RMT Analyse de données et modélisation
2014

Dynamic linear models for analyzing yield time series

David Makowski

INRA

Outline

1. Objective & main principles
2. Model specification
3. Filter and smoother using Gaussian dynamic linear models
4. Conclusion

Outline

1. Objective & main principles
2. Model specification
3. Filter and smoother using Gaussian dynamic linear models
4. Conclusion

1. Objective & main principles

Main principles

- Filtering: Updating state variable sequentially in time

Variable t	→	Variable t+1	→	Variable t+2	→	Variable t+3

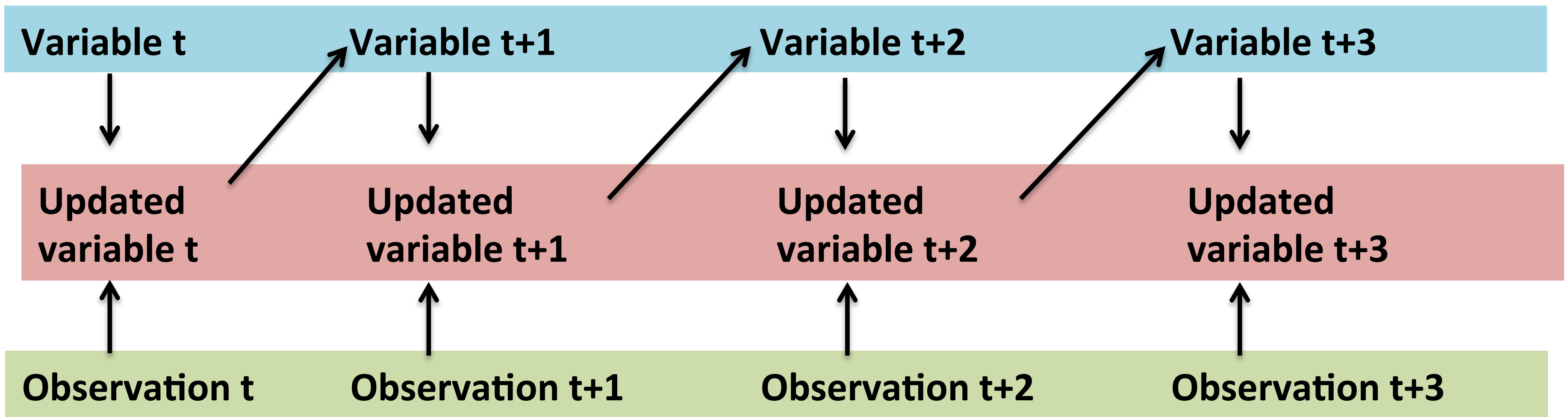
Main principles

- Filtering: Updating state variable sequentially in time

Variable t → **Variable t+1** → **Variable t+2** → **Variable t+3**

Observation t **Observation t+1** **Observation t+2** **Observation t+3**

Main principles



Main principles

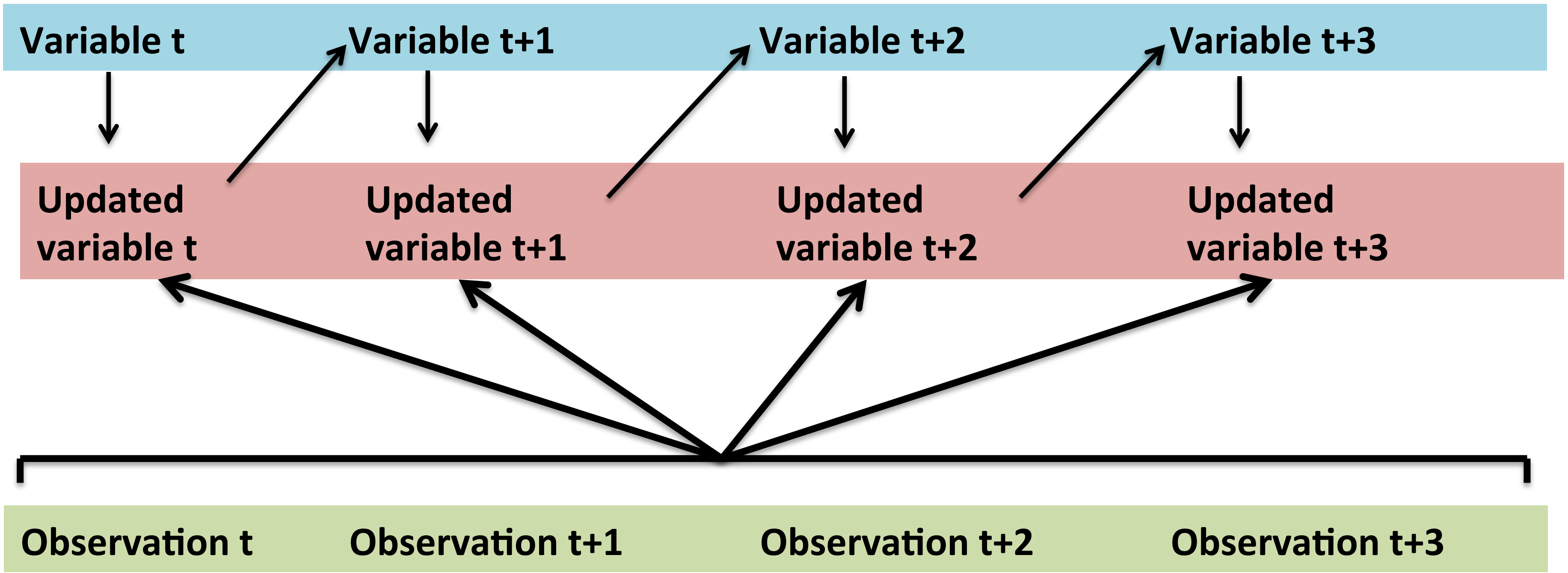
- Smoothing: Updating state variable using all observations

Variable t → **Variable t+1** → **Variable t+2** → **Variable t+3**

Observation t **Observation t+1** **Observation t+2** **Observation t+3**

Main principles

- Smoothing: Updating state variable using all observations



Summary

- **Filter and smoother** are two tools for updating dynamic state variables

- State variables are updated sequentially in time using data

- Measurement and model errors are both taken into account

Outline

1. Objective & main principles
2. Model specification
3. Filter and smoother using Gaussian dynamic linear models
4. Conclusion

Model specification

Two equations

- **Observation equation**

Relates an observation collected at time t to the model state variable(s)

- **System equation**

Describes the dynamic behavior of the state variables. It relates the values of the vector of the state variables at time t to the values at time t-1

Observation equation

$$Y_t = f\left(Z_t, X_t^{(y)}, \delta, \varepsilon_t\right)$$

State variable at time t

Input variable at time t

Fixed parameters

Random term accounting for the **imperfection** of the relationship

System equation

$$Z_t = g\left(Z_{t-1}, X_t^{(z)}, \theta, \eta_{t-1}\right)$$

State variable at time t

State variable at time t-1

Input variable at time t

Fixed parameters

Random term describing **model error**

Model specification

- **Observation equation**

$$Y_t = f\left(Z_t, X_t^{(y)}, \delta, \varepsilon_t\right)$$

- **System equation**

$$Z_t = g\left(Z_{t-1}, X_t^{(z)}, \theta, \eta_{t-1}\right)$$

Example 1: Random walk model

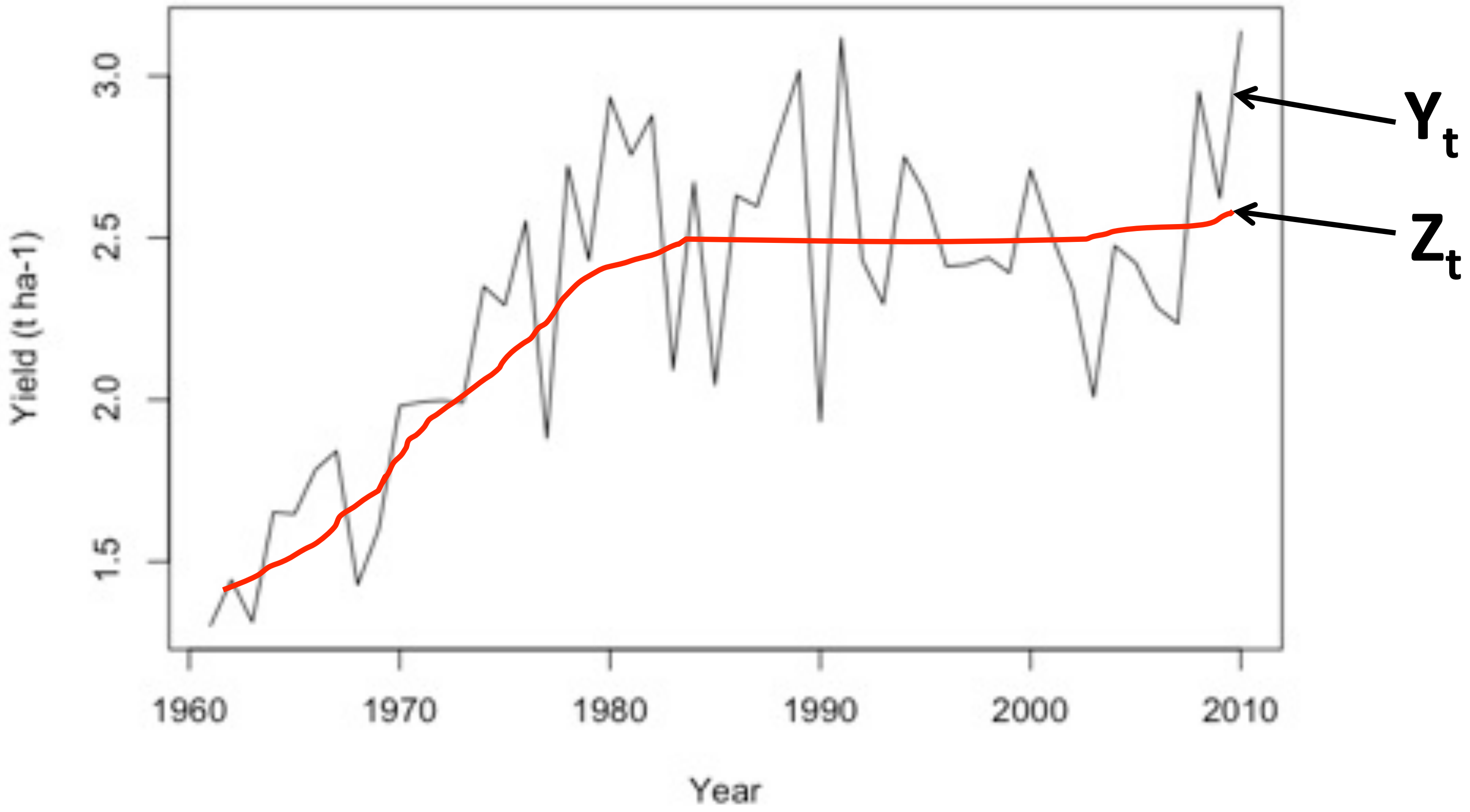
- **Observation equation**

$$Y_t = Z_t + \varepsilon_t \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

- **System equation**

$$Z_t = Z_{t-1} + \eta_{t-1} \qquad \eta_{t-1} \sim N\left(0, \sigma_\eta^2\right)$$

Example 1: Random walk model



Wheat yield data in Greece (FAO)

Summary

- Use of two equations
  - Observation equation
  - System equation
- Very flexible
  - Data: continuous, binary, count
  - One or several state variables
- From simple to complex models
  - Linear Gaussian models
  - Nonlinear models

Outline

1. Objective & main principles
2. Model specification
3. Filter and smoother using Gaussian dynamic linear models
4. Conclusion

Filter and smoother using Gaussian dynamic linear models

Gaussian linear model

- Observation equation

$$Y_t = f\left(Z_t, X_t^{(y)}, \delta, \varepsilon_t\right)$$

f **is linear**

ε_t **is Gaussian**

- System equation

$$Z_t = g\left(Z_{t-1}, X_t^{(z)}, \theta, \eta_{t-1}\right)$$

g **is linear**

η_{t-1} **is Gaussian**

Gaussian linear model

- Observation equation

$$Y_t = FZ_t + \varepsilon_t$$

F is a matrix and ε_t is a Gaussian random term. If Y_t includes N measurements and if Z_t includes m states variables, F is a (N x m) matrix, $\varepsilon_t \sim N(0,V)$, and V is a (N x N) variance-covariance matrix.

- System equation

$$Z_t = GZ_{t-1} + \eta_{t-1}$$

G is a (m x m) matrix, $\eta_t \sim N(0,W)$, and W is a (m x m) variance-covariance matrix.

Example 1: Random walk model

- Observation equation

$$Y_t = Z_t + \varepsilon_t \qquad \boldsymbol{f} = \textbf{identity} \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

- System equation

$$Z_t = Z_{t-1} + \eta_{t-1} \qquad \boldsymbol{g} = \textbf{identity} \qquad \eta_{t-1} \sim N\left(0, \sigma_\eta^2\right)$$

Kalman filter using Gaussian linear models

- Expected value and variance **before** update at time t

$$Y_{1:t-1} = \left(Y_1, \ldots, Y_{t-1}\right)$$

$$E\left(Z_t \middle| Y_{1:t-1}\right) \qquad V\left(Z_t \middle| Y_{1:t-1}\right)$$

- Expected value and variance **after** update at time t

$$Y_{1:t} = \left(Y_1, \ldots, Y_{t-1}, Y_t\right)$$

$$E\left(Z_t \middle| Y_{1:t}\right) \qquad V\left(Z_t \middle| Y_{1:t}\right)$$

Kalman filter using Gaussian linear models

- Expected value and variance **before** update at time t

$$Y_{1:t-1} = \left(Y_1, \dots, Y_{t-1}\right)$$

$$E\left(Z_t \middle| Y_{1:t-1}\right) \qquad V\left(Z_t \middle| Y_{1:t-1}\right)$$

- Expected value and variance **after** update at time t

$$Y_{1:t} = \left(Y_1, \dots, Y_{t-1}, Y_t\right)$$

$$E\left(Z_t \middle| Y_{1:t}\right) \qquad V\left(Z_t \middle| Y_{1:t}\right)$$

Kalman smoother using Gaussian linear models

$$Y_{1:N} = \left(Y_1, \ldots, Y_t, \ldots, Y_N \right)$$

$$E\left(Z_t \middle| Y_{1:N} \right) \qquad V\left(Z_t \middle| Y_{1:N} \right)$$

Example 1: Random walk model (t=1)

- Observation equation

$$Y_1 = Z_1 + \varepsilon_1 \qquad \varepsilon_1 \sim N\left(0, \sigma_\varepsilon^2\right)$$

- System equation

$$Z_1 = Z_0 + \eta_0 \qquad \eta_0 \sim N\left(0, \sigma_\eta^2\right)$$

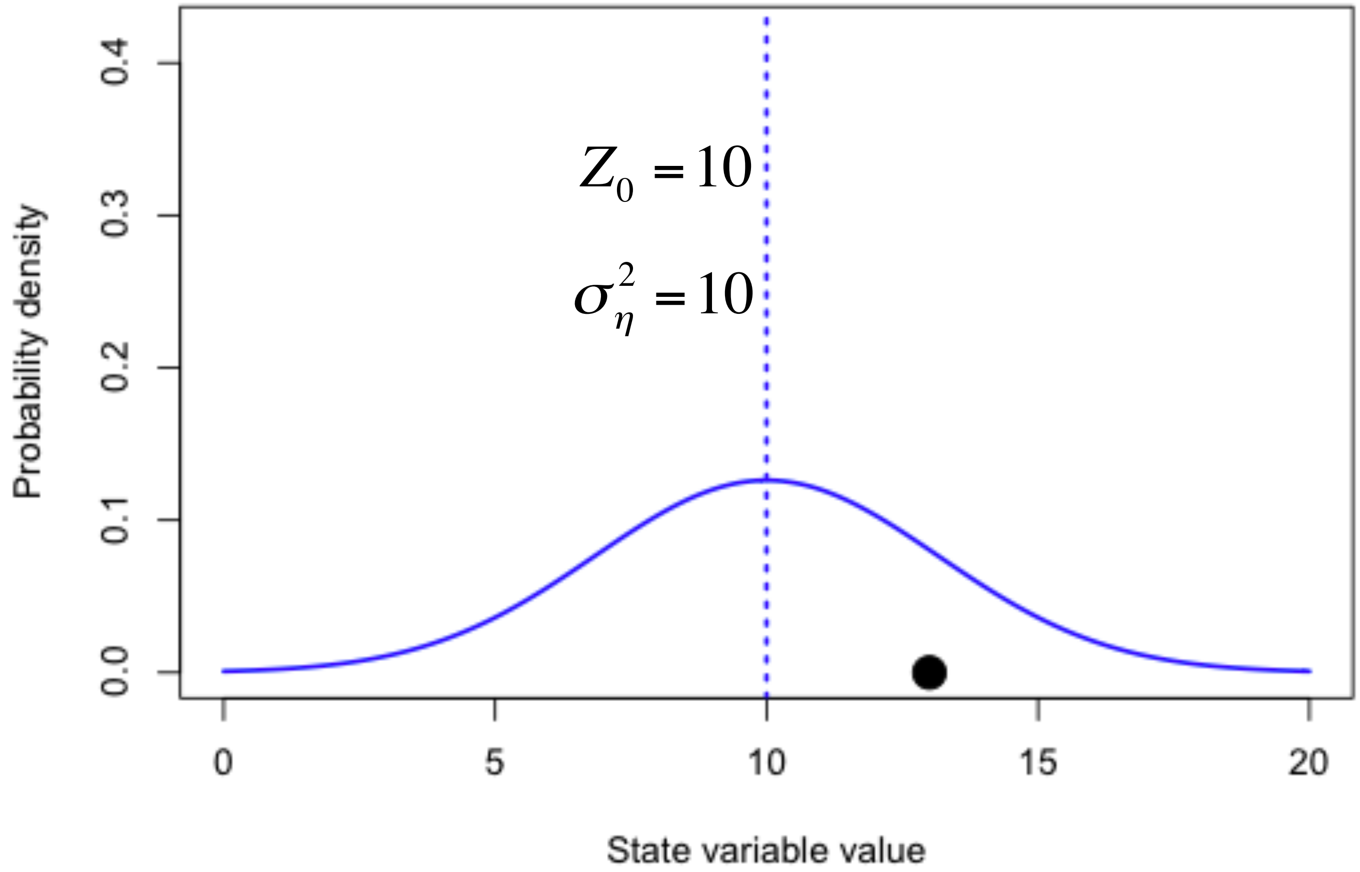
Example 1: Random walk model
Kalman filter (t=1)

$$E\left(Z_1 \mid Y_1\right) = Z_0 + K\left(Y_1 - Z_0\right)$$

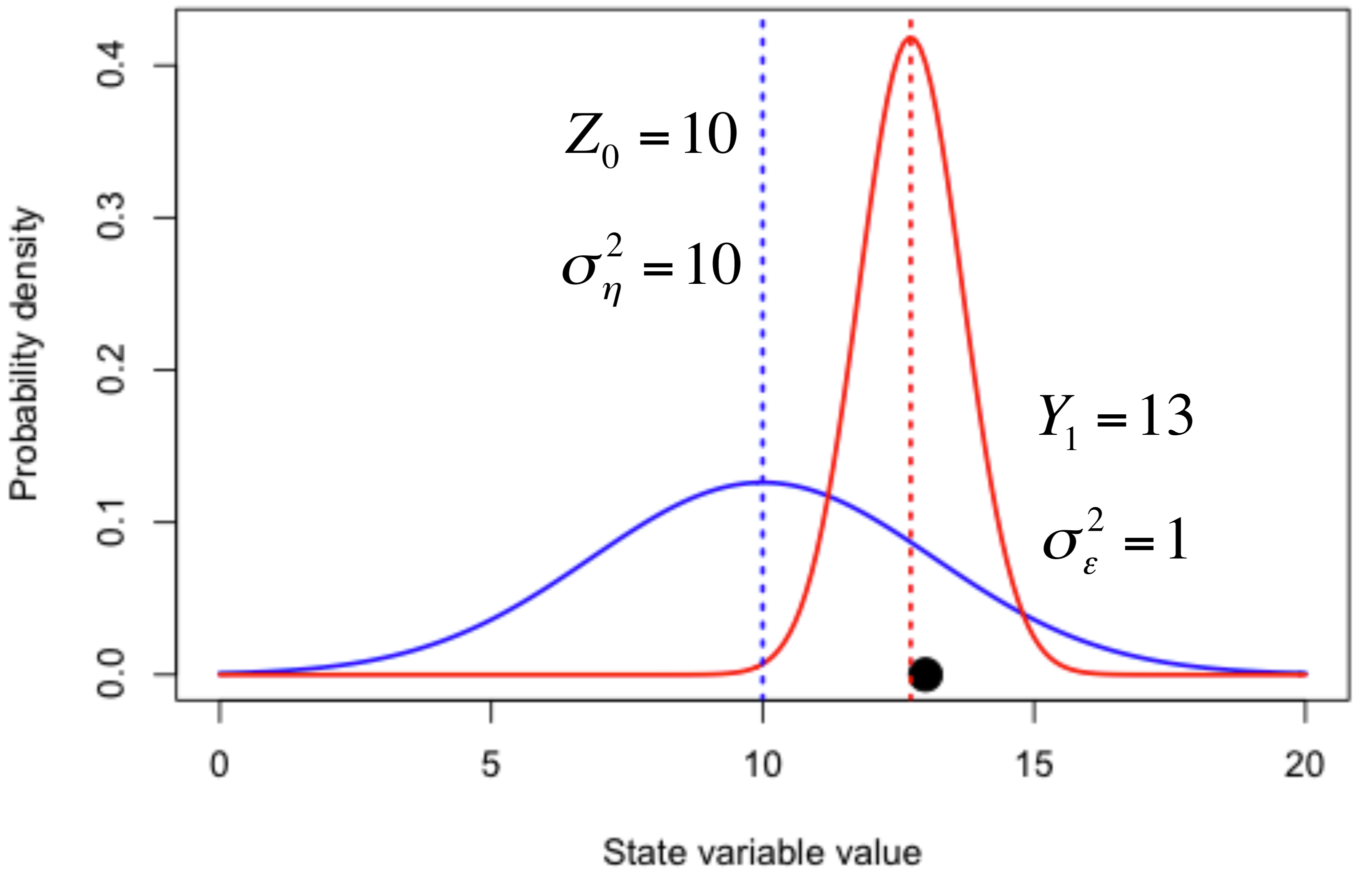
$$V\left(Z_1 \mid Y_1\right) = \left(1 - K\right)\sigma_\eta^2$$

$$K = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}$$

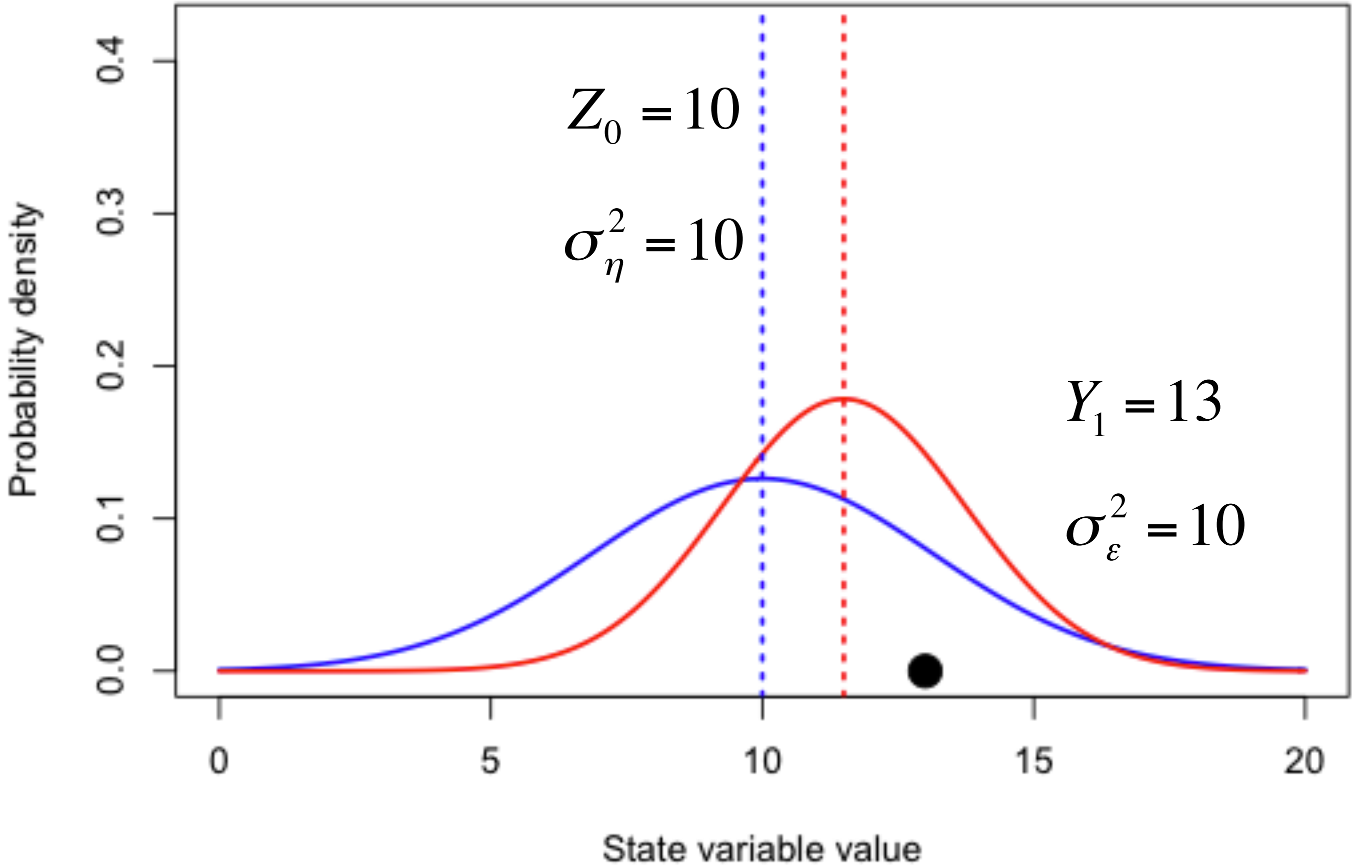
Example 1: Random walk model Kalman filter (t=1)



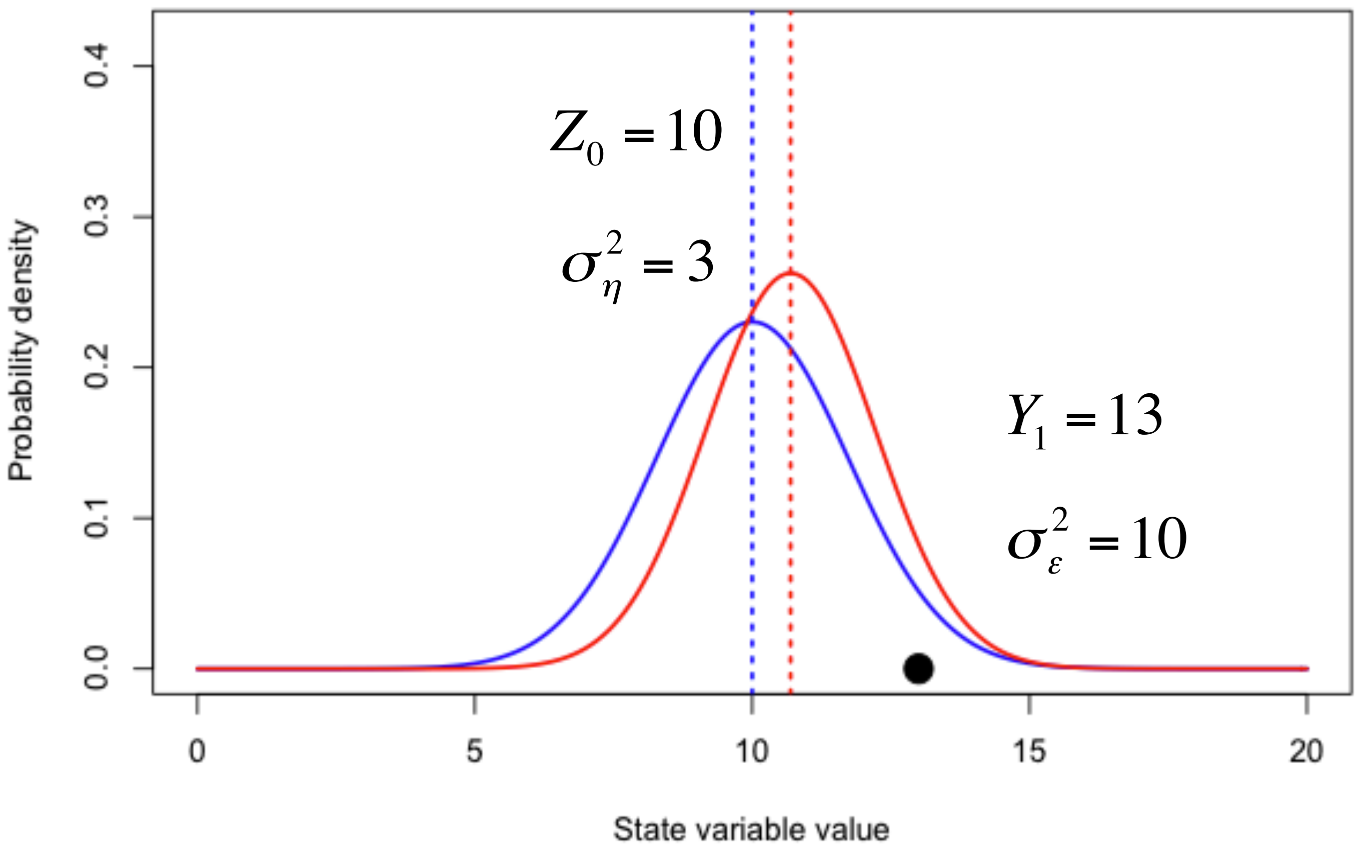
Example 1: Random walk model Kalman filter (t=1)



Example 1: Random walk model
Kalman filter (t=1)



Example 1: Random walk model Kalman filter (t=1)



Example 1: Random walk model (t=1, …, N)

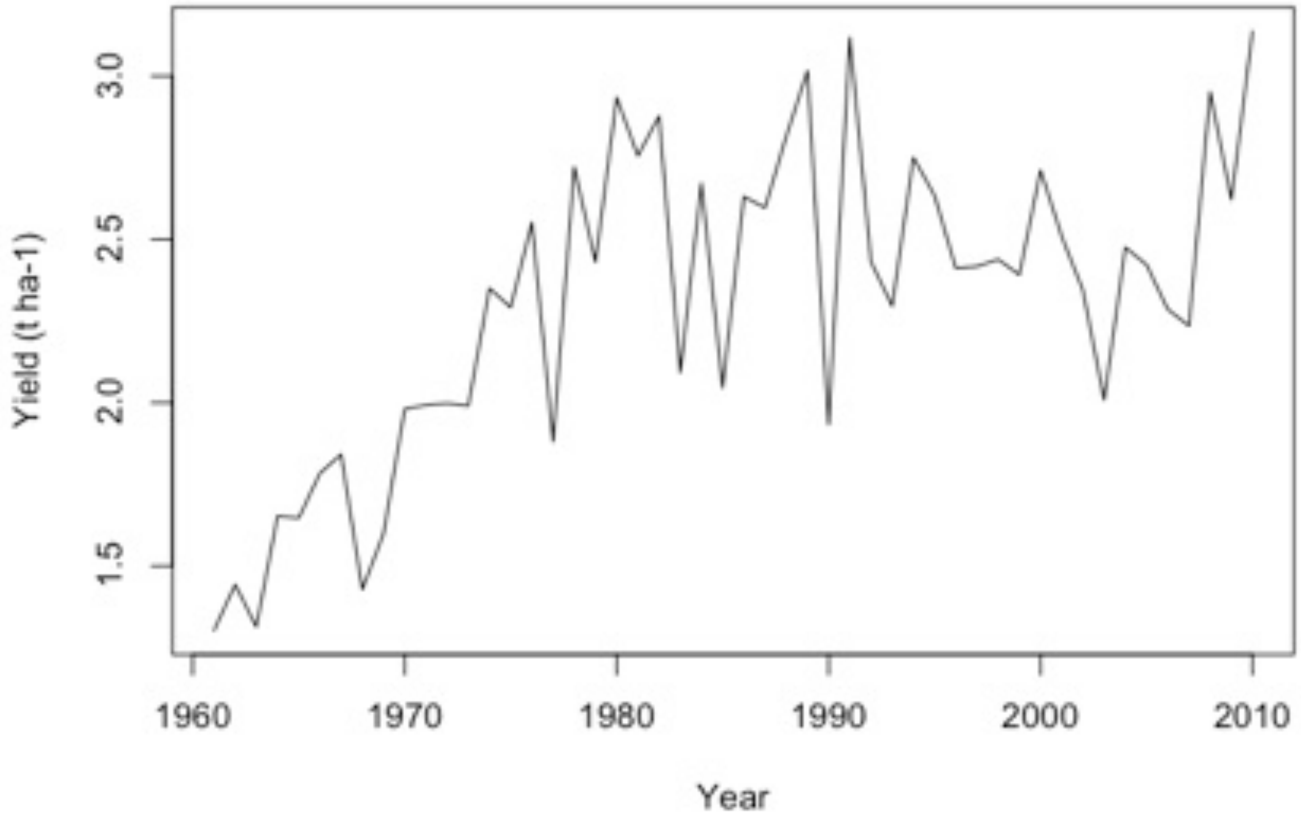
- Observation equation

$$Y_t = Z_t + \varepsilon_t \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

- System equation

$$Z_t = Z_{t-1} + \eta_{t-1} \qquad \eta_{t-1} \sim N\left(0, \sigma_\eta^2\right)$$

Example 1: Random walk model (t=1, ..., N)



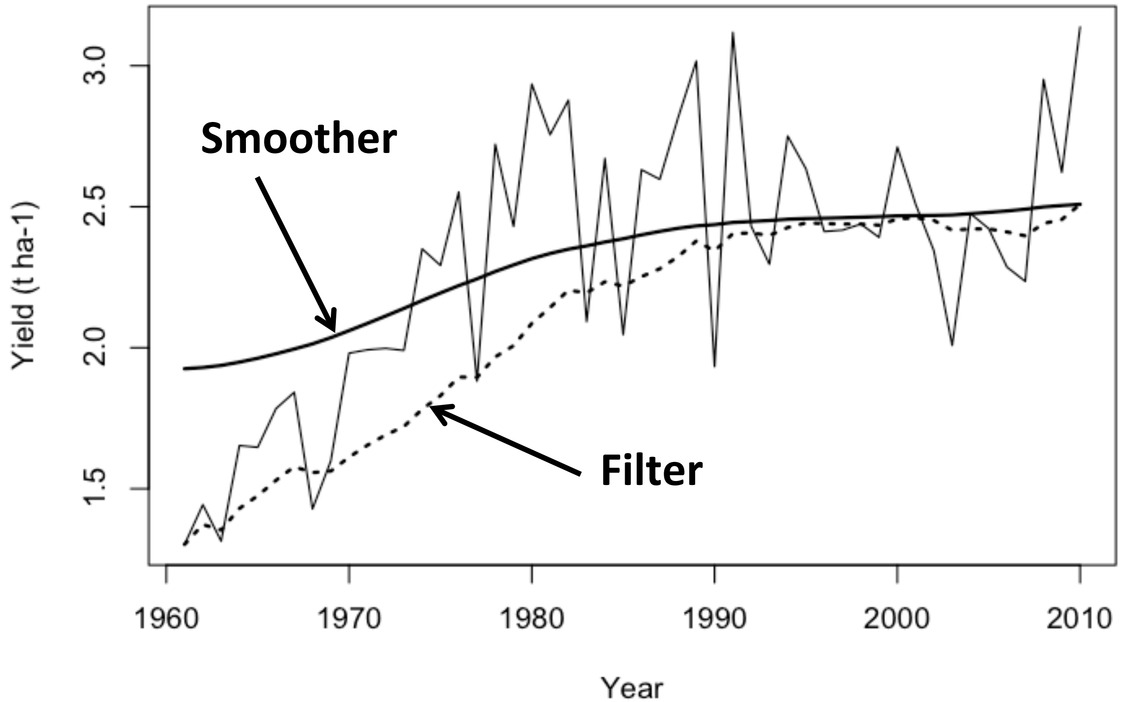
Example 1: Random walk model (t=1, …, N)

$$E\left(Z_t \mid Y_{1:t}\right) = E\left(Z_{t-1} \mid Y_{1:t-1}\right) + K\left(Y_t - E\left(Z_{t-1} \mid Y_{1:t-1}\right)\right)$$

$$V\left(Z_t \mid Y_{1:t}\right) = \left(1 - K\right)\sigma_\eta^2$$

$$K = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}$$

$$\sigma_{\eta}^{2} = 0.007 \quad \sigma_{\varepsilon}^{2} = 1$$



Parameter estimation for Gaussian linear models

- Results of the Kalman filter depends on key parameters
  - Variance of model errors
  - Variance of the observation equation

- These parameters can be estimated from data
  - Maximum likelihood
  - Bayesian method

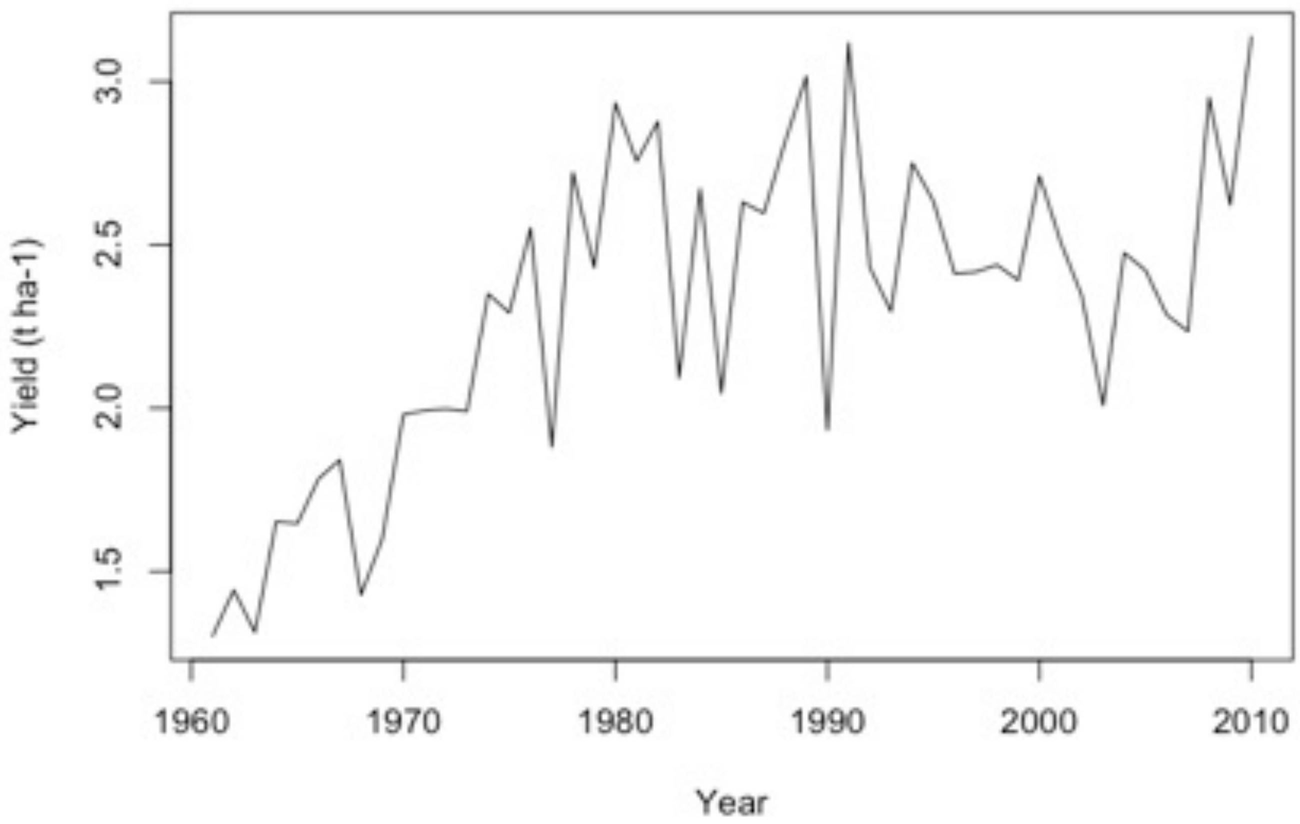
Example 1: Random walk model (t=1, …, N)

- Observation equation

$$Y_t = Z_t + \varepsilon_t \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

- System equation

$$Z_t = Z_{t-1} + \eta_{t-1} \qquad \eta_{t-1} \sim N\left(0, \sigma_\eta^2\right)$$



Yield (t ha-1)
1.5
2.0
2.5
3.0
1960
1970
1980
1990
2000
2010
Year

Example 1: Random walk model (t=1, …, N)

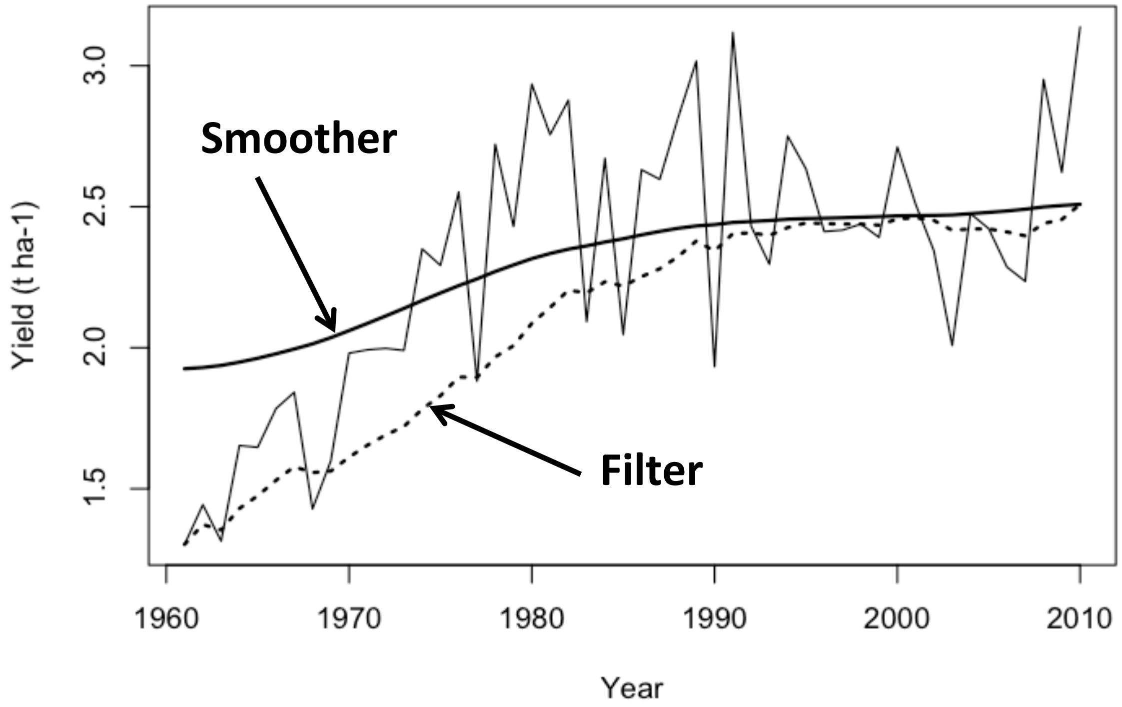
$$E\left(Z_t \mid Y_{1:t}\right) = E\left(Z_{t-1} \mid Y_{1:t-1}\right) + K\left(Y_t - E\left(Z_{t-1} \mid Y_{1:t-1}\right)\right)$$

$$V\left(Z_t \mid Y_{1:t}\right) = \left(1 - K\right)\sigma_\eta^2$$

$$K = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}$$

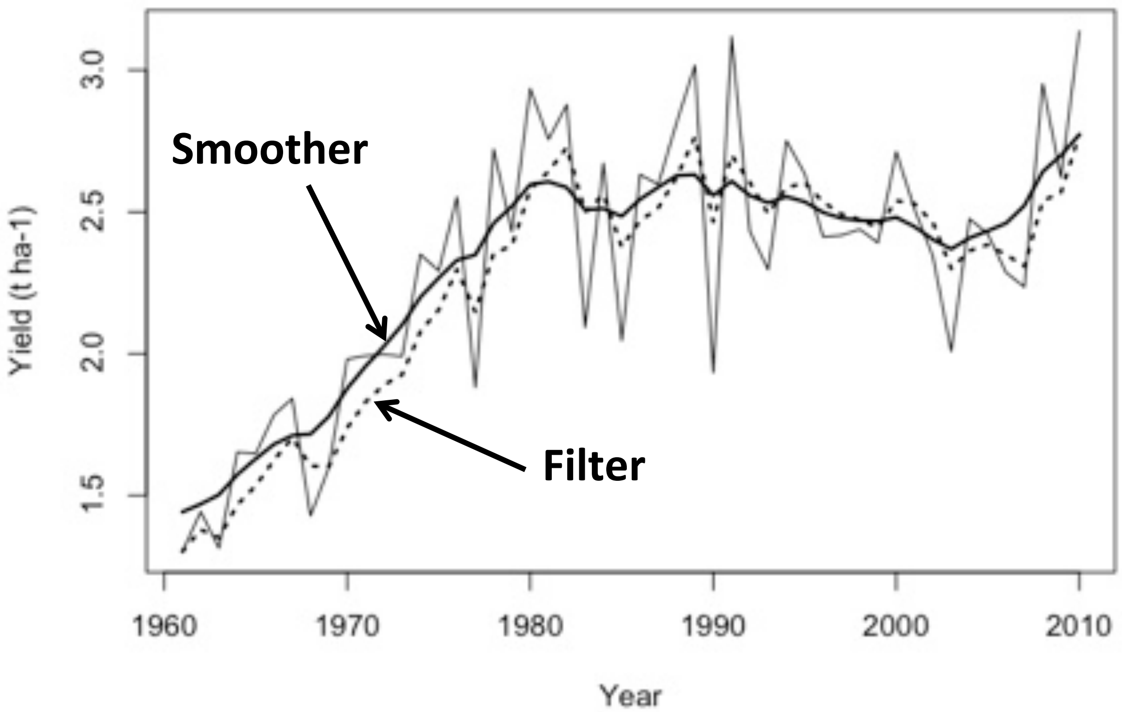
Arbitrary parameter values

$$\sigma_{\eta}^{2} = 0.007 \qquad \sigma_{\varepsilon}^{2} = 1$$



Maximum likelihood estimation

$$\sigma_\eta^2 = 0.014 \quad \sigma_\varepsilon^2 = 0.07$$



Example 2: Model with dynamic time trend

- **Observation equation**

$$Y_t = FZ_t + \varepsilon_t \qquad Z_t = \begin{pmatrix} a_t \\ b_t \end{pmatrix} \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$F = (1,0)$$

- **System equation**

$$Z_t = GZ_{t-1} + \eta_{t-1} \qquad \eta_{t-1} \sim N\left(0, \Sigma\right)$$

$$G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}$$

Example 2: Model with dynamic time trend

- **Observation equation**

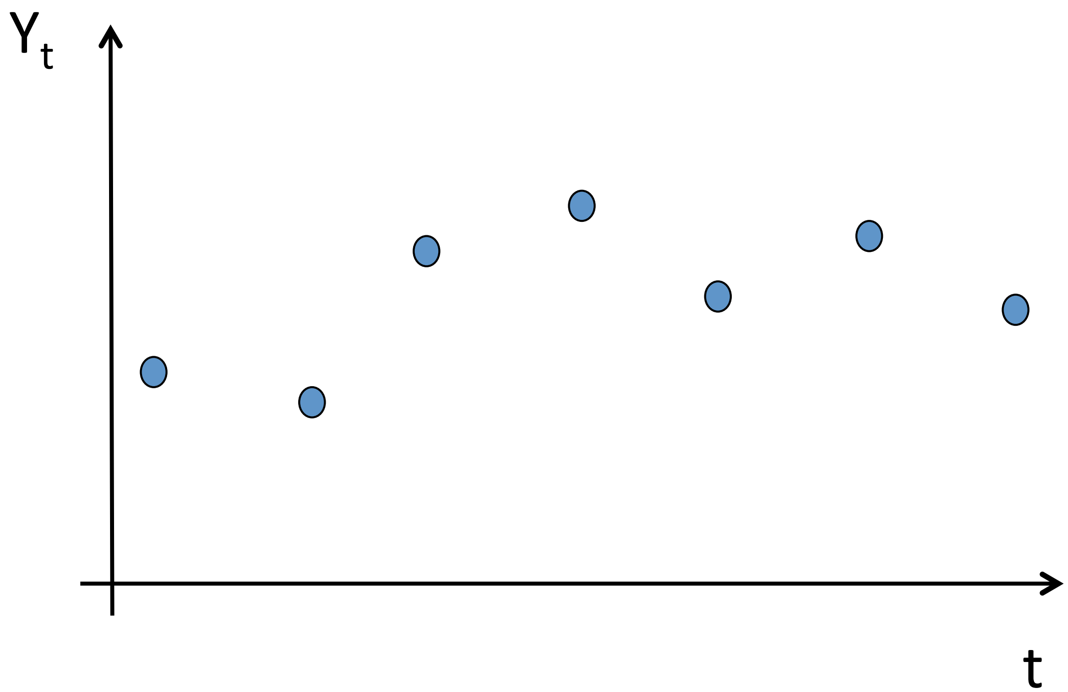
$$Y_t = a_t + \varepsilon_t \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

- **System equation**

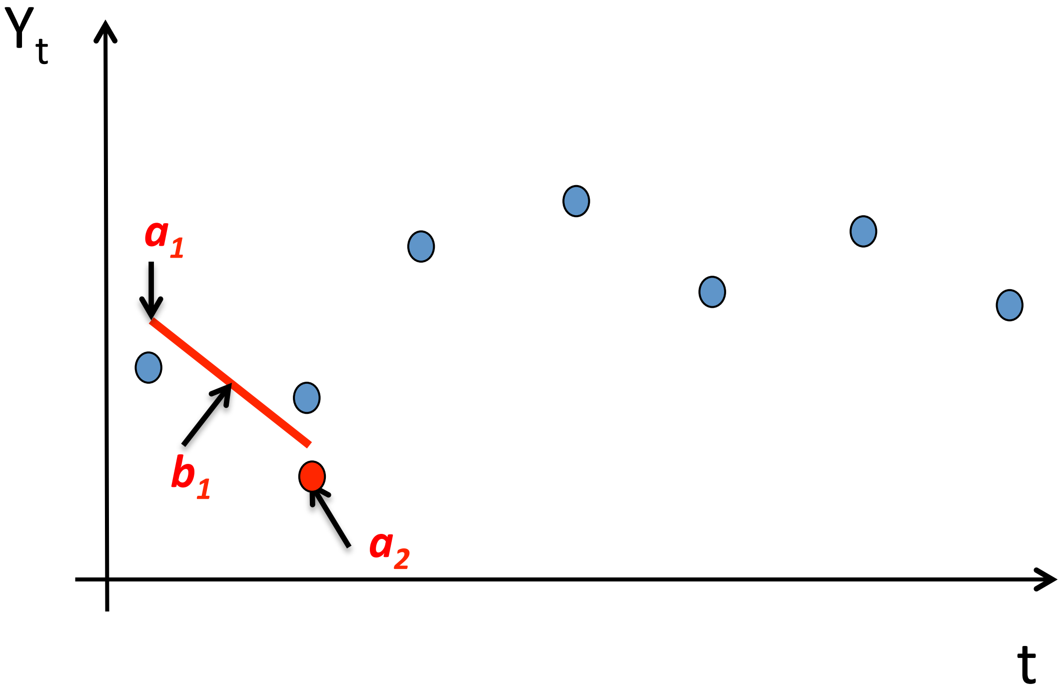
$$a_t = a_{t-1} + b_{t-1} + \eta_{t-1}^{(a)} \qquad \eta_{t-1}^{(a)} \sim N\left(0, \sigma_a^2\right)$$

$$b_t = b_{t-1} + \eta_{t-1}^{(b)} \qquad \eta_{t-1}^{(b)} \sim N\left(0, \sigma_b^2\right)$$

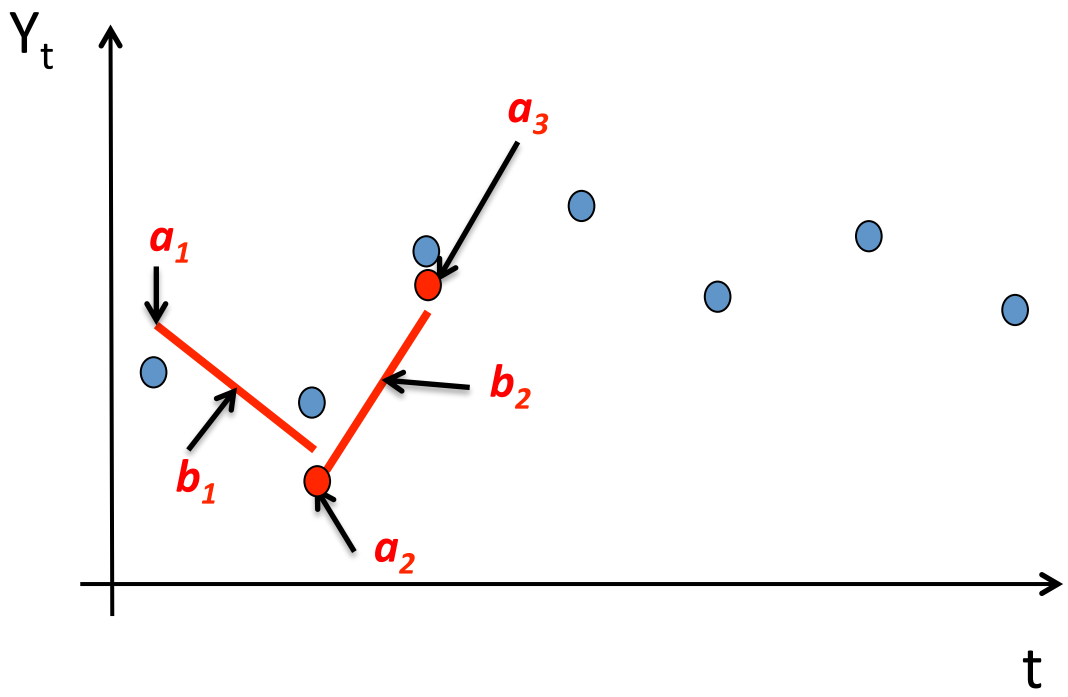
Example 2: Model with dynamic time trend



Example 2: Model with dynamic time trend



Example 2: Model with dynamic time trend

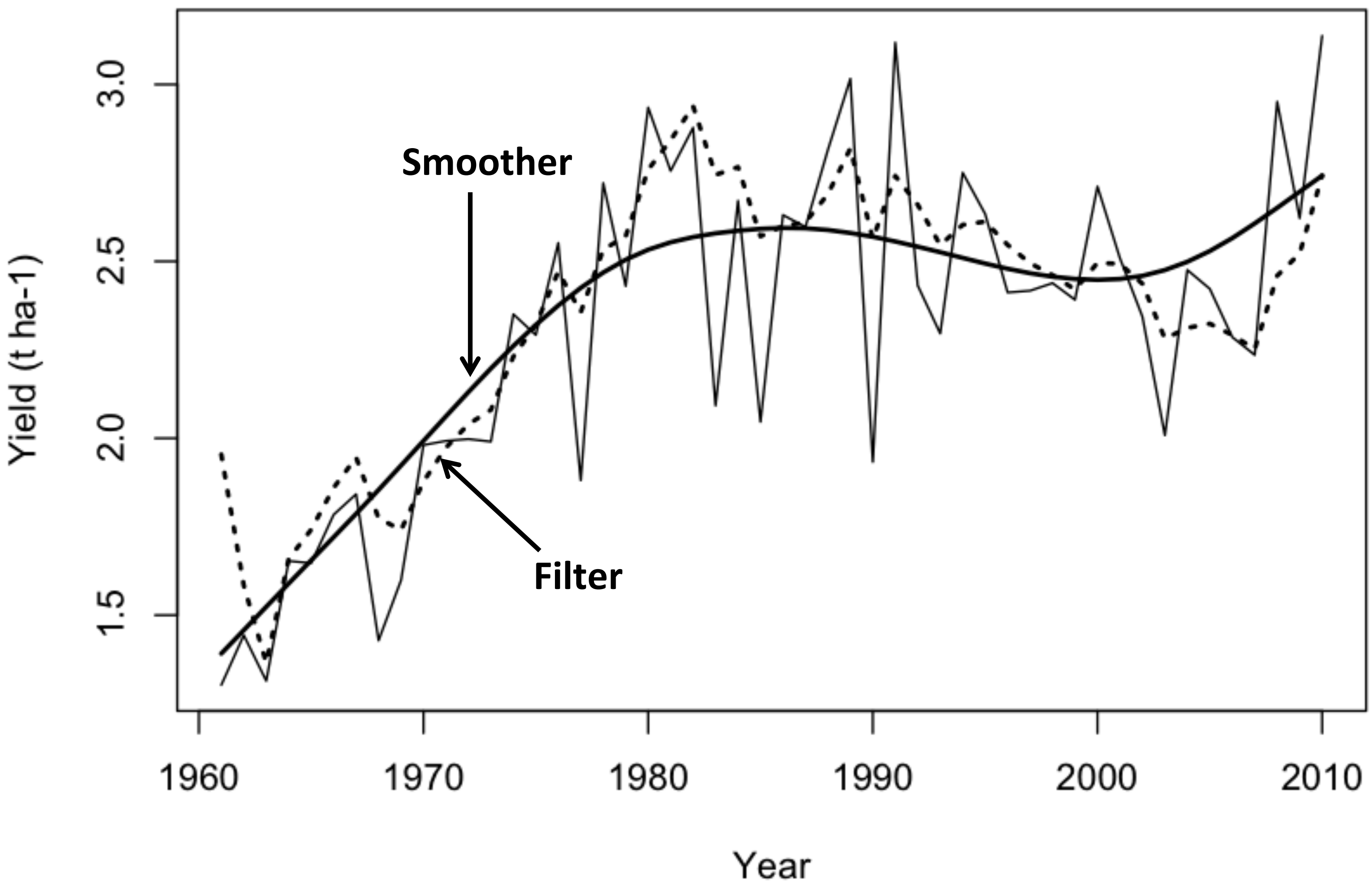


Example 2: Model with dynamic time trend

	Level	Trend
Filter	$E\left(a_t \mid Y_{1:t}\right)$ $V\left(a_t \mid Y_{1:t}\right)$	$E\left(b_t \mid Y_{1:t}\right)$ $V\left(b_t \mid Y_{1:t}\right)$
Smoother	$E\left(a_t \mid Y_{1:N}\right)$ $V\left(a_t \mid Y_{1:N}\right)$	$E\left(b_t \mid Y_{1:N}\right)$ $V\left(b_t \mid Y_{1:N}\right)$

Maximum likelihood estimation

$$\sigma_a^2 = 1.11E-09 \quad \sigma_b^2 = 2.38E-04 \quad \sigma_\varepsilon^2 = 0.077$$



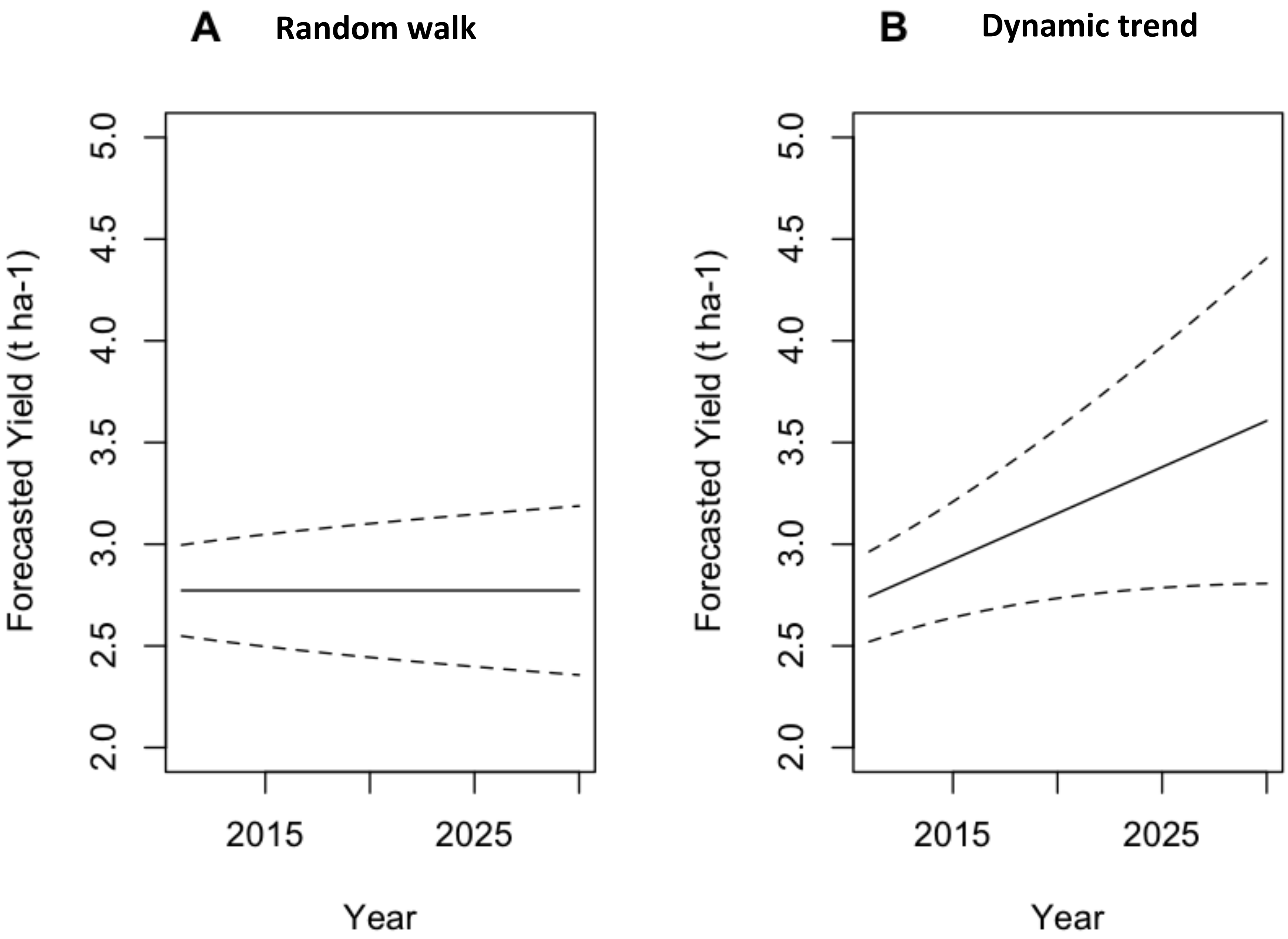
Predictions

- Random walk

$$Y_{t+K}^{(P)} = E\left(a_N \mid Y_{1:N}\right)$$

- Dynamic time trend

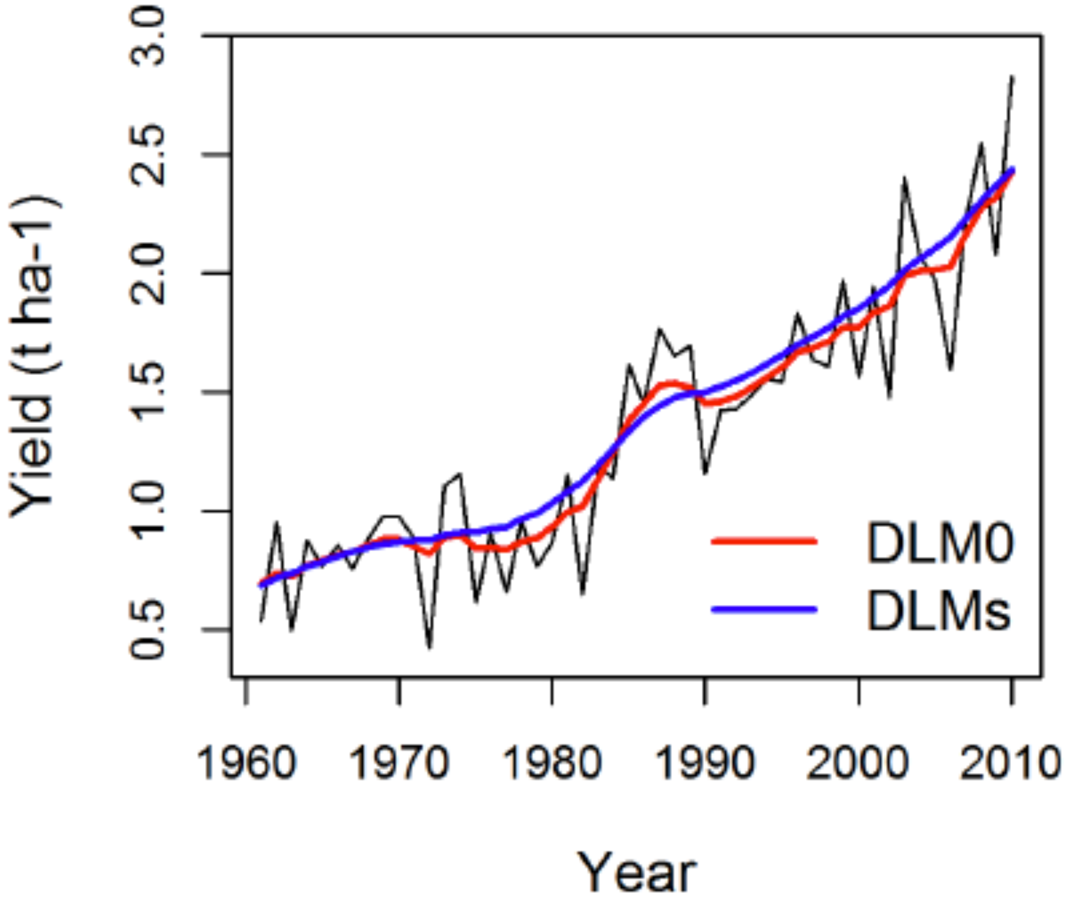
$$Y_{t+K}^{(P)} = E\left(a_N \mid Y_{1:N}\right) + E\left(b_N \mid Y_{1:N}\right) \times K$$


A
Random walk
B
Dynamic trend
Forecasted Yield (t ha-1)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
2015
2025
Year

A large scale application

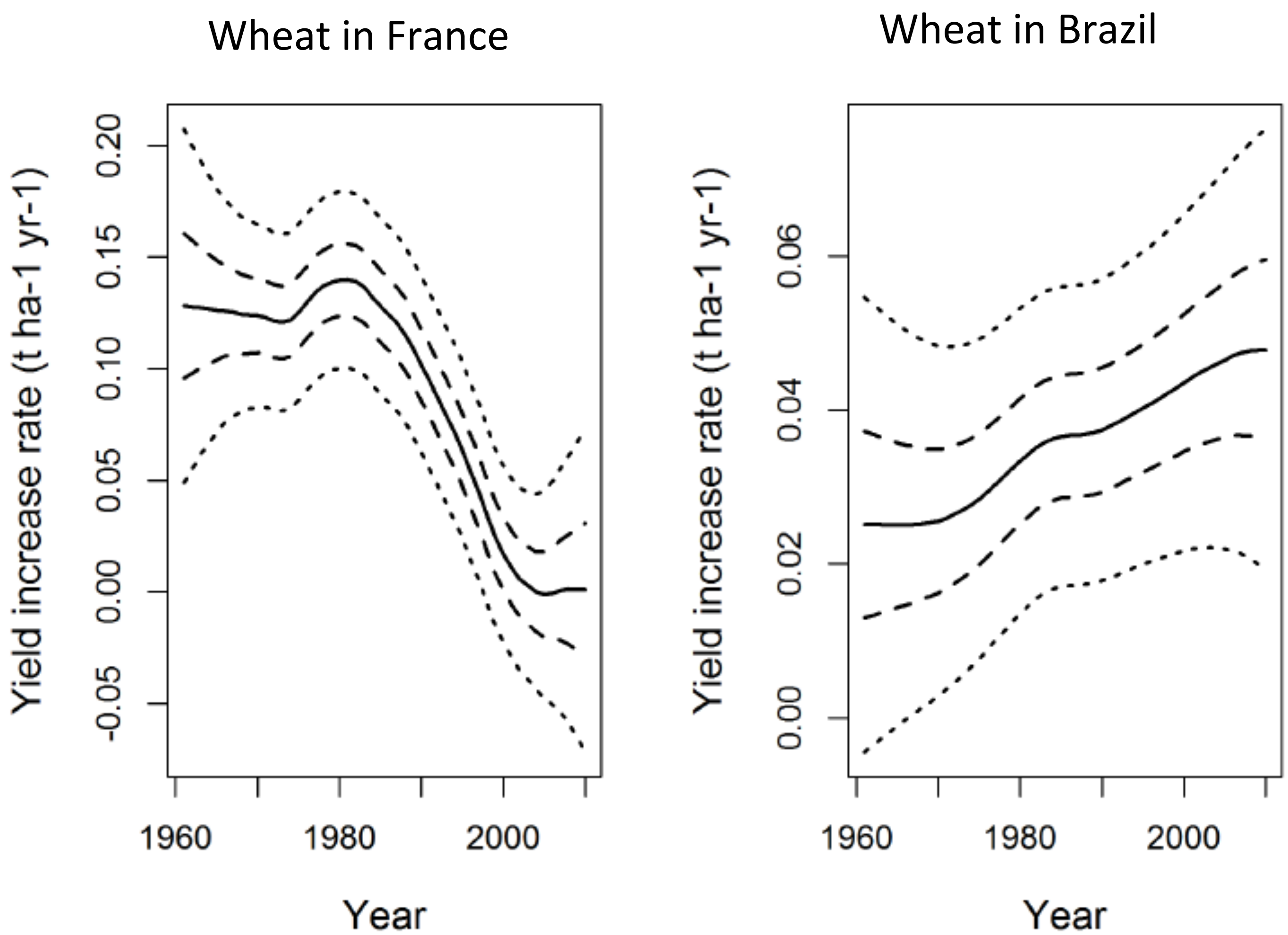
(Michel & Makowski, 2013)

Wheat in Brazil



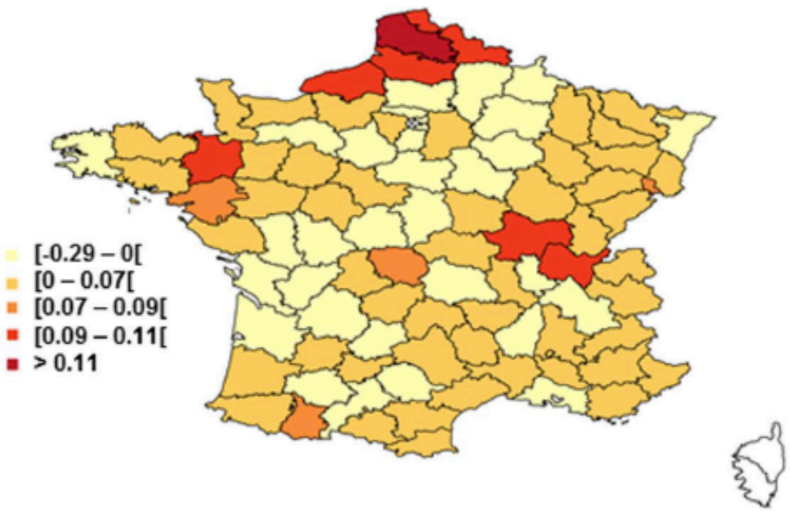
A large scale application

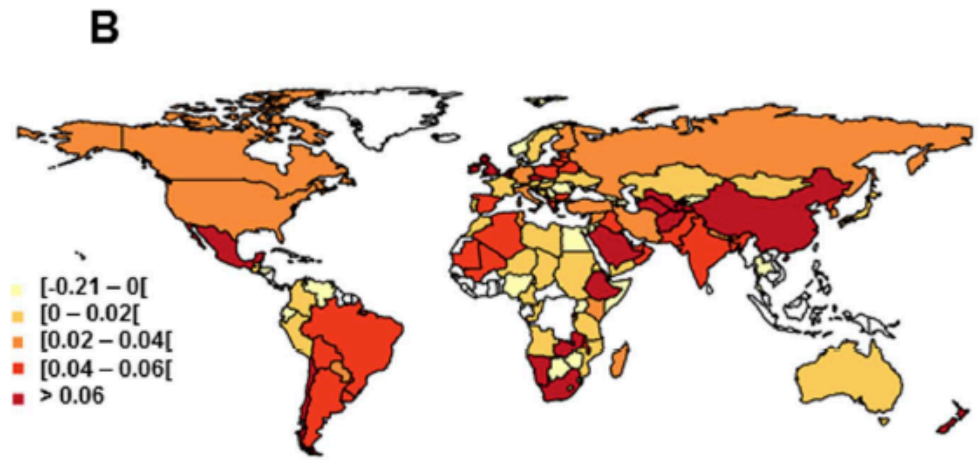
(Michel & Makowski, 2013)



A large scale application

(Michel & Makowski, 2013)





Package dlm (dynamic linear model)

- Petris (2010)
- Implement dynamic linear Gaussian models
- Estimate parameters by maximum likelihood
- Filtering and smoothing

Random walk model

- **Observation equation**

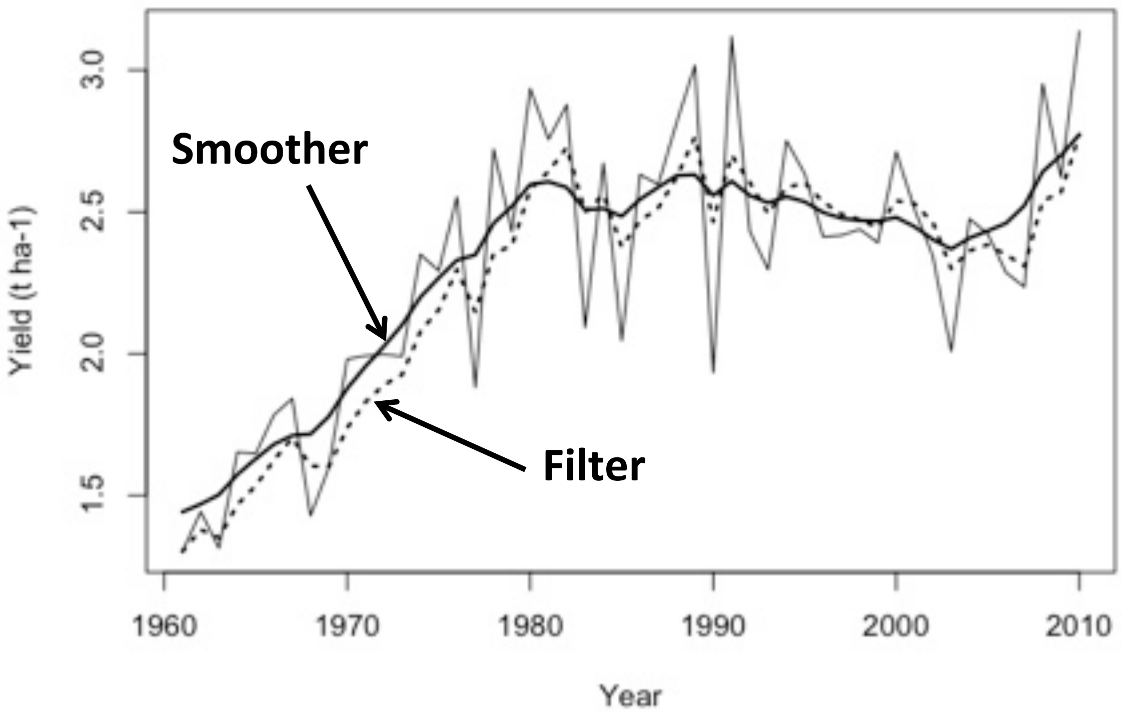
$$Y_t = Z_t + \varepsilon_t \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

- **System equation**

$$Z_t = Z_{t-1} + \eta_{t-1} \qquad \eta_{t-1} \sim N\left(0, \sigma_\eta^2\right)$$

Maximum likelihood estimation

$$\sigma_\eta^2 = 0.014 \qquad \sigma_\varepsilon^2 = 0.07$$



```
MyModel<-function(x) {
return(dlmModPoly(1, dV=exp(x[1]), dW=exp(x[2])))}
```

The first input of `dlmModPoly` defines the number of terms of the linear function used in the observation equation (i.e., only one term here). The R function `MyModel` defines a random-walk model including two parameters, namely σ_{ε}^{2} and σ_{η}^{2}. In `MyModel`, these two parameters are called dV and dW respectively. dV and dW are keywords for `dlm`.

```
FittedModel<-MyModel(c(0,-5))
```

The vector `c(0,-5)` includes two elements, `x[1]` and `x[2]`, related to dV and dW using an exponential function in order to constrain the variances to take positive values only. The vector is passed to the function `MyModel` to define a random walk model with specific value of σ_ε^2 and σ_η^2. This model is called `FittedModel`. Once the parameter values specified, the Kalman filter and smoother are implemented as follows:

```
YieldFilter<-dlmFilter(Yield, FittedModel)
YieldSmooth<-dlmSmooth(Yield, FittedModel)
```

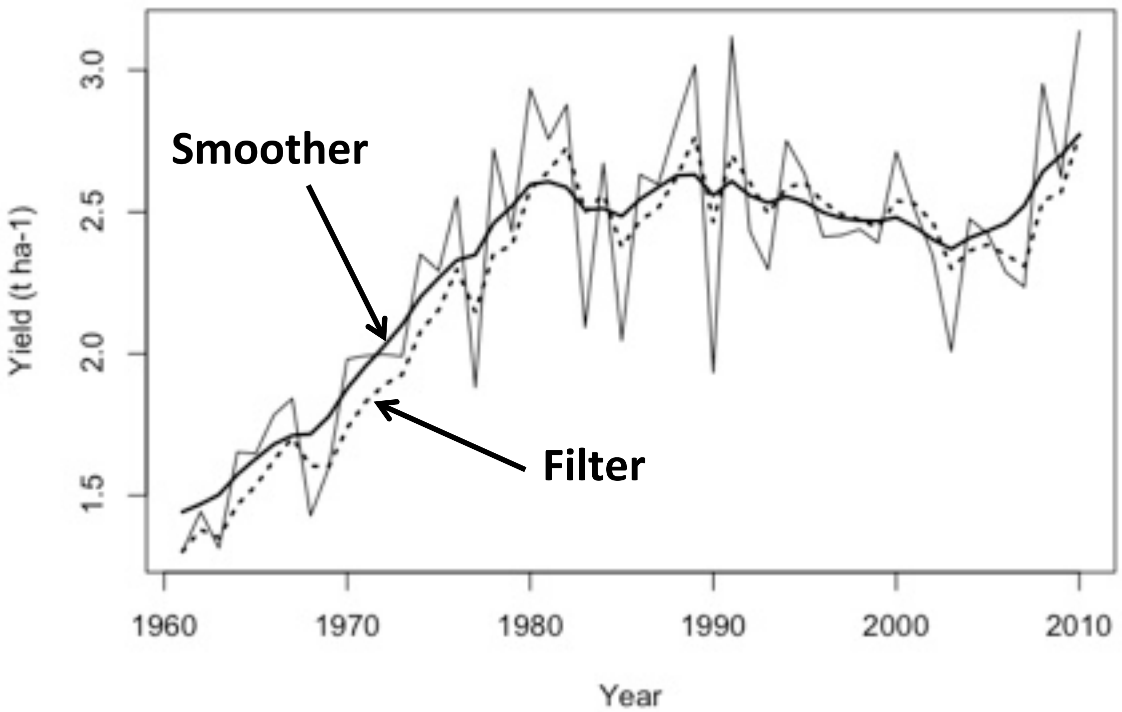
```
plot(Year,Yield,ylab="Yield (t ha-1)", type="l",lwd=1)
lines(Year,YieldFilter$m[-1],lwd=2, lty=3)
lines(Year,YieldSmooth$s[-1],lwd=2)
```

Maximum likelihood estimation

The two parameters of the random-walk model σ_ε^2 and σ_η^2 are estimated with the following R code `dlmMLE(Yield,parm=c(0,0), build=MyModel)`. The algorithm implemented by `dlmMLE` uses $\sigma_\varepsilon^2 = \sigma_\eta^2 = \exp(0) = 1$ as starting values and, after few iterations, returns two estimated parameter values: -2.65 and -4.26. These two values correspond to the estimated values of `x[1]` and `x[2]`. The estimated values for σ_ε^2 and σ_η^2 are equal to exp(-2.65) and exp(-4.26) respectively.

```
FittedModel<-MyModel(fitMyModel$par)
YieldFilter<-dlmFilter(Ym, FittedModel)
YieldSmooth<-dlmSmooth(Ym, FittedModel)
```

$$\sigma_{\eta}^{2} = 0.014 \qquad \sigma_{\varepsilon}^{2} = 0.07$$



Model with dynamic time trend

- **Observation equation**

$$Y_t = FZ_t + \varepsilon_t$$

$$F = (1,0)$$

$$\varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

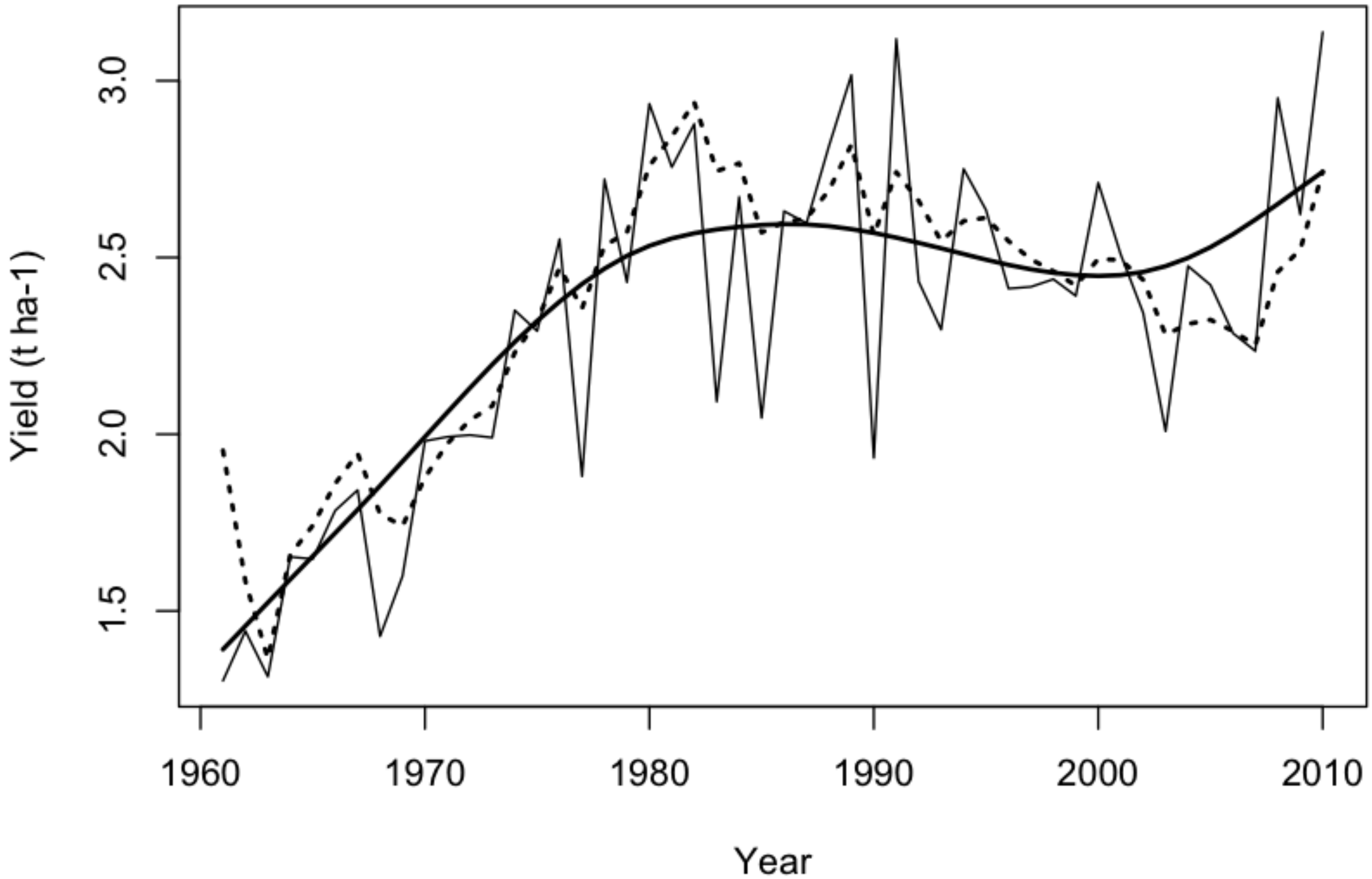
- **System equation**

$$Z_t = GZ_{t-1} + \eta_{t-1}$$

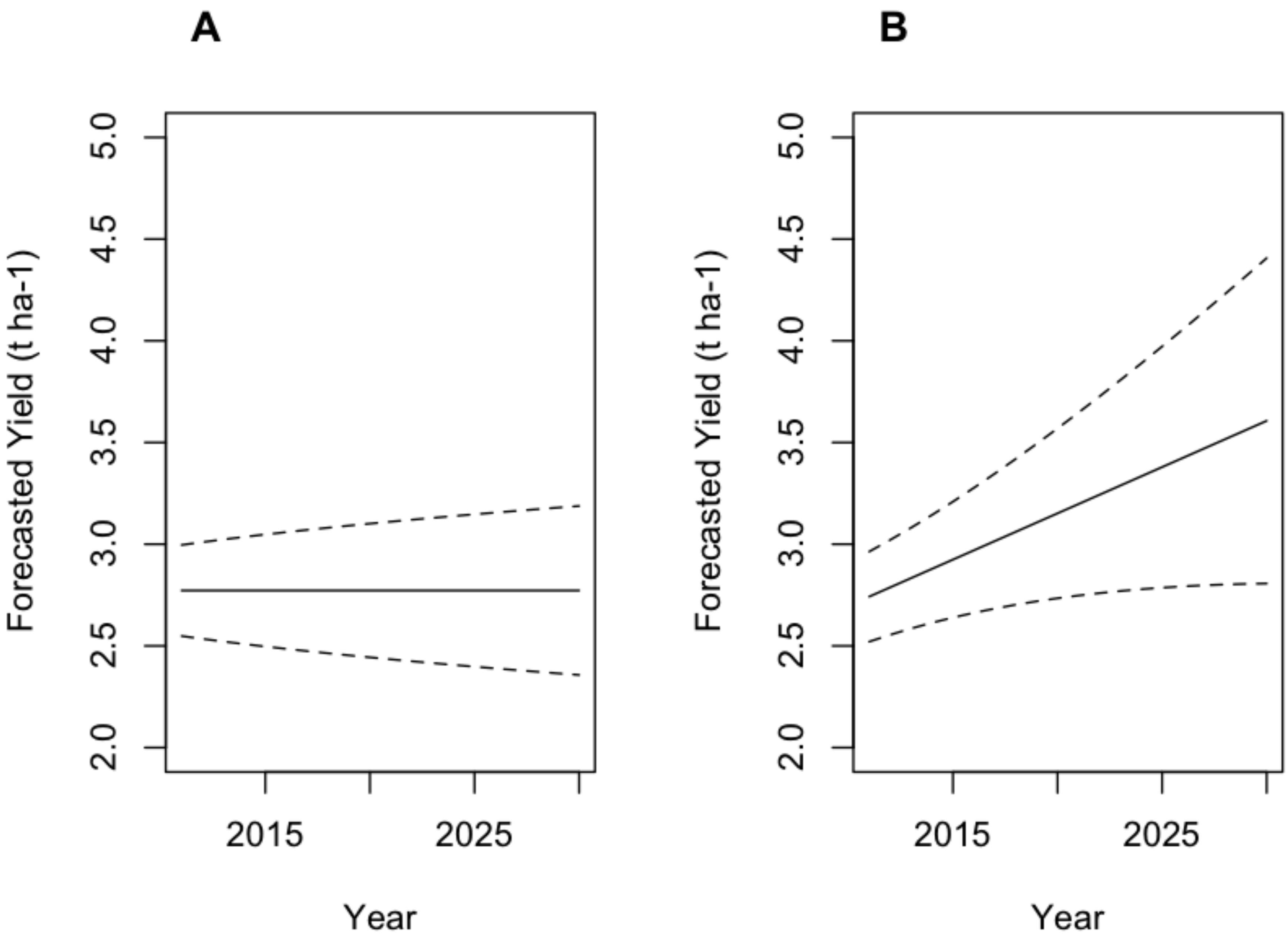
with $Z_t = \begin{pmatrix} a_t \\ b_t \end{pmatrix}$, $G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $\eta_{t-1} \sim N(0, \sum)$,

$$\sum = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}$$

```
#Definition of model 2
MyModel<-function(x) {
        return(dlmModPoly(2, dV=exp(x[1]), dW=c(exp(x[2]), exp(x[3]))))
        }

#Estimation of parameters of the model

fitMyModel<-dlmMLE(Yield,parm=c(0,0,0), build=MyModel)
print(fitMyModel)

#aVar<-solve(fitMyModel$hessian)
#sqrt(diag(aVar))

#Filtrage, Lissage, Prediction

FittedModel<-MyModel(fitMyModel$par)
#FittedModel<-MyModel(c(0,-5,-5))

YieldFilter<-dlmFilter(Yield, FittedModel)
YieldSmooth<-dlmSmooth(Yield, FittedModel)
YieldFilter_2<-YieldFilter


par(mfrow=c(1,1),oma=c(5,1,5,1))
plot(Year,Yield,ylab="Yield (t ha-1)", type="l",lwd=1)
lines(Year,YieldFilter$m[,1][-1]+YieldFilter$m[,2][-1],lwd=2, lty=3)
lines(Year,YieldSmooth$s[,1][-1]+YieldSmooth$s[,2][-1],lwd=2)
```


Yield (t ha-1)
1.5
2.0
2.5
3.0
1960
1970
1980
1990
2000
2010
Year

```
#Predictions for next 20 years
foreYield_1<-dlmForecast(YieldFilter_1,nAhead=20)
foreYield_2<-dlmForecast(YieldFilter_2,nAhead=20)

FuturYear<-seq(Year[50]+1,Year[50]+20)

par(mfrow=c(1,2),oma=c(5,1,5,1))
plot(FuturYear,foreYield_1$f,xlab="Year", ylab="Forecasted Yield (t ha-1)", ylim=c(2,5), type="l")
#Confidence intervals
lines(FuturYear,foreYield_1$f+qnorm(0.75)*sqrt(unlist(foreYield_1$Q)),lty=2)
lines(FuturYear,foreYield_1$f-qnorm(0.75)*sqrt(unlist(foreYield_1$Q)),lty=2)
title("A                              ")

plot(FuturYear,foreYield_2$f,xlab="Year", ylab="Forecasted Yield (t ha-1)", ylim=c(2,5), type="l")

#Confidence intervals
lines(FuturYear,foreYield_2$f+qnorm(0.75)*sqrt(unlist(foreYield_2$Q)),lty=2)
lines(FuturYear,foreYield_2$f-qnorm(0.75)*sqrt(unlist(foreYield_2$Q)),lty=2)
title("B                              ")
```


A
B
Forecasted Yield (t ha-1)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
2015
2025
Year
Forecasted Yield (t ha-1)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
2015
2025
Year

Summary

- Kalman filter and smoother can be applied using dynamic linear gaussian models
- These models can handle a great diversity of situations (see practical session)
- They can be implemented using R (see practical session)

Outline

1. Objective & main principles
2. Model specification
3. Filter and smoother using Gaussian dynamic linear models
4. Conclusion

Conclusion (1)

- Data assimilation is a powerful tool for updating dynamic models

- Filtering and smoothing allow one to combine a model and measurements in useful ways, taking into account the uncertainties in each.

- Filtering is useful for estimating sequentially in time the values of one or several state variables, whereas smoothing can be used to estimate past values of state variables using all available measurements.

Conclusion (2)

- To implement these methods, it is necessary to define the system models using two different equations; an observation equation (relating observation to state variables) and a system equation (describing the dynamic of the state variables).

Conclusion (3)

- Filtering and smoothing use these equations to calculate the expected values and variances of the state variables conditionally to one or several measurements.
- For linear Gaussian models, the expected values and variances can be computed analytically and the dlm R package makes the calculations very accessible.

References

Michel L., Makowski D. (2013). Comparison of statistical models for analyzing wheat yield time series. Plos one 8(10)

Petris G, Campagnoli P, Petrone S (2009) Dynamic Linear Models with R.

Springer. 258 p