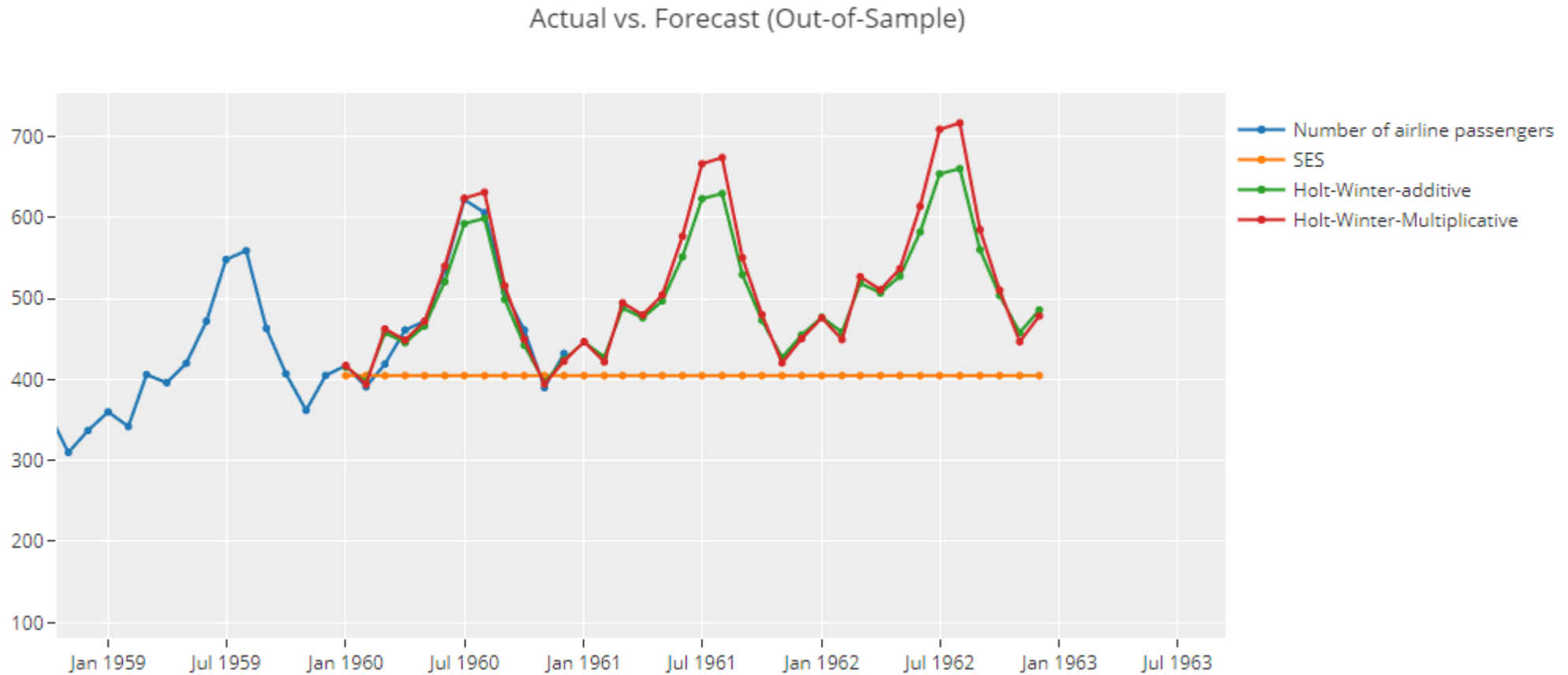
m : seasonality period

$$k = \text{int}\left(\frac{h-1}{m}\right)$$





Holt-Winters' additive vs multiplicative vs SES



→ Holt-Winters' damped methods

- Damping is possible with both additive and multiplicative Holt-Winters' methods.
- A method that often provides accurate and robust forecasts for seasonal data is the Holt-Winters method with a **damped trend** and **multiplicative seasonality**:

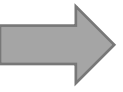
$$\begin{aligned}\hat{y}_{t+h|t} &= [\ell_t + (\phi + \phi^2 + \dots + \phi^h)b_t] s_{t+h-m(k+1)} \\ \ell_t &= \alpha(y_t/s_{t-m}) + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)\phi b_{t-1} \\ s_t &= \gamma \frac{y_t}{(\ell_{t-1} + \phi b_{t-1})} + (1 - \gamma)s_{t-m}.\end{aligned}$$



Summary

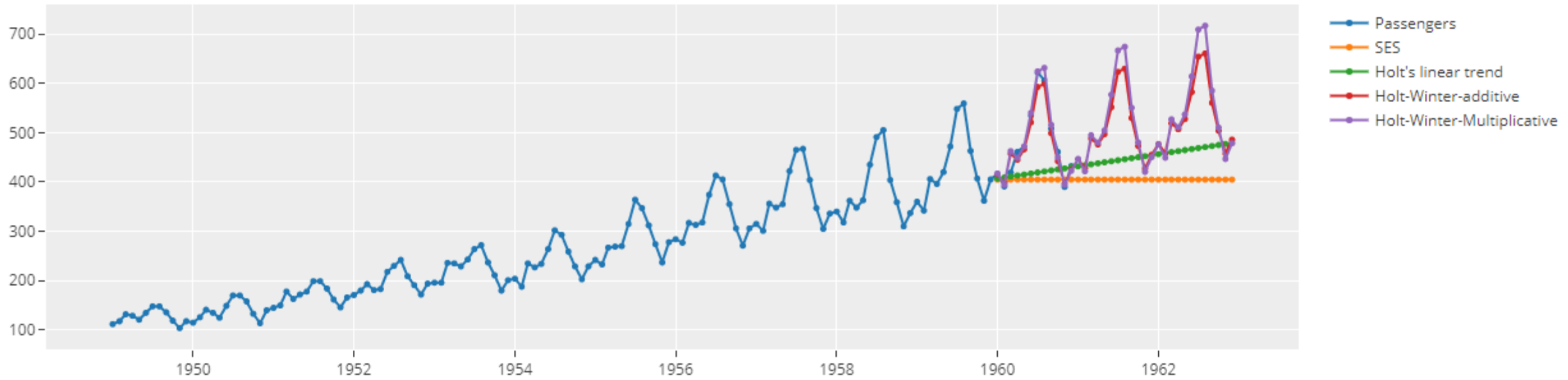
Method	Data Pattern		Forecast Equation
SES	No trend	No seasonality	$\hat{y}_{t+h t} = l_t$
Holt's linear trend	Trend	No seasonality	$\hat{y}_{t+h t} = l_t + hb_t$
Damped trend	Damped Trend	No seasonality	$\hat{y}_{t+h t} = l_t + (\phi + \phi^2 + \dots + \phi^h)b_t$
Holt Winter	Trend	Seasonality	$\hat{y}_{t+h t} = l_t + hb_t + s_{t+h-m(k+1)}$ $\hat{y}_{t+h t} = (l_t + hb_t) * s_{t+h-m(k+1)}$
Holt-Winter's Damped	Damped Trend	Seasonality	$\hat{y}_{t+h t} = [l_t + (\phi + \phi^2 + \dots + \phi^h)] b_t * s_{t+h-m(k+1)}$

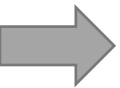
- we study the statistical models that underlie the exponential smoothing methods we have considered so far



Summary

Actual vs. Forecast (Out-of-Sample)





A taxonomy of exponential smoothing methods

Table 8.5: A two-way classification of exponential smoothing methods.

Trend Component	Seasonal Component		
	N	A	M
	(None)	(Additive)	(Multiplicative)
N (None)	(N,N)	(N,A)	(N,M)
A (Additive)	(A,N)	(A,A)	(A,M)
A_d (Additive damped)	(A_d ,N)	(A_d ,A)	(A_d ,M)

Some of these methods we have already seen using other names:

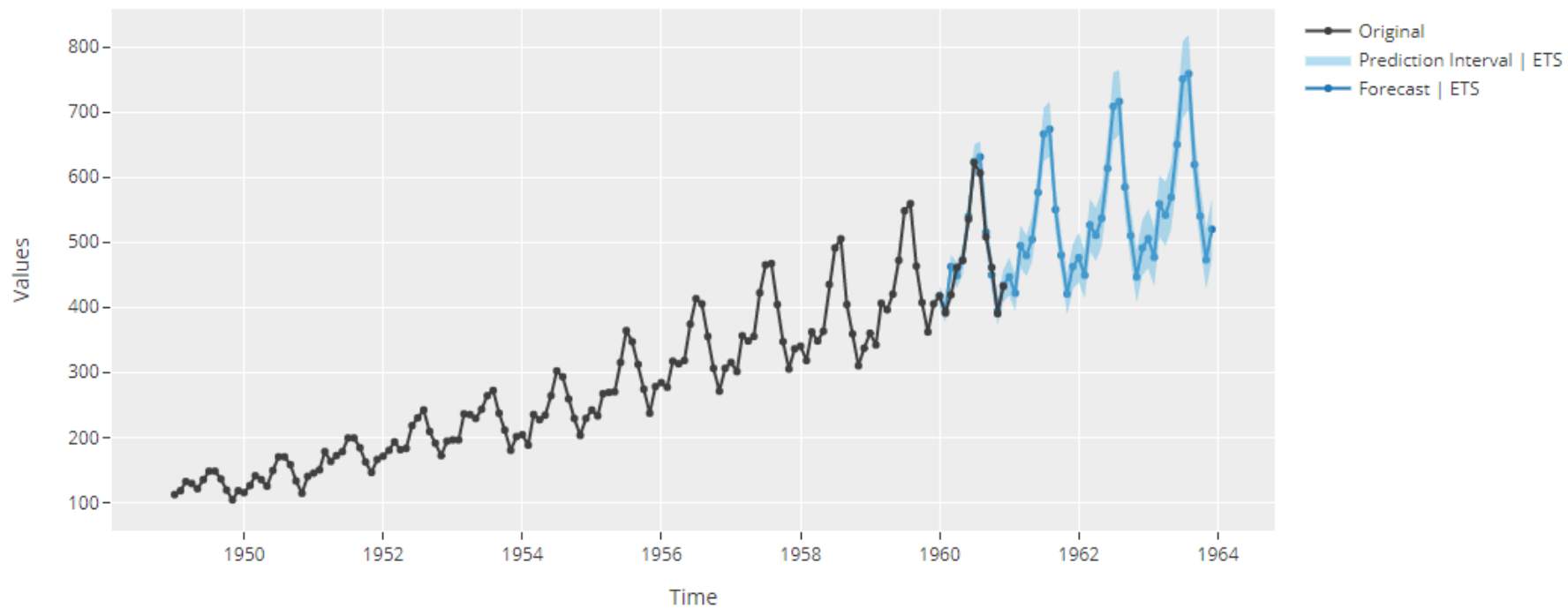
Short hand	Method
(N,N)	
(A,N)	
(A_d ,N)	
(A,A)	
(A,M)	
(A_d ,M)	

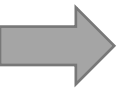


Module 3- Part II

Exponential Smoothing-based models

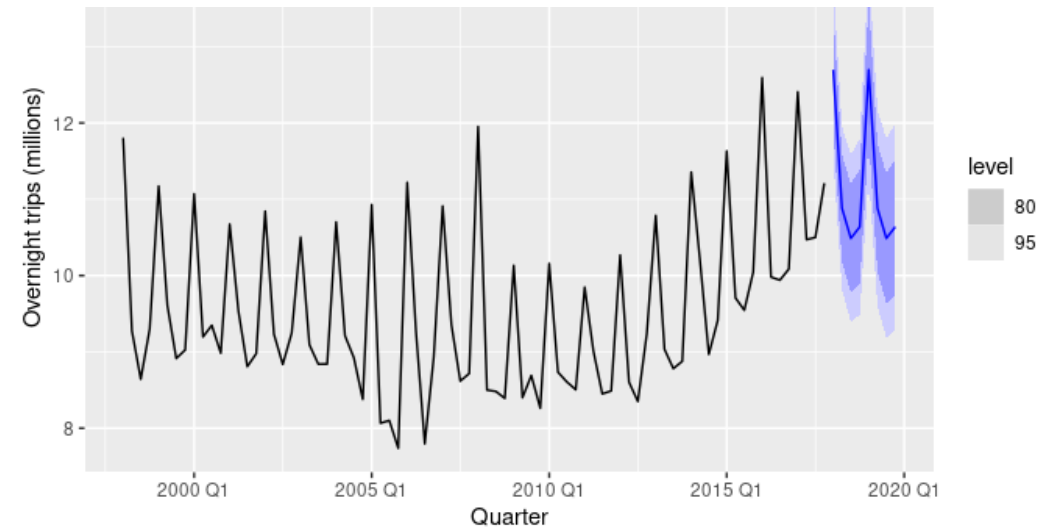
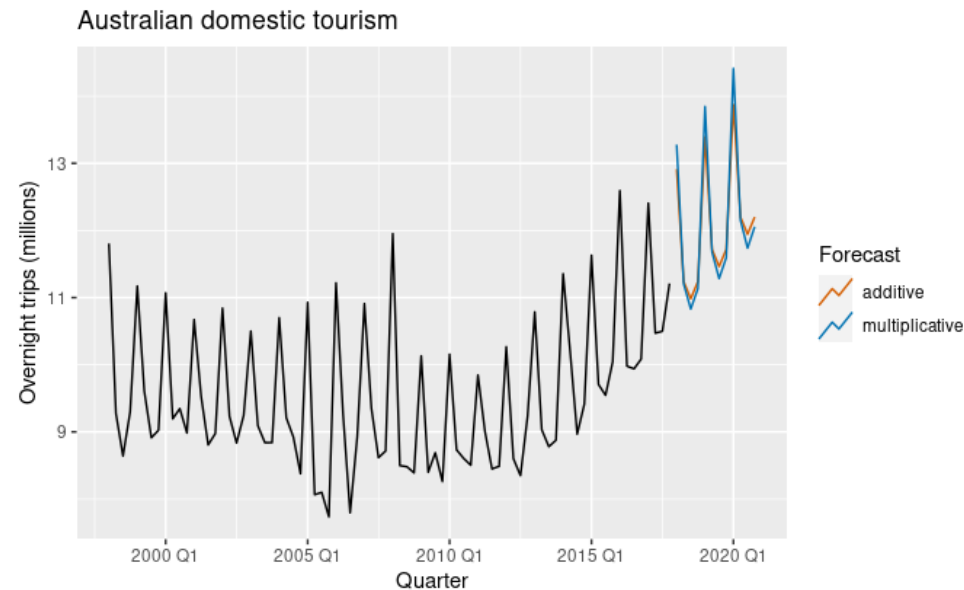
ETS





Exponential Smoothing and point forecasts

- The exponential smoothing **methods** are algorithms which generate **point** forecasts.
- The statistical **models** generate the **same point forecasts** but can also generate prediction (or forecast) **intervals**. i.e., producing the entire forecast distribution.





State Space Models for Exponential Smoothing

- These models are referred to as **state space** models because they describe how the unobserved components or **states** (level, trend, seasonal) change over time.
- The **error** part is the new part which enable us to produce **forecast distribution**.
- We label each state space model as ETS (. , . , .)!
- **ETS** stands for **E**rror, **T**rend, **S**easonality! Also thought of as **E**xponen**T**ial **S**moothing.



State Space Models for Exponential Smoothing

- For each **method** there exist two **models**: one with additive errors and one with multiplicative errors.
- Models with **multiplicative errors** are useful when the data are strictly positive but are not numerically stable when the data contain zeros or negative values.
- State possibilities notation:

Error= {A,M}

Trend={N, A, A_d }

Seasonal={N, A, M}

- We can write $2*3*3=18$ different state space model for each of the exponential smoothing methods.



ETS (A, N, N): Simple Exponential Smoothing with additive errors

- Recall SES components :

$$\begin{array}{ll}\text{Forecast equation} & \hat{y}_{t+h|t} = \ell_t \\ \text{Smoothing equation} & \ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1}\end{array}$$

- Re-arrange the smoothing equation for the level and get the **error correction** from.

$$\begin{aligned}\ell_t &= \ell_{t-1} + \alpha(y_t - \ell_{t-1}) \\ &= \ell_{t-1} + \alpha e_t,\end{aligned}$$

- Where $e_t = y_t - \ell_{t-1} = y_t - \hat{y}_{t|t-1}$ is the residual at time t. Remember, **level** is what the ETS(A,N,N) model predicts.
- If the model is over/under shooting, the level will **adjust** in the next period. The **magnitude** of adjustment depends on α . Smaller α means smoother adjustment.
- We can also write $y_t = \ell_{t-1} + e_t$, so **each observation = previous level + error**



ETS (A, N, N): Simple Exponential Smoothing with additive errors

- So far, we showed that $y_t = \ell_{t-1} + e_t$
- The only thing left out is to specify the probability distribution for errors. With that, we have our first **innovations state space model**.
- For a model with additive errors, we assume $e_t = \varepsilon_t \sim \text{NID}(0, \sigma^2)$, errors are normally and independently distributed. (**white noise**)
- The final model can be written as:

Measurement (observation) equation

$$y_t = \ell_{t-1} + \varepsilon_t$$

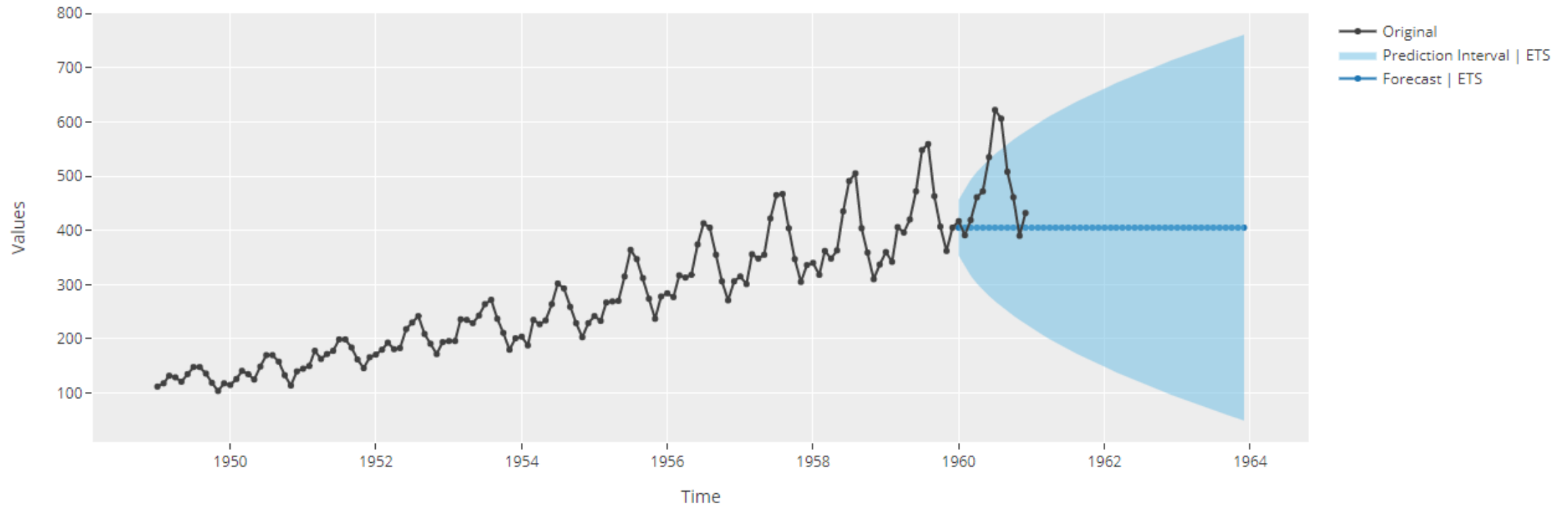
State (transition) equation

$$\ell_t = \ell_{t-1} + \alpha \varepsilon_t$$



ETS (A, N, N): Simple Exponential Smoothing with additive errors

Actual vs. 'Out-of-Sample' Forecast | Passengers





ETS(A,A,N): Holt's linear method with additive errors

- In this model, the training errors are given by $\varepsilon_t = y_t - \ell_{t-1} - b_{t-1} \sim \text{NID}(0, \sigma^2)$
- Substituting this into the level equation and trend equation, we get:

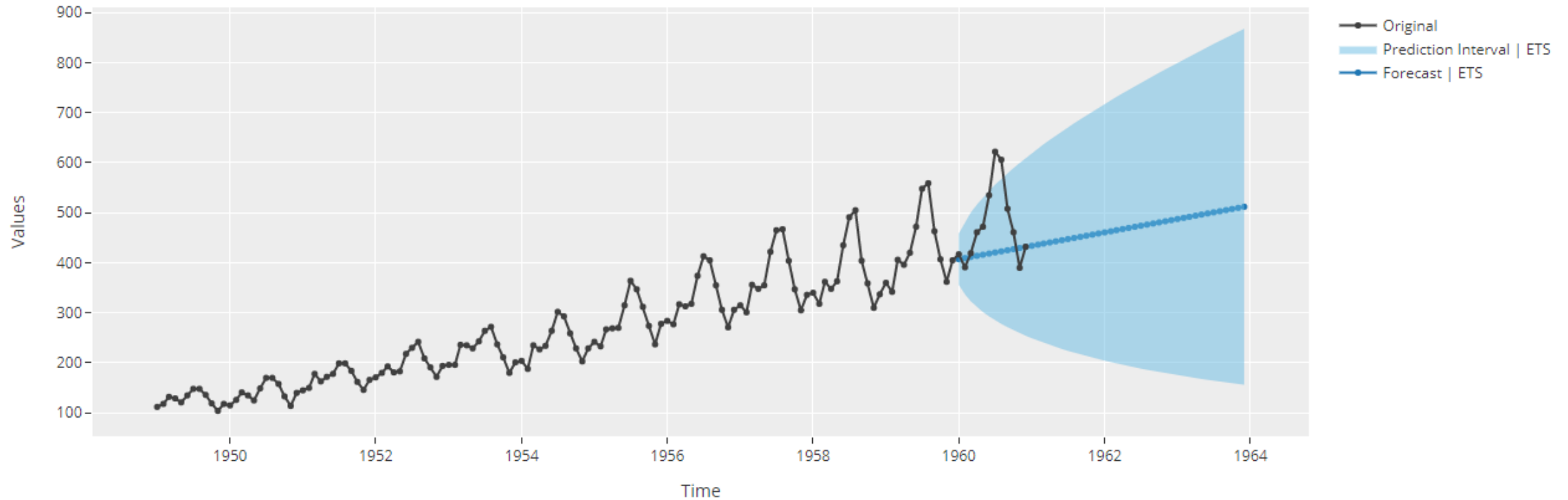
$$\begin{aligned}y_t &= \ell_{t-1} + b_{t-1} + \varepsilon_t \\ \ell_t &= \ell_{t-1} + b_{t-1} + \alpha \varepsilon_t \\ b_t &= b_{t-1} + \beta \varepsilon_t,\end{aligned}$$

- Note that $\beta = \alpha \beta^*$ where α and β^* are the smoothing parameters for the level and trend components, respectively.



ETS (A, A, N): Simple Exponential Smoothing with additive errors

Actual vs. 'Out-of-Sample' Forecast | Passengers





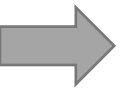
Other ETS models

- Recall: Error= {A,M}, Trend={N, A, A_d}, Seasonal={N, A, M}
- We can write $2*3*3=18$ different state space model for each of the exponential smoothing methods.

ADDITIVE ERROR MODELS

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha \varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$y_t = \ell_{t-1} s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / \ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha \varepsilon_t$ $b_t = b_{t-1} + \beta \varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha \varepsilon_t$ $b_t = b_{t-1} + \beta \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1}) s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $b_t = b_{t-1} + \beta \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / (\ell_{t-1} + b_{t-1})$
A _d	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t$ $b_t = \phi b_{t-1} + \beta \varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t$ $b_t = \phi b_{t-1} + \beta \varepsilon_t$ $s_t = s_{t-m} + \gamma \varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1}) s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t / s_{t-m}$ $b_t = \phi b_{t-1} + \beta \varepsilon_t / s_{t-m}$ $s_t = s_{t-m} + \gamma \varepsilon_t / (\ell_{t-1} + \phi b_{t-1})$





Other ETS models

- Recall: Error= {A,M}, Trend={N, A, A_d}, Seasonal={N, A, M}
- We can write $2*3*3=18$ different state space model for each of the exponential smoothing methods.

MULTIPLICATIVE ERROR MODELS

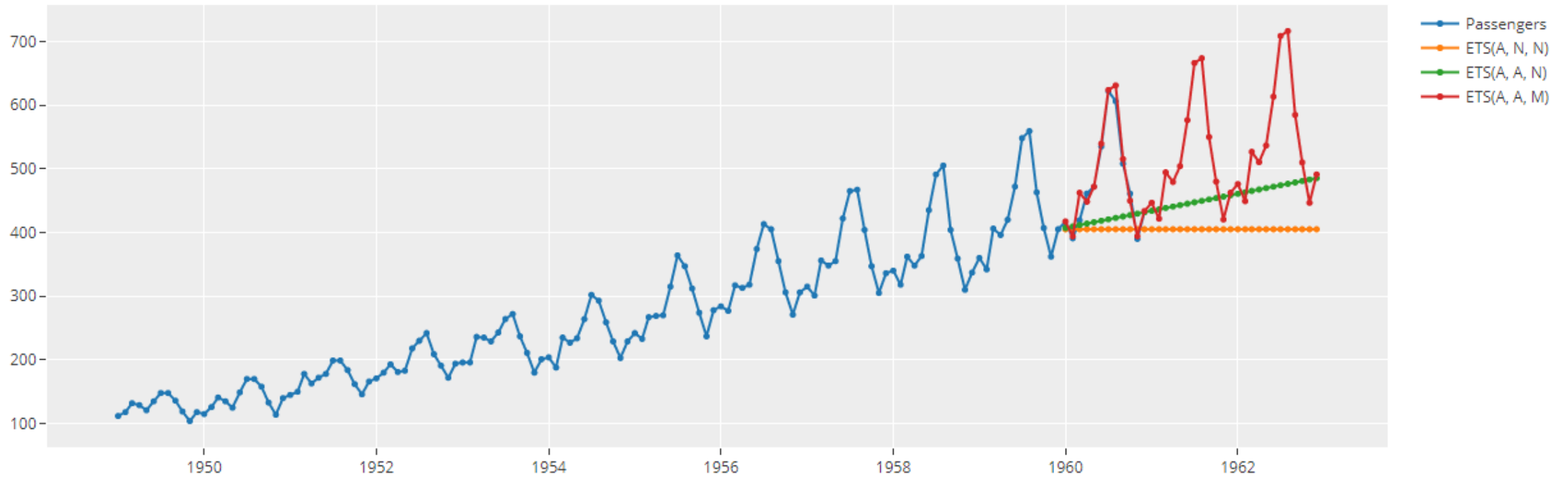
Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A _d	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$





ETS models

Actual vs. Forecast (Out-of-Sample)



→ ETS model estimation

- **Maximum Likelihood Estimation** is used to optimize the smoothing parameters and the initial values for level, trend and seasonal components.
- The smoothing parameters are restricted to be between 0 and 1. $0 < \alpha, \beta^*, \gamma^*, \phi < 1$ so that the equations can be **interpreted as weighted averages**.
- The parameters are constrained in order to **prevent** observations in the distant past having a continuing effect on current forecasts.
- Reminder: **Maximum likelihood estimation (MLE)** is a method used to estimate the parameters of a statistical model, given observations. It involves finding the parameter values that maximize the likelihood function, which is a function that describes the probability of a set of observations given the parameter values.

→ ETS model selection

- For model selection we can either use **information criteria** or any **cross validated** performance metrics like R^2 , MSE, RMSE, MAPE, sMAPE.

Information Criteria	Formula
Akaike's Information Criterion (AIC)	$AIC = -2 \log(L) + 2k$
AIC corrected for small sample bias (AICc)	$AIC_c = AIC + \frac{2k(k+1)}{T-k-1}$
Bayesian Information Criterion (BIC)	$BIC = AIC + k[\log(T) - 2]$

- **L** is the likelihood of the model and **K** is the total number of parameters and initial states that have been estimated (including the residual variance)
- The model with the **minimum information criteria** is often the best model for forecasting

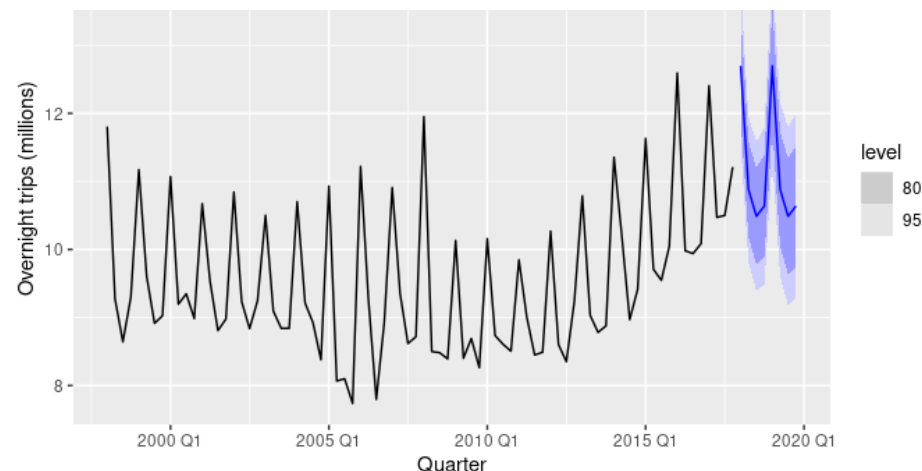
Forecasting with ETS models

- **Point forecasts** can be obtained from the models by iterating the equations for the forecasting horizon. $t = T + 1, \dots, T + h$
- Setting all $\epsilon_t = 0$ for $t > T$
- These point forecasts are identical to the forecasts from the exponential smoothing **methods**.
- Prediction intervals: for most ETS models, a **prediction interval** can be written as:

Critical values

$$\hat{y}_{T+h|T} \pm c\sigma_h$$

Forecast standard error



➔ Forecast variance: σ_h^2

Table 8.8: Forecast variance expressions for each additive state space model, where σ^2 is the residual variance, m is the seasonal period, and k is the integer part of $(h - 1)/m$ (i.e., the number of complete years in the forecast period prior to time $T + h$).

Model	Forecast variance: σ_h^2
(A,N,N)	$\sigma_h^2 = \sigma^2 [1 + \alpha^2(h - 1)]$
(A,A,N)	$\sigma_h^2 = \sigma^2 \left[1 + (h - 1) \left\{ \alpha^2 + \alpha\beta h + \frac{1}{6}\beta^2 h(2h - 1) \right\} \right]$
(A,A _d ,N)	$\sigma_h^2 = \sigma^2 \left[1 + \alpha^2(h - 1) + \frac{\beta\phi h}{(1-\phi)^2} \{2\alpha(1 - \phi) + \beta\phi\} \right. \\ \left. - \frac{\beta\phi(1-\phi^h)}{(1-\phi)^2(1-\phi^2)} \{2\alpha(1 - \phi^2) + \beta\phi(1 + 2\phi - \phi^h)\} \right]$
(A,N,A)	$\sigma_h^2 = \sigma^2 [1 + \alpha^2(h - 1) + \gamma k(2\alpha + \gamma)]$
(A,A,A)	$\sigma_h^2 = \sigma^2 \left[1 + (h - 1) \left\{ \alpha^2 + \alpha\beta h + \frac{1}{6}\beta^2 h(2h - 1) \right\} \right. \\ \left. + \gamma k \{2\alpha + \gamma + \beta m(k + 1)\} \right]$
(A,A _d ,A)	$\sigma_h^2 = \sigma^2 \left[1 + \alpha^2(h - 1) + \gamma k(2\alpha + \gamma) \right. \\ \left. + \frac{\beta\phi h}{(1-\phi)^2} \{2\alpha(1 - \phi) + \beta\phi\} \right. \\ \left. - \frac{\beta\phi(1-\phi^h)}{(1-\phi)^2(1-\phi^2)} \{2\alpha(1 - \phi^2) + \beta\phi(1 + 2\phi - \phi^h)\} \right. \\ \left. + \frac{2\beta\gamma\phi}{(1-\phi)(1-\phi^m)} \{k(1 - \phi^m) - \phi^m(1 - \phi^{mk})\} \right]$

➔ Road map!

- ✓ Module 1- Introduction to Deep Forecasting
- ✓ Module 2- Setting up Deep Forecasting Environment
- ✓ Module 3- Exponential Smoothing
- Module 4- ARIMA models
- Module 5- Machine Learning for Time series Forecasting
- Module 6- Deep Neural Networks
- Module 7- Deep Sequence Modeling (RNN, LSTM)
- Module 8- Transformers (Attention is all you need!)
- Module 9- Prophet and Neural Prophet

