Time Series Forecasting.
1. Exponential Smoothin Models. Compare TSF Models

Alexey Romanenko
alexromsput@gmail.com

FIVT MIPT, September 2018

Содержание

Components of Time Series

Level — average level of values;
Trend — monotonic long-term changes of Level;

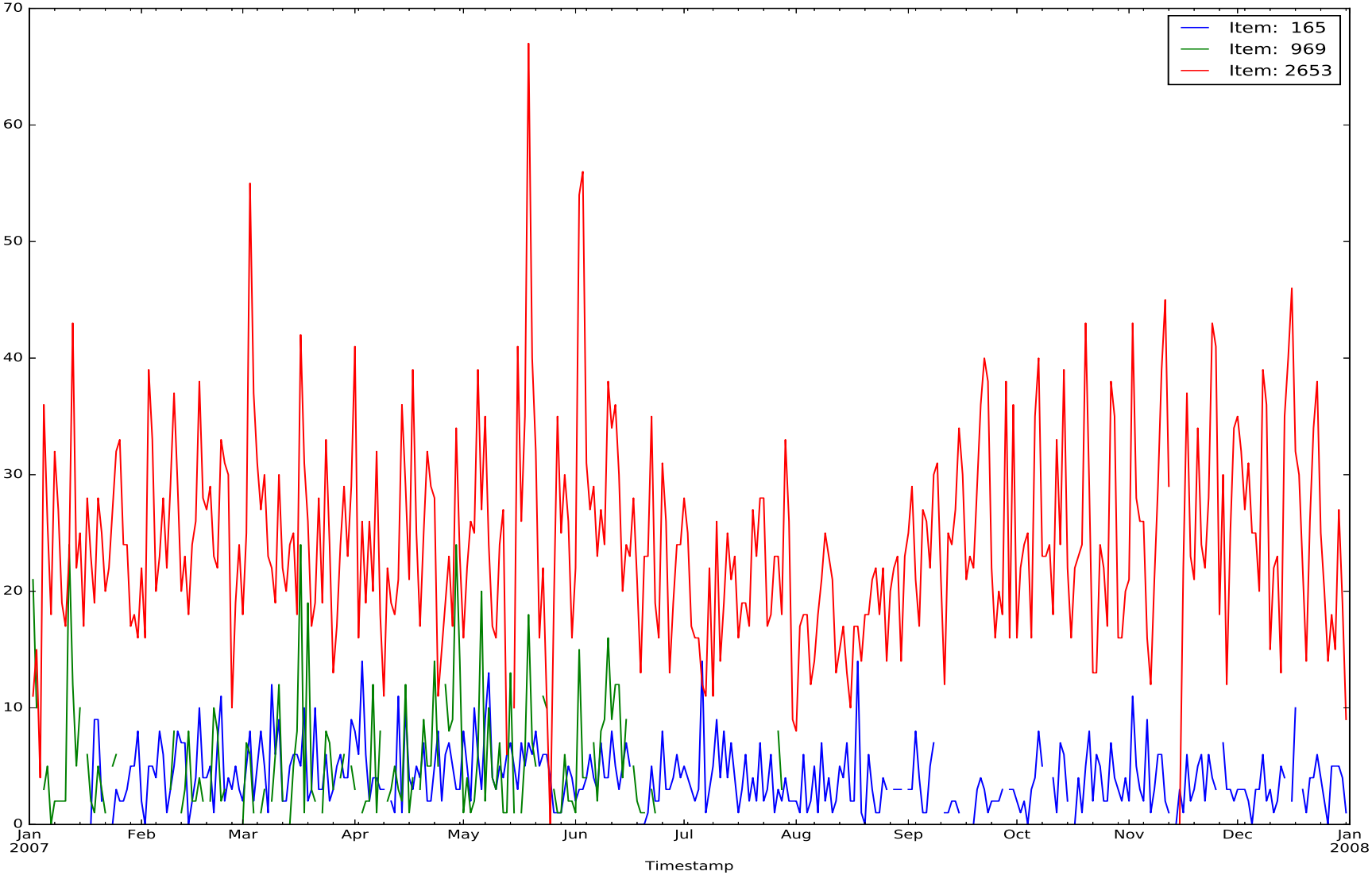
Seasonality — periodical changes of values with constant period;

Cycle — changes in time series values (economical cycles, solar activity cycles).

Error — random (unbiased) component of time series.

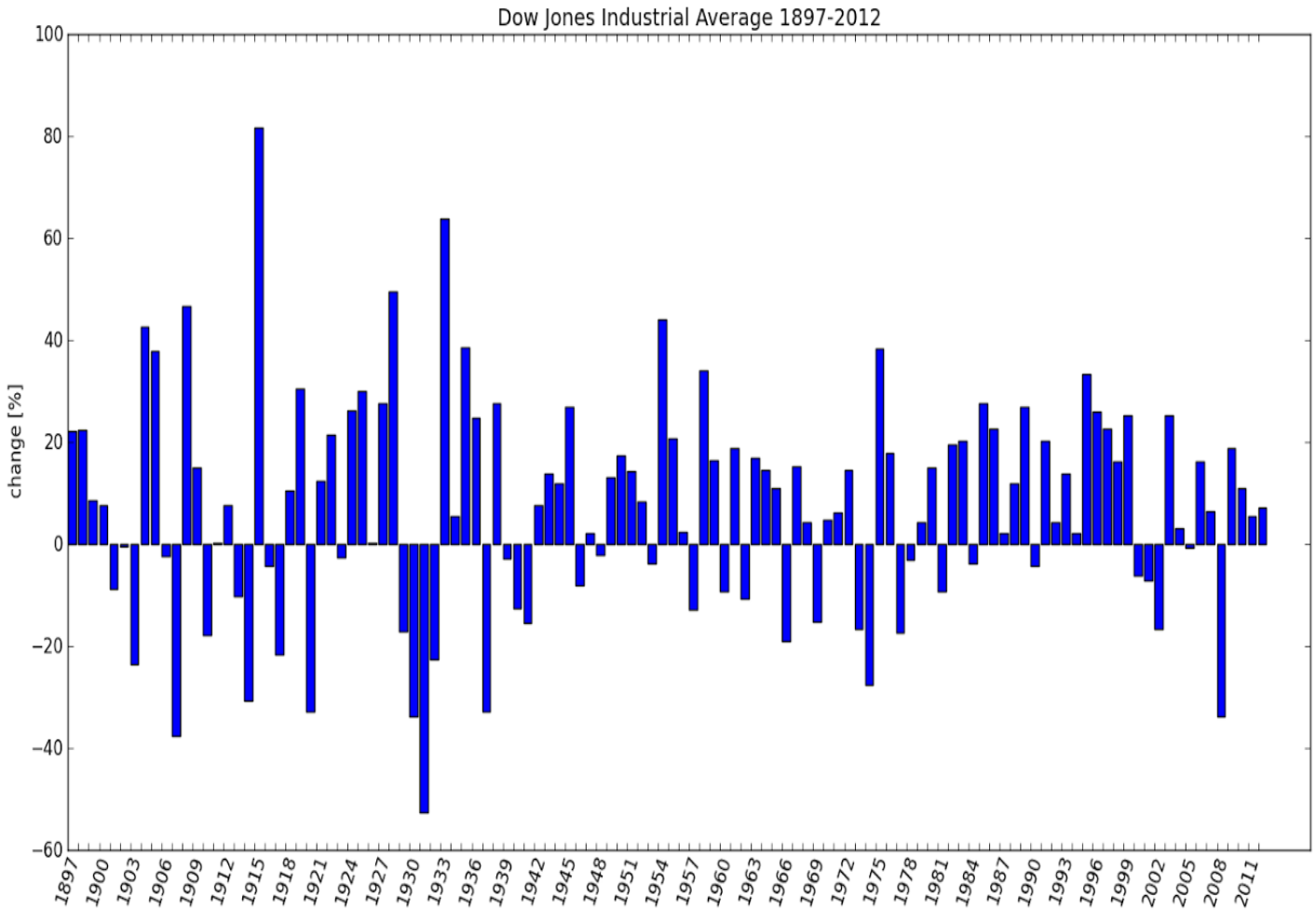
Time Series Model

Real time series of in Retail Chain:



Time Series Model

Index Dow-Jones:



Time Series Model

$y_0, y_1, \ldots, y_t, \ldots$ — is a time series, $y_i \in \mathbb{R}$
The model of time series:

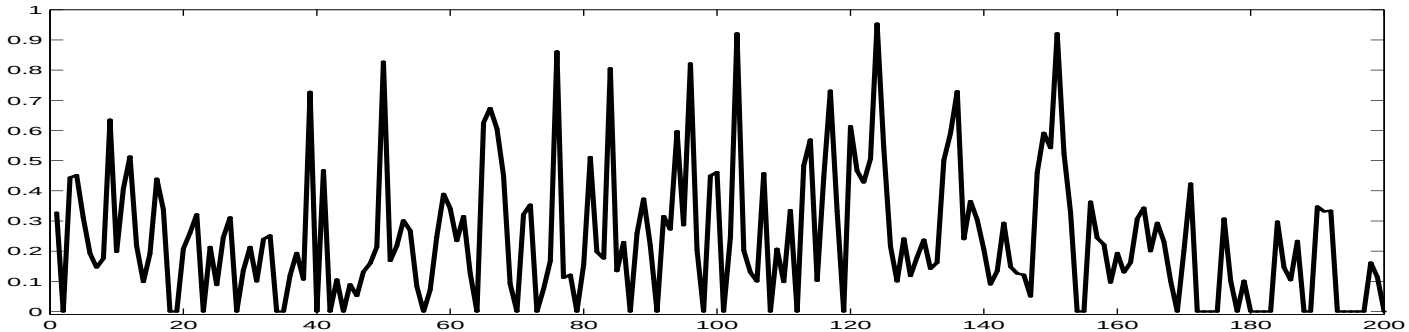
$$y_t = l_t + \varepsilon_t$$

where l_t — level of time series (changing slowly),
ε_t — (unobserved) error component (noise),

Forecasting model:

$$\hat{y}_{t+d} = \hat{l}_t$$

where $\hat{l}_t$ — an estimation of level,



Simple Exponential Smoothing

Weighted average with exponentially diminishing weights forecast:

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

$\alpha \uparrow 1 \Rightarrow$ greater weight to last points,
$\alpha \downarrow 0 \Rightarrow$ greater smoothing.

Time point	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
y_T	0.2	0.4	0.6	0.8
y_{T-1}	0.16	0.24	0.24	0.16
y_{T-2}	0.128	0.144	0.096	0.032
y_{T-3}	0.1024	0.0864	0.0384	0.0064
y_{T-4}	0.08192	0.05184	0.01536	0.00128
y_{T-5}	0.065536	0.031104	0.006144	0.000256

We find the optimal α^* using moving control:

$$Q(\alpha) = \sum_{t=T_0}^{T_1} \big(\hat{y}_t(\alpha) - y_t\big)^2 \to \min_{\alpha}$$

Empirical rules:
if $\alpha^* \in (0, 0.3)$ the series is stationary, ES works;
if $\alpha^* \in (0.3, 1)$ the series is non-stationary, we need a trend model.

Simple Exponential Smoothing

- The method suits forecasting of time series without trend and seasonality:

$$\hat{y}_{t+1|t} = l_t,$$
$$l_t = \alpha y_t + (1-\alpha)\, l_{t-1} = \hat{y}_{t|t-1} + \alpha \cdot e_t.$$

where $e_t = y_t - \hat{y}_{t|t-1}$ — forecast error at time point t
Proof:

$$\hat{y}_{t+1} := \alpha y_t + (1-\alpha)\hat{y}_t = \hat{y}_t + \alpha \cdot e_t$$

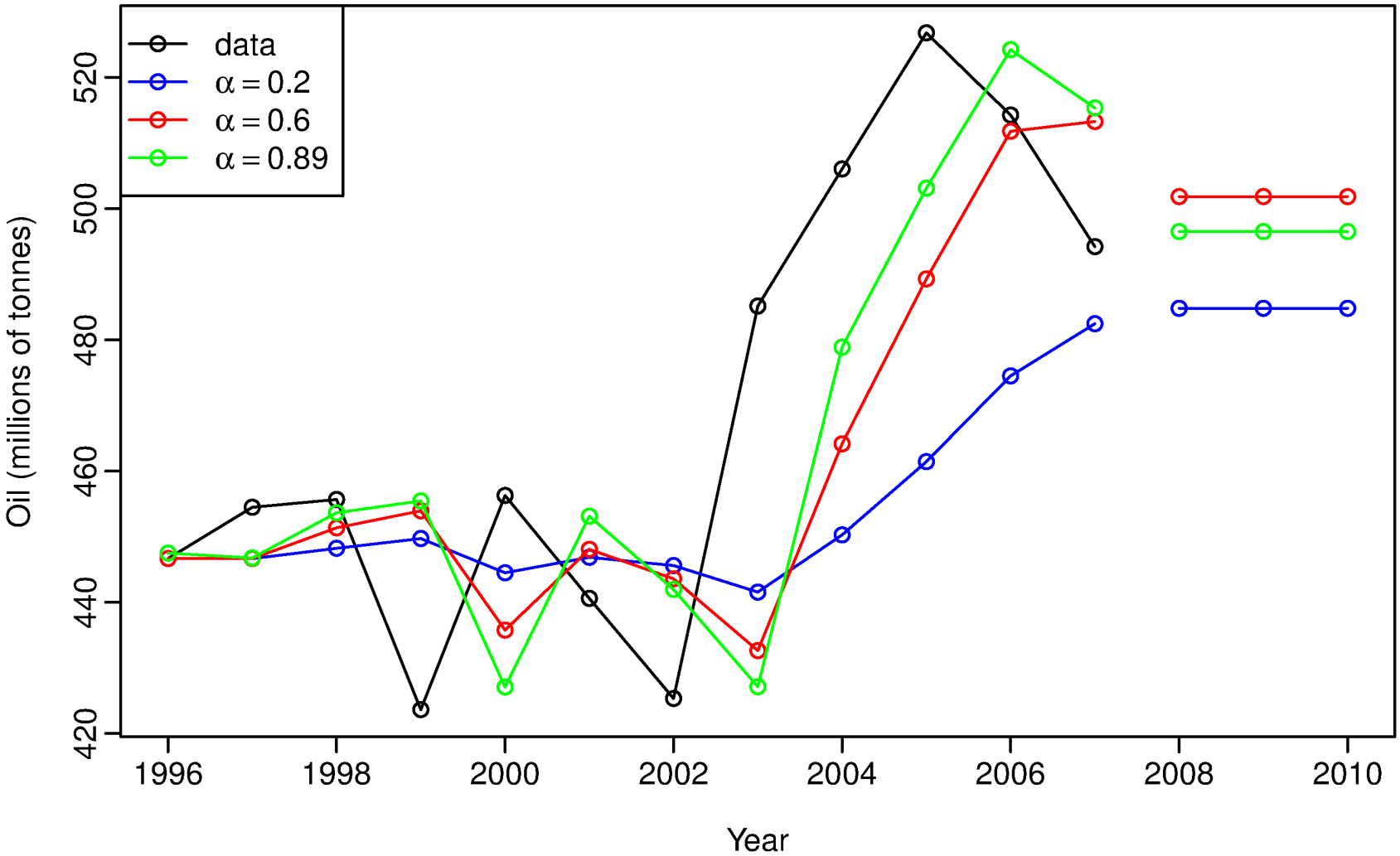
- The forecast depends on l_0:

$$\hat{y}_{T+1|T} = \sum_{j=0}^{T-1} \alpha\,(1-\alpha)^j\, y_{T-j} + (1-\alpha)^T\, l_0.$$

We can take $l_0 = y_1$ or optimize it.

- Forecast turns out flat, i.e. $\hat{y}_{t+d|t} = \hat{y}_{t+1|t}$.

Simple Exponential Smoothing



Simple ES applied to data on oil production in Saudi Arabia (1996–2007).

Tracking Signal

Tracking signal [Trigg, 1964]

$$K_t = \frac{\hat{e}_t}{\tilde{e}_t} \qquad \begin{aligned} \hat{e}_{t+1} &:= \gamma e_t + (1-\gamma)\hat{e}_t; \\ \tilde{e}_{t+1} &:= \gamma |e_t| + (1-\gamma)\tilde{e}_t. \end{aligned}$$

Recommendation: $\gamma = 0.05 \ldots 0.1$

Statistics adequacy test (at $\gamma \geq 0.1$, $t \to \infty$):

hypotheses H_0: $\mathsf{E}\varepsilon_t = 0$, $\mathsf{E}\varepsilon_t\varepsilon_{t+d} = 0$
is accepted at significance level δ if

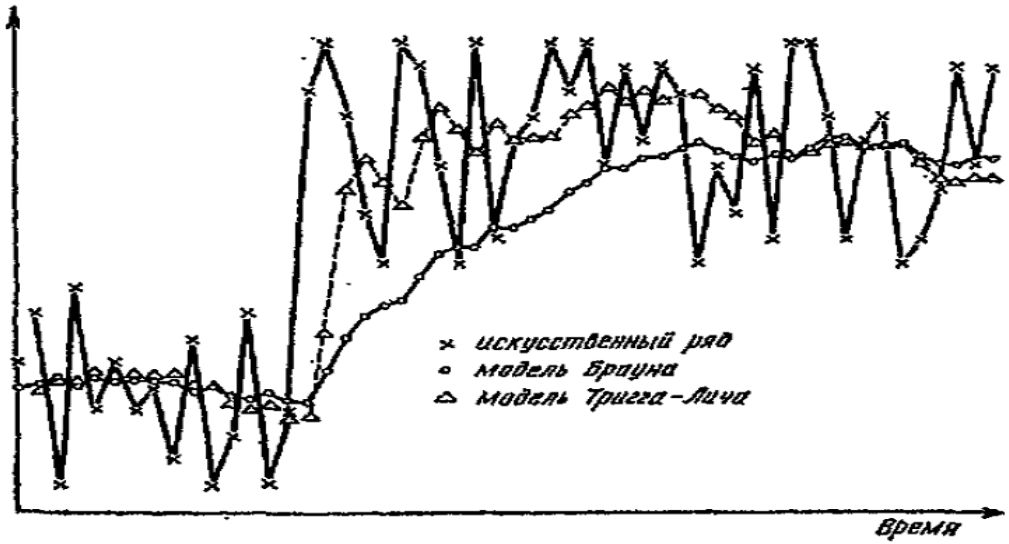
$$|K_t| \leq 1.2\Phi_{1-\delta/2}\sqrt{1-\gamma/(1+\gamma)},$$

$\Phi_{1-\delta/2}$ — normal distribution quantile,
$\Phi_{1-\delta/2} = \Phi_{0.975} = 1.96$ at $\delta = 0.05$

Trigg-Leach Model [Trigg, Leach, 1967]

Problem: adaptive models adjust poorly to sharp changes of structure

Solution: $\alpha = |K_t|$

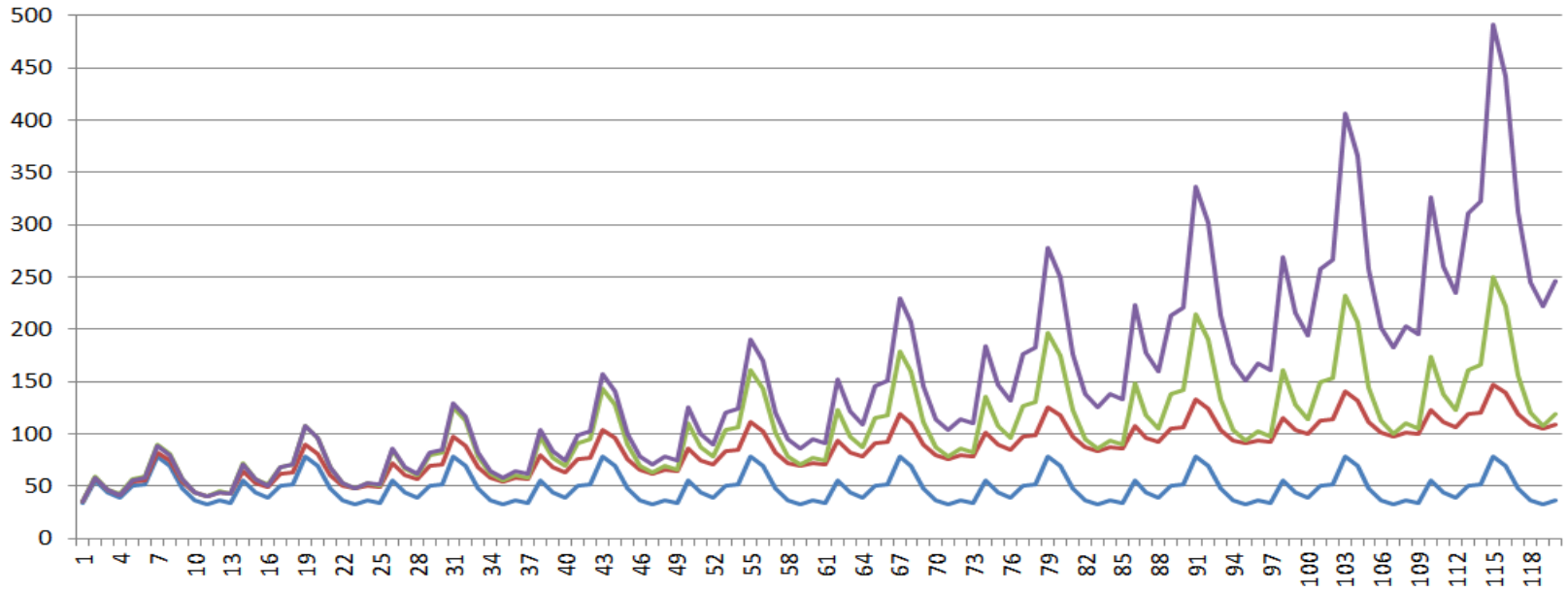


Drawbacks:

1) reacts poorly to single outliers; ($\alpha_t = |K_{t-1}|$)
2) requires fitting γ given recommended $\gamma = 0.05 \ldots 0.1$.

Examples of Trend and Seasonality

Example: Combination of trend and seasonality (model data)



Ряд 1 — seasonality and no trend
Ряд 2 — linear trend, additive seasonality
Ряд 3 — linear trend, multiplicative seasonality
Ряд 4 — exponential trend, multiplicative seasonality

Holt Model = Additive Trend

Additive (linear) trend without seasonality effect:

$$\hat{y}_{t+d} = l_t + b_t d,$$

where l_t, b_t — estimations of unopbserved components

Recursive formula:

$$l_t := \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) = \hat{y}_t + \alpha e_t;$$
$$b_t := \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1} = b_{t-1} + \alpha\beta e_t.$$

Particular case — Brown linear growth model:

$$\alpha = \alpha, \quad \beta = \alpha$$

Other Methods that Account Trend

Multiplicative (exponential) trend:

$$\begin{aligned}\hat{y}_{t+d|t} &= l_t b_t^d, \\ l_t &= \alpha y_t + (1-\alpha)\left(l_{t-1} b_{t-1}\right), \\ b_t &= \beta \frac{l_t}{l_{t-1}} + (1-\beta)\, b_{t-1}.\end{aligned}$$

$\alpha, \beta \in [0, 1]$.

Other Methods that Account Trend

Additive damped trend:

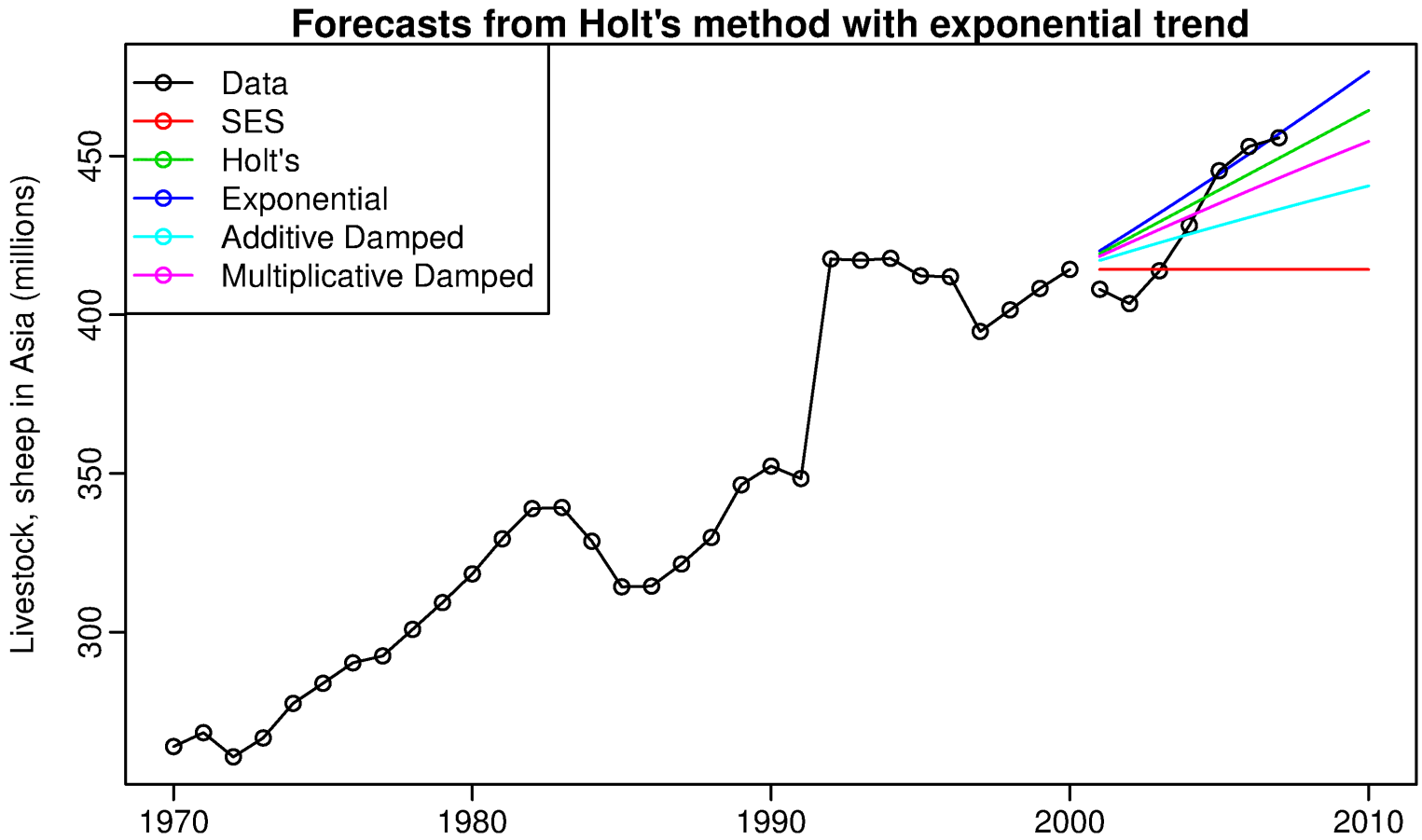
$$\begin{aligned}\hat{y}_{t+d|t} &= l_t + \left(\phi + \phi^2 + \cdots + \phi^d\right) b_t, \\ l_t &= \alpha y_t + (1-\alpha)\left(l_{t-1} + \phi b_{t-1}\right), \\ b_t &= \beta\left(l_t - l_{t-1}\right) + (1-\beta)\phi b_{t-1}.\end{aligned}$$

Multiplicative damped trend:

$$\begin{aligned}\hat{y}_{t+d|t} &= l_t b_t^{\left(\phi+\phi^2+\cdots+\phi^d\right)}, \\ l_t &= \alpha y_t + (1-\alpha) l_{t-1} b_{t-1}^{\phi}, \\ b_t &= \beta\frac{l_t}{l_{t-1}} + (1-\beta) b_{t-1}^{\phi}.\end{aligned}$$

$\alpha, \beta \in [0,1]\,,\ \phi \in (0,1)\,.$

Other Methods that Account Trend



Forecast of sheep population in Asia with regard for trend.

	SES	Holt's	Exponential	Additive damped	Multiplicative damped
α	1	0.98	0.98	0.99	0.98
β		0	0	0	0.00
ϕ				0.98	0.98

Winters Model = Multiplicative Seasonality

Multiplicative Seasonality of Period p:

$$\hat{y}_{t+d} = l_t \cdot s_{t-p+(d \bmod p)},$$

$s_0, \ldots, s_{p-1}$ — seasonality profile of period p.
Recursive formula:

$$l_t := \alpha(y_t / s_{t-p}) + (1-\alpha) l_{t-1} = l_{t-1} + \alpha e_t / s_{t-p};$$
$$s_t := \beta(y_t / l_t) + (1-\beta) s_{t-p} = s_{t-p} + \beta(1-\alpha) e_t / l_t.$$

Proof of the last equation:

$$s_t := s_{t-p} + \beta\left(y_t/l_t - s_{t-p}\right) = s_{t-p} + \beta\left(y_t - s_{t-p} l_t\right)/l_t =$$
$$s_{t-p} + \beta\big(y_t - s_{t-p}(l_{t-1} + \alpha e_t / s_{t-p})\big)/l_t = s_{t-p} +$$
$$+\beta\left(\underbrace{y_t - s_{t-p} l_{t-1}}_{e_t} - \alpha e_t\right)/l_t$$

Additive Seasonality ES Model

Additive seasonality with period of length p:

$$\begin{aligned}
\hat{y}_{t+d|t} &= l_t + s_{t-p+(d \mod p)}, \\
l_t &= \alpha\left(y_t - s_{t-p}\right) + (1-\alpha)\left(l_{t-1}\right) = l_{t-1} + \alpha e_t; \\
s_t &= \gamma\left(y_t - l_{t-1}\right) + (1-\gamma)\, s_{t-p} = s_{t-p} + \gamma(1-\alpha)e_t.
\end{aligned}$$

Theil-Wage Model

Linear trend with additive seasonality of period s:

$$\hat{y}_{t+d} = (l_t + b_t d) + s_{t+(d \bmod s)-p}.$$

$l_t + b_t d$ — trend cleaned of seasonality,
$s_0, \ldots, s_{p-1}$ — seasonality profile of period p.

Recursive formula:

$$\begin{aligned}
l_t &:= \alpha(y_t - s_{t-p}) + (1-\alpha)(l_{t-1} + b_{t-1}) = l_{t-1} + b_{t-1} + \alpha e_t; \\
b_t &:= \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1} = b_{t-1} + \alpha\beta e_t; \\
s_t &:= \gamma(y_t - l_t) + (1-\gamma) s_{t-p} = s_{t-p} + \gamma(1-\alpha) e_t.
\end{aligned}$$

Winters Model with Linear Trend

Multiplicative seasonality of period s with a linear trend:

$$\hat{y}_{t+d} = (l_t + b_t d) \cdot s_{t+(d \bmod p)-p},$$

$l_t + b_t d$ — trend cleaned of seasonality,
$s_0, \ldots, s_{p-1}$ — seasonality profile of period s.

Recursive formula:

$$\begin{aligned}
l_t &:= \alpha(y_t/s_{t-p}) + (1-\alpha)(l_{t-1} + b_{t-1}) = l_{t-1} + b_{t-1} + \alpha e_t/s_{t-p};\\
b_t &:= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} = b_{t-1} + \alpha\beta e_t/s_{t-p};\\
s_t &:= \gamma(y_t/l_t) + (1-\gamma)s_{t-p} = s_{t-p} + \gamma(1-\alpha)e_t/l_t.
\end{aligned}$$

Winters Model with multiplicative trend

Multiplicative (exponential) trend model exponential trend:

$$\hat{y}_{t+d} = l_t(b_t)^d \cdot s_{t+(d \bmod p)-p},$$

$l_t(b_t)^d$ — exponential trend without seasonality,
$s_0, \ldots, s_{p-1}$ — seasonal trend p.

Recurrent version:

$$\begin{aligned}
l_t &:= \alpha(y_t/s_{t-p}) + (1-\alpha)l_{t-1}b_{t-1} = l_{t-1}b_{t-1} + \alpha e_t/s_{t-1};\\
b_t &:= \beta(l_t/l_{t-1}) + (1-\beta)b_{t-1} = b_{t-1} + \alpha\beta e_t/st-1;\\
s_t &:= \gamma(y_t/l_t) + (1-\gamma)s_{t-p} = s_{t-p} + \gamma(1-\alpha)e_t/l_t.
\end{aligned}$$

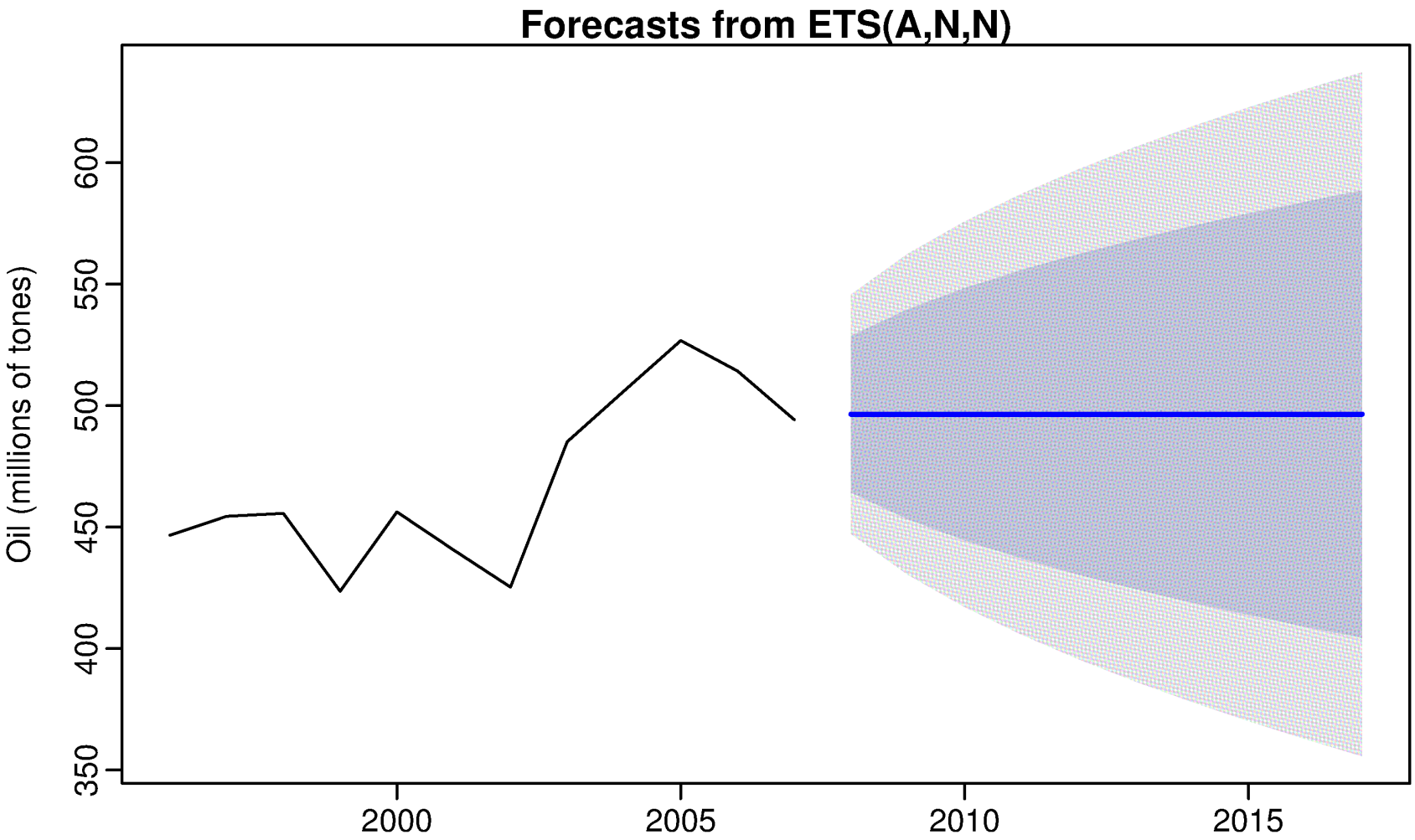
Types of ES Models

	Seasonality		
Trend	N (None)	A (Additive)	M (Multiplicative)
N (None)	(N,N)	(N,A)	(N,M)
A (Additive)	(A,N)	(A,A)	(A,M)
Ad (Additive damped)	(Ad,N)	(Ad,A)	(Ad,M)
M (Multiplicative)	(M,N)	(M,A)	(M,M)
Md (Multiplicative damped)	(Md,N)	(Md,A)	(Md,M)

We may additionally suggest an additive (A) or a multiplicative (M) error (the type of error does not influence single-value prediction). Multiplicative error is suitable only for strictly positive time series.

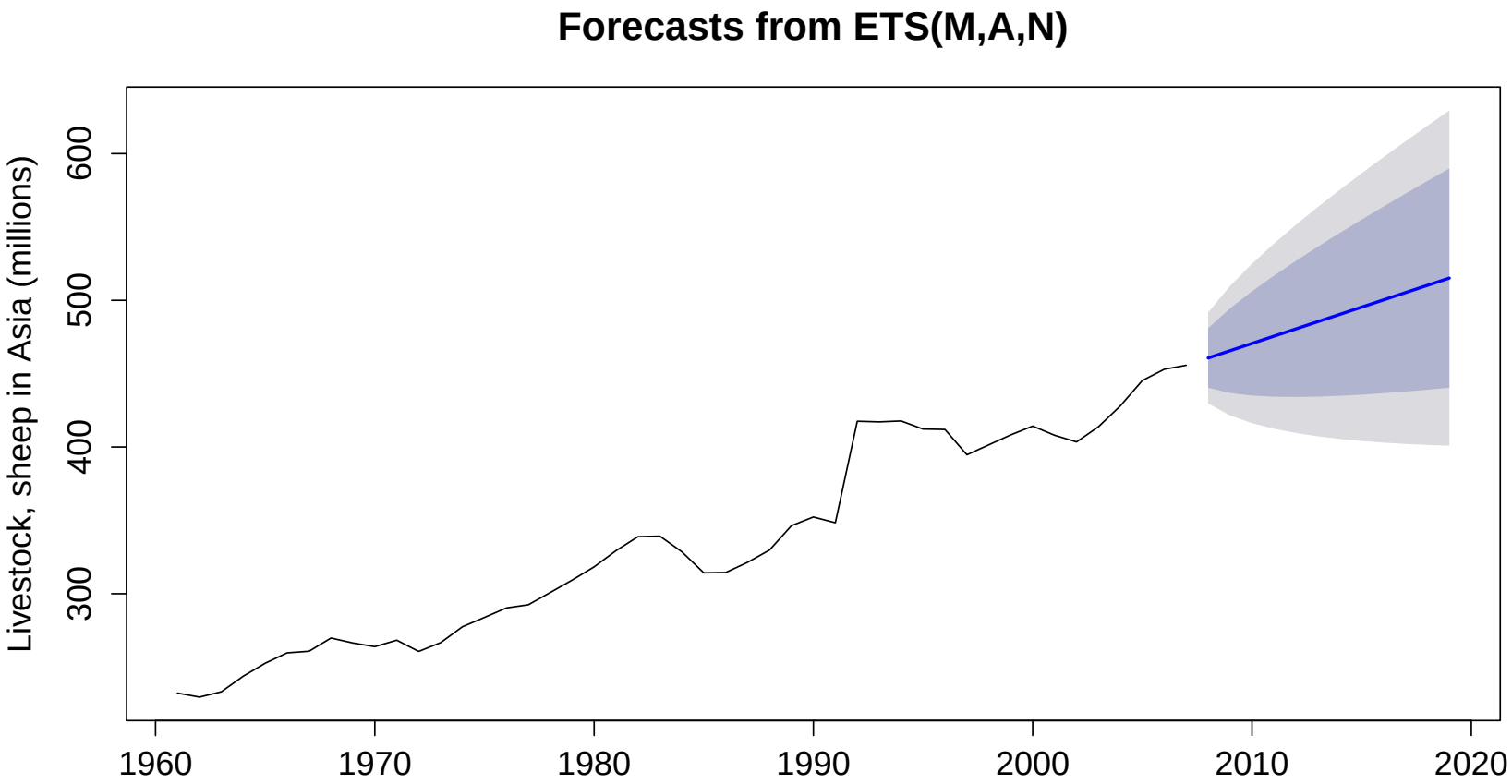
The final model may be written as $ESM(\cdot,\cdot,\cdot)$.

Examples of Forecast



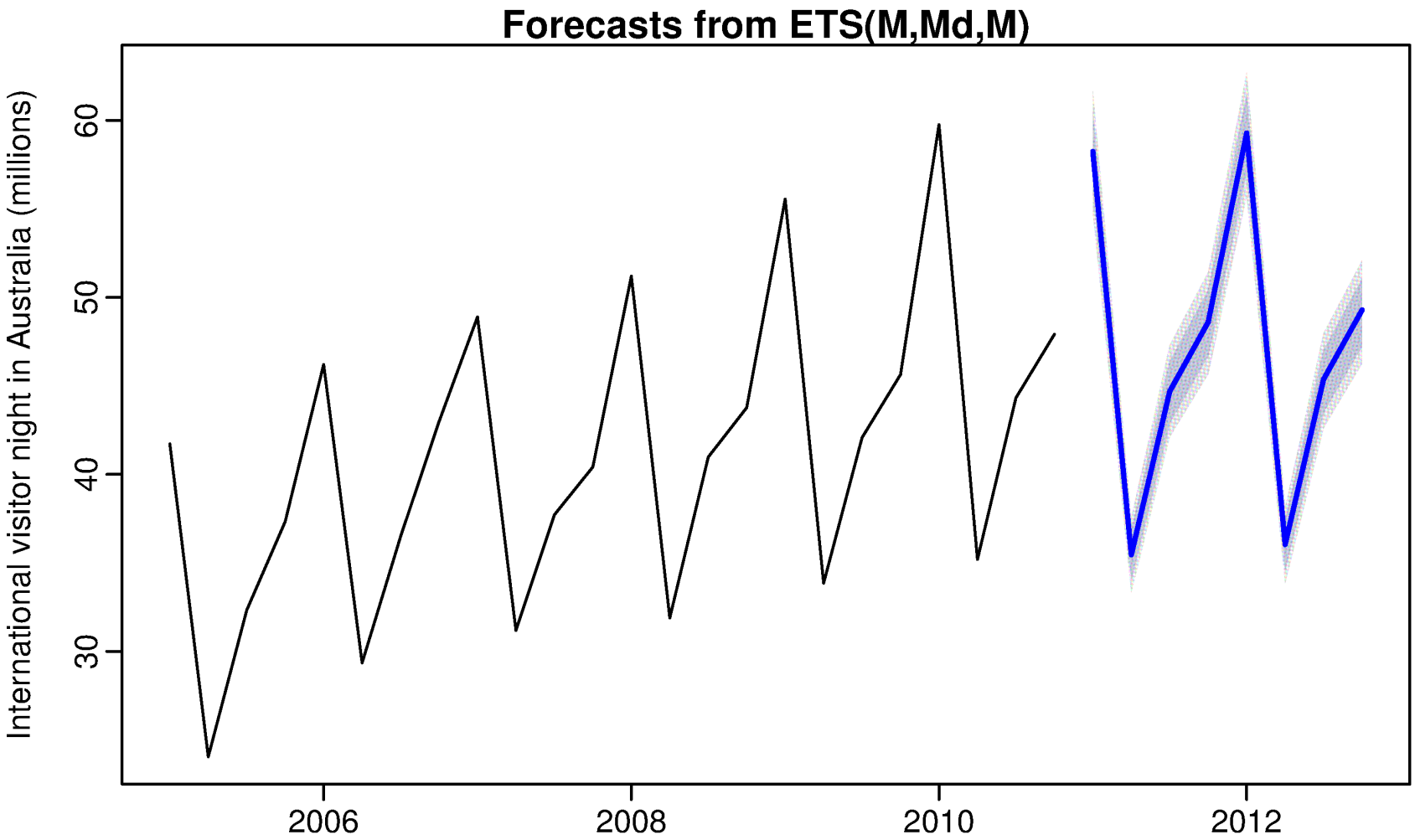
For the data on oil production in Saudi Arabia function ESM selects simple ES.

Examples of Forecast



For the sheep population in Asia function ESM selects the model with multiplicative error and an additive linear trend.

Examples of Forecast



For the quantity of nights spent by tourists in Australia function ESM selects a model with multiplicative error, seasonality and a damped trend.

Loss Functions of dotted forecasts

Mean squared error:

$$MSE = \frac{1}{T-R+1}\sum_{t=R}^{T}\left(\hat{y}_t - y_t\right)^2.$$

Mean absolute error:

$$MAE = \frac{1}{T-R+1}\sum_{t=R}^{T}\left|\hat{y}_t - y_t\right|.$$

Mean absolute percentage error:

$$MAPE = \frac{100}{T-R+1}\sum_{t=R}^{T}\left|\frac{\hat{y}_t - y_t}{y_t}\right|.$$

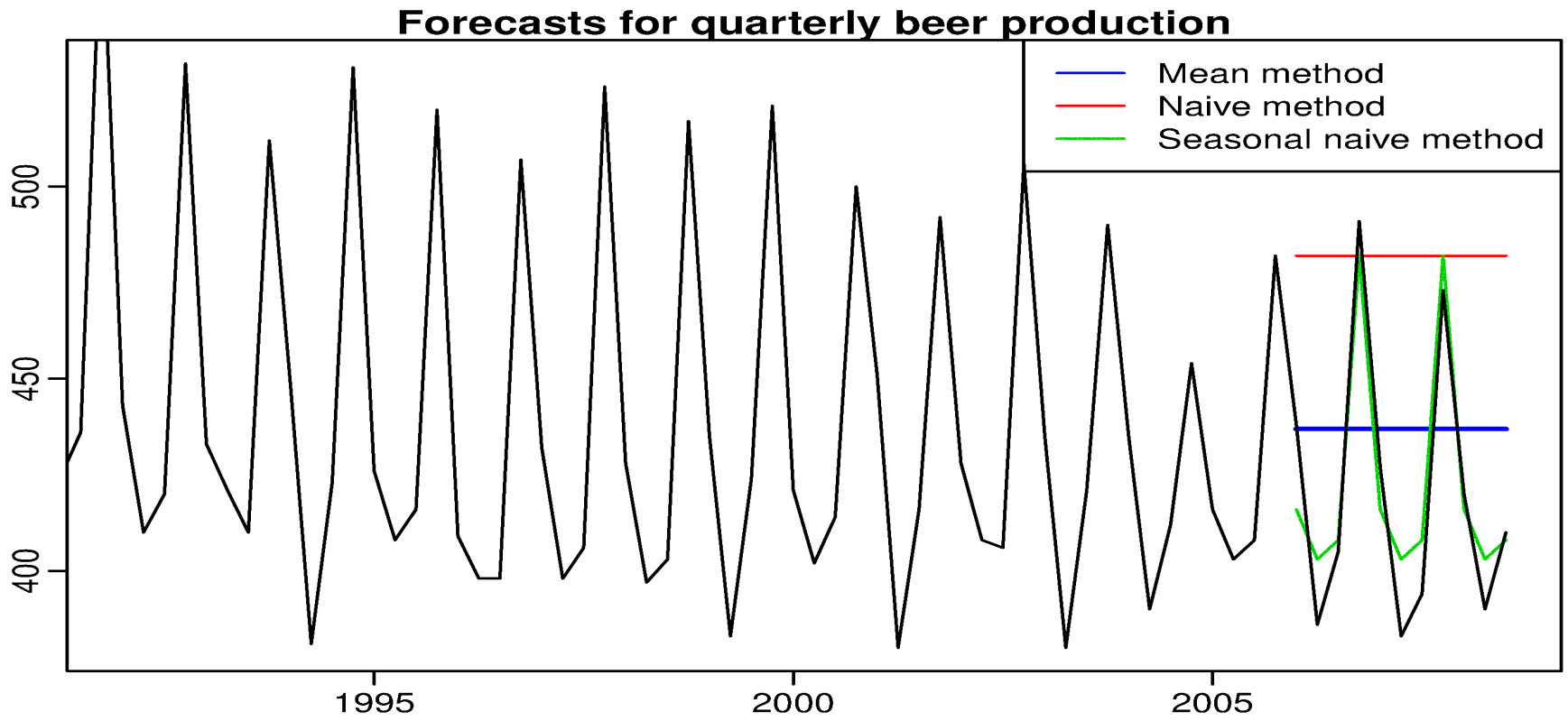
Symmetric mean absolute percentage error (MACAPE):

$$SMAPE = \frac{200}{T-R+1}\sum_{t=R}^{T}\left|\frac{\hat{y}_t - y_t}{\hat{y}_t + y_t}\right|.$$

Mean absolute scaled error:

$$MASE = \frac{1}{T-R+1}\sum_{t=R}^{T}\left|\hat{y}_t - y_t\right| \Bigg/ \frac{1}{T-1}\sum_{t=2}^{T}\left|y_t - y_{t-1}\right|.$$

Loss Functions of dotted forecasts



Algorithm	RMSE	MAE	MAPE	MASE
MA	38.01	33.78	8.17	2.30
Naive	70.91	63.91	15.88	4.35
Seasonal Naive	12.97	11.27	2.73	0.77

Relative Measures

Uncertainty coefficient (Theil's coefficient) estimates accuracy of forecast with respect to naive forecast :

$$U(d) = \sqrt{\frac{\sum_{t=R}^{T-d} \left(\hat{y}_{t+d|t} - y_{t+d}\right)^2}{\sum_{t=R}^{T-d} \left(y_t - y_{t+d}\right)^2}}, \quad d = 1, \ldots, D.$$

If $U(d) = 1$, then $\hat{y}_{t+d|t}$ is close to naive forecast ; if $U(d) < 1$, then forecast $\hat{y}_{t+d|t}$ is better than naive forecast, $U(d) > 1$ — naive forecast is better.

Comparing of two algorithms

$y_1, \ldots, y_T$ — time series,

$\hat{y}_{1R}, \ldots, \hat{y}_{1T}$ — forecasts of the first algorithm for period $R, \ldots, T$

$\hat{\varepsilon}_{1R}, \ldots, \hat{\varepsilon}_{1T}$ — residuals of the first algorithms,

$\hat{y}_{2R}, \ldots, \hat{y}_{2T}$ — forecasts of the second algorithm for period $R, \ldots, T$,

$\hat{\varepsilon}_{2R}, \ldots, \hat{\varepsilon}_{2T}$ — residals of the second algorithm;

$g\left(y_t, \hat{y}_{it}\right)$ — some loss function,

(for example, $|\hat{\varepsilon}_{it}|$ or $\hat{\varepsilon}_{it}^2$),

$$d_t = g\left(y_t, \hat{y}_{1t}\right) - g\left(y_t, \hat{y}_{2t}\right).$$

$$H_0: \text{ average } d_t = 0,$$
$$H_1: \text{ average } d_t <\neq> 0.$$

Wilcoxon signed-rank test:

$$W = \sum_{t=R}^{T} \operatorname{rank}\left(|d_t|\right) \operatorname{sign}\left(d_t\right).$$

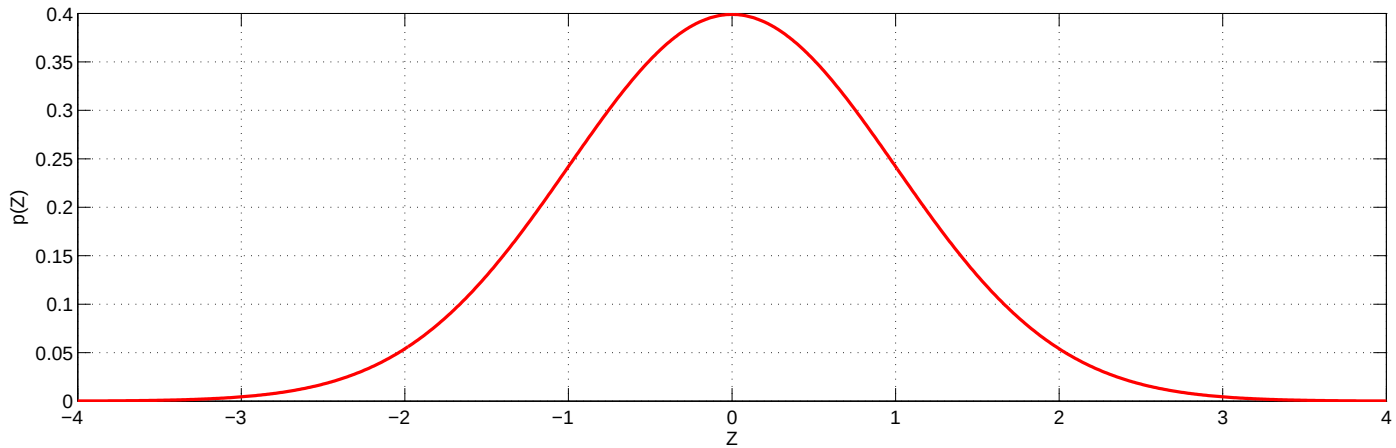
Diebold-Mariano test

null hypothesis: $H_0 : \mathbb{E}d_t = 0$

alternative hypothesis: $H_1 : \mathbb{E}d_t <\neq> 0$

statistics: $B = \frac{\bar{d}}{\sqrt{\hat{f}/T}}, \quad \hat{f} = \sum_{\tau=-M}^{M} \hat{r}_\tau, \quad M = T^{1/3}$

zero distribution: $N(0,1)$



Modification for short time series(Harvey, Leybourne, Newbold):

$$B^* = \frac{B}{\sqrt{\frac{T+1-2d+\frac{d(d-1)}{T}}{T}}}.$$

External Parameters Fitting and Model Selection for TS

- X — feature space ($\mathbb{R}^n$); Y — answer space ($\mathbb{R}$);
  $X^\ell = (x_i, y_i)_{i=1}^\ell$ — train samples;
  $y_i = y(x_i)$, $y\colon X \to Y$ — unknown function;
  Loss function $\lambda(y_i, \hat{y}_i)$
- Learning method is a function: $\mu\colon 2^{X \times Y} \to \mathfrak{A}$
- loss of algorithm $A \in \mathfrak{A}$:

$$Loss_A = \mathsf{E}_{x,y}\left[\lambda\left(y, \mu\left(X^\ell\right)\right)^2\right]$$

- loss of learning method μ:

$$Q_\mu = \mathsf{E}_{X^\ell}\left[Loss_A\right]$$

Main problem of ML — is to minimize Q_μ
Estimation of loss of algorithm $A = \mu(X^\ell)$

$$Loss_A(X^\ell) = \sum_{i=1}^{\ell} \lambda\left(y_i, A(x_i)\right)$$

Cross Validation: Train, Validate and Test

CV approach: $X^\ell = X^L \bigcup X^K$

$$Q_\mu^{CV}(X^\ell) = Q\left(\mu(X^L), X^K\right) = Loss_{\mu(X^L)}\left(X^K\right)$$

Cross Validation: Train, Validate and Test

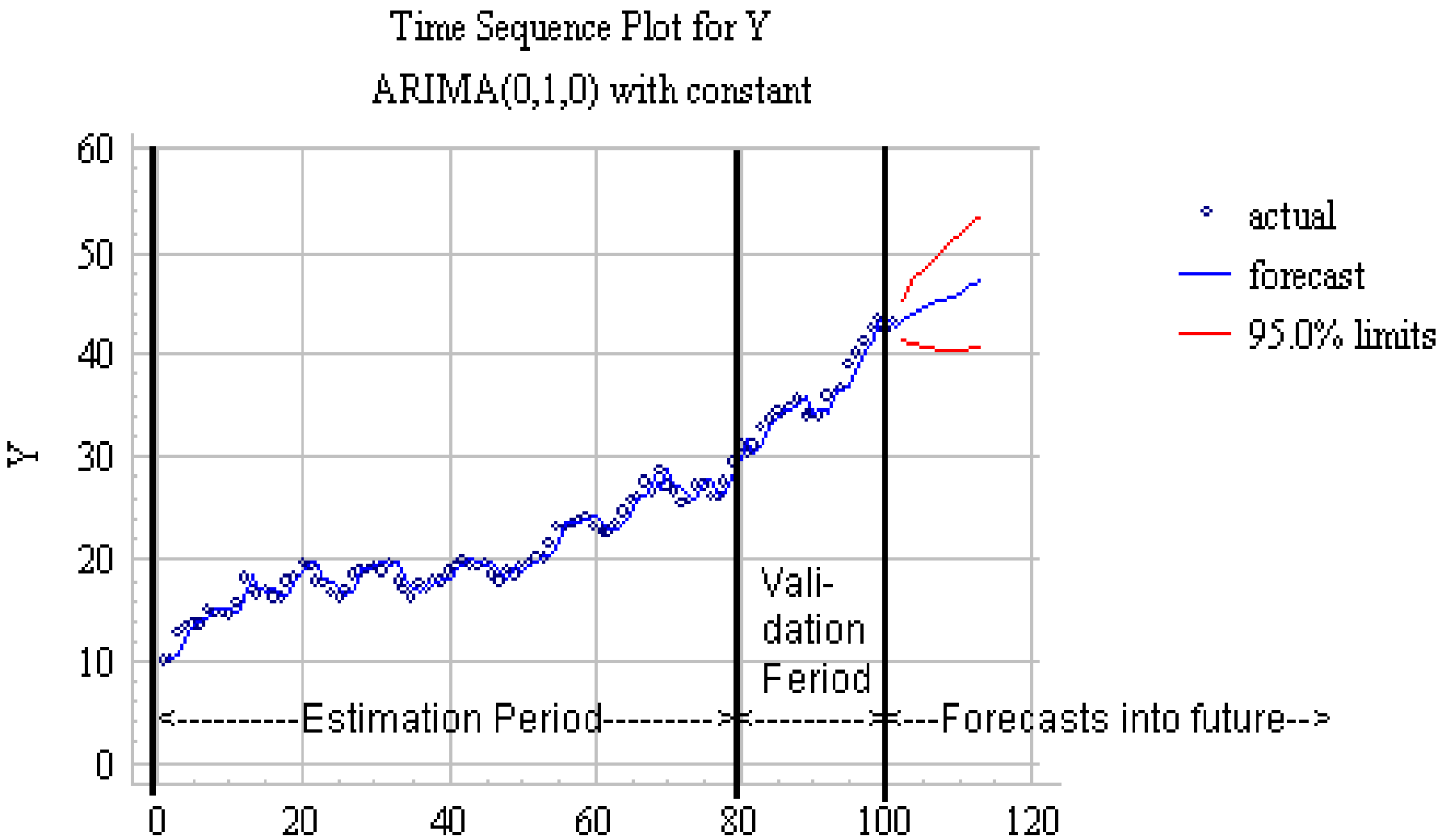
Most popular in ML q-fold CV: $X^\ell = X_1^{\ell_1} \bigcup \cdots \bigcup X_q^{\ell_q}$

$$\mathrm{Q}_\mu^{\mathrm{q-fold}}(X^\ell) = \frac{1}{q}\sum_{i=1}^{q} Q\left(\mu\left(X^\ell \backslash \{X_i^{\ell_i}\}\right), X_i^{\ell_i}\right)$$

Question: why test sample is needed?

Cross Validation: Train, Validate and Test

TS Forecasting:



What if there is no Validation Period?

AIC (Akaike Information Criterion)

$$Q_{\mu}^{AIC}(X^{\ell}) = Q\left(\mu\left(X^{\ell}\right), X^{\ell}\right) + \frac{2\hat{\sigma}^2}{\ell} \cdot n$$

BIC (Bayes Information Criterion)

$$Q_{\mu}^{BIC}(X^{\ell}) = \frac{\ell}{\hat{\sigma}^2} Q\left(\mu\left(X^{\ell}\right), X^{\ell}\right) + \ln(\ell) \cdot n$$

HQIC (Hannan–Quinn information criterion)

$$Q_{\mu}^{HQIC}(X^{\ell}) = \frac{\ell}{\hat{\sigma}^2} Q\left(\mu\left(X^{\ell}\right), X^{\ell}\right) + \ln\ln(\ell) \cdot n$$

AIC, BIC: `http://www.stat.cmu.edu/~larry/=stat705/Lecture16.pdf`

Conclusion

- there are a lot of measures for accuracy of forecasts;
- to fit parameters of ES models TS should be divided into Train Period and Validation Period

Conclusion

- there are a lot of measures for accuracy of forecasts;
- to fit parameters of ES models TS should be divided into Train Period and Validation Period

Pro ES:

- SES is very simple model
- very useful for short TS or simple TS
- can be easily modified for different components of TS
- forecast of ES models can be easliy explained

Conclusion

- there are a lot of measures for accuracy of forecasts;
- to fit parameters of ES models TS should be divided into Train Period and Validation Period

Pro ES:

- SES is very simple model
- very useful for short TS or simple TS
- can be easily modified for different components of TS
- forecast of ES models can be easliy explained

Cons ES:

- ES does not take into account independent variables
- heuristic method (there is no theoretical guaranties about it's work)
- Forecast of ES depends on initialization (l_0)

Your fitback

Leave your fitback here `https://goo.gl/forms/TpY4aaojXLszGPQy2`

If you have questions or suggestions you can write me: alexromsput@gmail.com

Literature

Hyndman R.J., Athanasopoulos G. Forecasting: principles and practice. — OTexts, https://www.otexts.org/book/fpp

Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. Финансы и статистика, 2003, http://www.arshinov74.ru/files/files/3.pdf