The Abdus Salam International Centre for Theoretical Physics

2453-11

School on Modelling Tools and Capacity Building in Climate and Public Health

15 - 26 April 2013

Modelling Time

SA CARVALHO Marilia

PROCC FIOCRUZ
Avenida Brasil 4365, Rio De Janeiro 21040360
BRAZIL

Modelling Time

Marilia Sá Carvalho

Fundação Oswaldo Cruz

Outline

Outline

1. Introduction
   - Motivating Example – Leptospirosis

Statistical Analysis

- Exploratory Analysis to:
  - describe the data
  - support the selection of appropriate statistical techniques
- Hypothesis testing:
  - Does this observed pattern differ from... ?
- Modelling:
  - What is the effect of rainfall, humidity and temperature on the number of cases of malaria?

References

- Cryer, J.D.; Chan, K-S. *Time Series Analysis: With Applications in R*. Springer Texts in Statistics, 2010, 2nd Ed.
- Hastie, T.; Tibshirani, R. *Generalized Additive Models*. Chapman & Hall, 1990.
- Wood, S.N. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science Series, 2006.
- Faraway, J.J. *Extending the Linear Model with R*. Chapman & Hall/CRC Texts in Statistical Science Series, 2006.

Time Series

- A sequence of data points, measured typically at successive points in time spaced at uniform time intervals
- Time series analysis → methods for analysing time series data in order to extract meaningful statistics
- Natural temporal ordering can result in serial dependence → dependence of each time point on previous points
- Components:
  - Trend
  - Seasonality and cyclical patterns
  - Time dependence structure

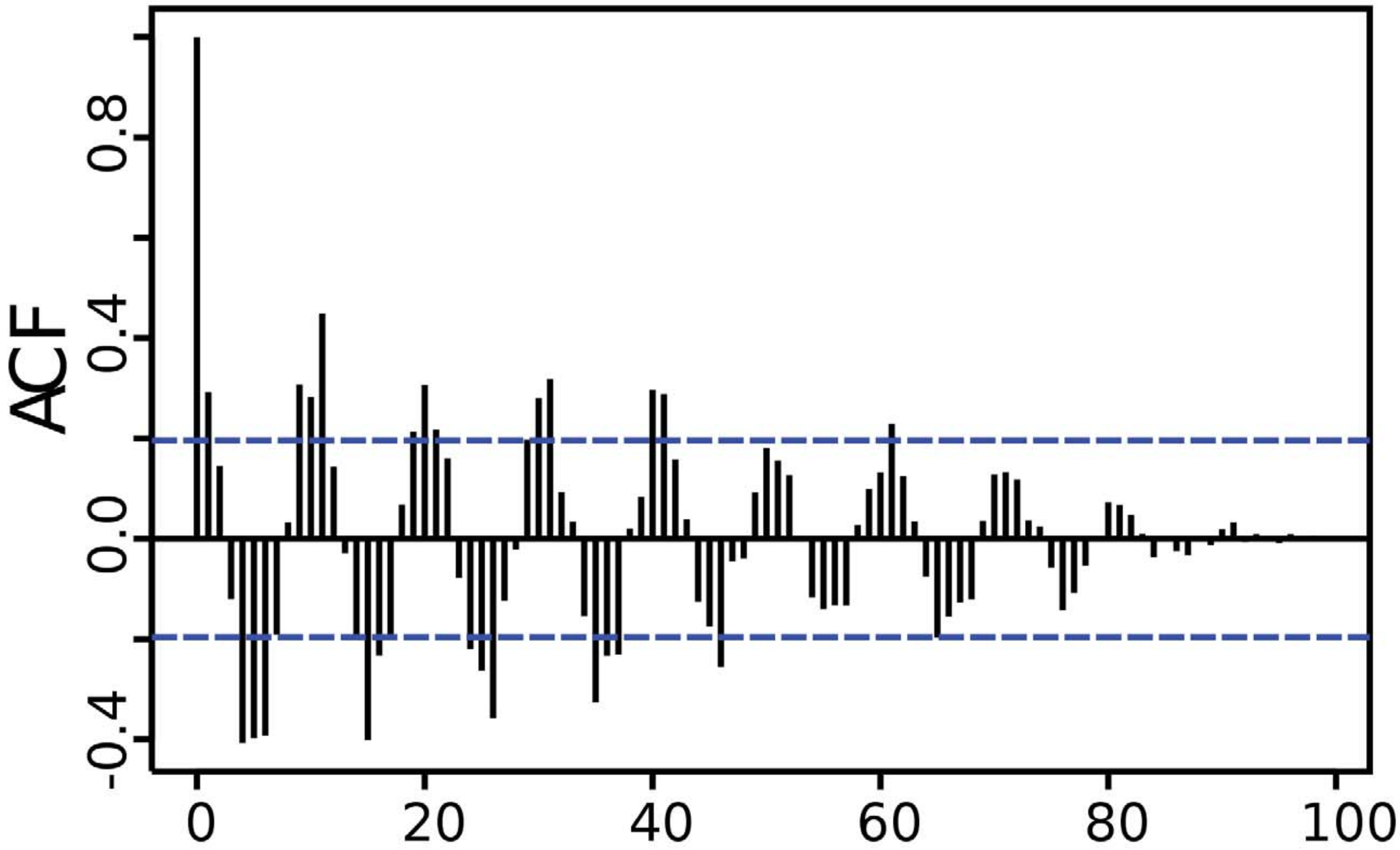
Time dependence structure: autocorrelation

- Autocorrelation (or autocovariance) is a measure of similarity of the event over time with itself previously
- It is the correlation between values of a random process at different times

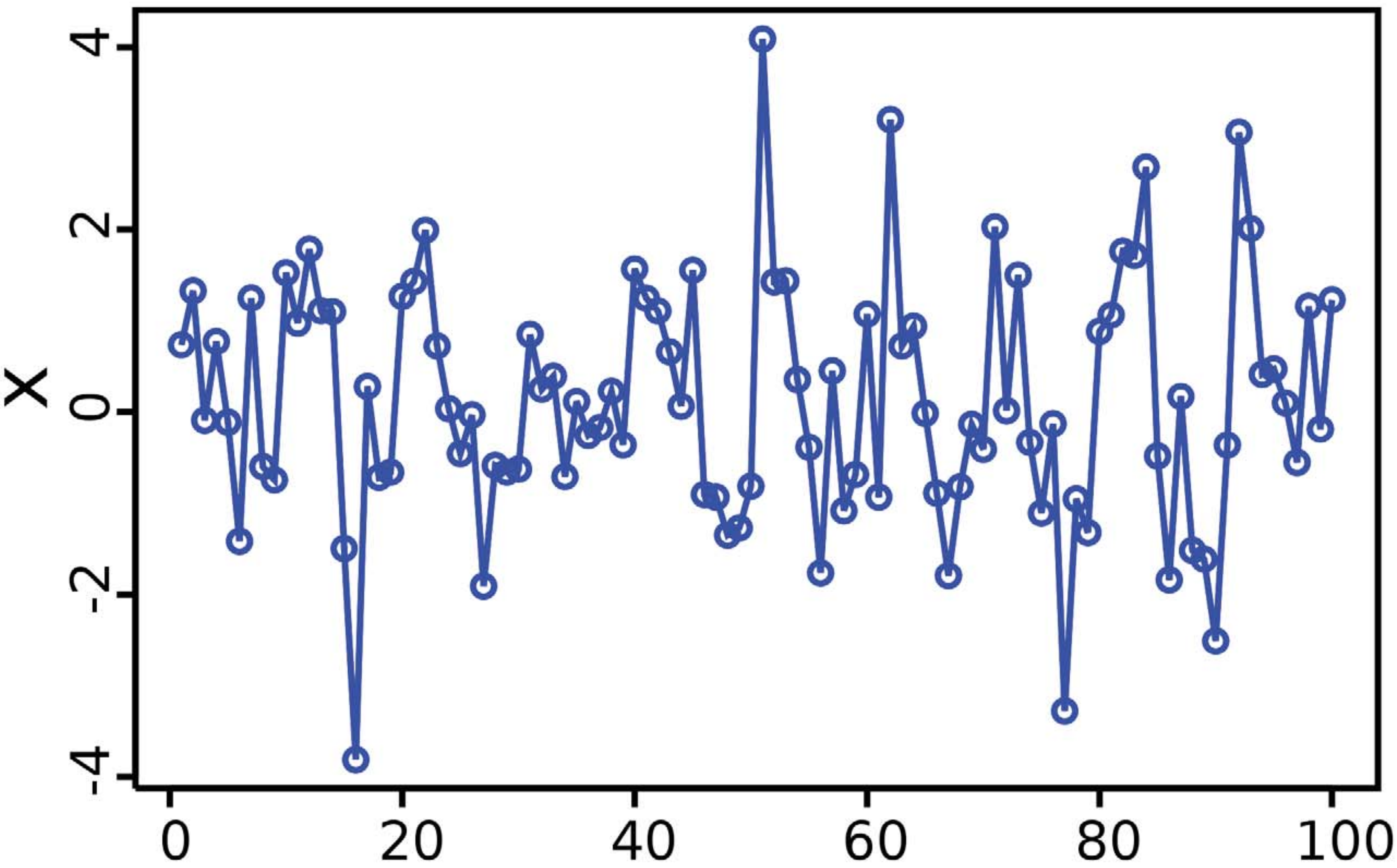
$$r_k = \frac{\sum_k (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum (x_t - \bar{x})^2}$$

- It is a tool to depict the structure of the time series

ACF – example



ACF – example



Outline

Motivating Example – Leptospirosis Epidemics

- Bacterial zoonosis (*Leptospira sp*)
- Transmitted to humans through contact with urine from infected animals (rats in urban setting)
- Clinical manifestations:
  - self-limiting fever, with headache and muscle pain $\rightarrow$ easily taken for a bad cold or dengue fever
  - life-threatening disease $\rightarrow$ kidney failure, pulmonary hemorrhage, Weil's syndrome
  - early treatment! (dialysis mainly)
- Globally spread, affecting people on all continents – 5-10% mortality of severe cases; about 607 deaths in 2014
  - Sporadic disease, related with specific occupational exposures and recreational activities
  - Slums and flooding in urban areas

Leptospirosis & Climate

- People living in slums $\rightarrow$ a seroprevalence survey at Pau-da-Lima (Salvador/BA) indicates 23% at 50 years of age
- However, not many severe cases (three in 8 years)
- Severe cases numbers increase during the tropical storms season
- Reasoning: heavy rainfall cleans out the rats holes, bringing the *Leptospira* to the soil surface
- People clean mud after flooding $\rightarrow$ large inoculant dose

The environment

The people

Leptospirosis & Climate: main questions

- Does rainfall really lead to severe leptospirosis epidemics?
- Are other environment factors – humidity & temperature – involved?
- Is there a threshold?
- What is the time delay between tropical storms and increase in the number of cases?
    - duration of incubation period
    - survival of *Leptospira* on the soil, possibly related to temperature, sun and moisture

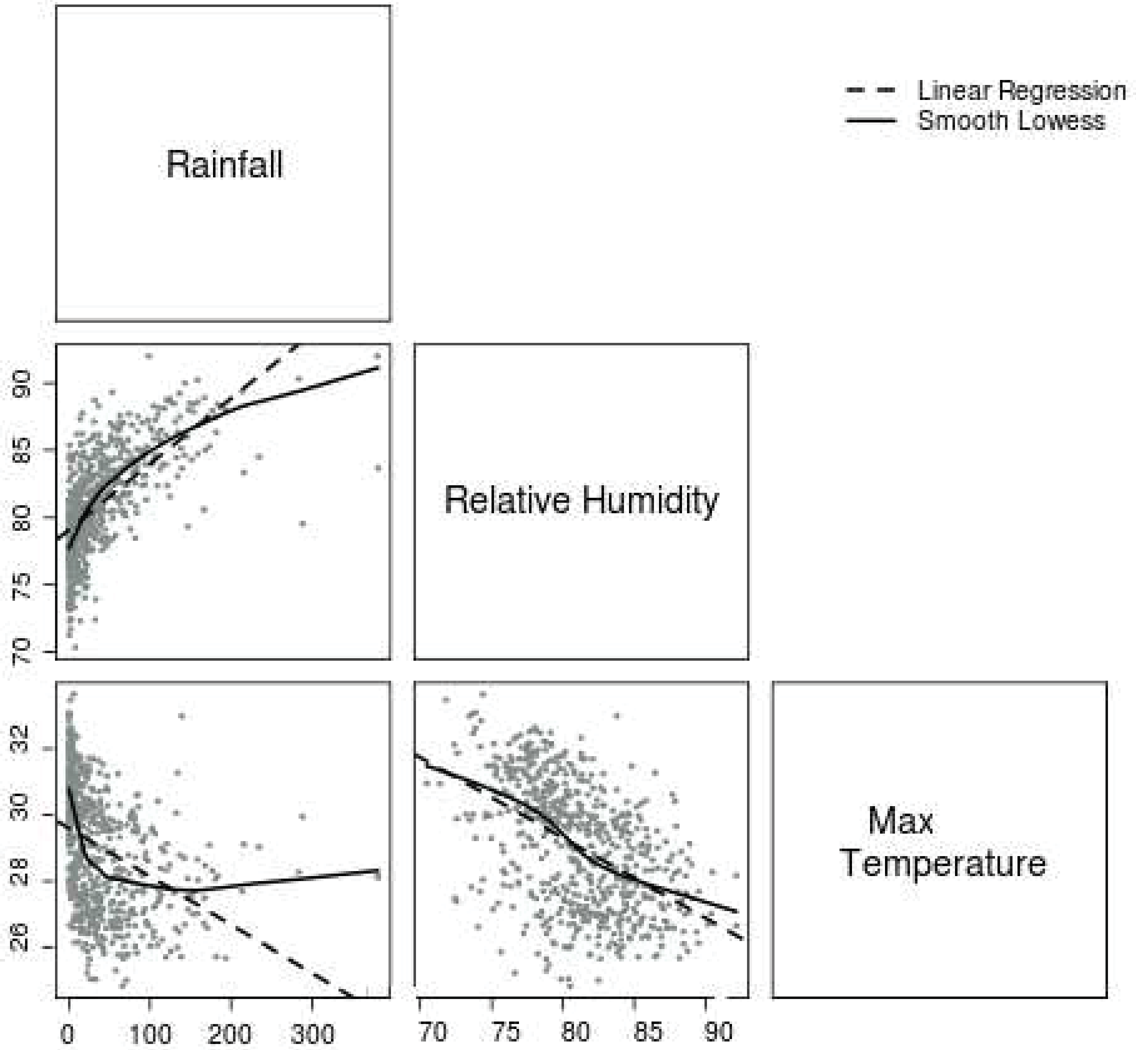
Data

- Local epidemiology surveillance system
- Weekly aggregated cases
- Climate covariates (per week):
  - temperature (mean and maximum)
  - mean relative humidity
  - accumulated rainfall

Outline

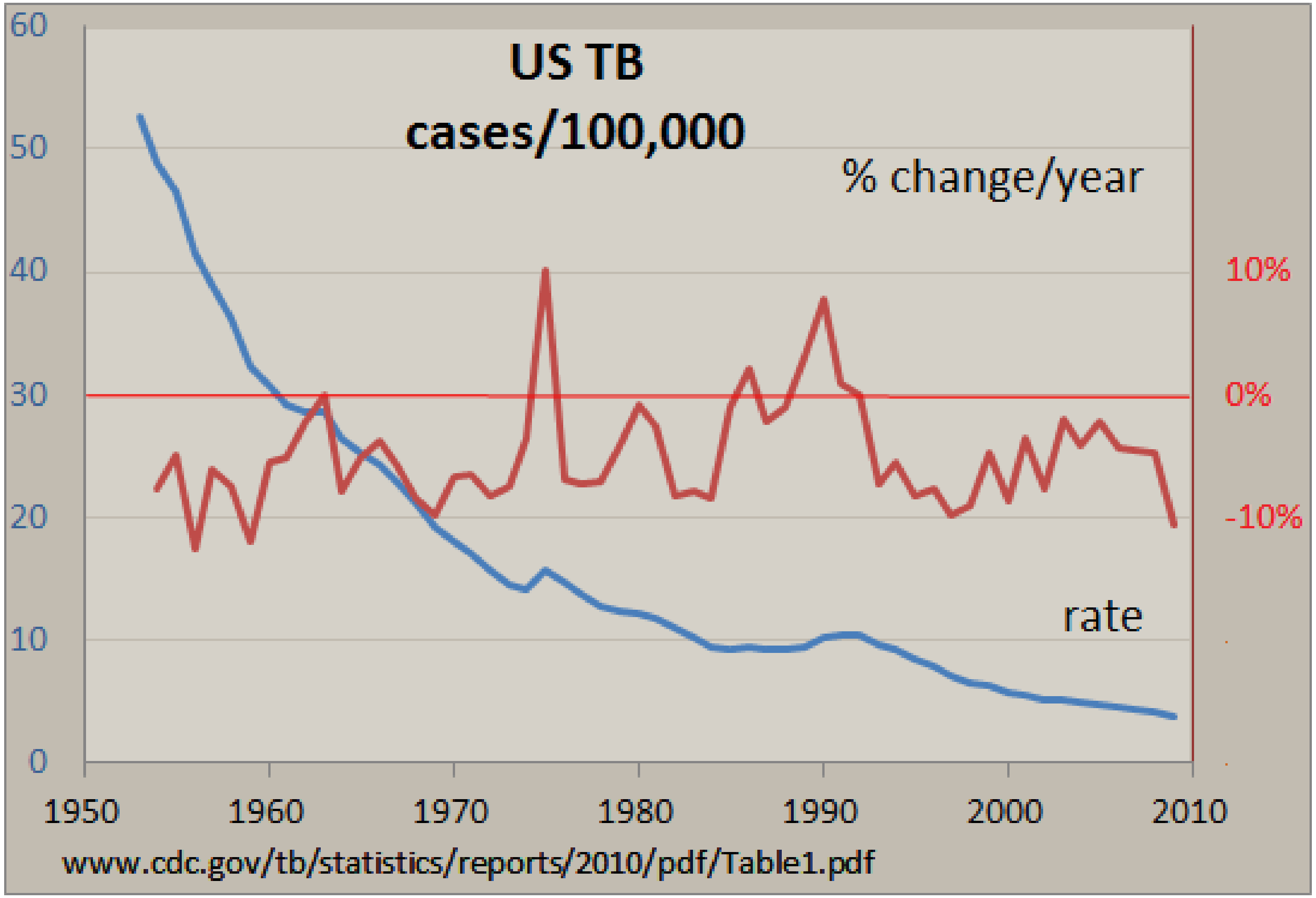
2. Exploratory Analysis
   - Kernel
   - Loess
   - Splines

Exploratory analysis – The usual



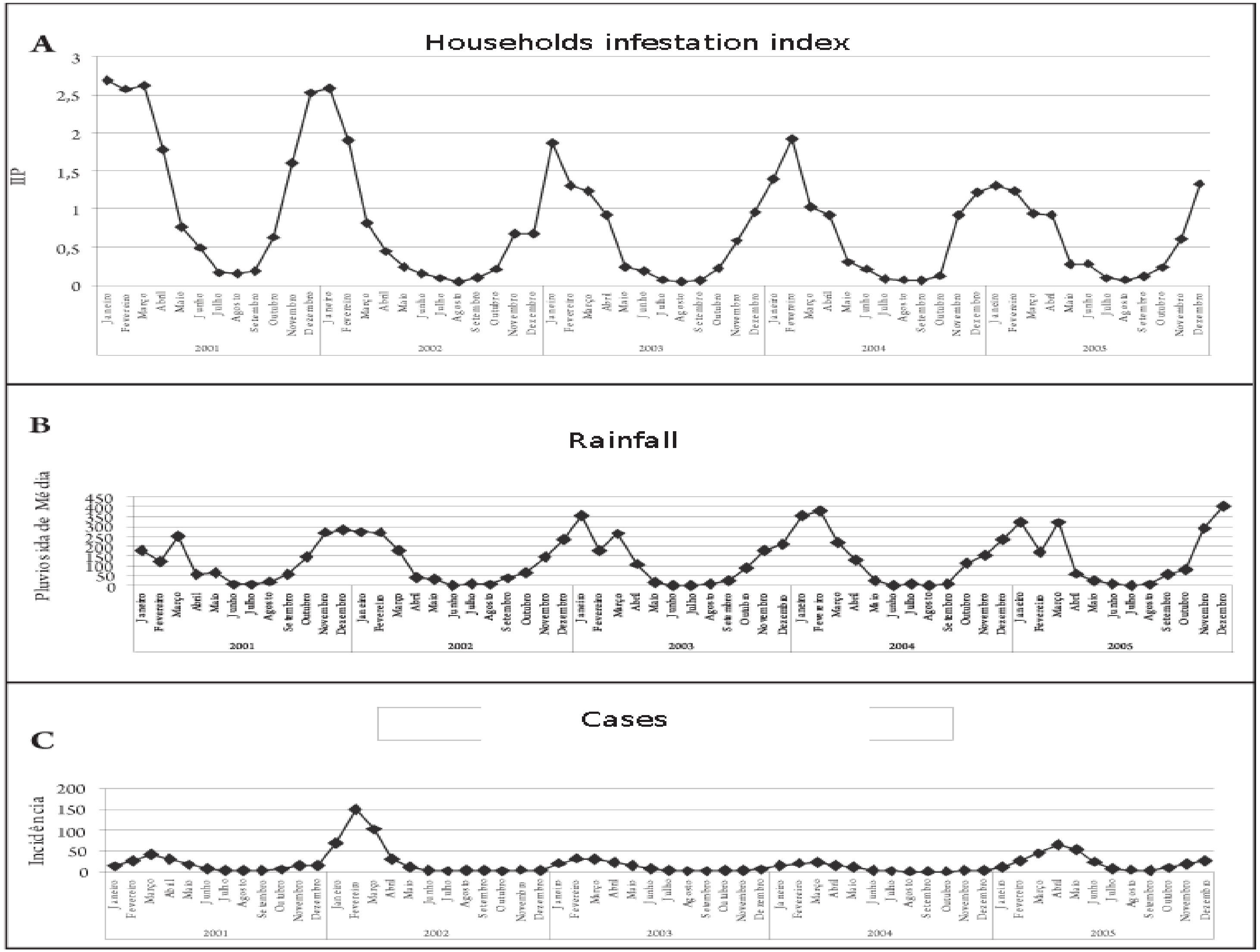
Exploratory analysis – Line Charts

Tuberculosis



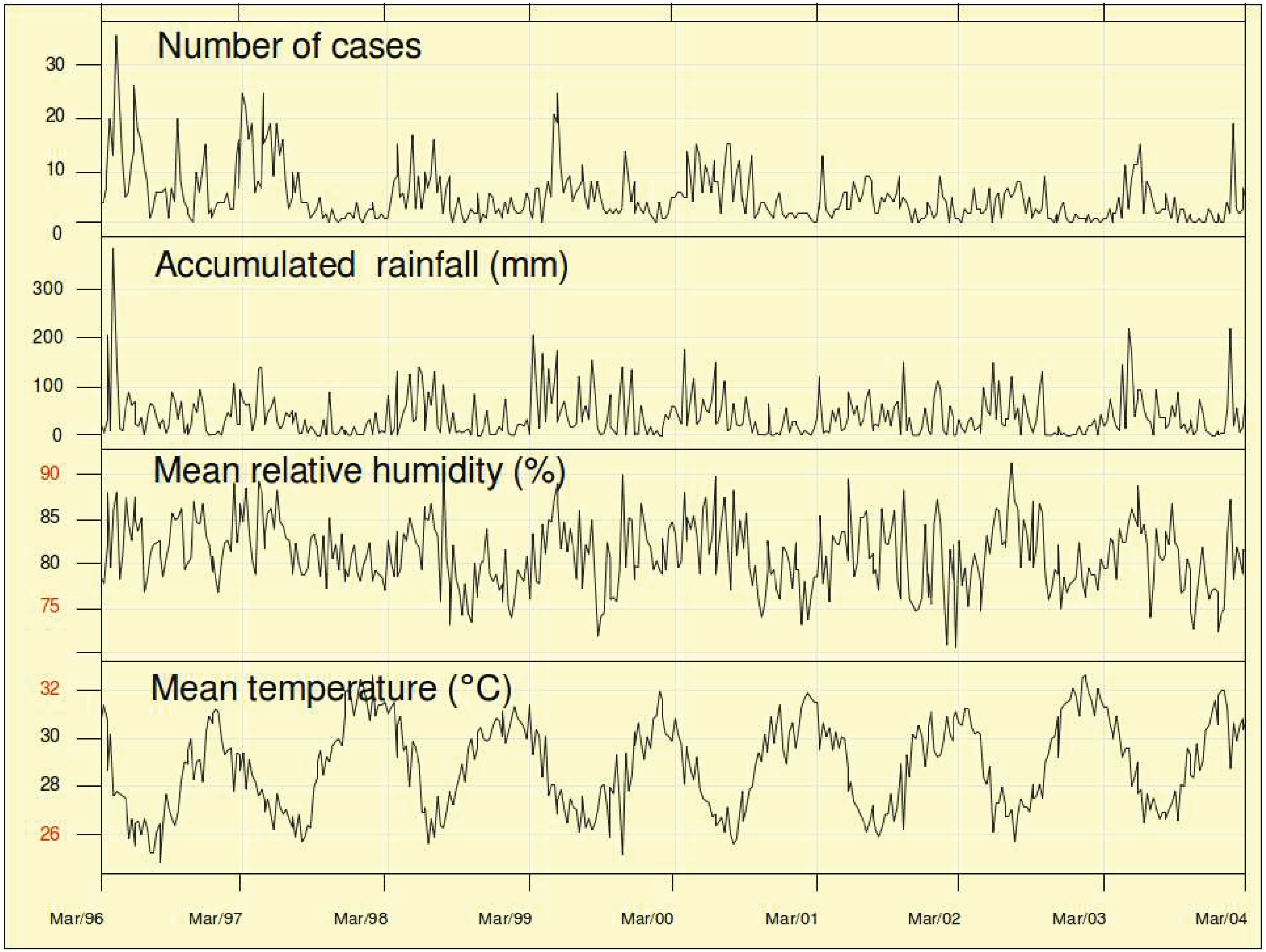
Exploratory analysis – Line Charts

Aedes aegypti & rainfall



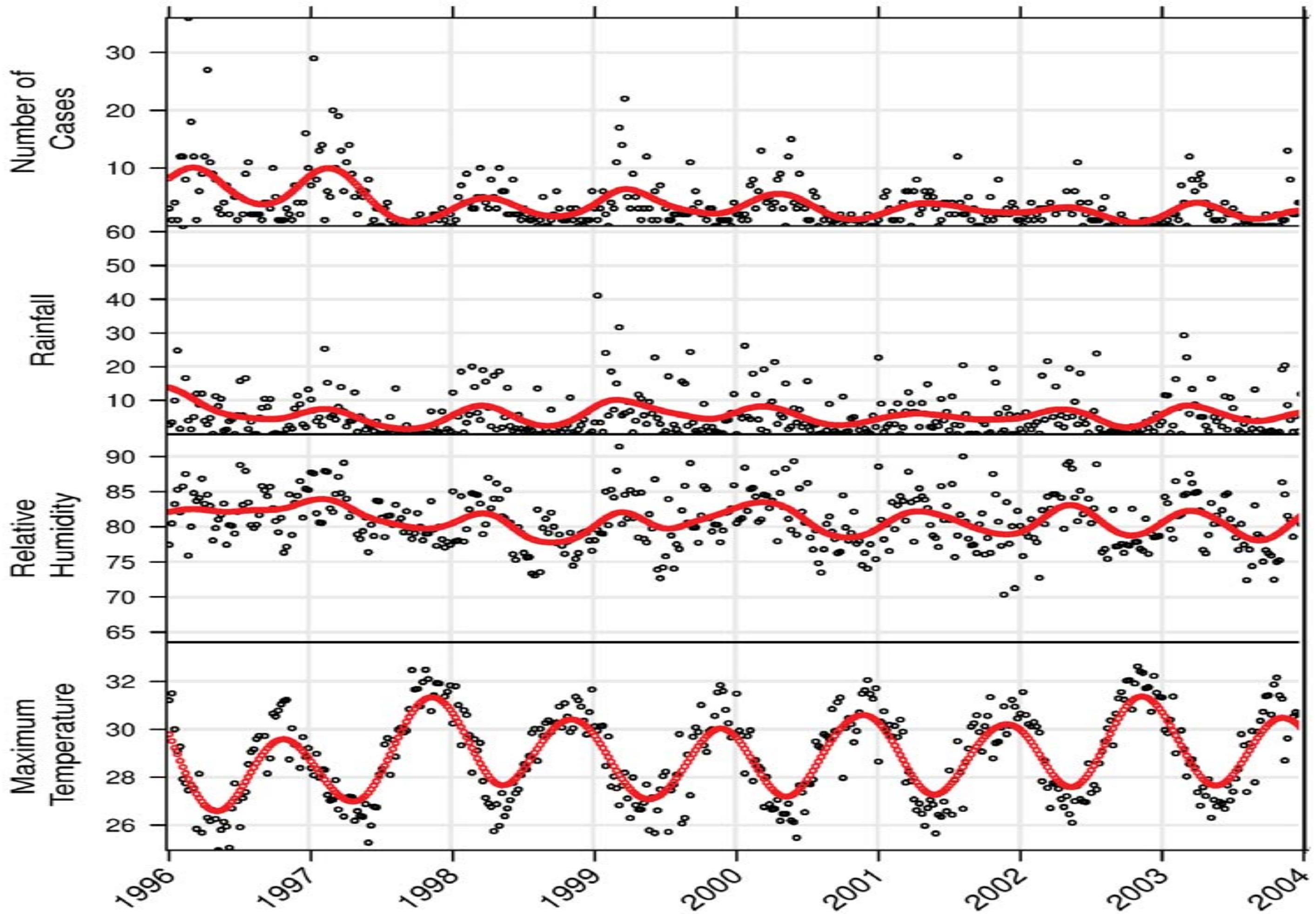
Exploratory analysis – Line Charts

Leptospirosis data

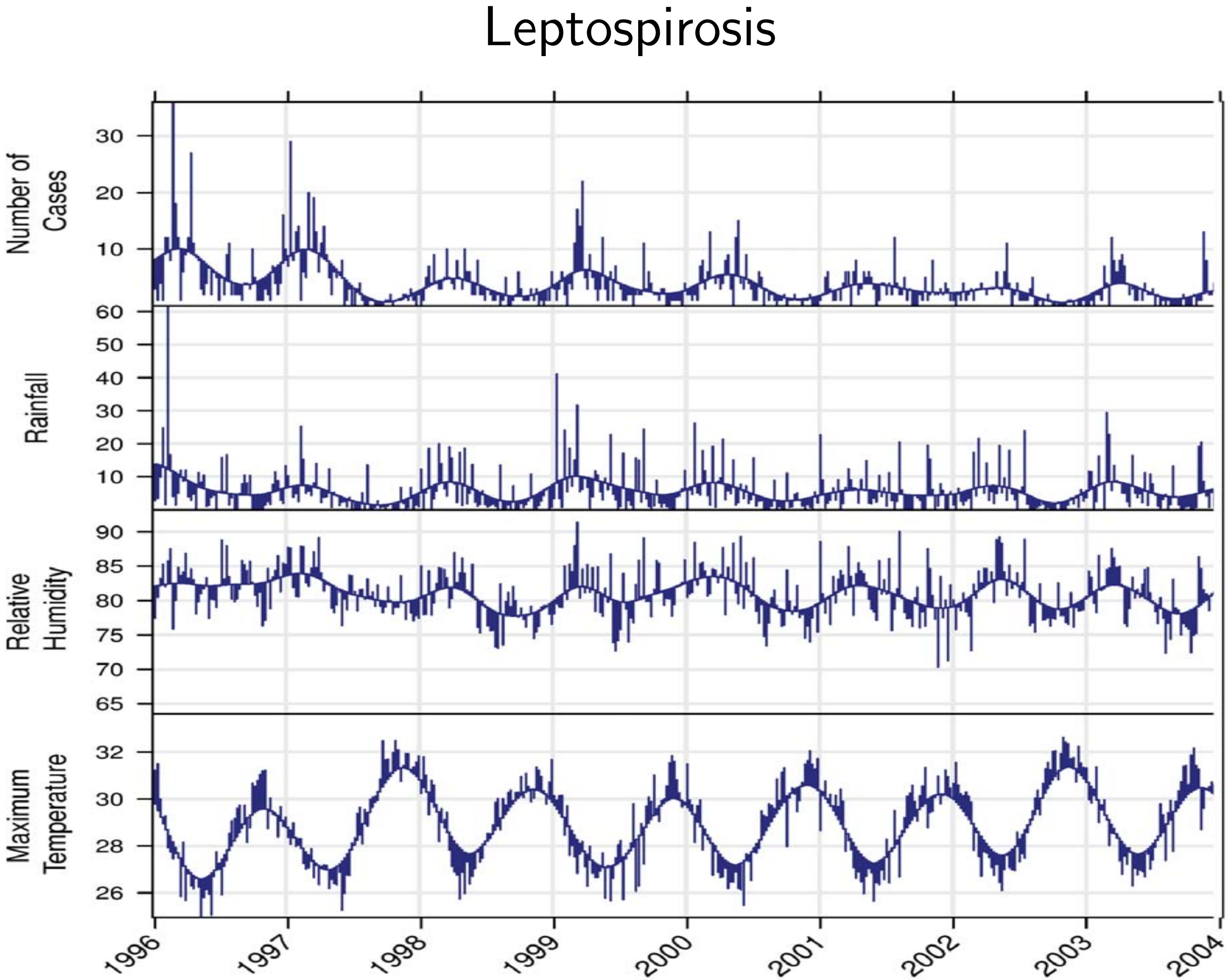


Exploratory analysis – Smoothing

Leptospirosis



Exploratory analysis – Smoothing



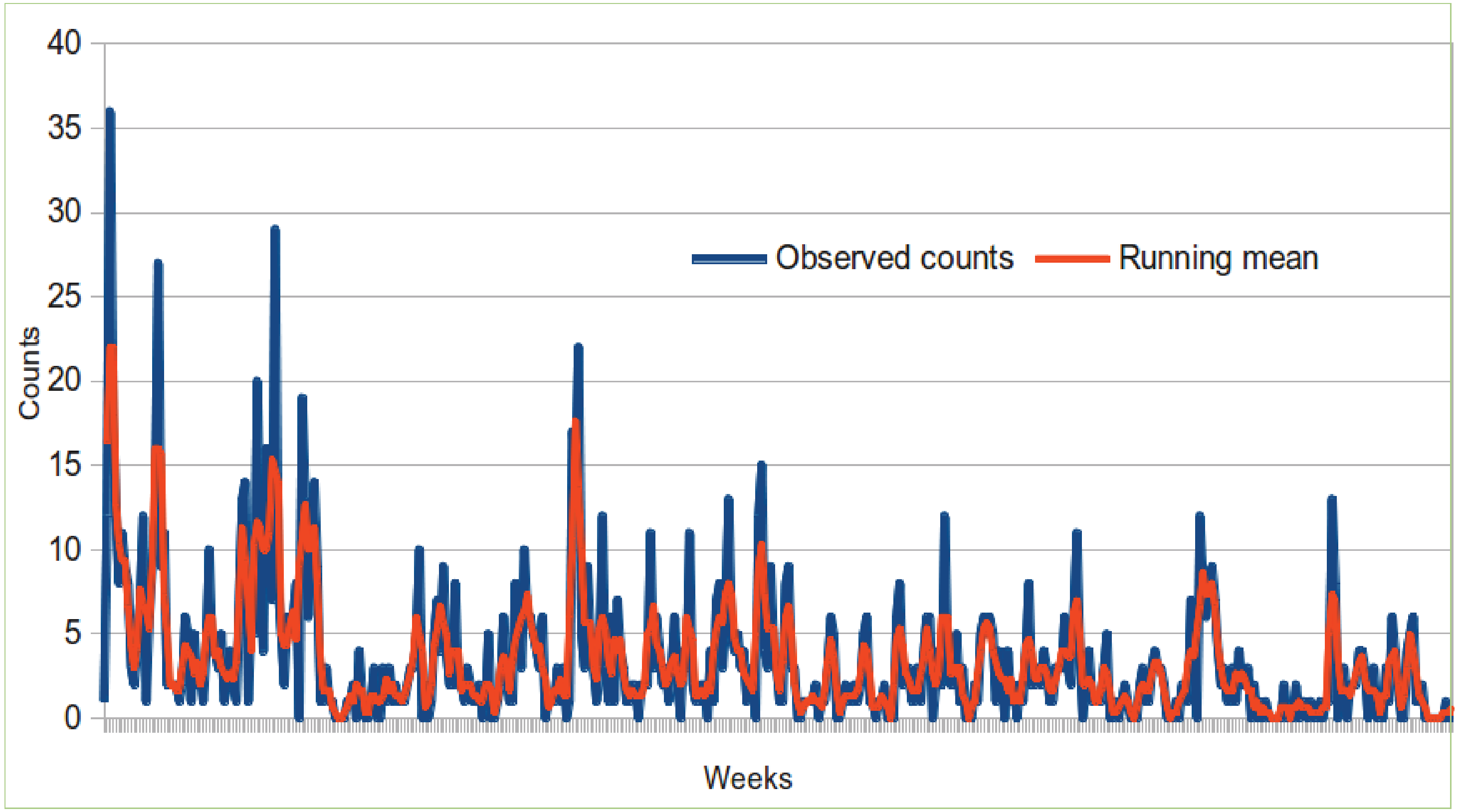
Smoothing

- Moving average – very simple
- Kernel density – a non-parametric way to estimate the probability density function of a random variable
- LOESS or LOWESS – locally weighted scatterplot smoothing
- Splines – minimisation of an objective function where a trade-off between fidelity to the data and roughness of the function estimate is explicit

Outline

2 Exploratory Analysis
- Kernel

Running average



Kernel – the algorithm

1. Define the kernel function:
    - symmetric
    - unimodal
    - centred on (x)
    - going to zero at the edge – neighbourhood
2. Let (x) be the point where to estimate $f(.)$
3. Define the limit of the area of influence of each point $\rightarrow$ window or bandwidth
4. This range controls the smoothing parameter of the kernel function
5. Calculate the value of $f(x)$ for each point and connect them.

Kernel – the function

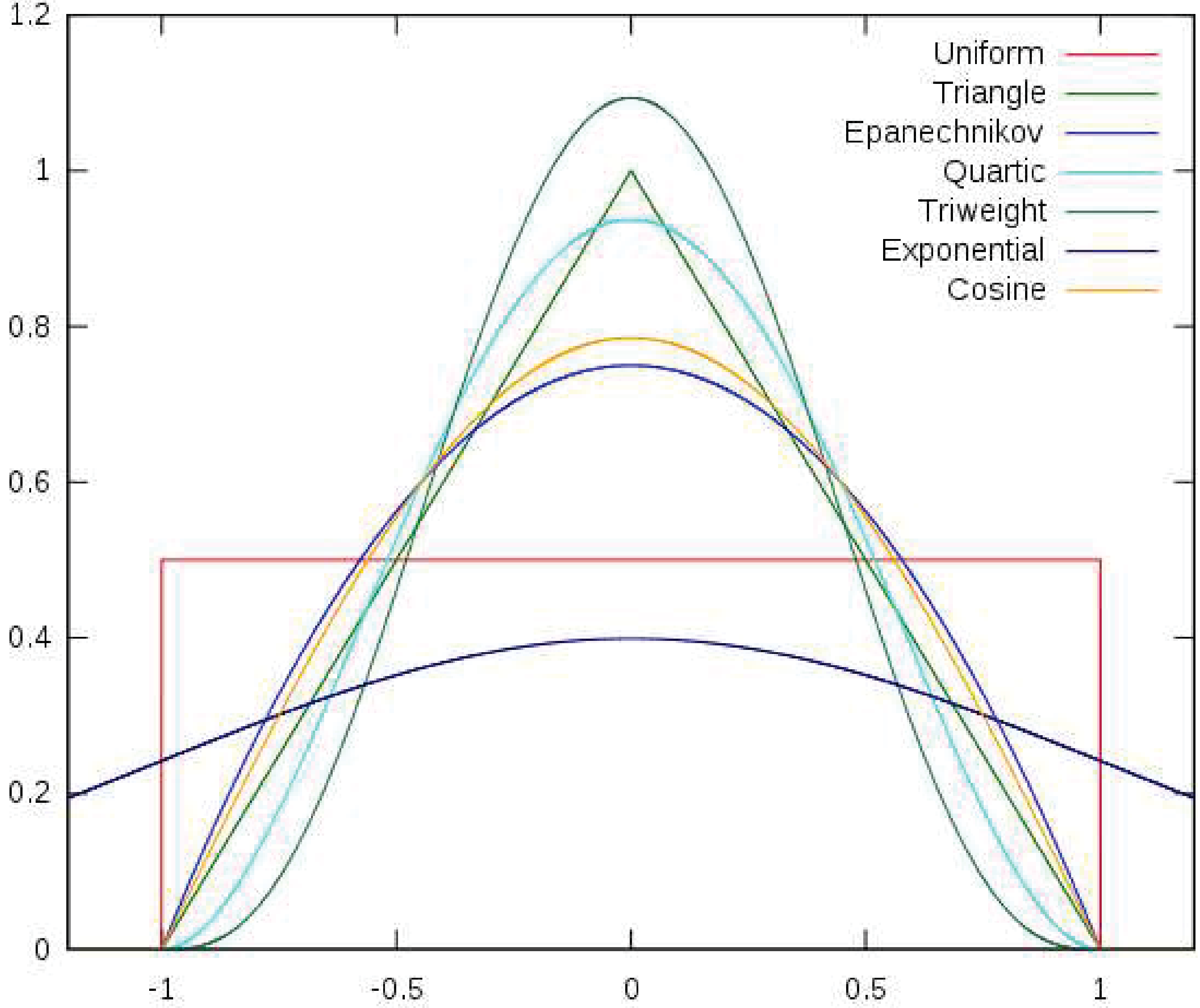
$$\hat{f}_h(x) = \frac{1}{Nh} \sum K\left(\frac{x - x_i}{h}\right)$$

$h \rightarrow$ bandwidth – can be estimated by cross validation

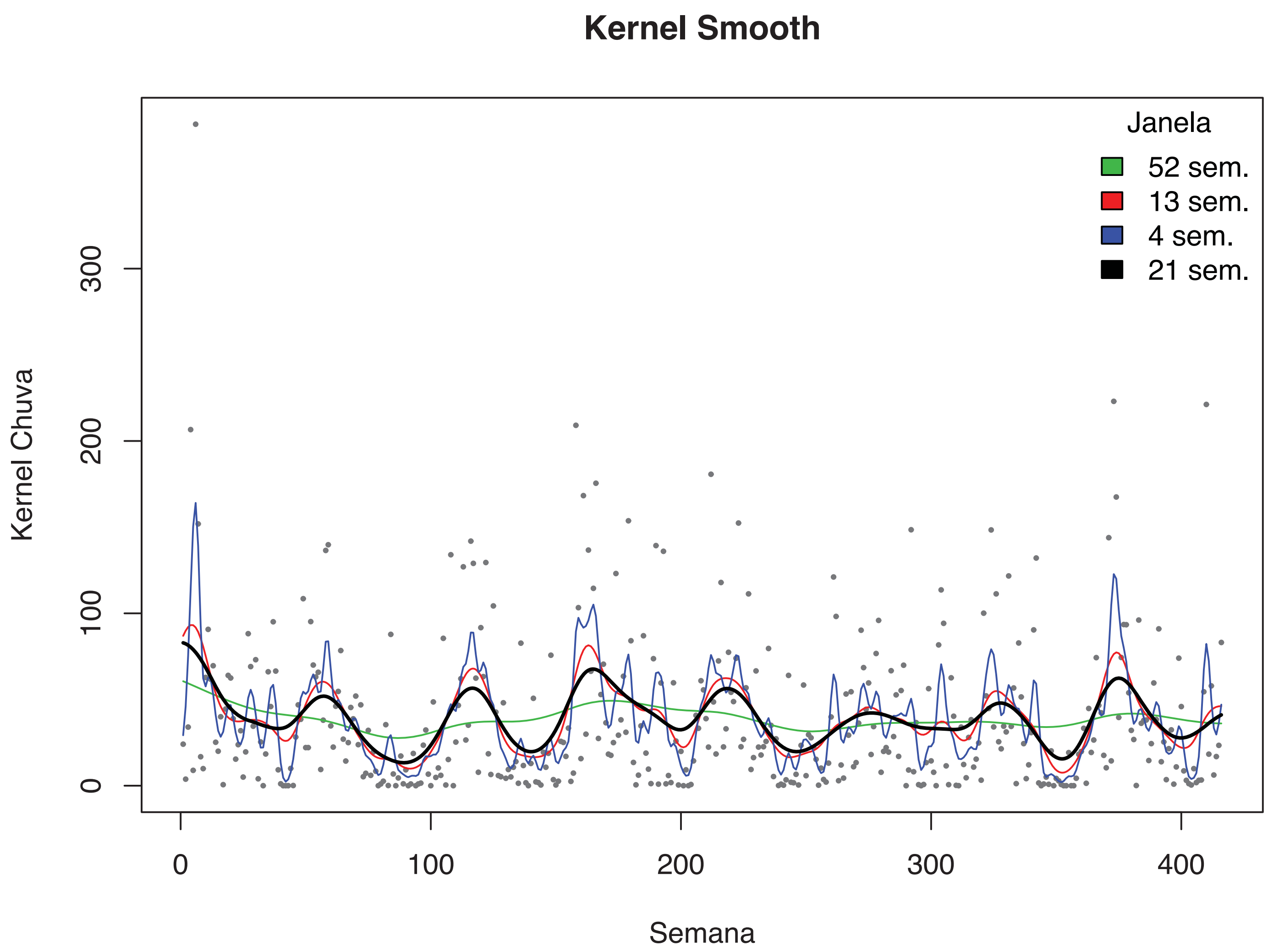
$K \rightarrow$ smoothing function

Gaussian Kernel: $k(x) = \frac{1}{\sqrt{2\pi}} exp(1/2x^2)$

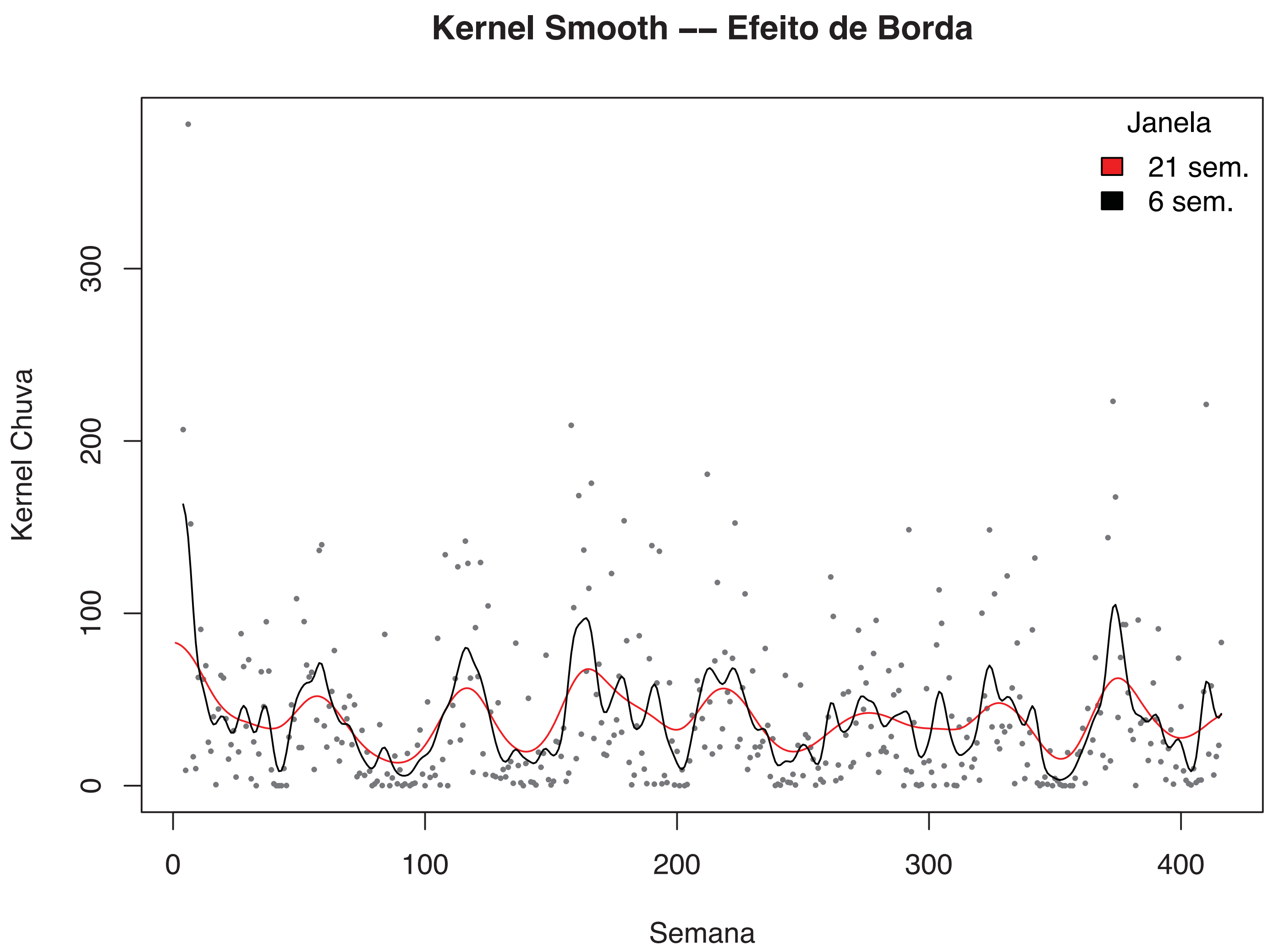
Kernel – several functions



Kernel – Example



Kernel – Border effect



Kernel

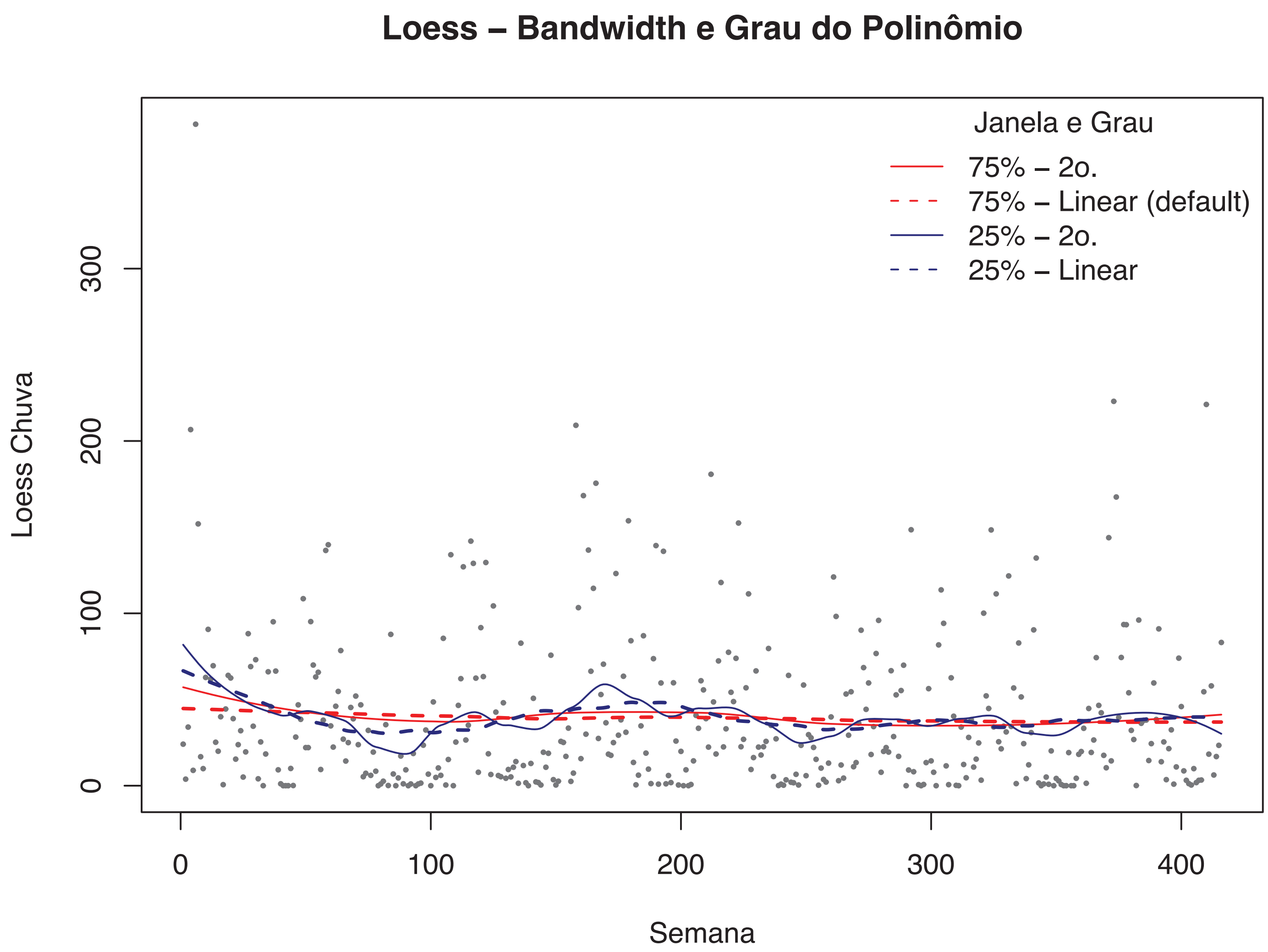
- Advantages: simple, great for exploratory analysis.
- Problem: border effect.
- Very sensitive to bandwidth.
- Automatic choice of bandwidth may not be desirable.
- Not very sensitive to function shape, as long as it is smooth.

Outline

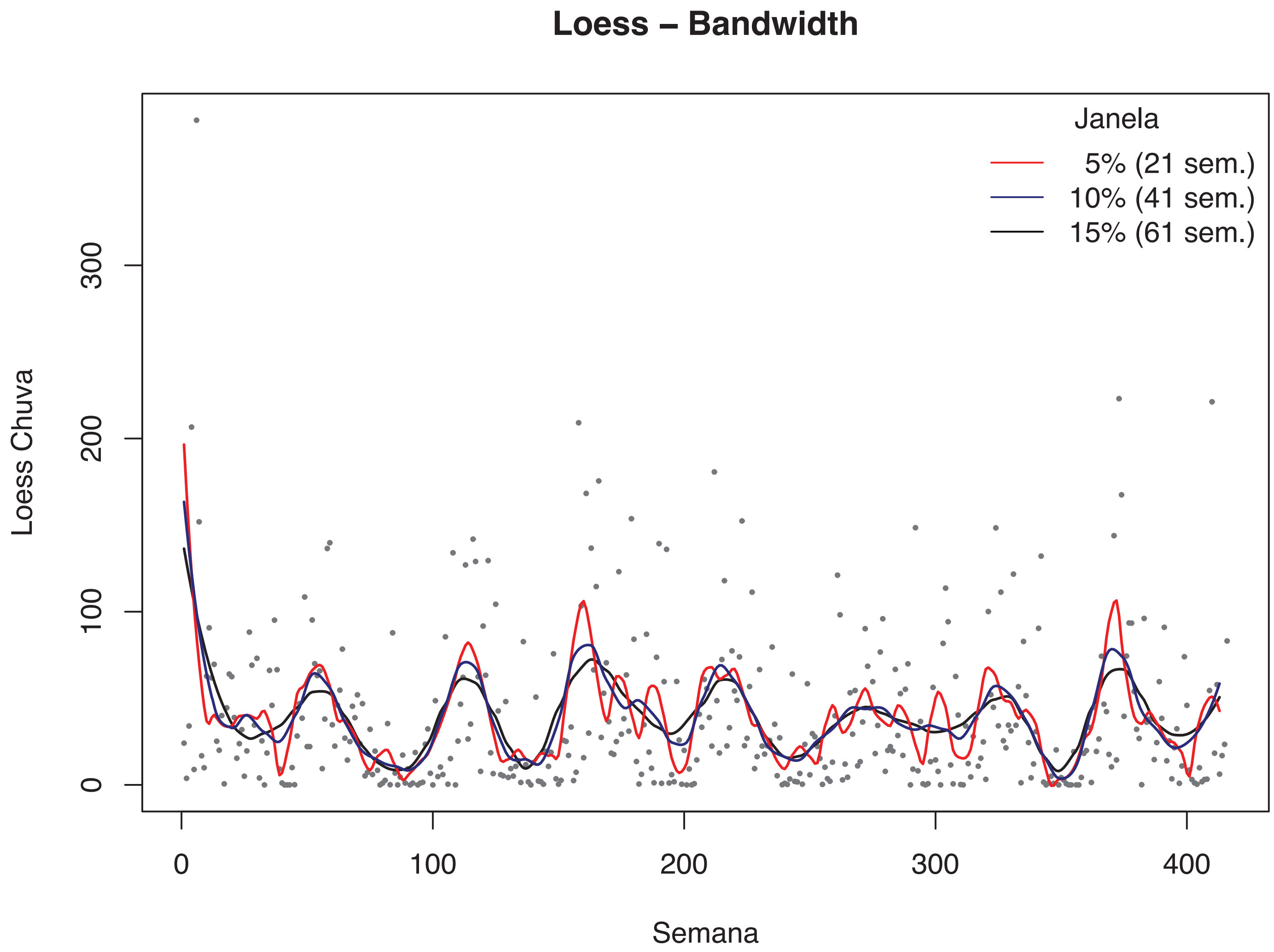
Loess

- Similar to the kernel, but the base is a local regression instead of a weighted average
- At each point (x) and neighbouring points (window or bandwidth) a polynomial is fitted using weighted least squares, where closer points are given larger weight
- The bandwidth or smoothing parameter controls the flexibility of the regression
- The degree of the polynomial regression is in general low:
  - A polynomial of degree $0 =$ running average;
  - First degree $=$ local linear regression

Loess – Span & Degree



Loess – Span & Border



Loess

- Advantages:
  - simple, great for exploratory analysis.
  - Less sensitive to border effect
- Disadvantages: sensitive to extreme values

Comparing

http://en.wikipedia.org/wiki/Kernel_smoothing

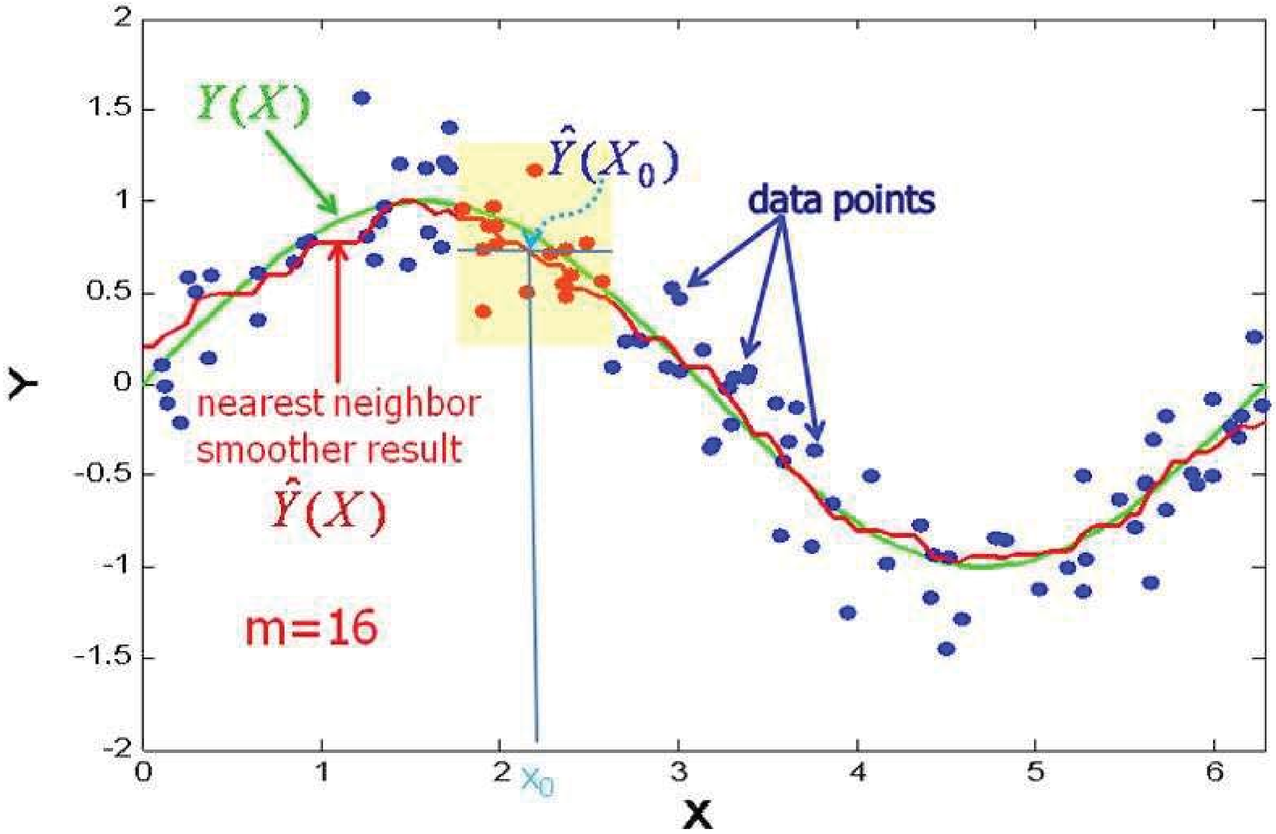


Fig.: Nearest neighbour

Comparing

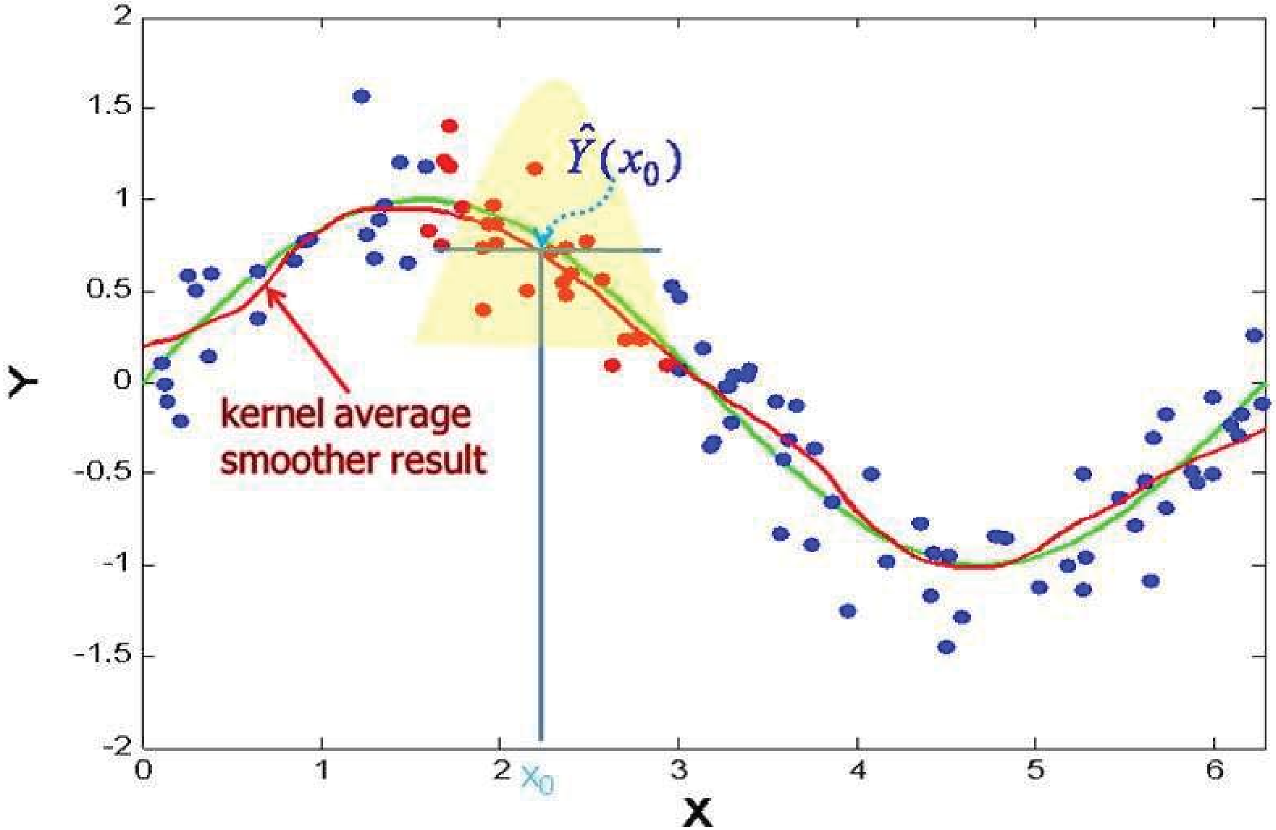


Fig.: Weighted average

Comparing

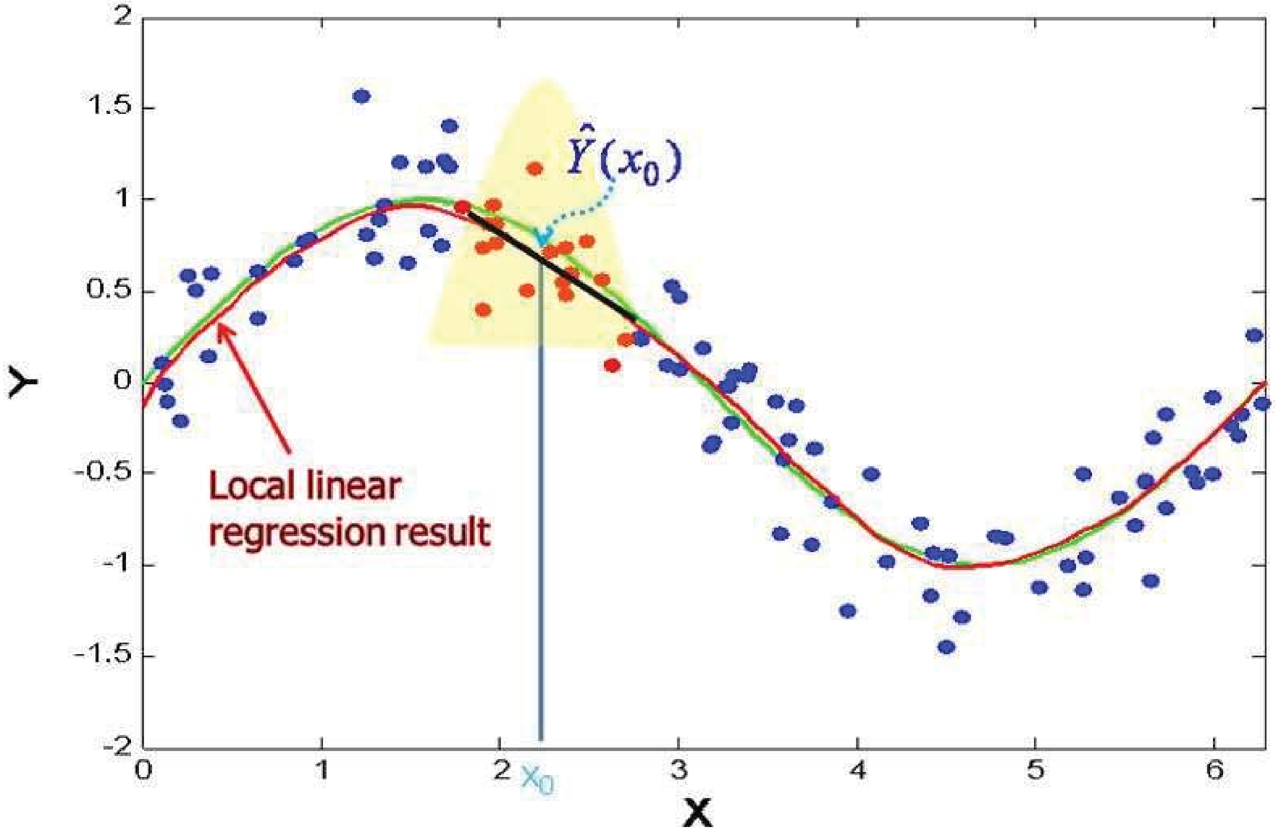


Fig.: Loess

Outline

2 Exploratory Analysis
- Kernel
- Loess
- Splines

Splines

- Splines are smooth polynomial function piecewise-defined
- Very smooth, including the places where the polynomial pieces or knots connect
- Splines do not oscillate ate the edges (Runge's phenomenon present when using high degree polynomial interpolation)

Splines

- A problem of penalised regression: a solution for $\hat{f}(x)$ that minimises:

$$\sum [y_i - f(x_i)]^2 + \tau \int [f''(x)]^2 dx$$

where τ is the smoothing parameter: controls the trade-off between fidelity to the data and roughness of the function estimate

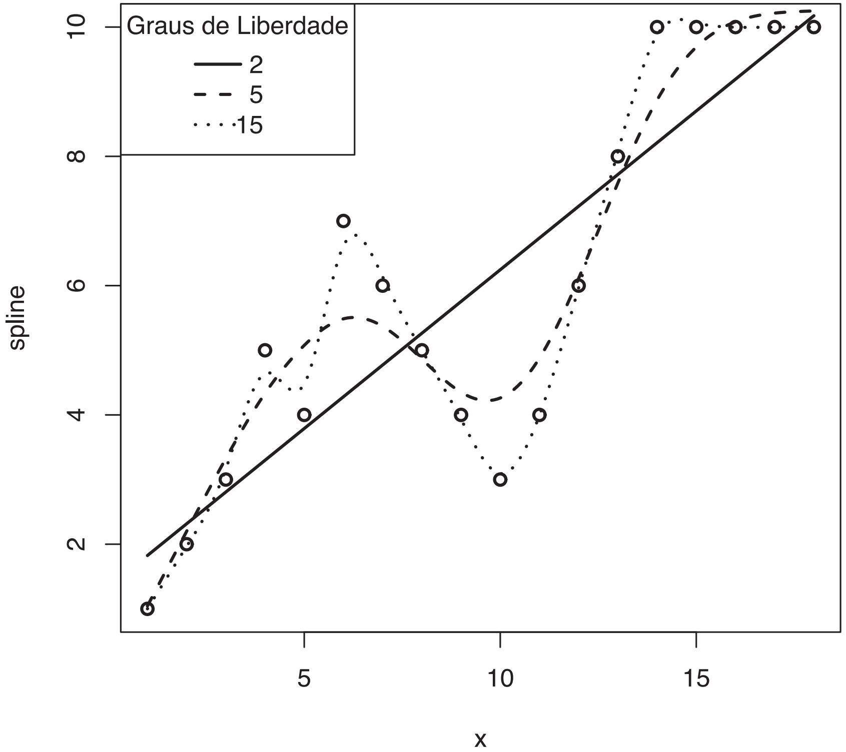
  - If $\tau = 0 \rightarrow \hat{f}(x)$ interpolating spline
  - If τ is very large, $\int [f''(x)]^2 dx$ needs to approach zero $\rightarrow$ linear least squares estimate

- When $\sum [y_i - f(x_i)]^2$ is replaced by a log-likelihood $\rightarrow$ penalised likelihood
- The smoothing spline is the special case of penalised likelihood resulting from a Gaussian likelihood

Splines

- The choice of the smoothing parameter can be visual or via some automatic algorithm (e.g. cross validation)
- The results of splines and loess are similar for similar degrees of freedom
- Multivariate splines: $\eta = \beta_0 + f_1(x_{i1}, x_{i2}, \ldots, x_{ip}) + \ldots$
- Several applications to temporal and spatial models

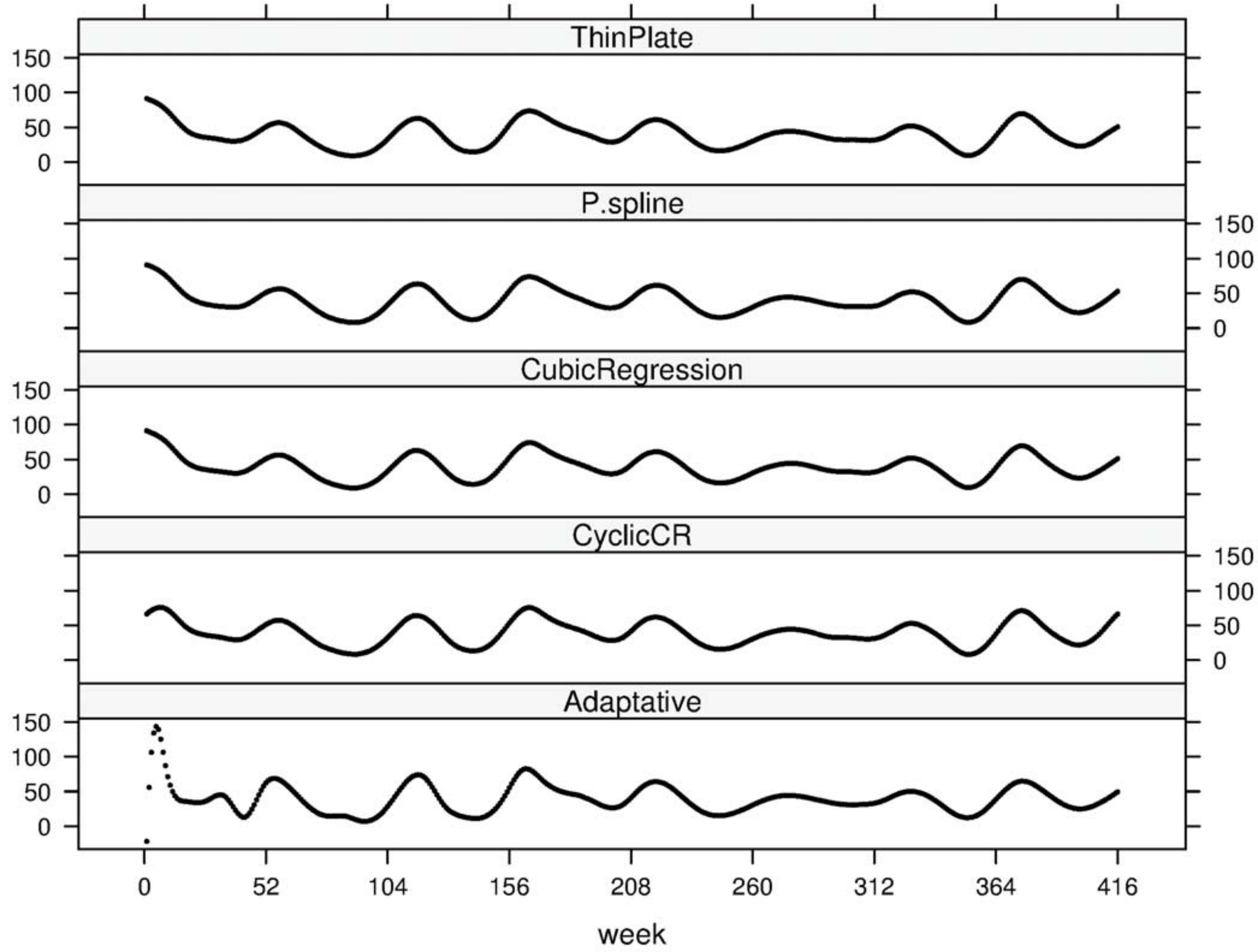
Splines – bandwidth



Splines functions

- Cubic regression spline – 3^{rd} degree polynomial fitted to knots distributed over the data range
- Cyclic cubic regression spline – imposes the first and last values to be equal (interesting for seasonal time series)
- P-splines – with a differential penalty for adjacent parameters, to control "wiggliness"
- Thin plate – the smallest mean square error, smallest number of parameters, considered the optimal estimator, easily adapted to two dimensions (space!)
- Tensor Product – Similar to Thin Plate, better when scale of each dimension is not the same

Splines functions



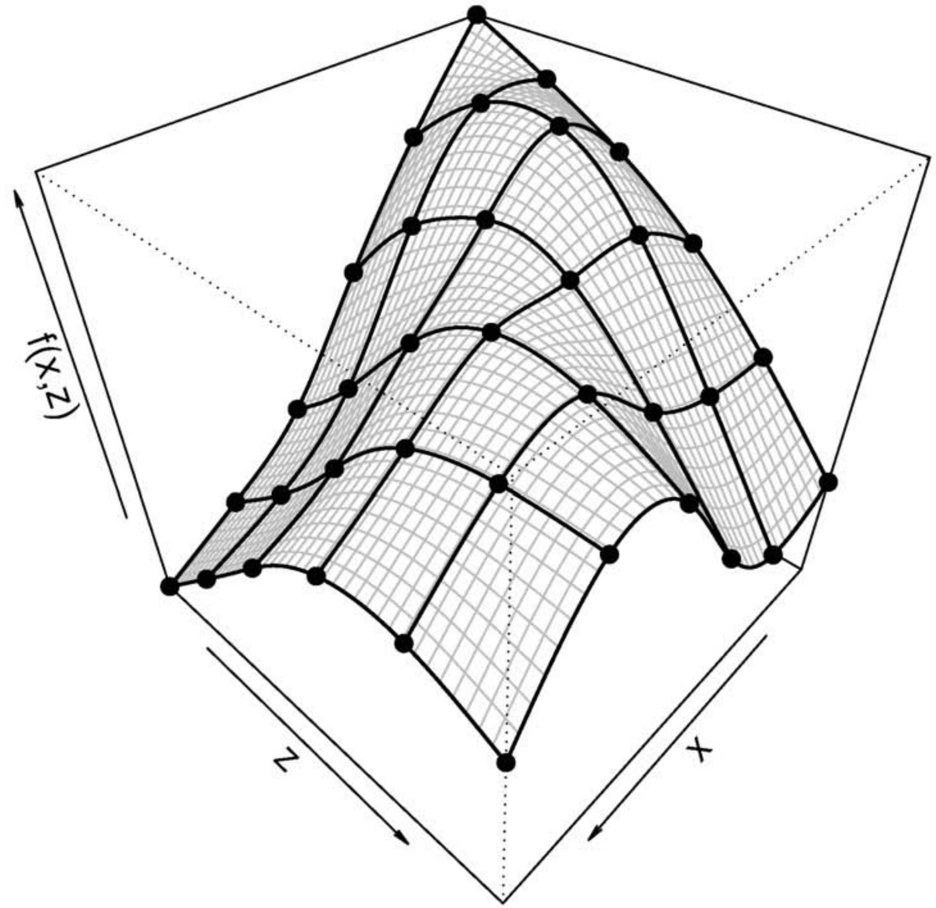
Choice of function

- Modelling just one variable – time – not much difference
- For more then one variable – space – choose carefully:
- Thin plate:
  - isotropic,
  - invariant to rotation
  - smaller square error
  - smaller number of parameters, considered the optimal estimator
  - HOWEVER: sensitive to changes in scale
- Tensor Product:
  - possible to have different scales

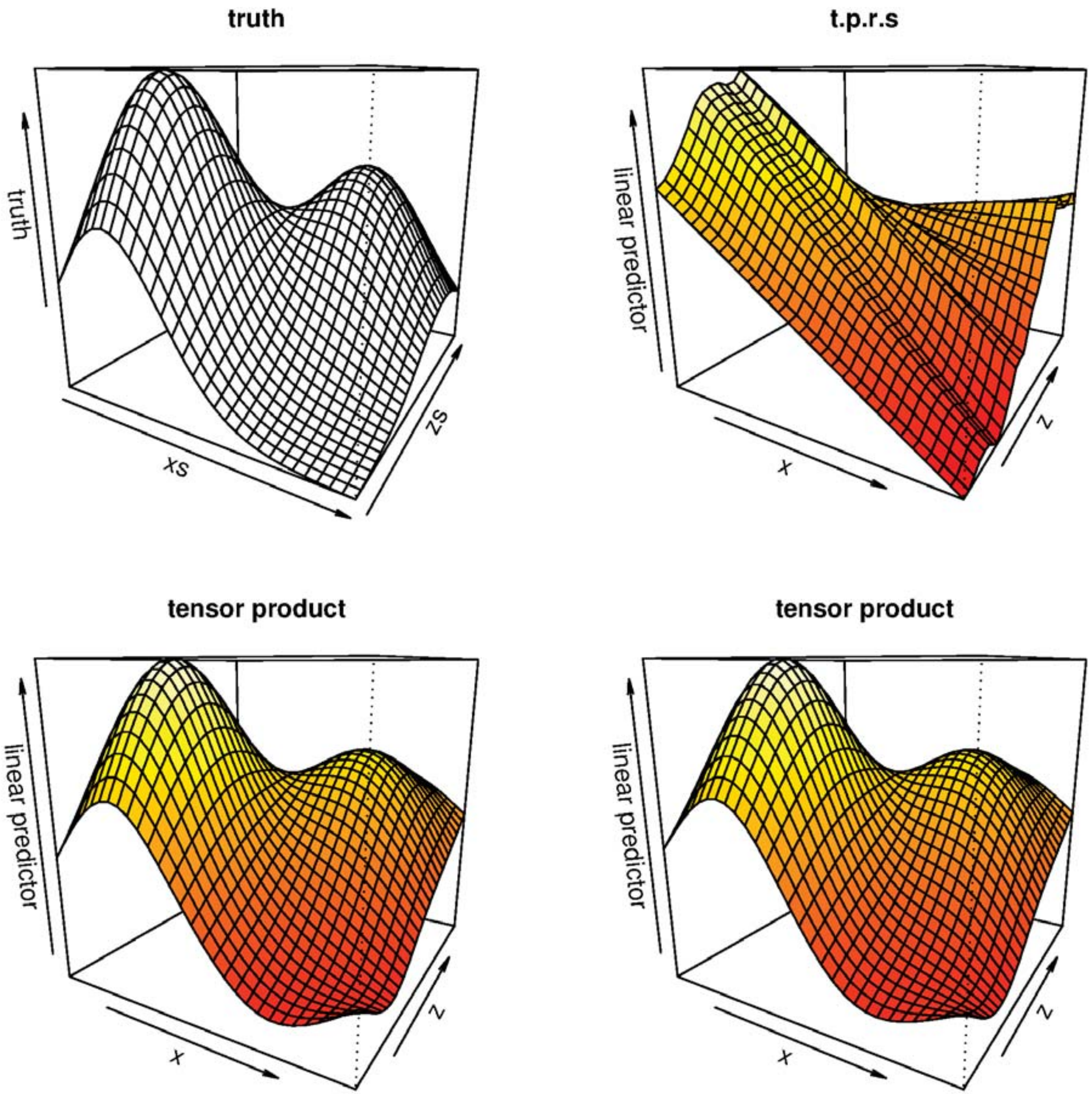
Splines functions– summary

bs=	Description	Advantages	Disadvantages
"tp"	Thin Plate	Multiple covariates	Computationally intensive
		Rotational invariant	Varies with
		Optimal estimator	scale
"tpr'	Tensor Product	Multiple covariates	Varies with
		Scale invariant	rotation
"cr"	Cubic	Computational cheap	Only one variable
	Regression	Parameters directly	Based on knots choice
		interpretable	Non-optimal estimator
"cc"	Cyclic CRS	Beginning and end ='s	the same
"ps"	P-splines	Any combination of	Evenly spaced knots
		base and *order*	Not easily interpretable
			Non-optimal estimator

Bivariate spline



Changing the scale



Outline

The problem

How do these variables behave in relation to each other?

- Age → external causes deaths from 5 to 45 years
- Income → cardiovascular diseases
- Distance to health services → mammography
- Adherence to HIV treatment → development of virus resistance
- ...
- Time → transmissible diseases
- Space → vector-borne diseases

GAM – definition

- extension of GLM, where the linear predictor η is not limited to linear regression
- the model includes any function of the independent covariates (x_i):

$$\eta \ = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots$$

- $f(x) \rightarrow$ can be a non-parametric function such as lowess
- When to use? When the covariate effect changes depending upon its value

Why not to use

- Statistical models aim to explain the observed data, not to simply reproduce it – overfitting
- Parametric models in general are better to estimate standard errors or confidence intervals
- Parametric models are more efficient, if correctly specified (smaller number of observations)

Outline

The problem

- Going back to the leptospirosis example.
- To estimate the effect of rainfall, humidity and temperature on the number of cases of leptospirosis
- Why not just apply a regression model?
    - Trend
    - Seasonality
    - Autocorrelation

Autocorrelation

- Autocovariance is the covariance of the variable against a time-shifted version of itself

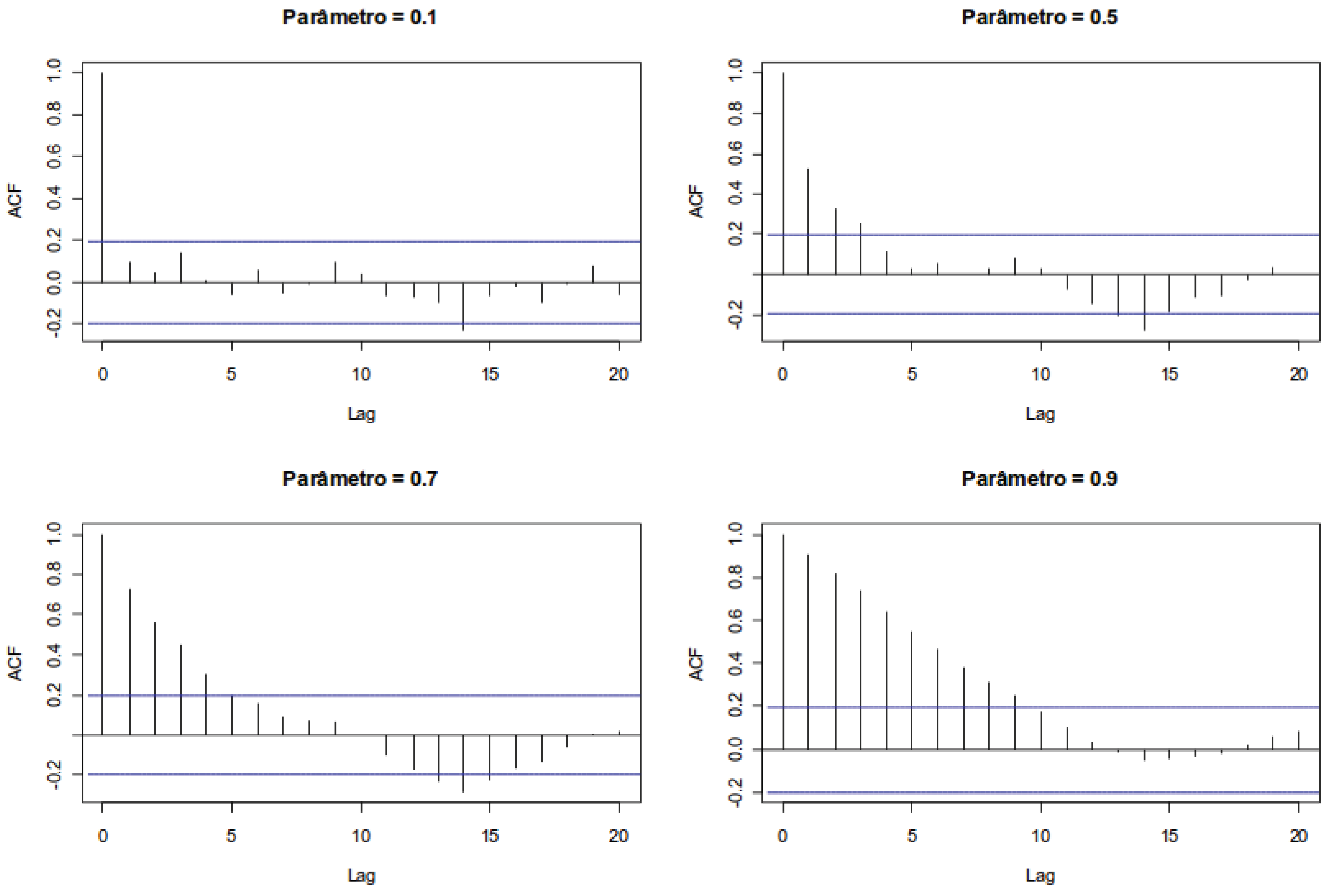
$$C_{xx}(t, s) = E[(X_t - \mu_t)(X_s - \mu_s)] - \mu_t\mu_s$$

- If $X(t)$ is stationary $\rightarrow \mu_t = \mu_s = \mu$ and

$$C_{xx}(t, s) = C_{xx}(t, s) = C_{xx}(\tau)$$

- Autocorrelation $c_{xx}(\tau) = C_{xx}(\tau)/\sigma^2$
  $\tau \rightarrow$ the lag
  $\sigma^2 \rightarrow$ the variance
- It is a measure of how similar a series is to a time-shifted version of itself
- Range: $[-1, 1]$

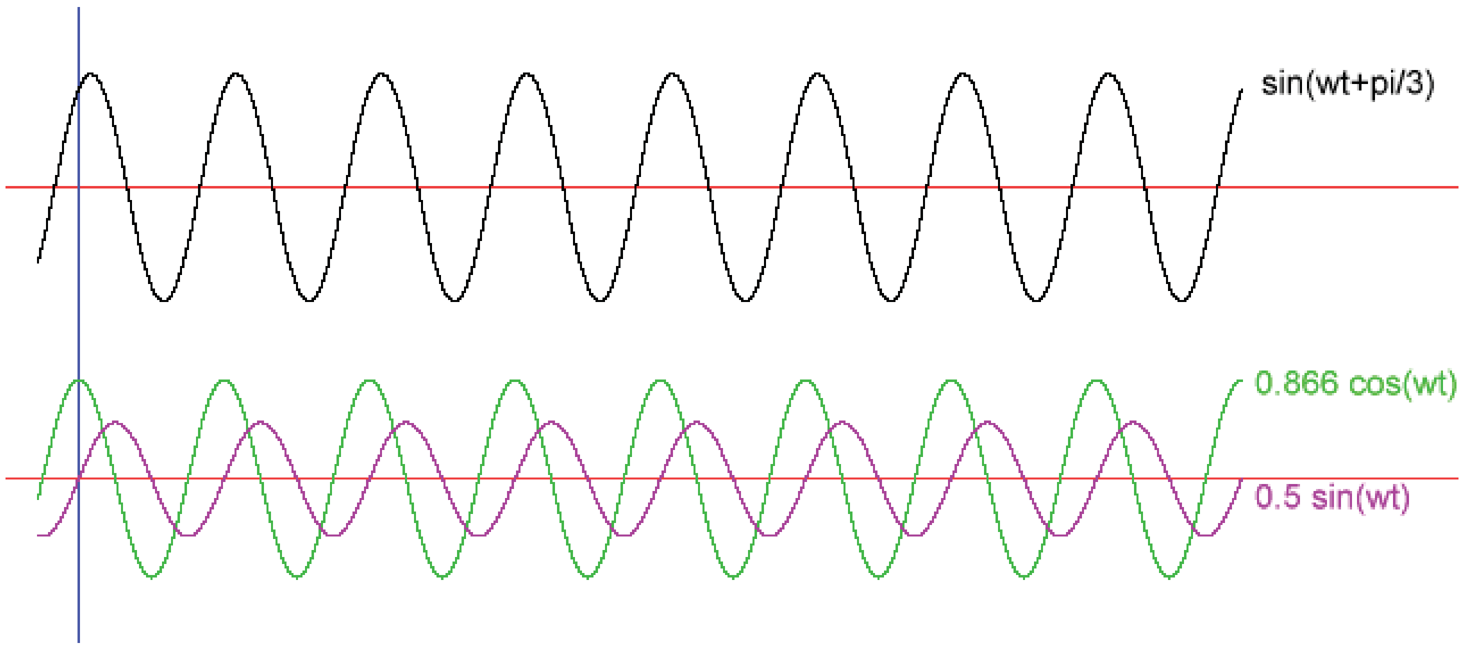
Autocorrelation



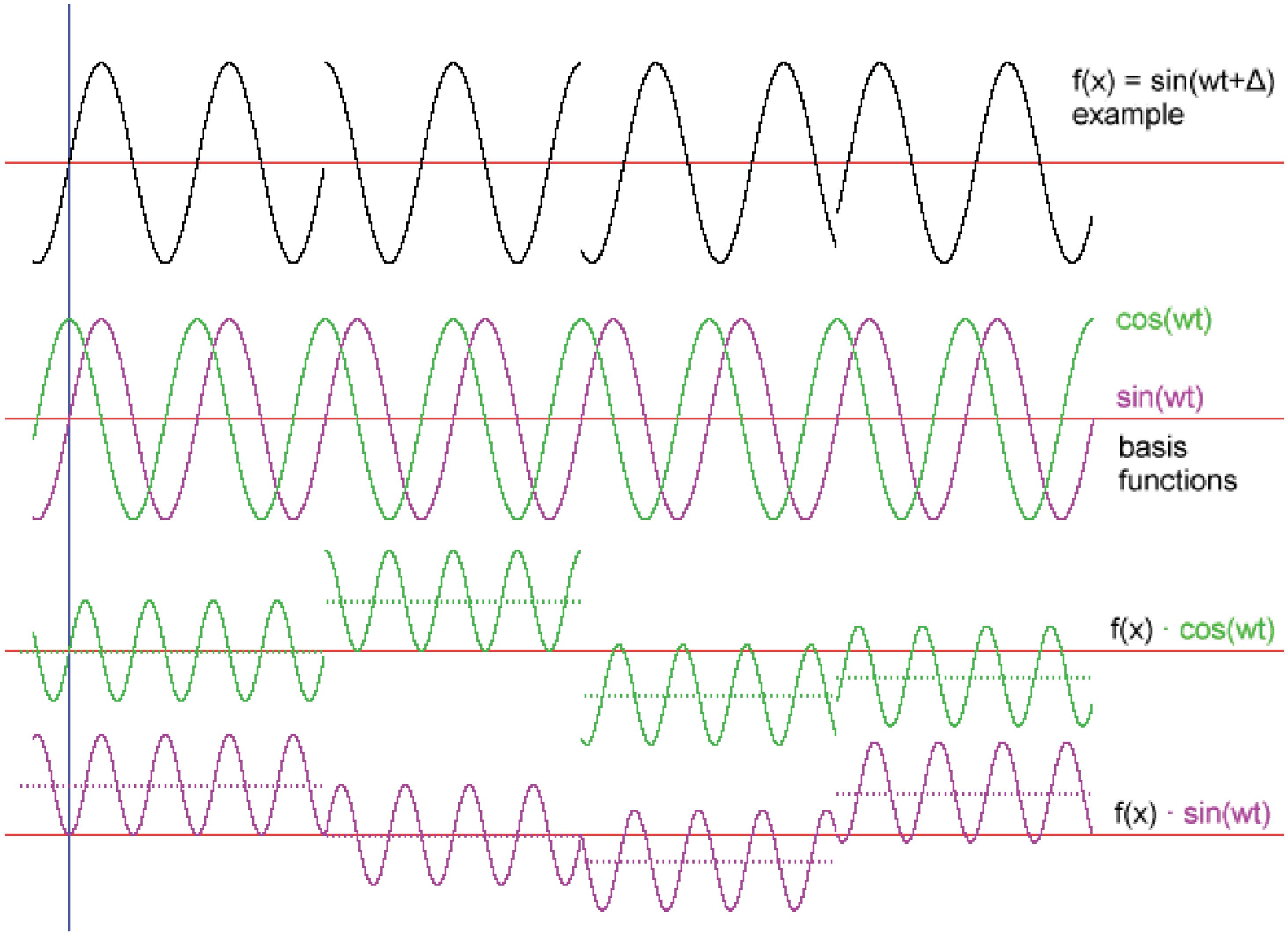
Seasonality

- Component of a time series which is defined as the repetitive and predictable movement around the trend line
- Not necessarily related to climate seasons
- Can be either removed or modelled:
  - sinusoid
  - including each month (or season) as a categorical variable

Seasonality: sinusoid



Seasonality: sinusoid



Trend

- A stationary process is a stochastic process whose joint probability distribution does not change when shifted in time (or space)
- Mean and variance, if they exist, are constant
- Trend model: linear (?!?), polynomial, splines
- Do we really want to remove the trend?

Modelling time series

- Time series books – ARIMA models
- Not much used in epidemiology:
    - Intervention
    - Explanation
    - “Causes”
- Regression models including (if needed) AR components
- Emphasis on covariates

GAM for Time Series

- The main idea is to model the effect of covariates on some health event over time
- Reasons:
  - allow the inclusion of time dependence
  - non-linear relationship
  - trend and seasonality can be easily incorporated

GAM for Time Series

- Considering the response variable a count, the best choices in GLMs are:
  - Poisson: $\lambda =$ expected values and $=$ variance $\rightarrow$ overdispersion
  - Quasipoisson – it is not a distribution, but a way to relax the previous assumption and allow for overdispersion. It does not present AIC.
- Other models, very often used:
  - Negative Binomial – has a mean μ, scale parameter θ and variance function $V(\mu) = \mu + \mu^2/\theta$.
  - Zero-inflated models – mixture models combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial – are available as well (package VGAM)

GAM for Time Series

$$\text{Lepto}(t) = \text{rain}(t-?) + \text{humidity}(t-?) + AR(t, t-1) + trend + seasonality + \varepsilon$$

- Trend and seasonality $\rightarrow$ smooth function
- Covariates – time lag
- It is possible to include the variation on the population at risk (offset)

Outline

Why Distributed Lags?

- When risk factors and health events are measured on populations:
  - asthma & air pollution
  - cold weather & heart attack
  - flooding & leptospirosis
- Between climate and health event $\rightarrow$ time interval – lag
- Questions:
  - How much time after?
  - How long does the effect last?
  - When does the effect disappear?
  - Is there a threshold?

Recommended reading

- Schwartz J. The distributed lag between air pollution and daily deaths.*Epidemiology*, 2000;11(3):320-326.
- Welty, LJ. & Zeger, SL. Are the Acute Effects of Particulate Matter on Mortality in the National Morbidity, Mortality, and Air Pollution Study the Result of Inadequate Control for Weather and Season? A Sensitivity Analysis using Flexible Distributed Lag Models. *American Journal of Epidemiology*, 2005;162:(1):80-88.
- Gasparrini A., Armstrong, B., Kenward M. G. Distributed lag non-linear models. *Statistics in Medicine*. 2010; 29(21):2224-2234.
- Armstrong B. Models for the relationship between ambient temperature and daily mortality. *Epidemiology*. 2010, 17(6):624-631.

Problems

- Effects change over time – increasing and decreasing
- Covariates – temperature, humidity, rainfall and pollution – highly correlated
- Possible non-linear structure

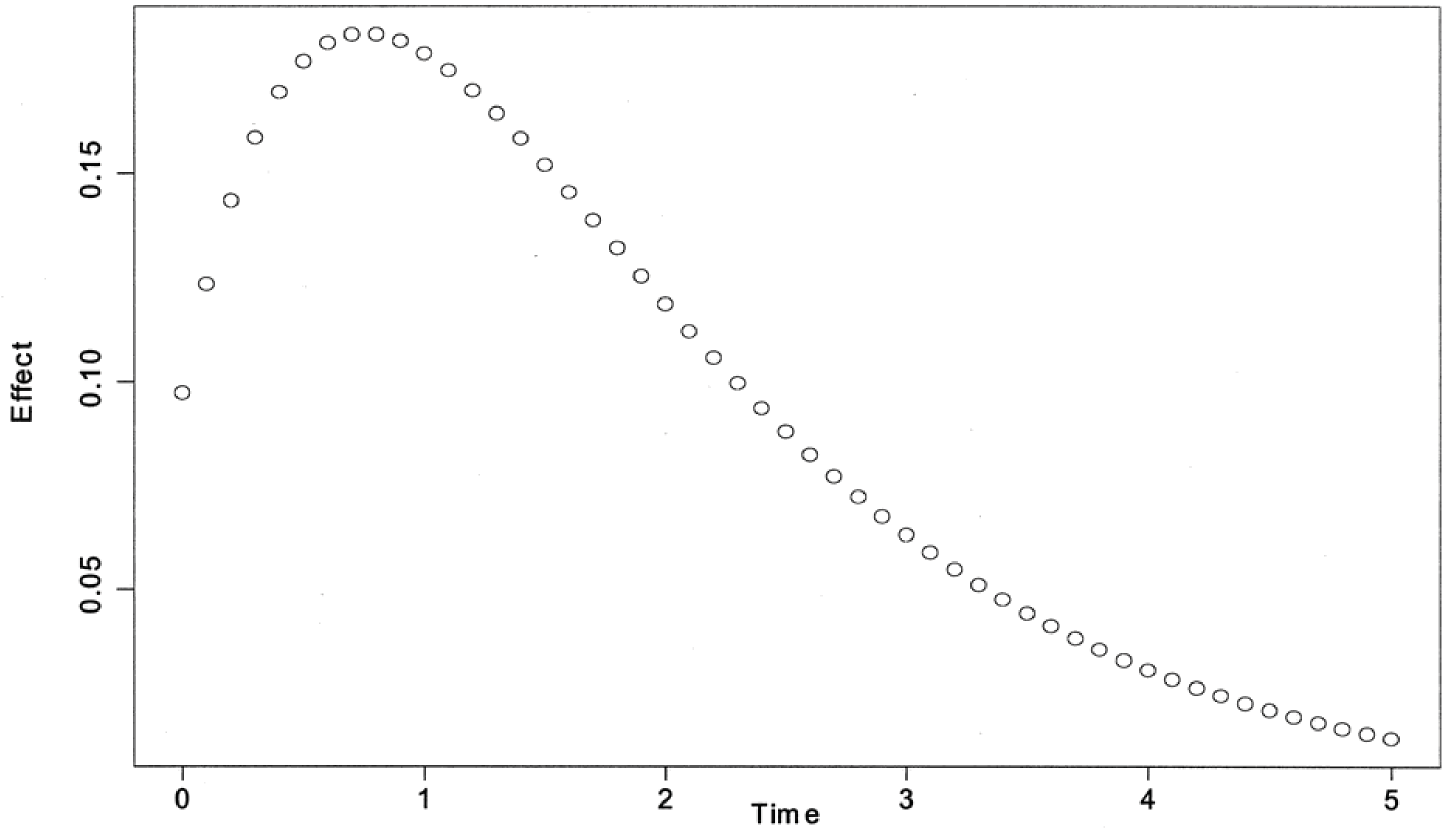
Effect throughout time

- In a linear model the sum of the effect of all independent variables, shifted by each time lag, is associated with the outcome

$$y(t) = \nu + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots + \beta_k x_{t-k} + \varepsilon_t \quad (1)$$

- Supposing that the number of events in a week follows the rainfall one week before
- This number increases up to two weeks after, and decreases smoothly up to the 5^{th} lag, the graphic of the β's of the model would present a curve such as:

Effect throughout time



A hypothesized curve showing the impact of an environmental toxin over time. The effect rises, and then falls, possibly with a long tail. The goal of this analysis is to determine what the actual shape of the curve representing the time course of deaths after exposure to PM10 is.

From: Schwartz: Epidemiology, Volume 11(3).May 2000.320-326

Alternative models

- Running average of the predictor → the shape of increase and decrease cannot be observed
- One parameter for each lag → no supposition about the shape of the curve
- To restrict the parameters to a specific shape → PDL (*Polynomial Distributed Lag*)
- To combine possible non linear effects with lag → DLNM (*Distributed lag non-linear models*)

Effect throughout time

- We use a transformation to represent the accumulated effect of X, weighted by a polynomial (2°degree)
- With this transformation of $X \Rightarrow Z$:
  - colinearity disappears
  - the shape induced on the relationship (in the example quadratic), imposes a restriction on the parameters
- After estimation of the parameters α of z, parameters β for X are obtained via back transformation
- The error of α goes back as well to β

When the effect is non-linear

- The solution is a combination of splines and lags
- cross-basis: a bidimensional space of functions describing simultaneously the shape and the effect distributed over time
- The idea is to specify two independent set of base functions
- PDL is a particular cases of DLNM, with a linear predictor

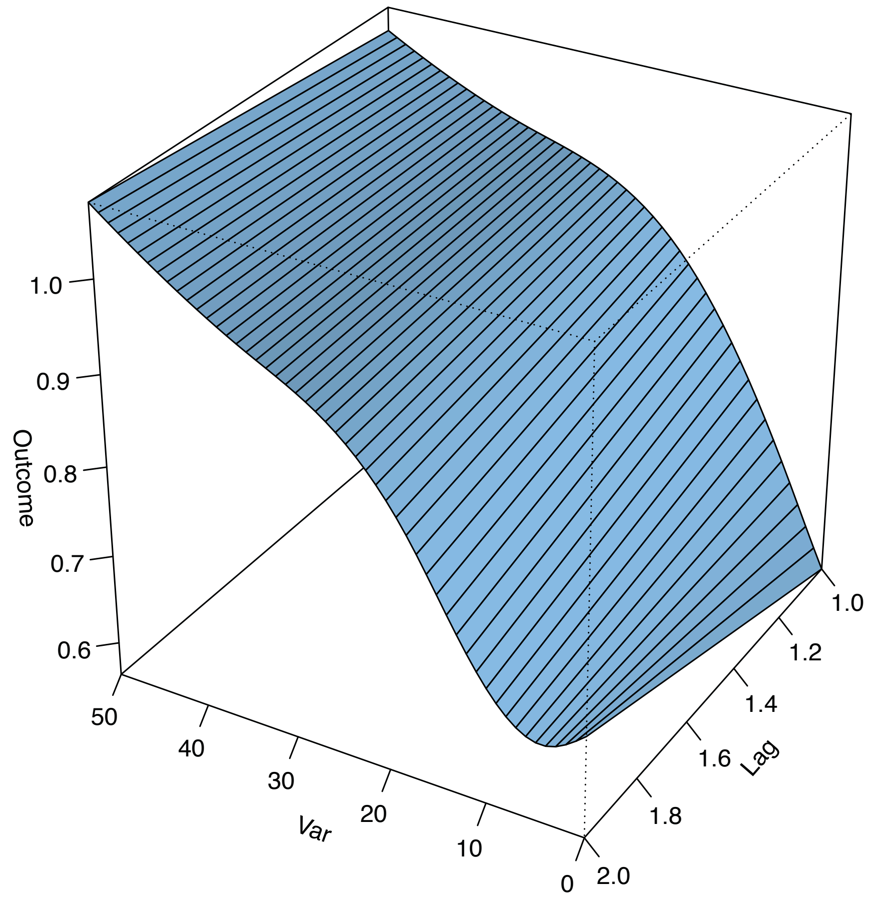
When the effect is non-linear

- The solution is a combination of splines e lags
- cross-basis: a bidimensional space of functions describing simultaneously the shape and the effect distributed over time
- The idea is to specify two independent set of base functions
- PDL is a particular cases of DLNM, with a linear predictor

How to interpret

- A grid is built on possible predicted values over time
- It is possible to evaluate the effect of a given value of the predictor over time $\rightarrow$ cut-points
- Or observe on each lag the shape of the relationship between predictor and outcome
- It is also possible to estimate the cumulative effect over time for values of the predictor

Rainfall & Leptospirosis



Outline

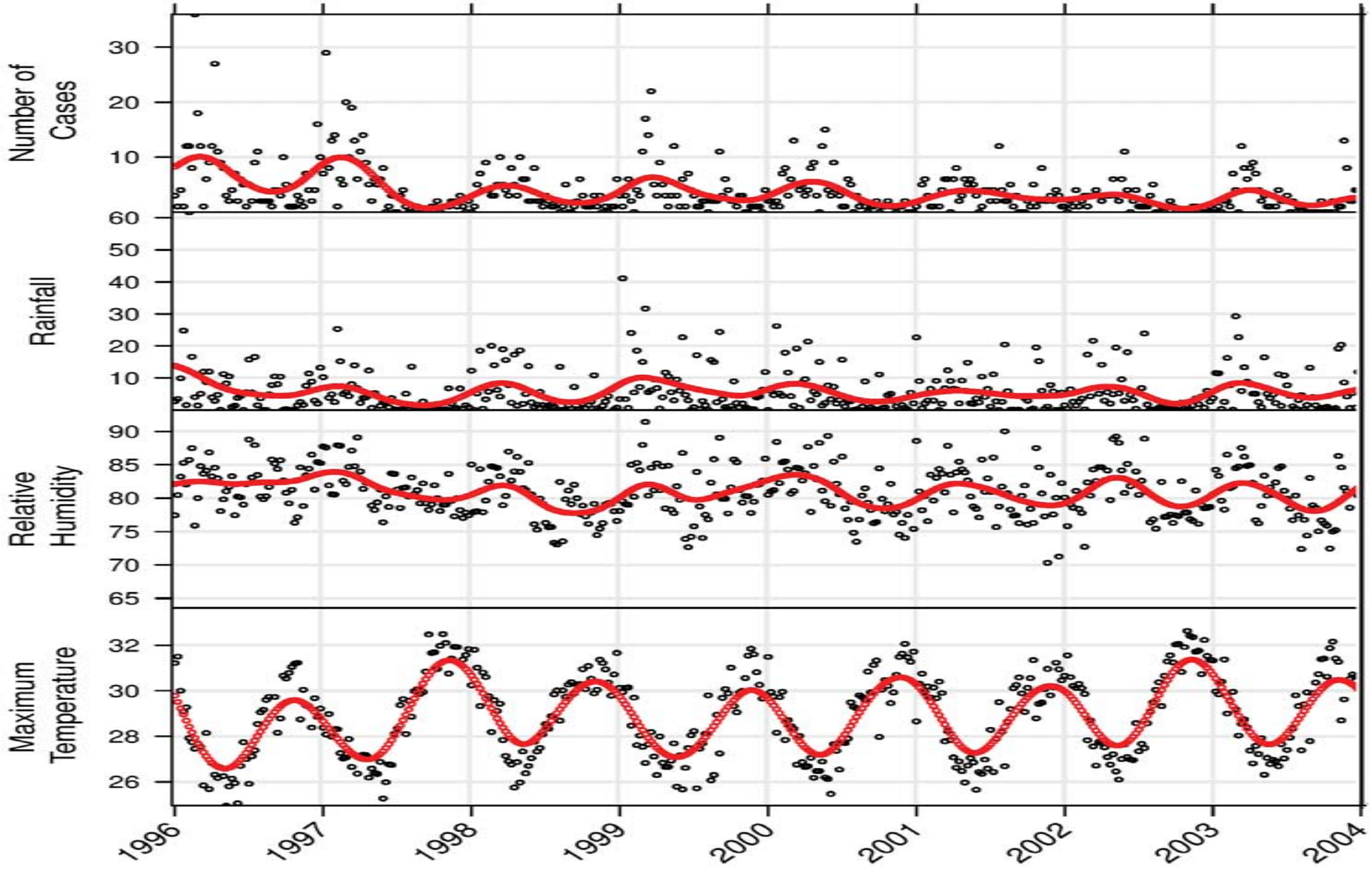
Models for Time Series

- ARIMA or SARIMA models: regression models where independent variables are just a shifted version of the dependent variable.
- Stationary time series:
  - stochastic process whose joint probability distribution does not change when shifted in time or space
  - mean and variance do not change over time or position
  - removing trend and seasonality (S and I terms)
- detection of order of autoregressive and moving average terms
- fit, evaluation, ...
- prediction
- Dynamic models!

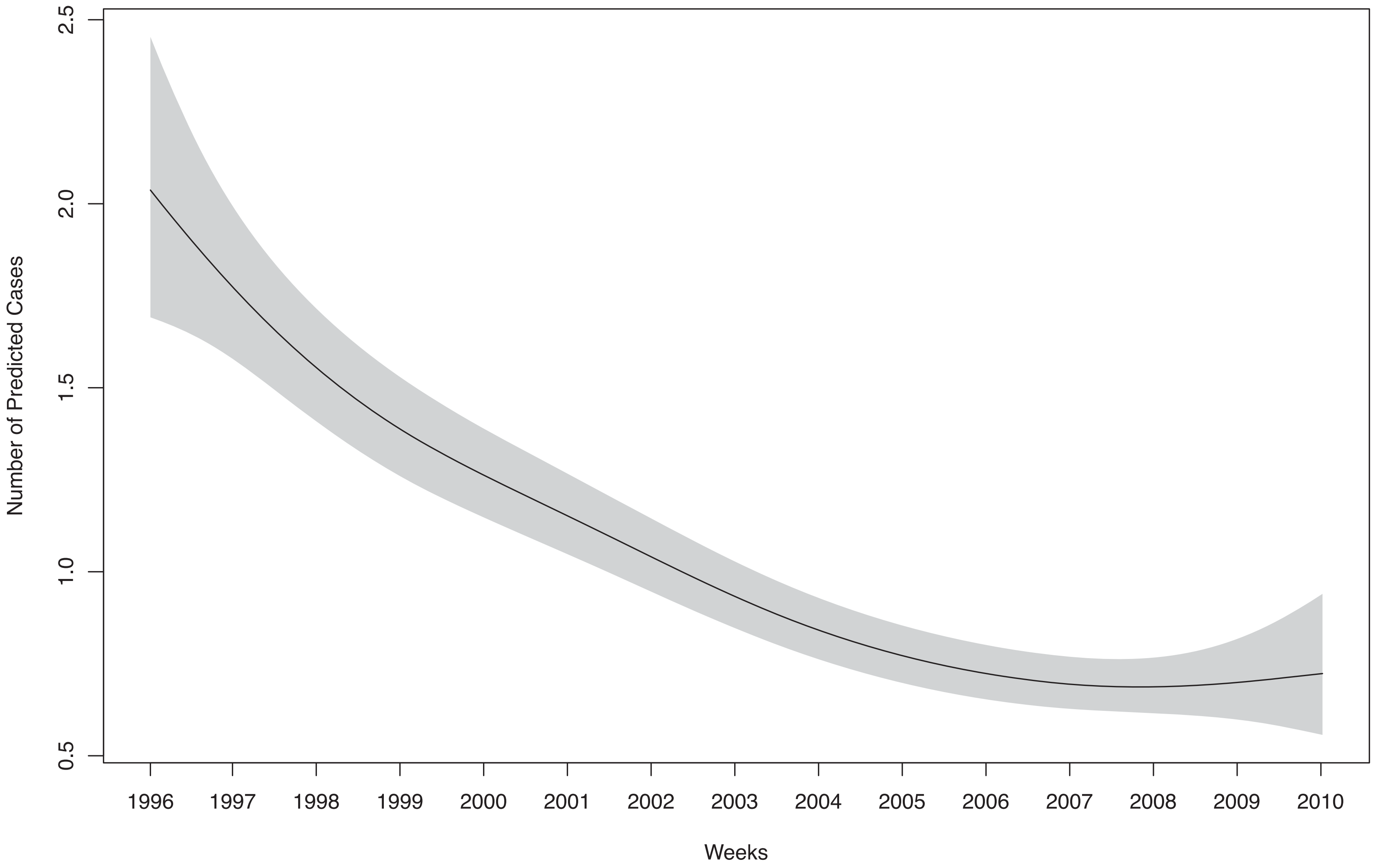
Modelling for:

- Explaining why events happen this way over time:
  - Independent variables are associated with events y in $t \rightarrow$ regression
  - Past events are “cause” of present events $t \rightarrow$ dynamic models
- How to predict $t + k$?

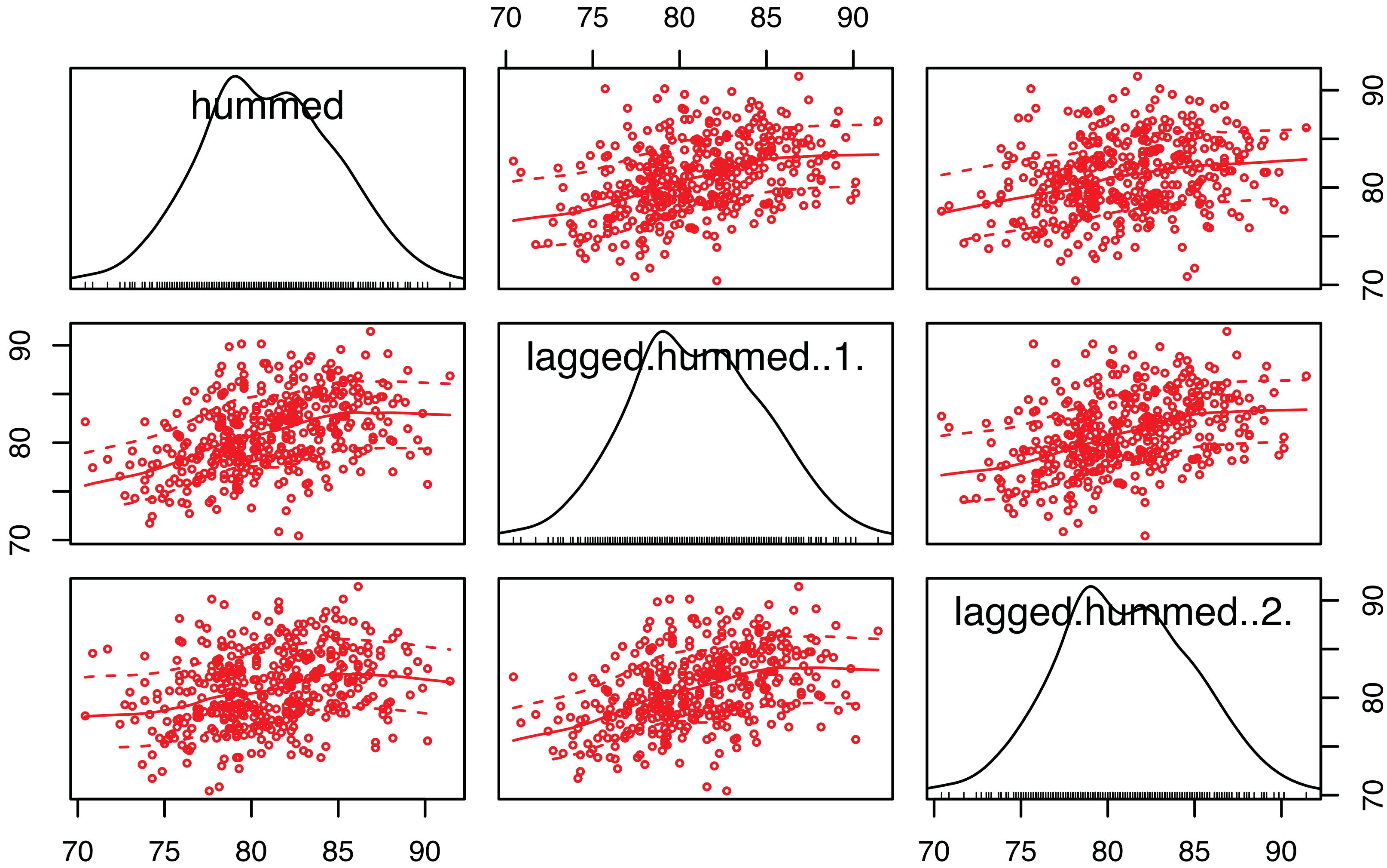
Exploratory analysis



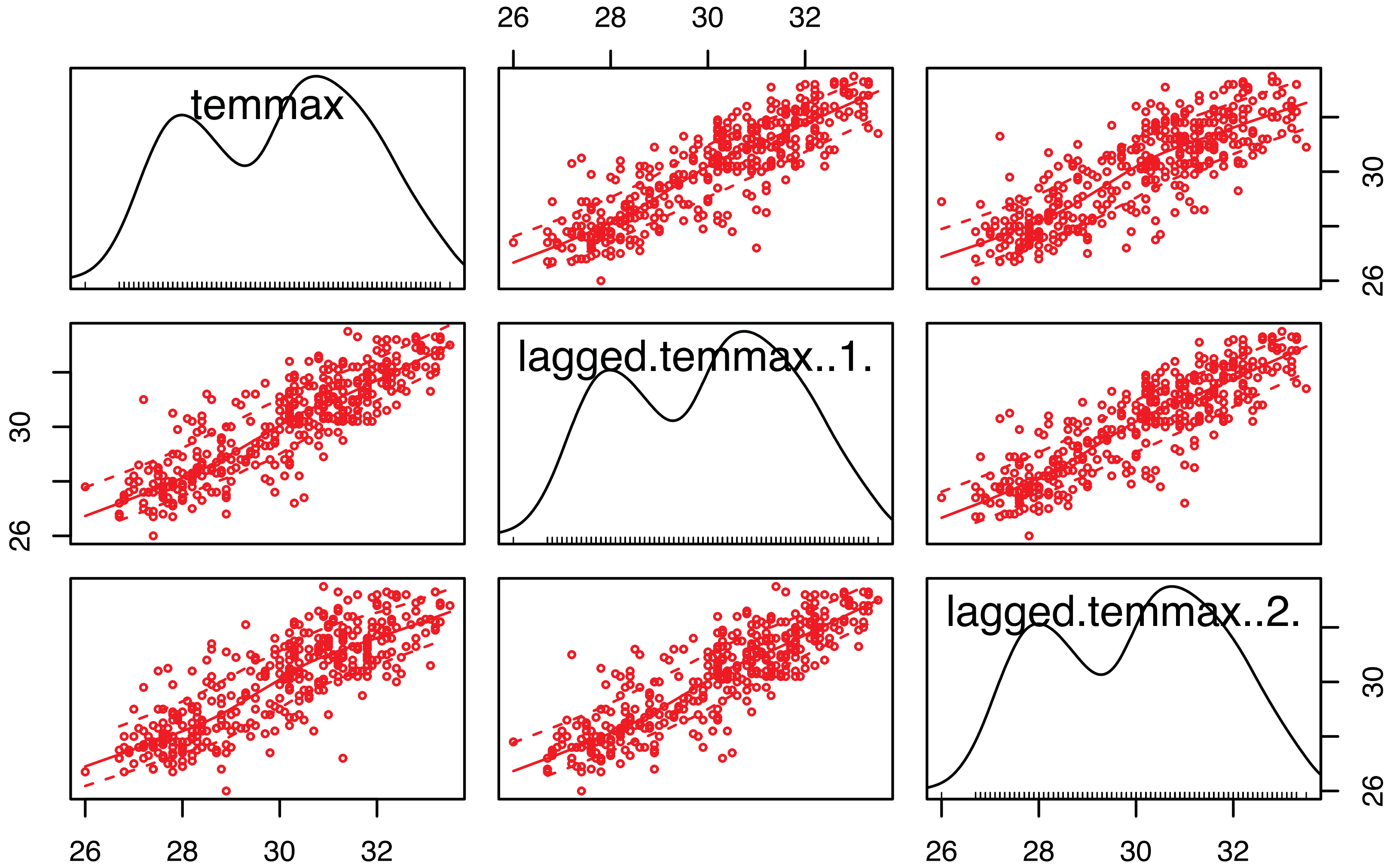
Exploratory analysis – trend



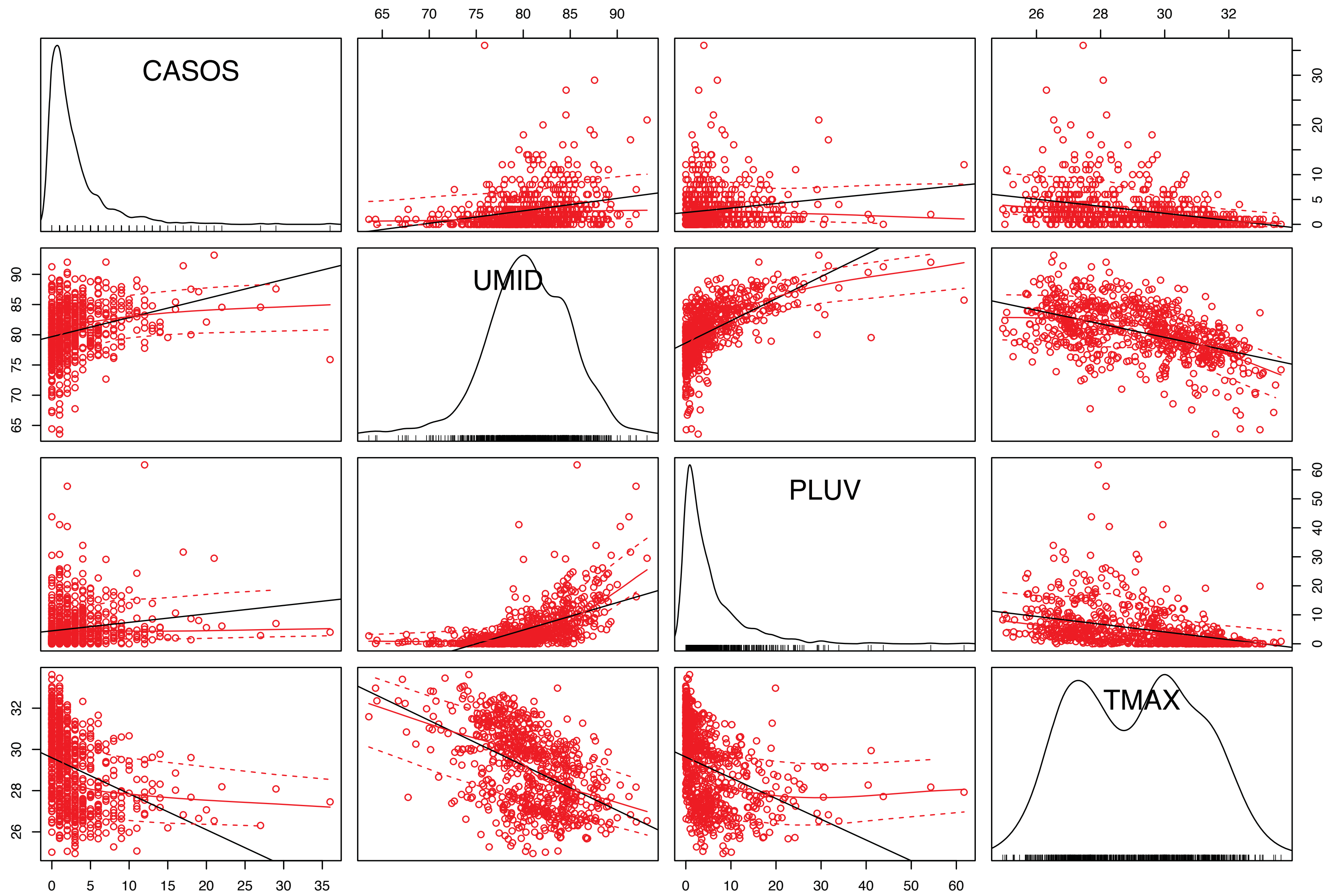
Colinearity



Colinearity



Colinearity



Structure

- Autocorrelation
- Components

TS Components

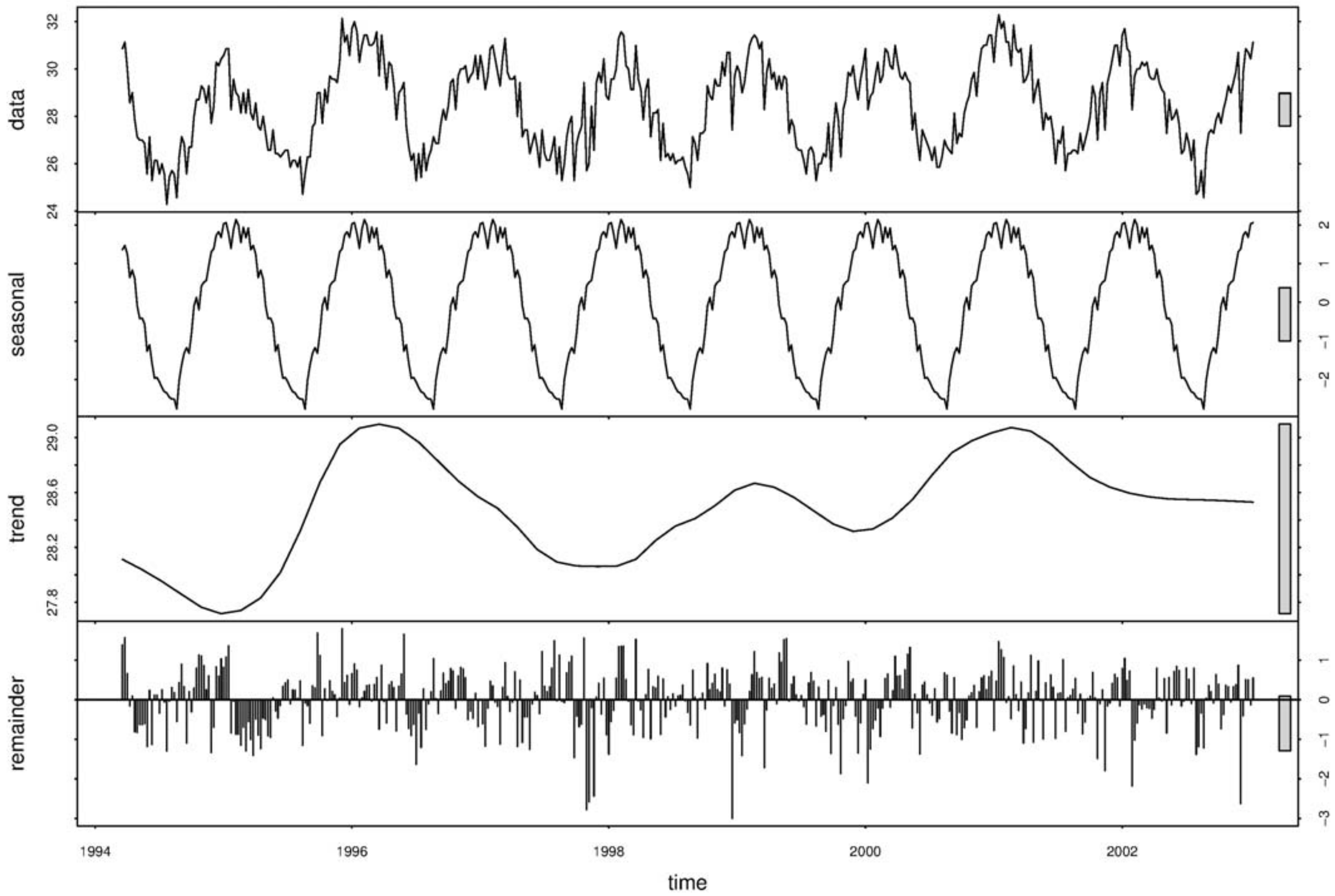
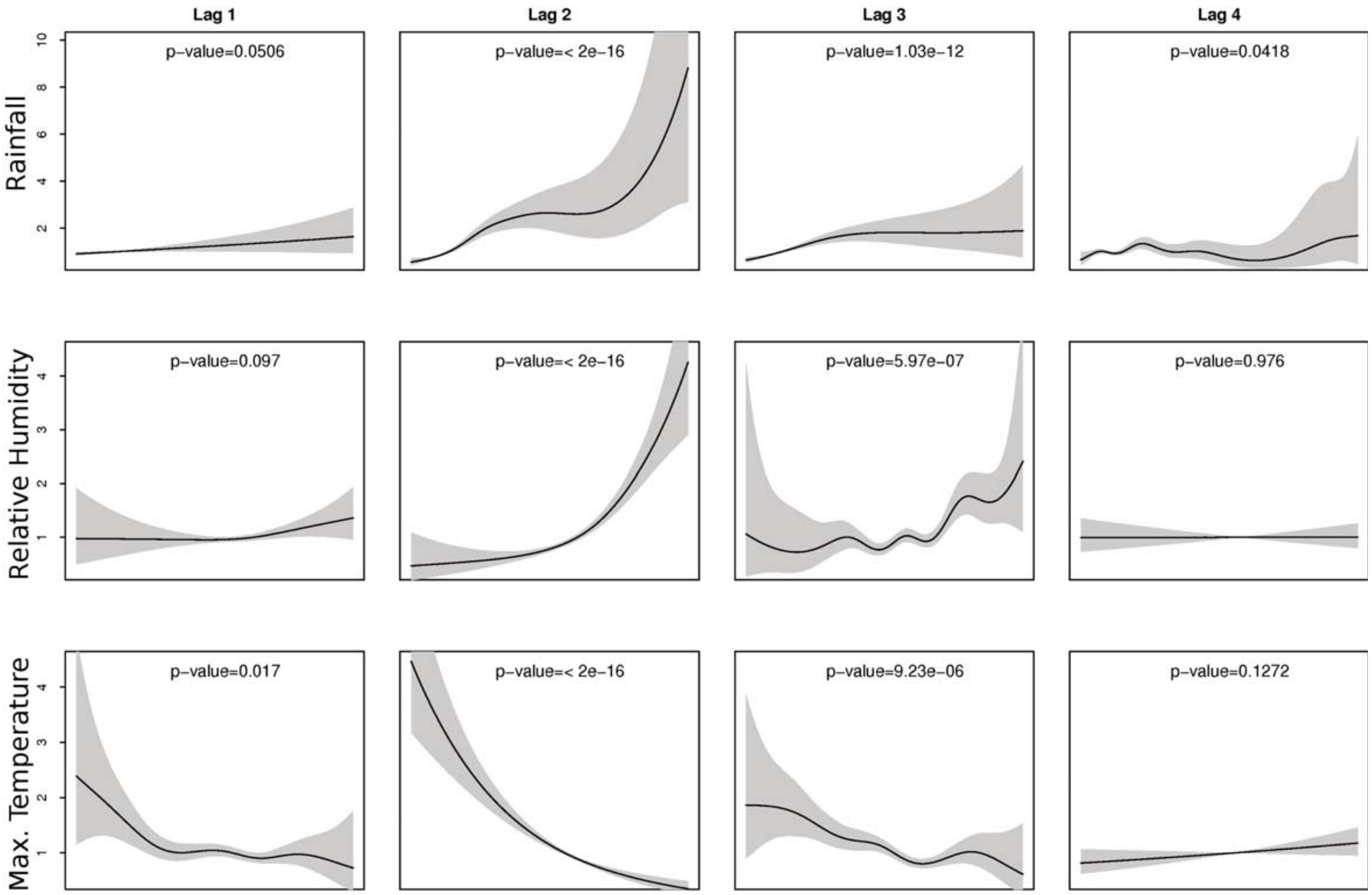


Fig.: Maximum temperature

Functional form



Seasonality

- Exclude the seasonality of the independent variables using sinusoid functions.
- Use the residuals of this seasonal model as independent variables
- Include a seasonal term in the complete model
- Interpretation is the same, as the residuals keep the same measure unit: the meaning of the parameter estimated is the same

Multiple model

- Test the significance of each time lag, respecting the functional form
- Join all lags and covariates
- When the functional form is not linear $\rightarrow$ categorise, segmented regression, CART model (Classification and regression trees)
- Splines & PDL

Residuals

- ACF of residuals again
- still trend?
- inclusion of AR term

Summary

- Counts: Poisson, Quasipoisson or Negative Binomial → overdispersion!
- Trend and seasonality → `s(tempo)` e `s(tempo, k=52)`
- Removal of seasonality of independent variables
- Regression model

```
gam(cases ~ offset(log(pop)) + s(time) +
          sin(2*pi*(1:\text{length(dataset)}/52.14) +
          covs + lag(cases, 1),
          family=negbin(c(1,10), data=dataset)
```

Modelling Time

Marilia Sá Carvalho

Fundação Oswaldo Cruz

Outline

Outline

1. Introduction
   - Motivating Example – Leptospirosis

Statistical Analysis

- Exploratory Analysis to:
  - describe the data
  - support the selection of appropriate statistical techniques
- Hypothesis testing:
  - Does this observed pattern differ from... ?
- Modelling:
  - What is the effect of rainfall, humidity and temperature on the number of cases of malaria?

References

- Cryer, J.D.; Chan, K-S. *Time Series Analysis: With Applications in R*. Springer Texts in Statistics, 2010, 2nd Ed.
- Hastie, T.; Tibshirani, R. *Generalized Additive Models*. Chapman & Hall, 1990.
- Wood, S.N. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science Series, 2006.
- Faraway, J.J. *Extending the Linear Model with R*. Chapman & Hall/CRC Texts in Statistical Science Series, 2006.

Time Series

- A sequence of data points, measured typically at successive points in time spaced at uniform time intervals
- Time series analysis $\rightarrow$ methods for analysing time series data in order to extract meaningful statistics
- Natural temporal ordering can result in serial dependence $\rightarrow$ dependence of each time point on previous points
- Components:
  - Trend
  - Seasonality and cyclical patterns
  - Time dependence structure

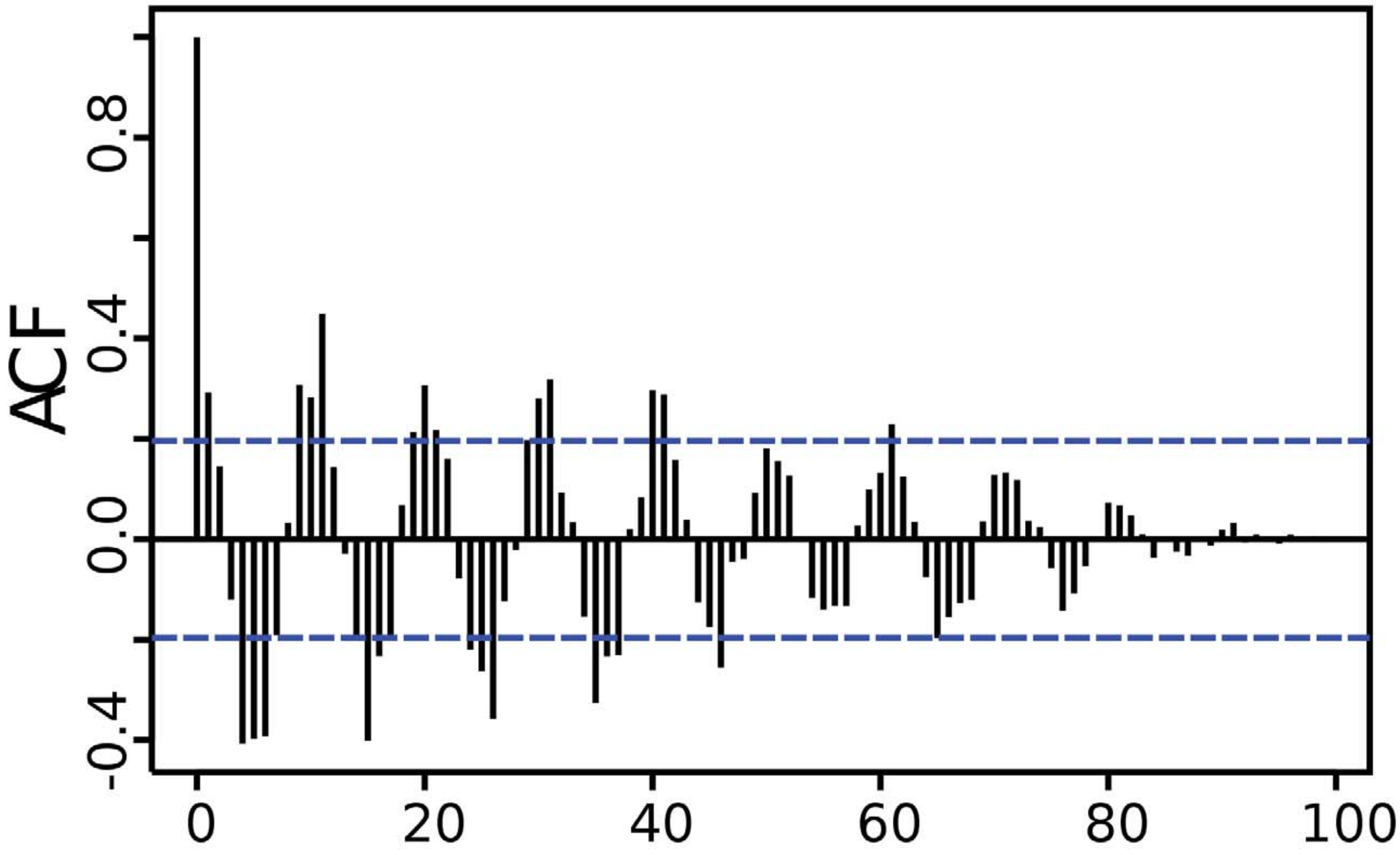
Time dependence structure: autocorrelation

- Autocorrelation (or autocovariance) is a measure of similarity of the event over time with itself previously
- It is the correlation between values of a random process at different times

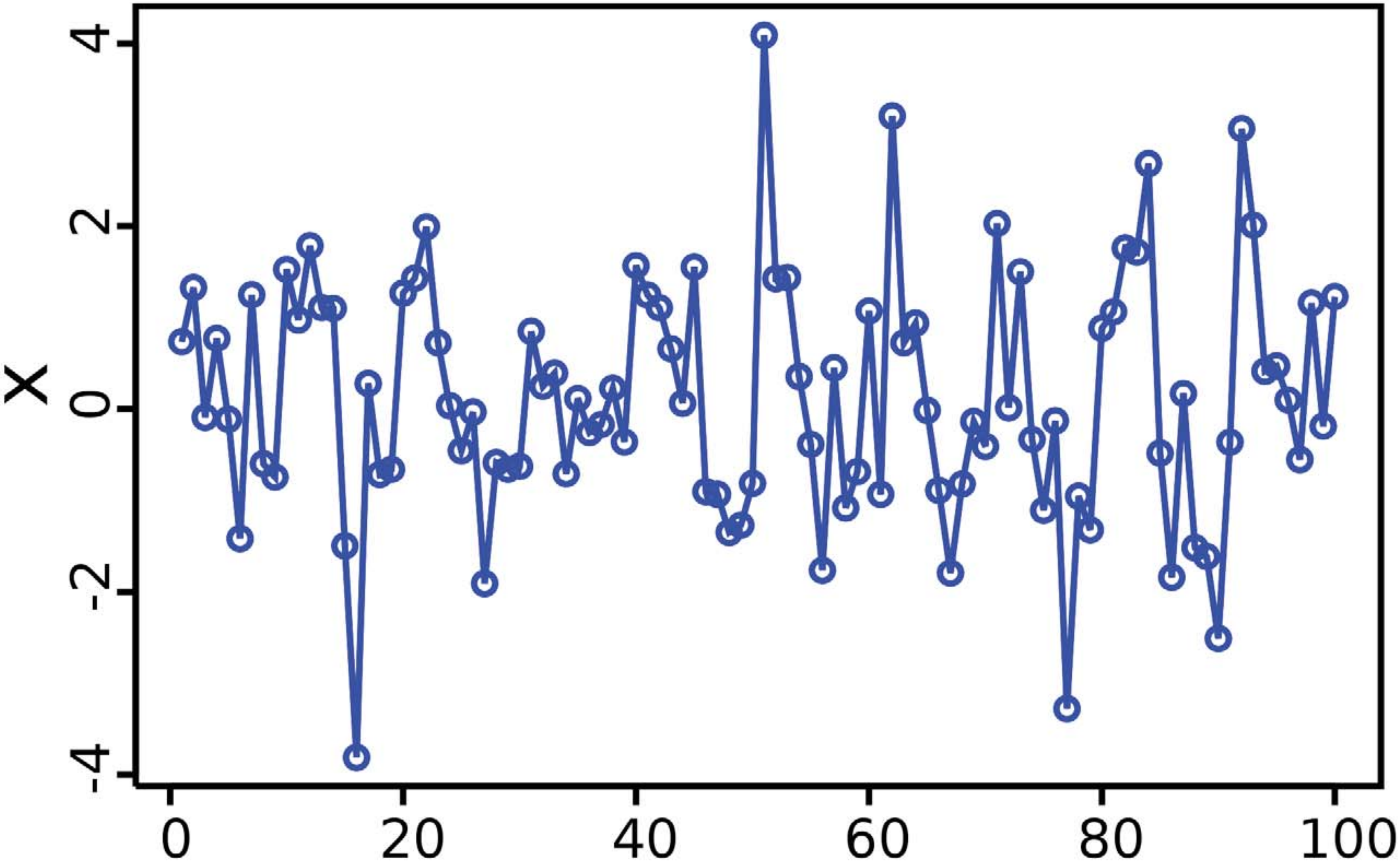
$$r_k = \frac{\sum_k (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum (x_t - \bar{x})^2}$$

- It is a tool to depict the structure of the time series

ACF – example



ACF – example



Outline

Motivating Example – Leptospirosis Epidemics

- Bacterial zoonosis (*Leptospira sp*)
- Transmitted to humans through contact with urine from infected animals (rats in urban setting)
- Clinical manifestations:
  - self-limiting fever, with headache and muscle pain $\rightarrow$ easily taken for a bad cold or dengue fever
  - life-threatening disease $\rightarrow$ kidney failure, pulmonary hemorrhage, Weil's syndrome
  - early treatment! (dialysis mainly)
- Globally spread, affecting people on all continents – 5-10% mortality of severe cases; about 607 deaths in 2014
  - Sporadic disease, related with specific occupational exposures and recreational activities
  - Slums and flooding in urban areas

Leptospirosis & Climate

- People living in slums $\rightarrow$ a seroprevalence survey at Pau-da-Lima (Salvador/BA) indicates 23% at 50 years of age
- However, not many severe cases (three in 8 years)
- Severe cases numbers increase during the tropical storms season
- Reasoning: heavy rainfall cleans out the rats holes, bringing the *Leptospira* to the soil surface
- People clean mud after flooding $\rightarrow$ large inoculant dose

The environment

The people

Leptospirosis & Climate: main questions

- Does rainfall really lead to severe leptospirosis epidemics?
- Are other environment factors – humidity & temperature – involved?
- Is there a threshold?
- What is the time delay between tropical storms and increase in the number of cases?
  - duration of incubation period
  - survival of *Leptospira* on the soil, possibly related to temperature, sun and moisture

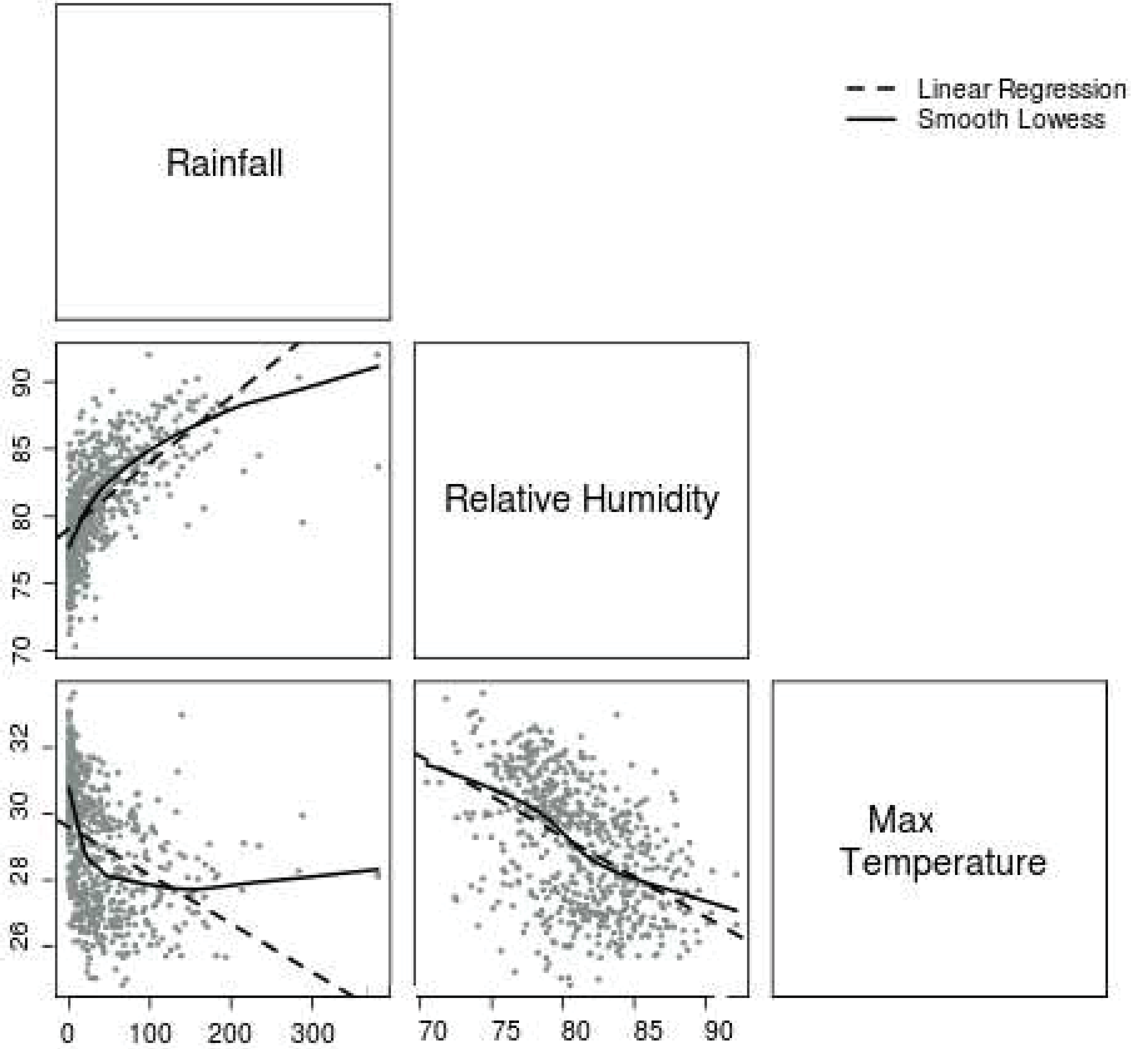
Data

- Local epidemiology surveillance system
- Weekly aggregated cases
- Climate covariates (per week):
  - temperature (mean and maximum)
  - mean relative humidity
  - accumulated rainfall

Outline

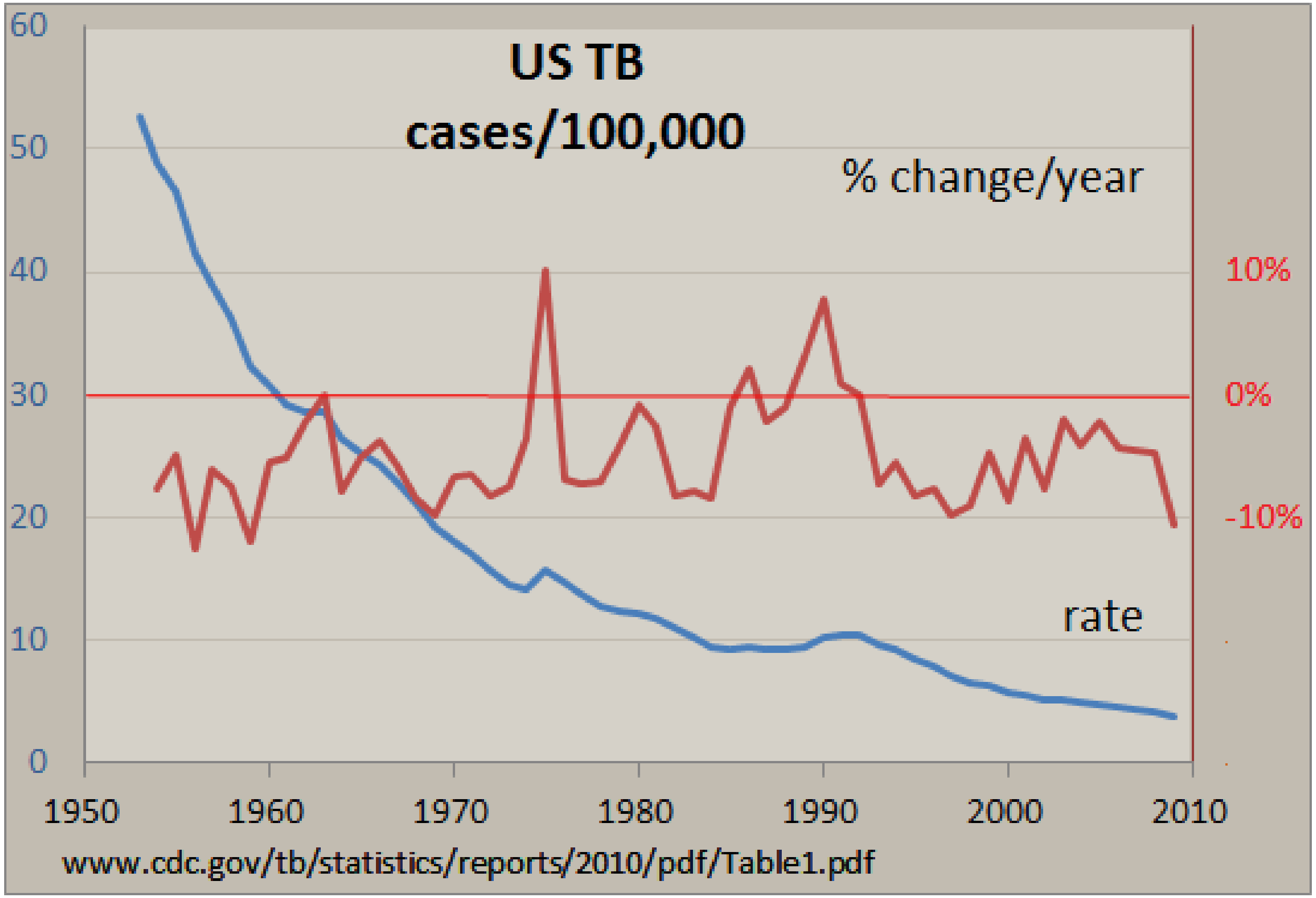
2. Exploratory Analysis
    - Kernel
    - Loess
    - Splines

Exploratory analysis – The usual



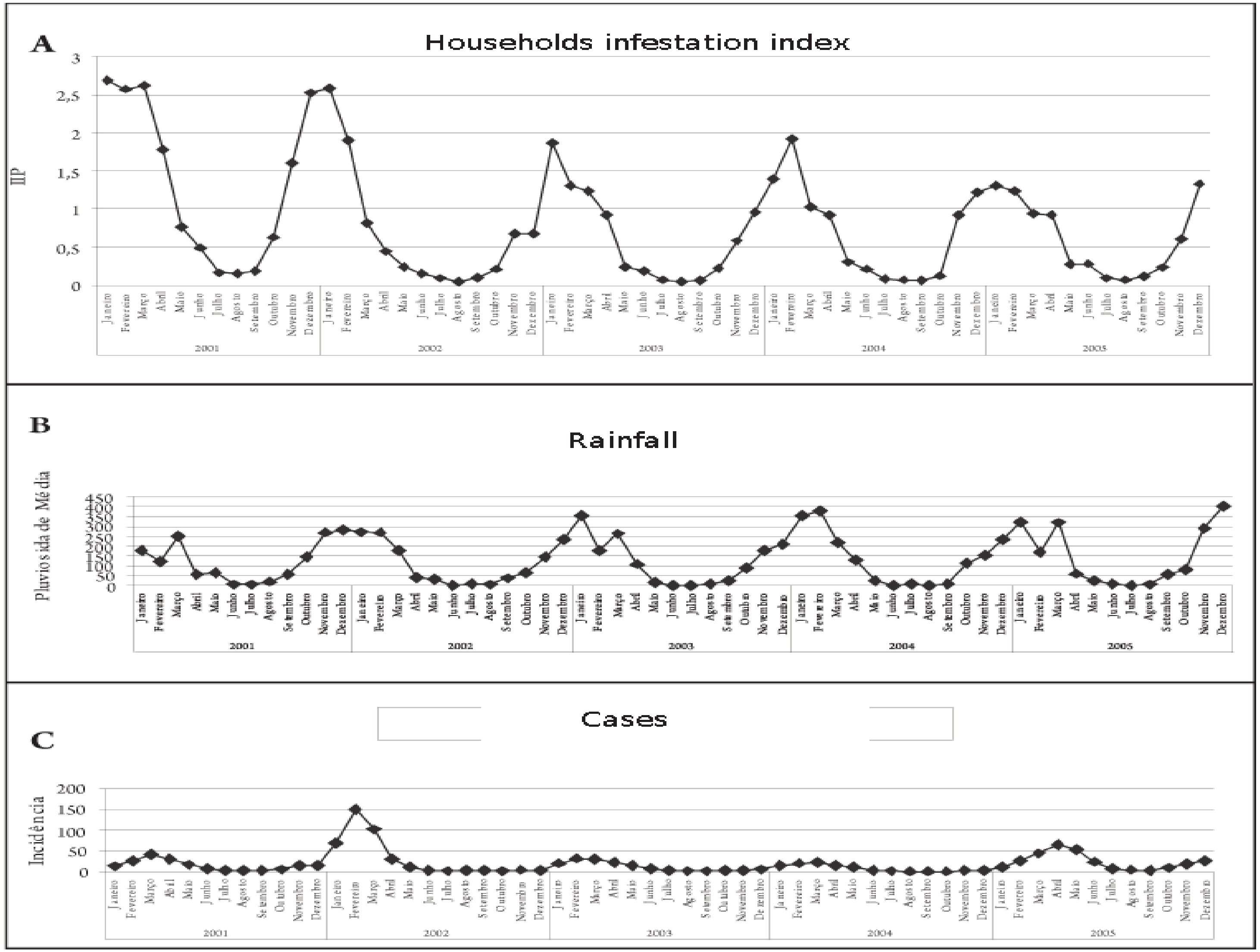
Exploratory analysis – Line Charts

Tuberculosis



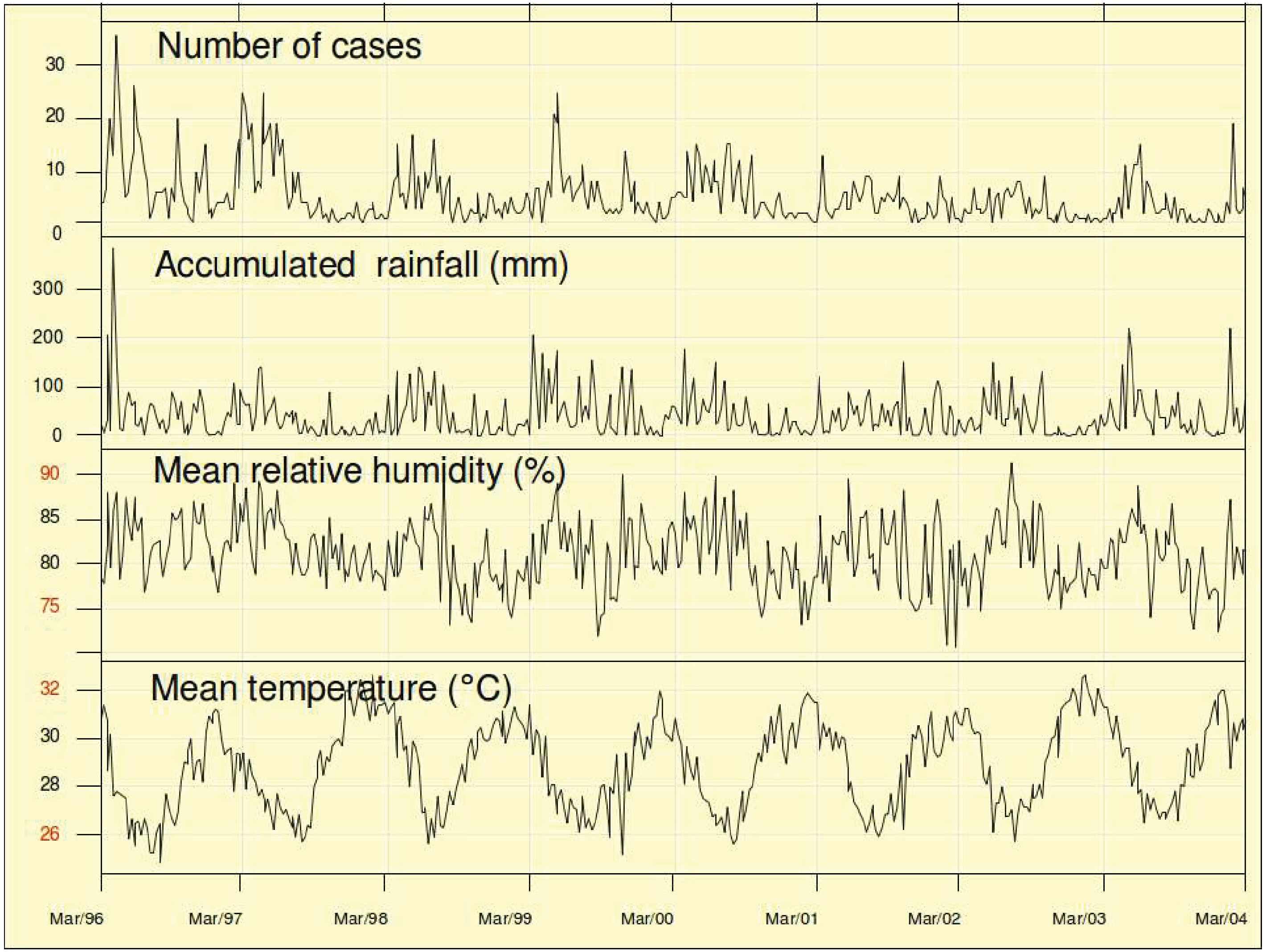
Exploratory analysis – Line Charts

Aedes aegypti & rainfall



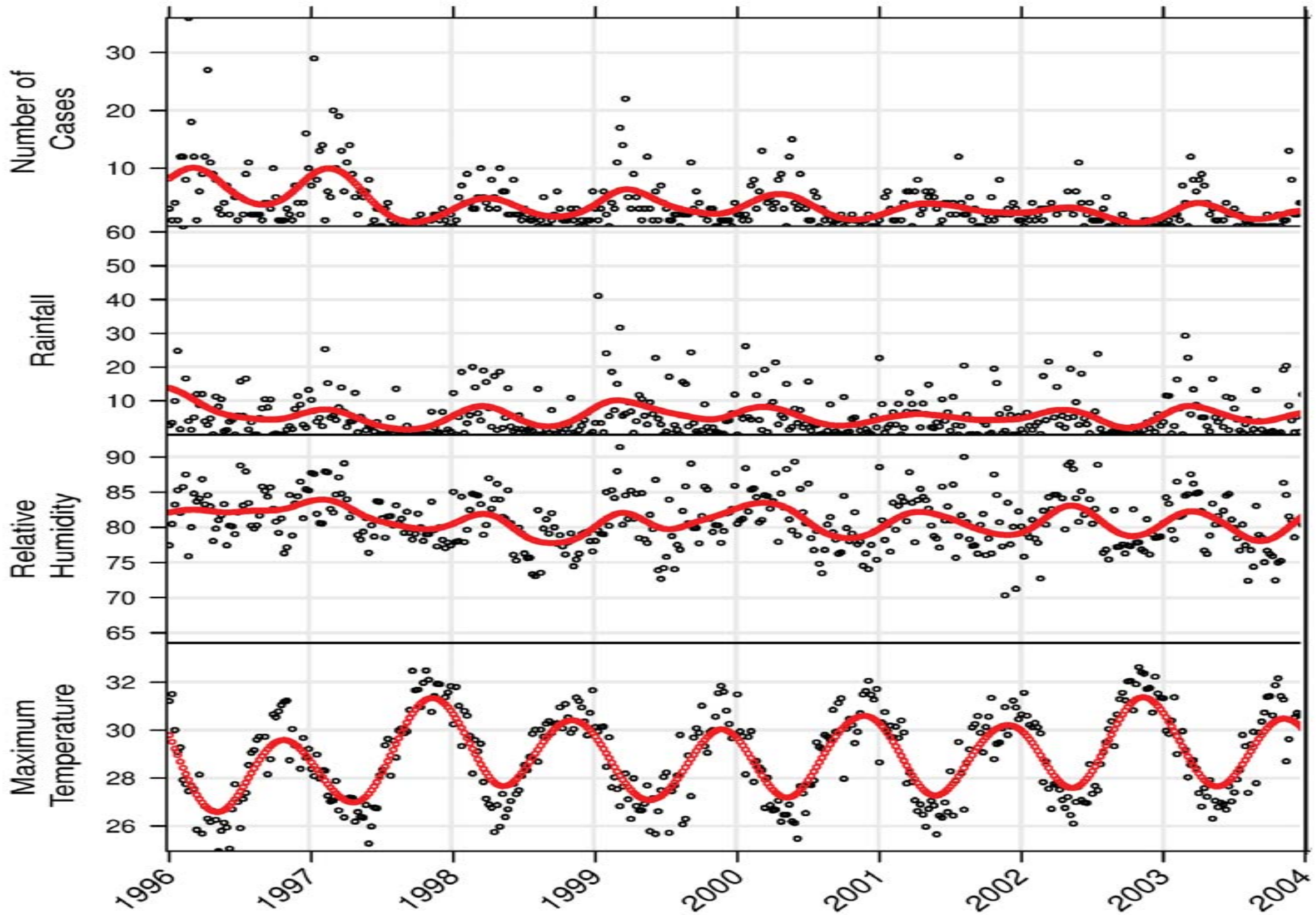
Exploratory analysis – Line Charts

Leptospirosis data



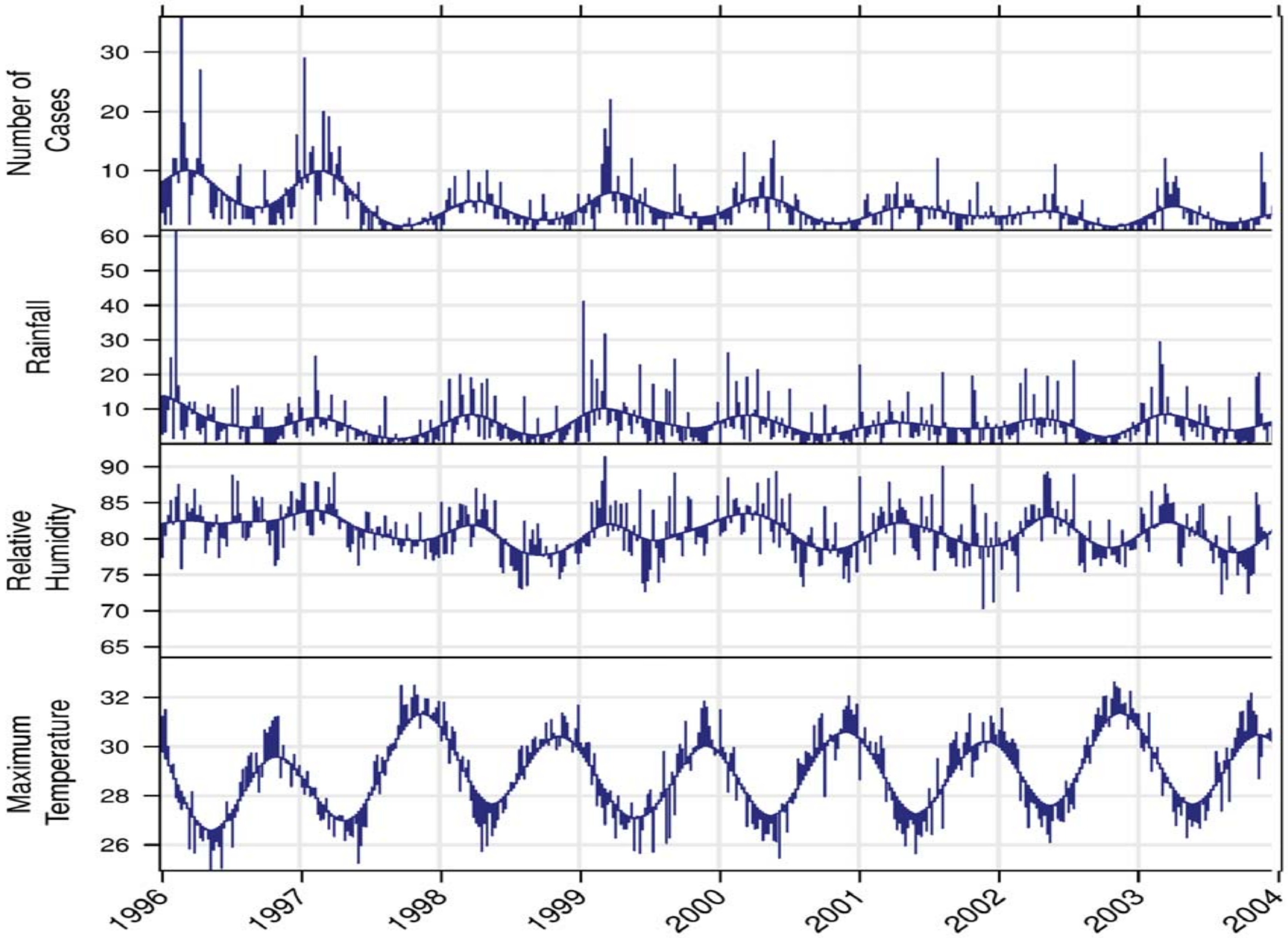
Exploratory analysis – Smoothing

Leptospirosis



Exploratory analysis – Smoothing

Leptospirosis



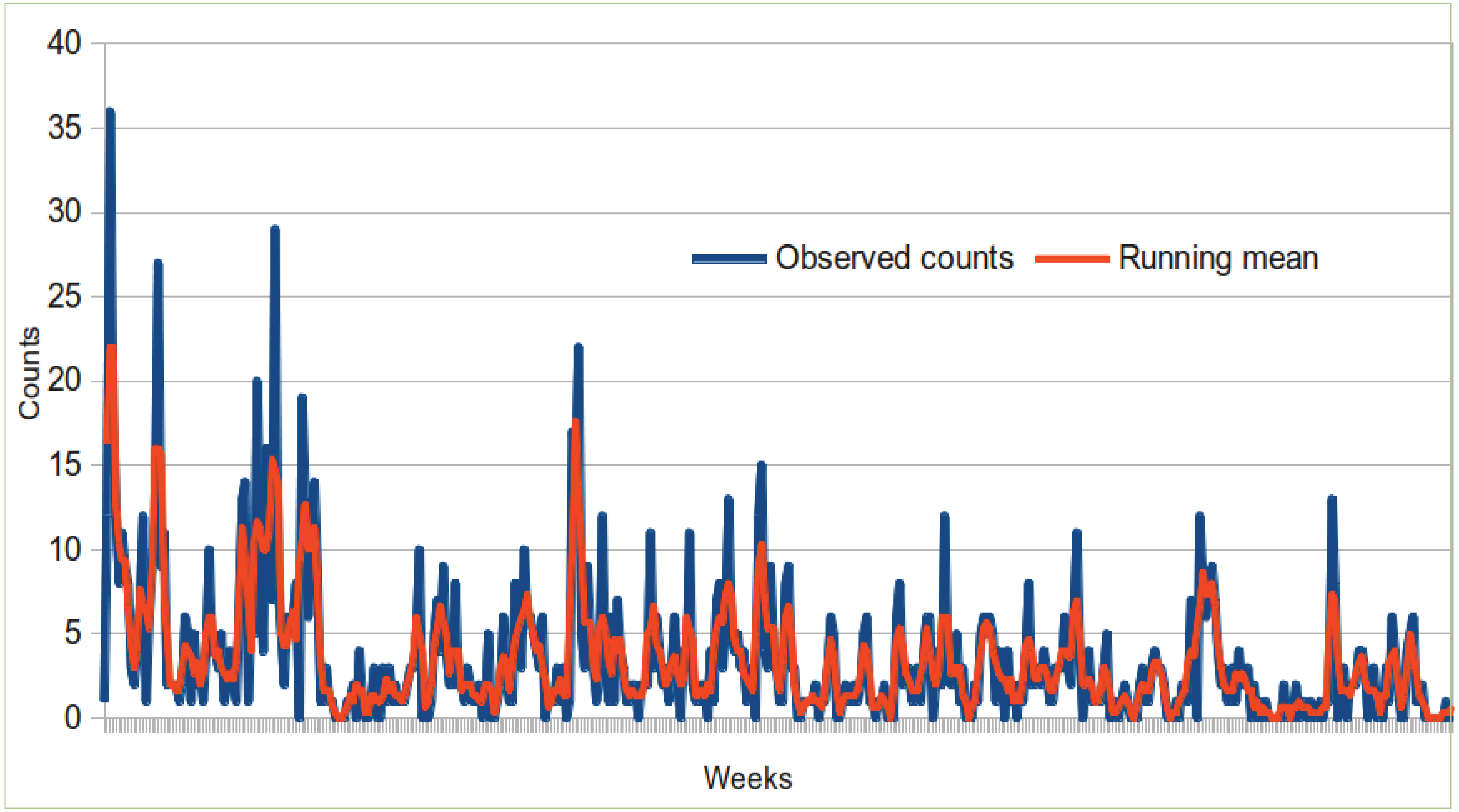
Smoothing

- Moving average – very simple
- Kernel density – a non-parametric way to estimate the probability density function of a random variable
- LOESS or LOWESS – locally weighted scatterplot smoothing
- Splines – minimisation of an objective function where a trade-off between fidelity to the data and roughness of the function estimate is explicit

Outline

2 Exploratory Analysis
- Kernel

Running average



Kernel – the algorithm

1. Define the kernel function:
    - symmetric
    - unimodal
    - centred on (x)
    - going to zero at the edge – neighbourhood
2. Let (x) be the point where to estimate $f(.)$
3. Define the limit of the area of influence of each point $\rightarrow$ window or bandwidth
4. This range controls the smoothing parameter of the kernel function
5. Calculate the value of $f(x)$ for each point and connect them.

Kernel – the function

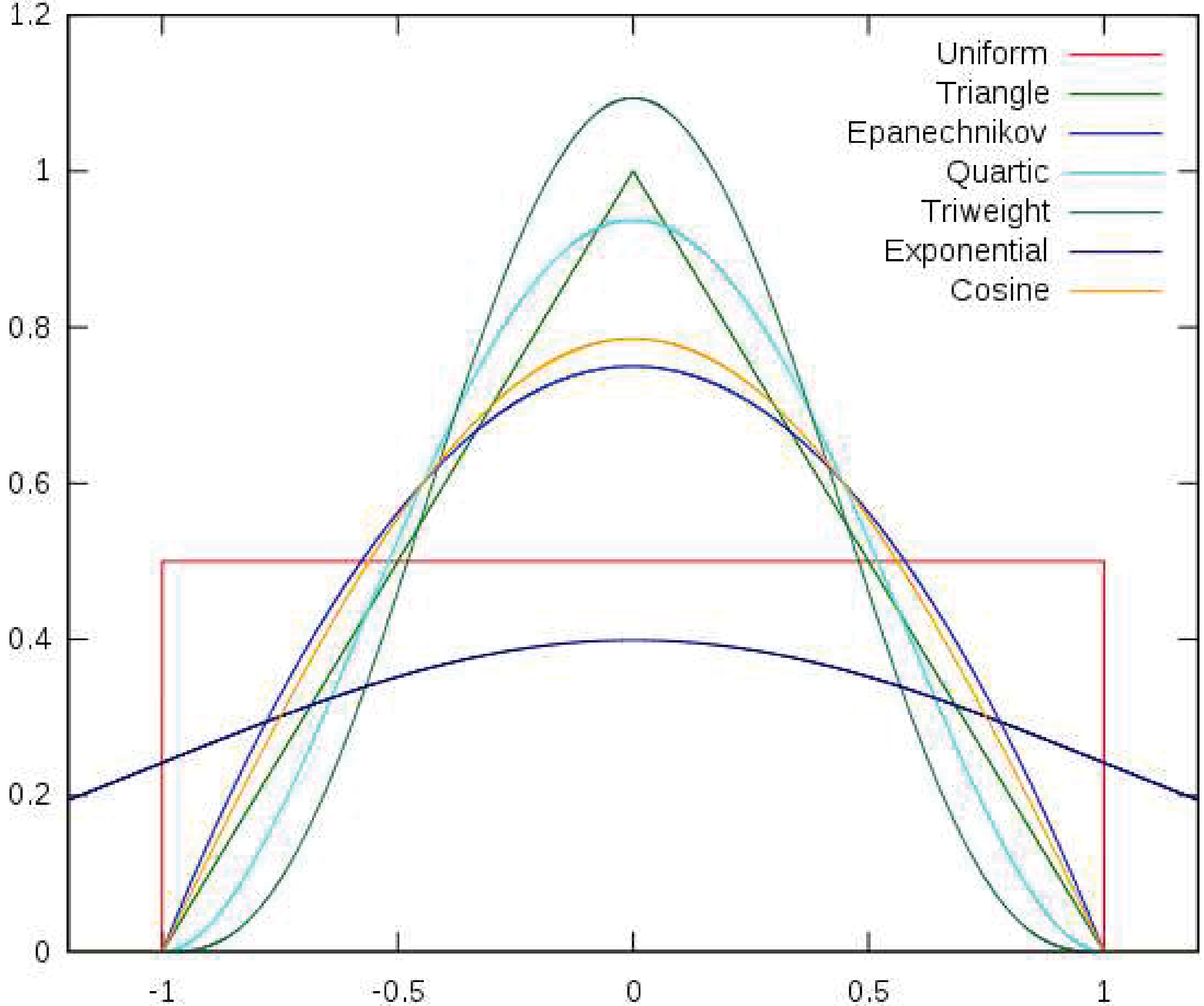
$$\hat{f}_h(x) = \frac{1}{Nh} \sum K\left(\frac{x - x_i}{h}\right)$$

h $\rightarrow$ bandwidth – can be estimated by cross validation

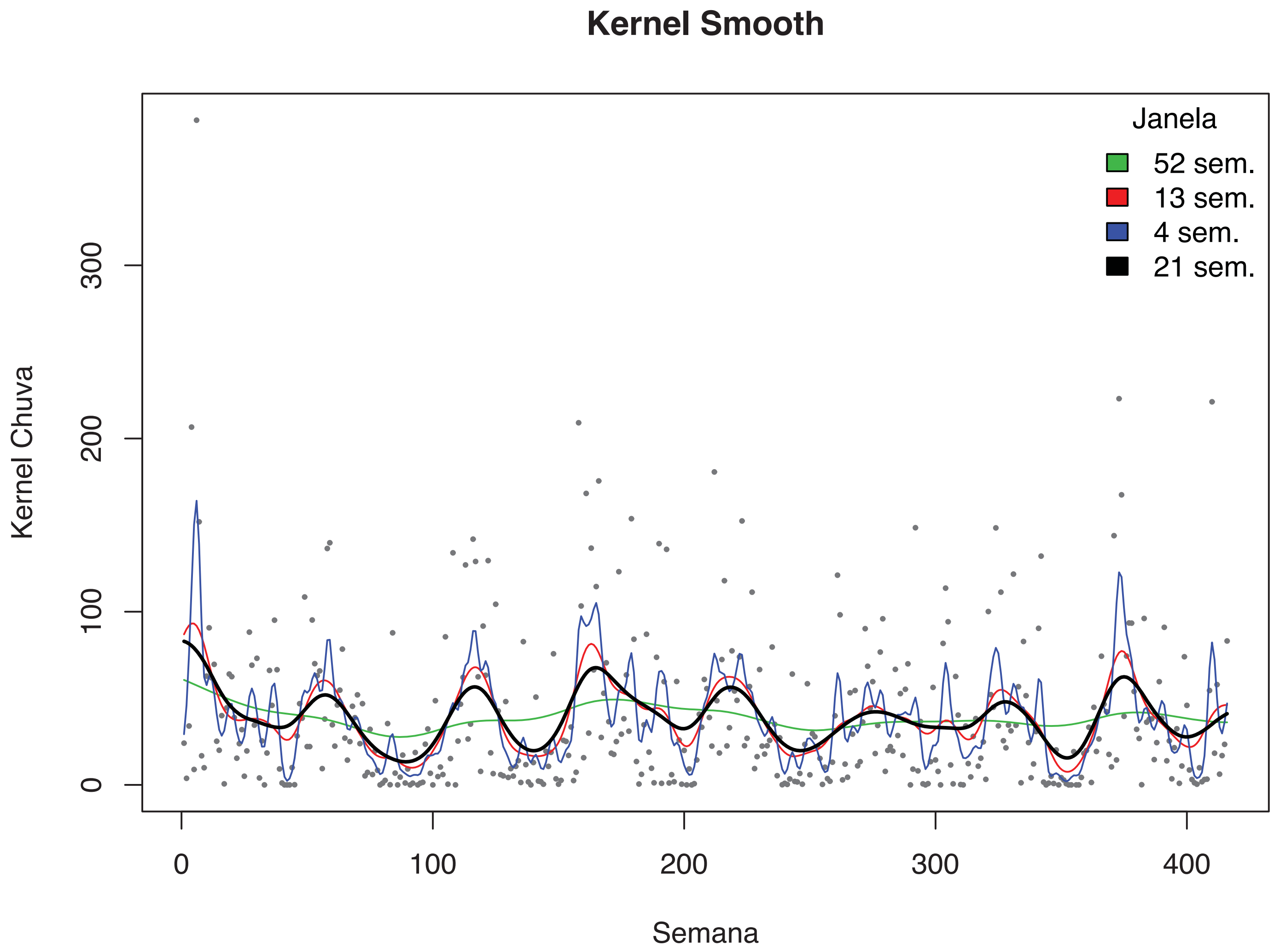
K $\rightarrow$ smoothing function

Gaussian Kernel: $k(x) = \frac{1}{\sqrt{2\pi}} exp(1/2x^2)$

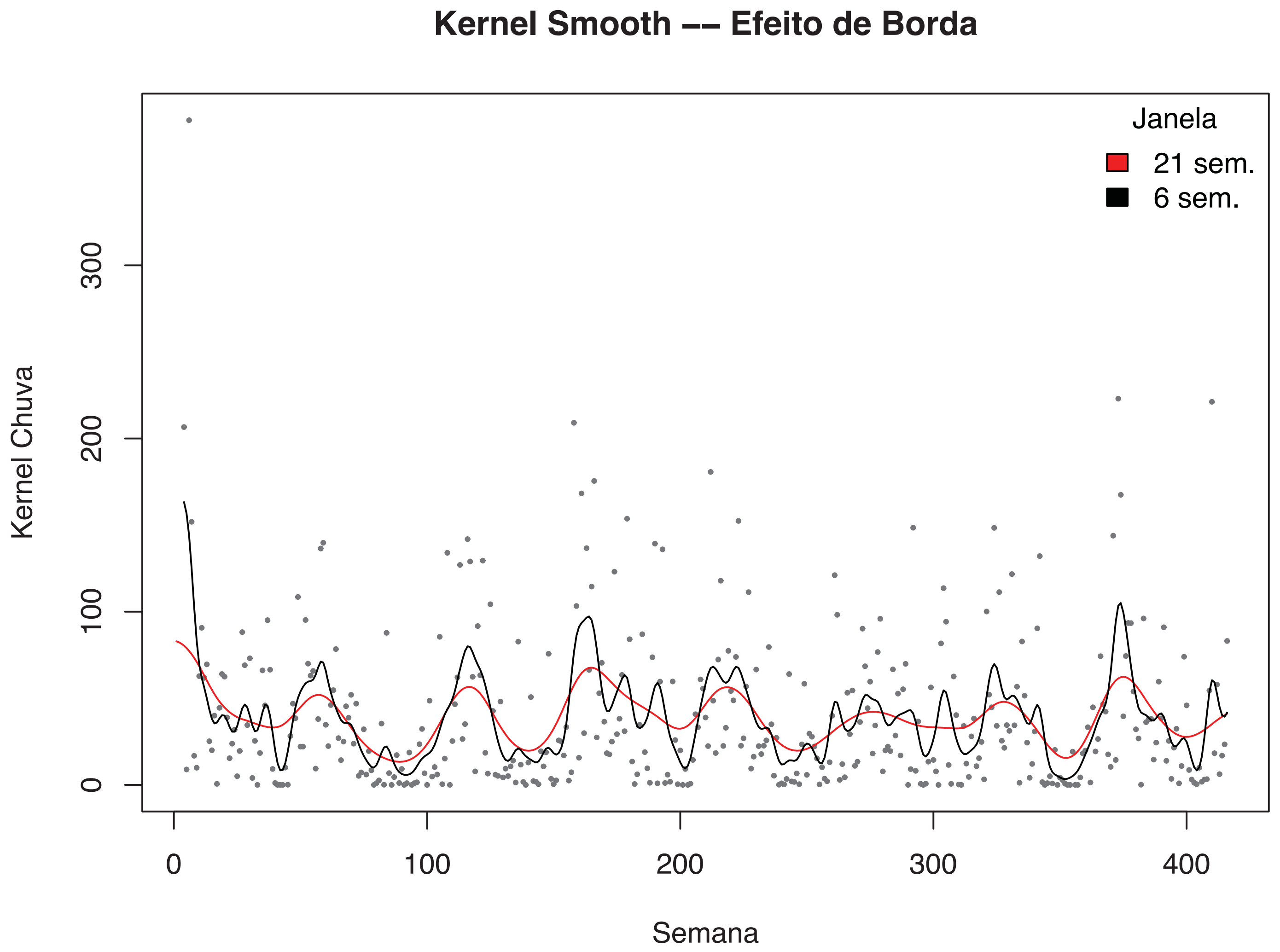
Kernel – several functions



Kernel – Example



Kernel – Border effect



Kernel

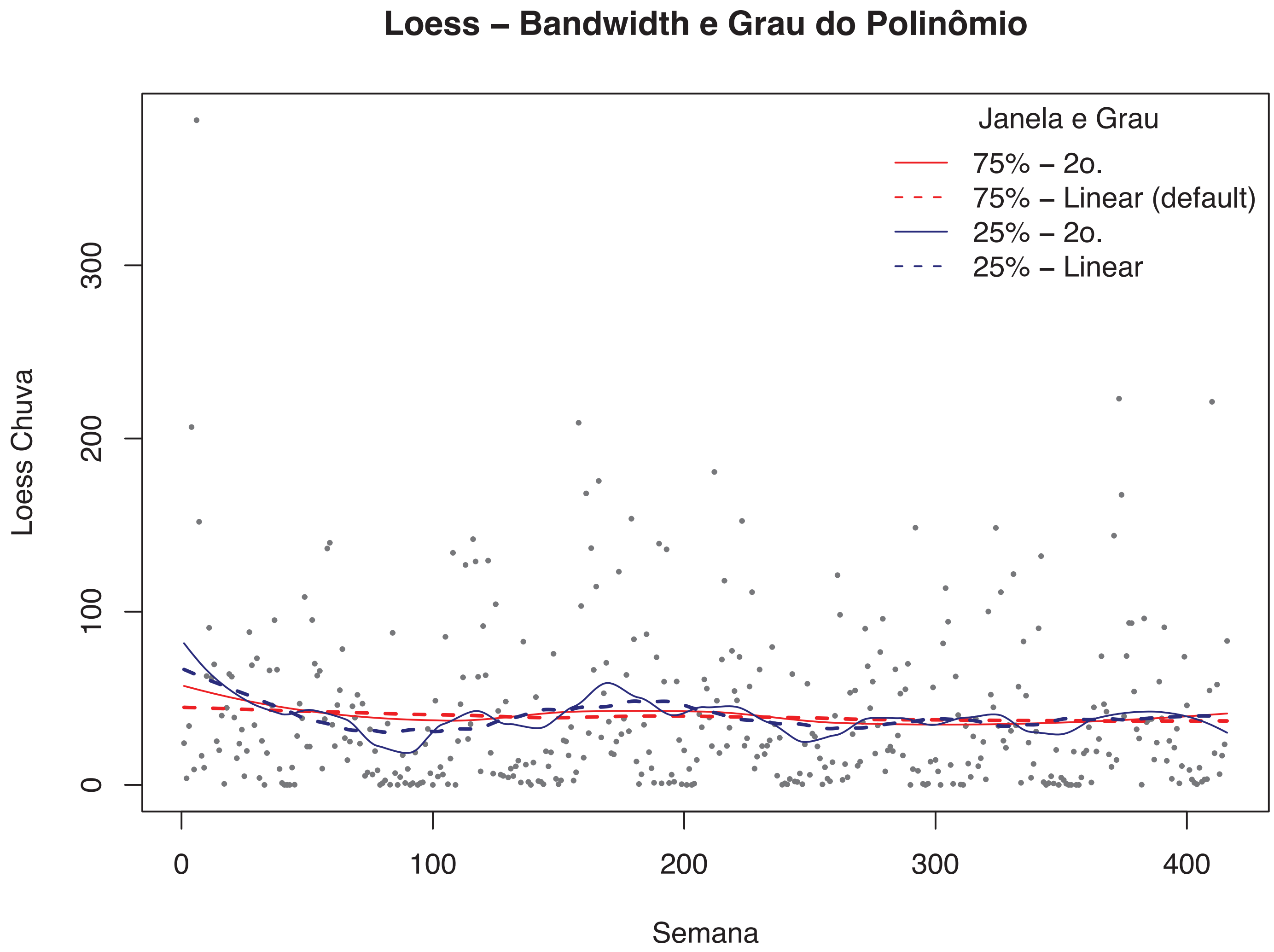
- Advantages: simple, great for exploratory analysis.
- Problem: border effect.
- Very sensitive to bandwidth.
- Automatic choice of bandwidth may not be desirable.
- Not very sensitive to function shape, as long as it is smooth.

Outline

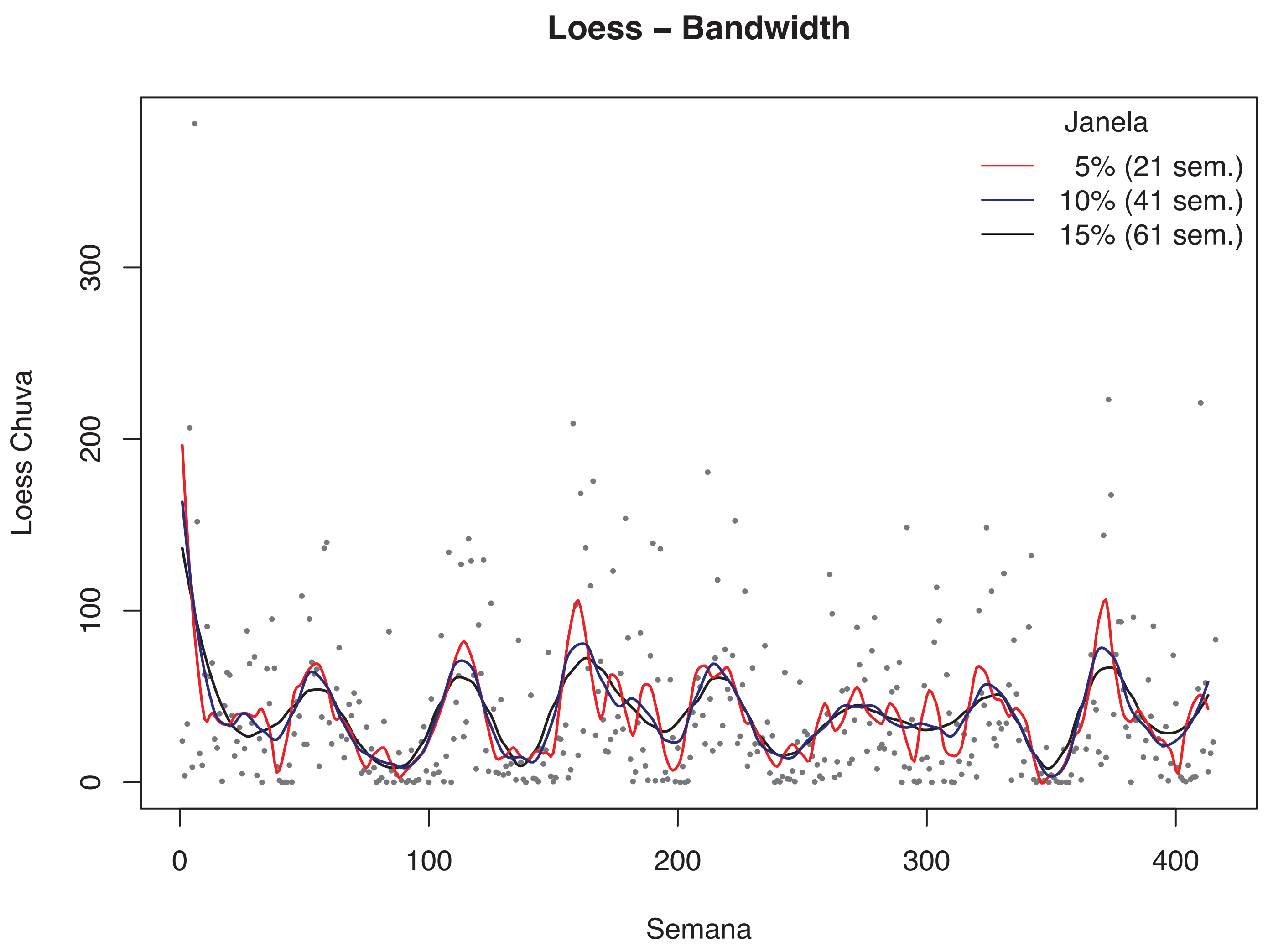
Loess

- Similar to the kernel, but the base is a local regression instead of a weighted average
- At each point (x) and neighbouring points (window or bandwidth) a polynomial is fitted using weighted least squares, where closer points are given larger weight
- The bandwidth or smoothing parameter controls the flexibility of the regression
- The degree of the polynomial regression is in general low:
  - A polynomial of degree 0 = running average;
  - First degree = local linear regression

Loess – Span & Degree



Loess – Span & Border



Loess

- Advantages:
  - simple, great for exploratory analysis.
  - Less sensitive to border effect
- Disadvantages: sensitive to extreme values

Comparing

http://en.wikipedia.org/wiki/Kernel_smoothing

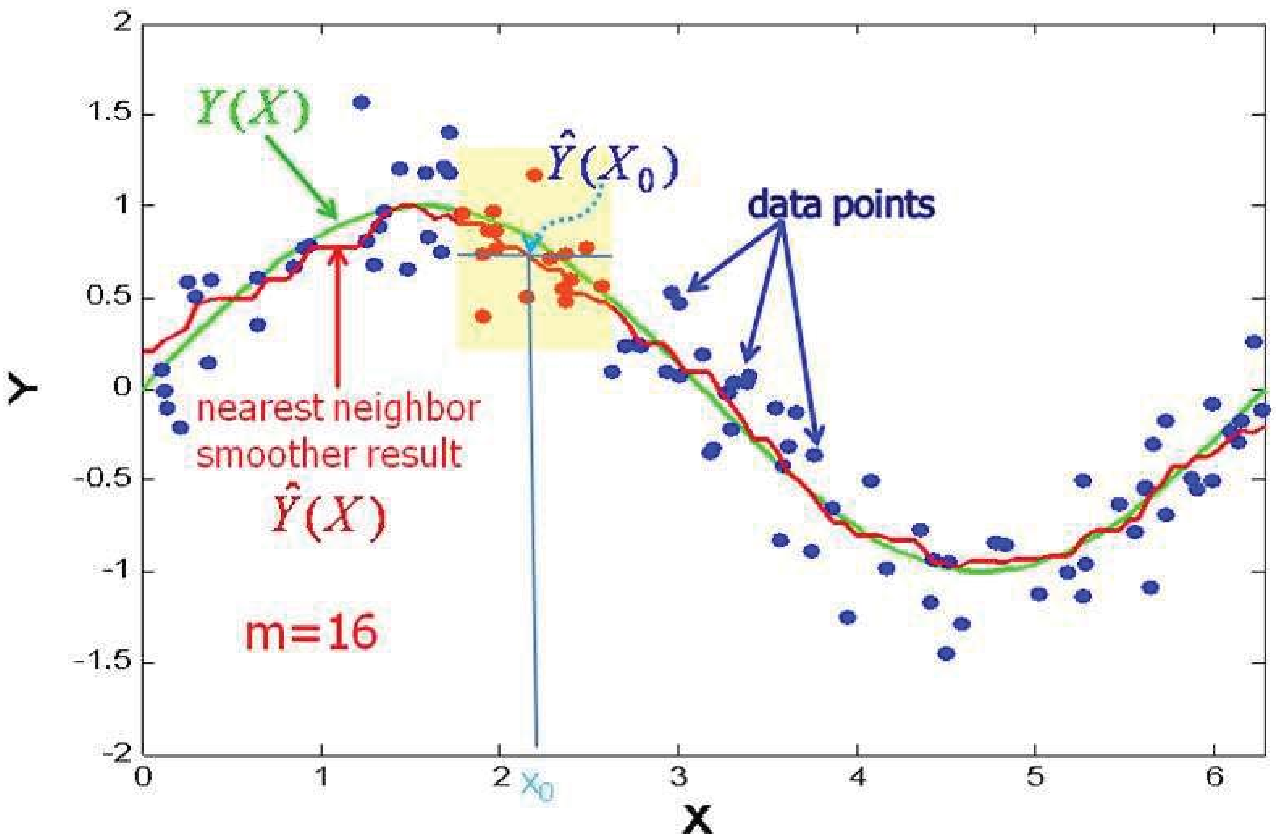


Fig.: Nearest neighbour

Comparing

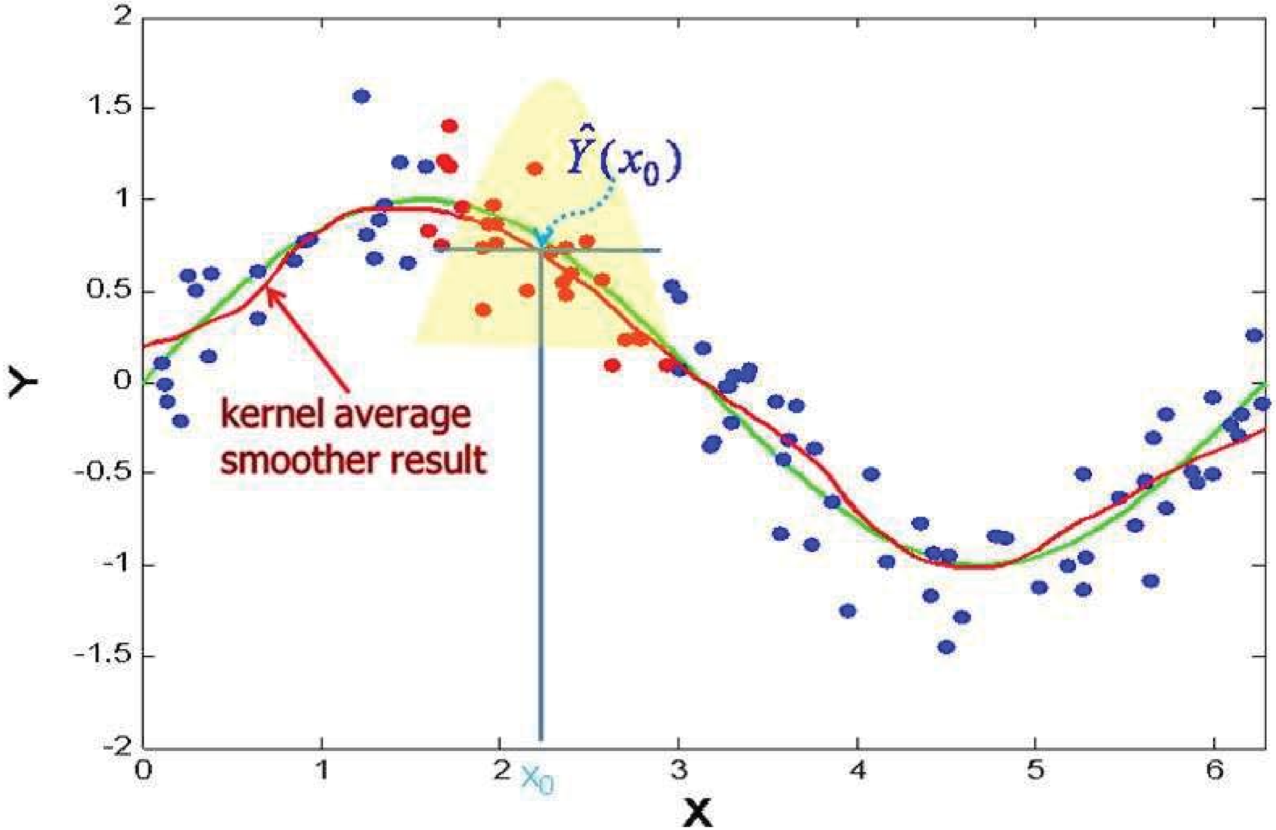


Fig.: Weighted average

Comparing

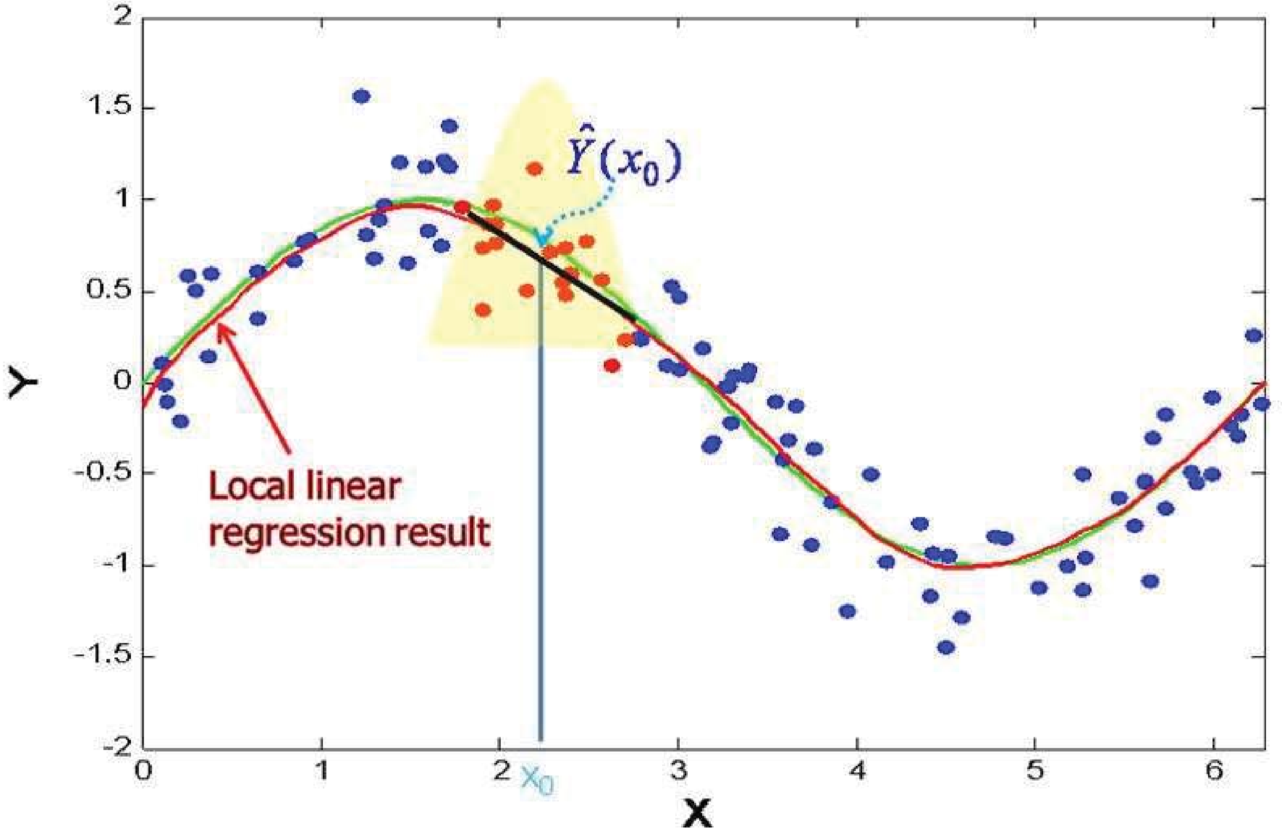


Fig.: Loess

Outline

2. Exploratory Analysis
   - Kernel
   - Loess
   - Splines

Splines

- Splines are smooth polynomial function piecewise-defined
- Very smooth, including the places where the polynomial pieces or knots connect
- Splines do not oscillate ate the edges (Runge's phenomenon present when using high degree polynomial interpolation)

Splines

- A problem of penalised regression: a solution for $\hat{f}(x)$ that minimises:

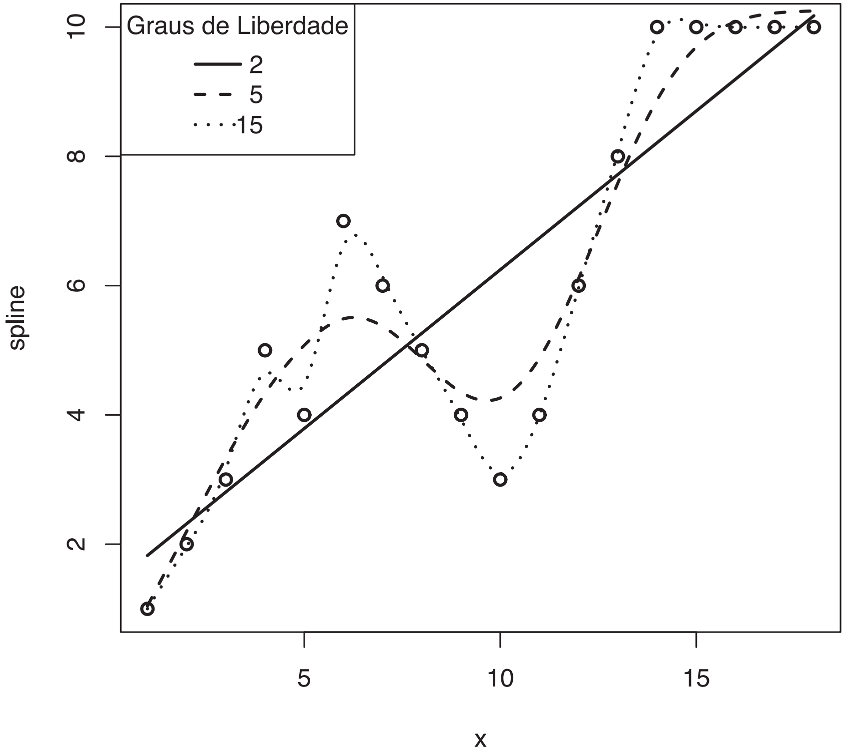
$$\sum [y_i - f(x_i)]^2 + \tau \int [f''(x)]^2 dx$$

  where τ is the smoothing parameter: controls the trade-off between fidelity to the data and roughness of the function estimate
  - If $\tau = 0 \rightarrow \hat{f}(x)$ interpolating spline
  - If τ is very large, $\int [f''(x)]^2 dx$ needs to approach zero $\rightarrow$ linear least squares estimate
- When $\sum [y_i - f(x_i)]^2$ is replaced by a log-likelihood $\rightarrow$ penalised likelihood
- The smoothing spline is the special case of penalised likelihood resulting from a Gaussian likelihood

Splines

- The choice of the smoothing parameter can be visual or via some automatic algorithm (e.g. cross validation)
- The results of splines and loess are similar for similar degrees of freedom
- Multivariate splines: $\eta = \beta_0 + f_1(x_{i1}, x_{i2}, \ldots, x_{ip}) + \ldots$
- Several applications to temporal and spatial models

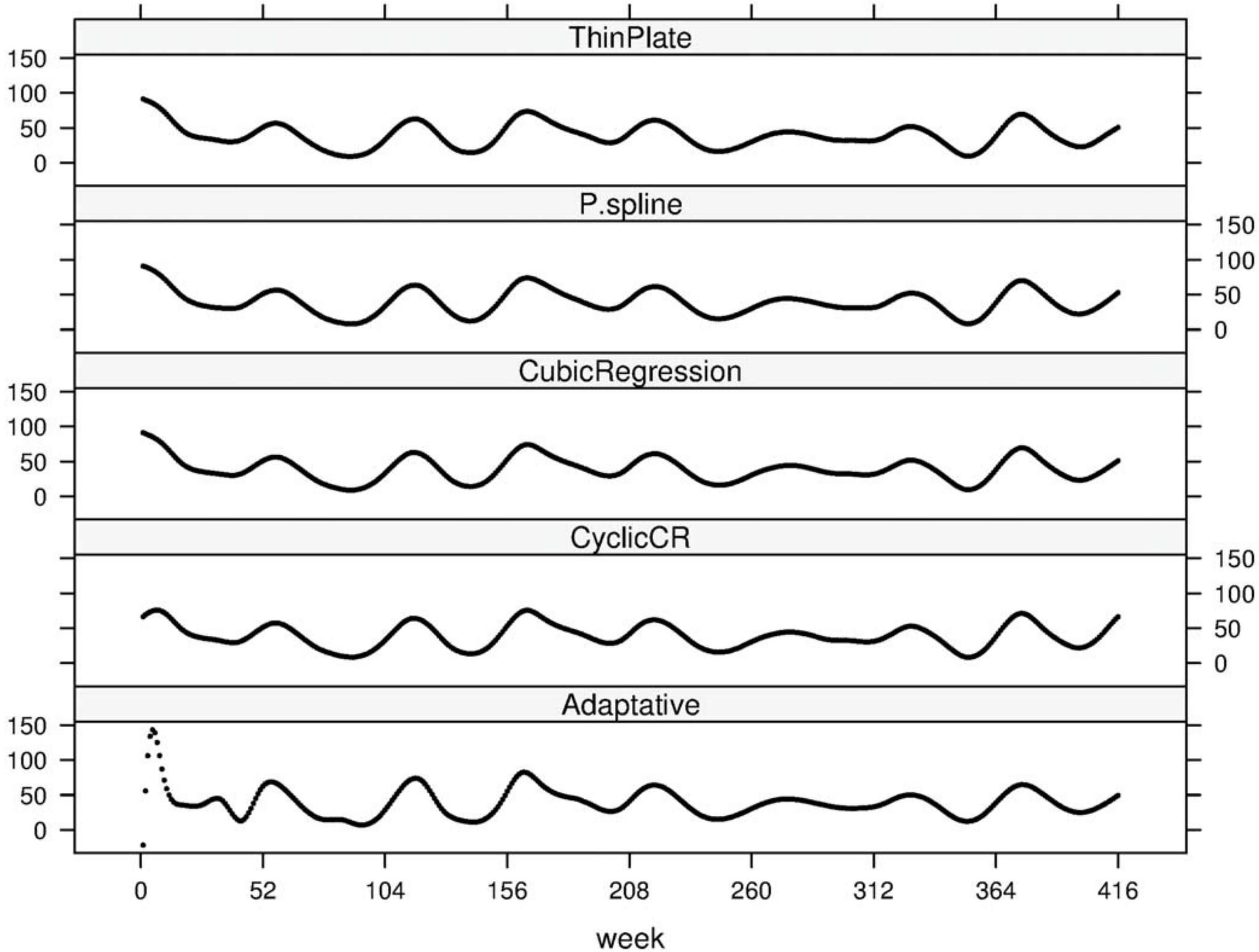
Splines – bandwidth



Splines functions

- Cubic regression spline – 3^{rd} degree polynomial fitted to knots distributed over the data range
- Cyclic cubic regression spline – imposes the first and last values to be equal (interesting for seasonal time series)
- P-splines – with a differential penalty for adjacent parameters, to control “wiggliness”
- Thin plate – the smallest mean square error, smallest number of parameters, considered the optimal estimator, easily adapted to two dimensions (space!)
- Tensor Product – Similar to Thin Plate, better when scale of each dimension is not the same

Splines functions



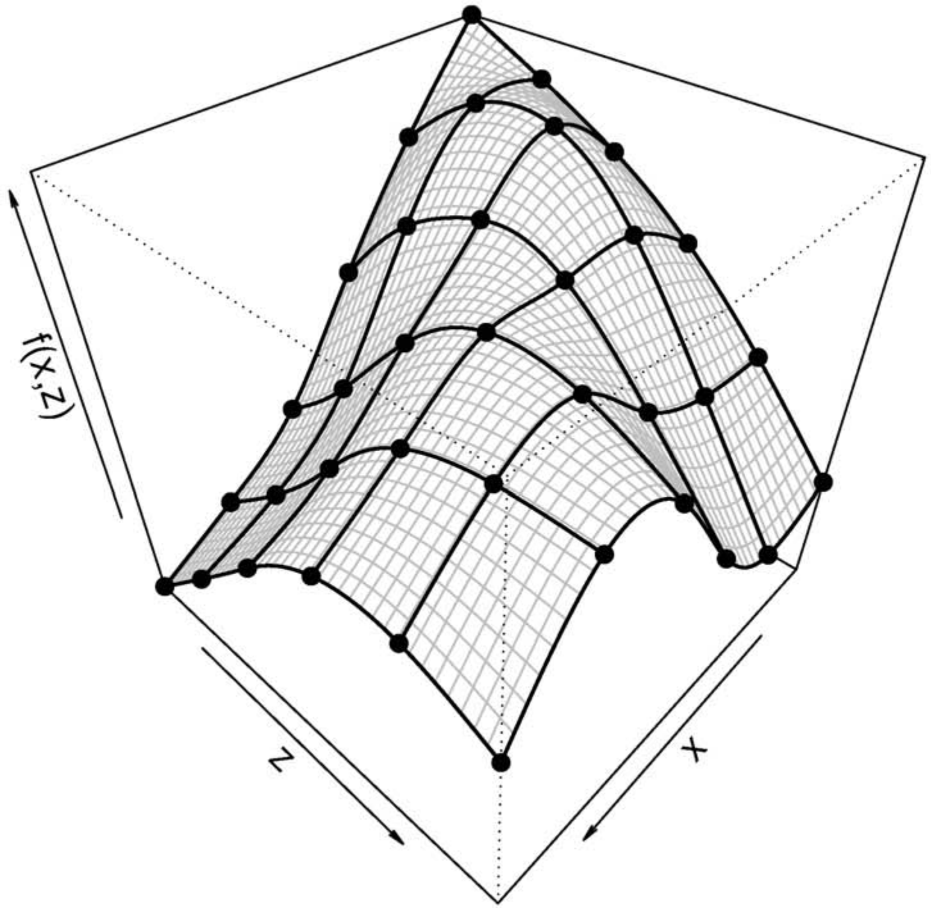
Choice of function

- Modelling just one variable – time – not much difference
- For more then one variable – space – choose carefully:
- Thin plate:
  - isotropic,
  - invariant to rotation
  - smaller square error
  - smaller number of parameters, considered the optimal estimator
  - HOWEVER: sensitive to changes in scale
- Tensor Product:
  - possible to have different scales

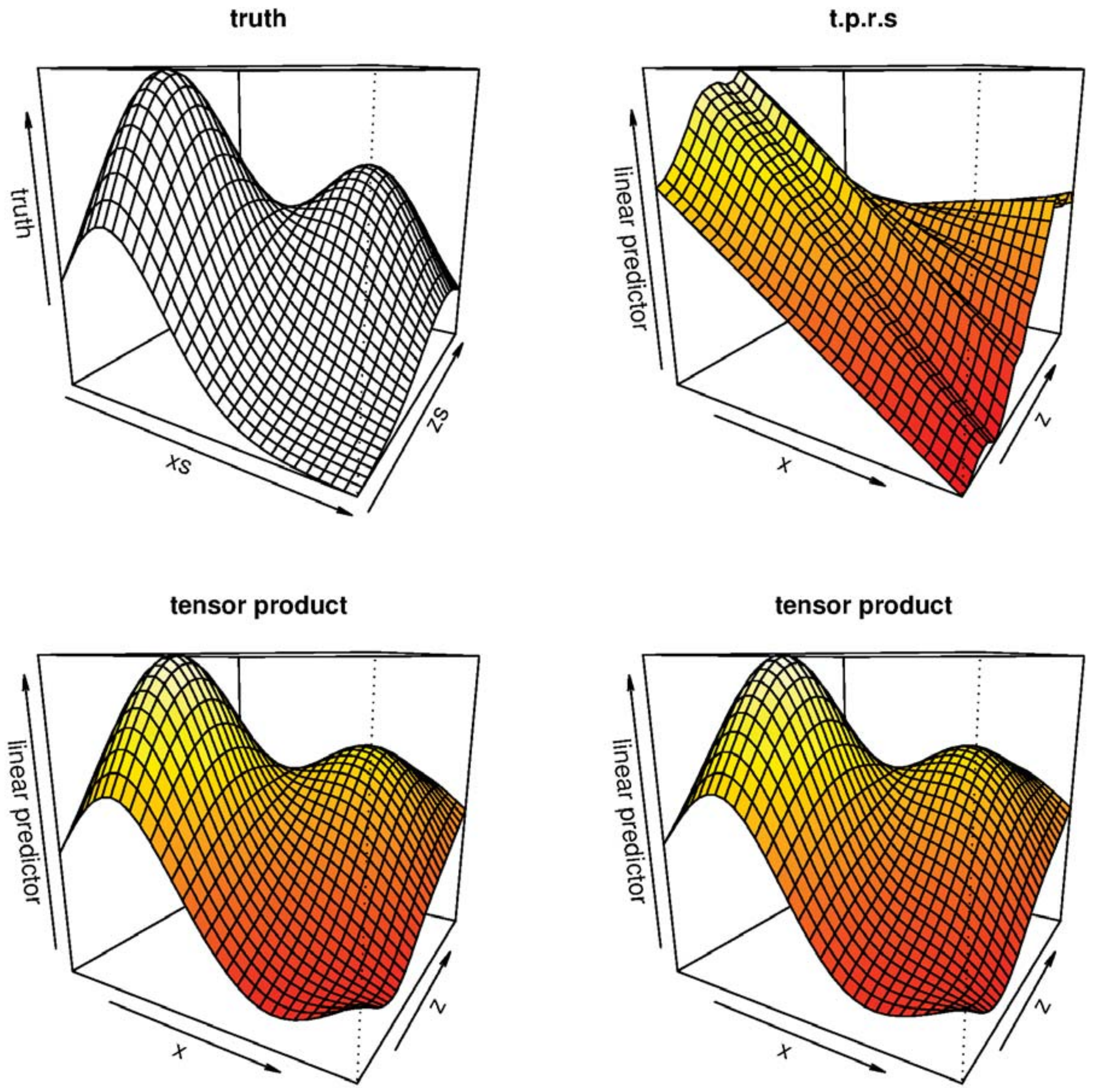
Splines functions– summary

bs=	Description	Advantages	Disadvantages
"tp"	Thin Plate	Multiple covariates	Computationally intensive
		Rotational invariant	Varies with
		Optimal estimator	scale
"tpr'	Tensor Product	Multiple covariates	Varies with
		Scale invariant	rotation
"cr"	Cubic	Computational cheap	Only one variable
	Regression	Parameters directly	Based on knots choice
		interpretable	Non-optimal estimator
"cc"	Cyclic CRS	Beginning and end ='s	the same
"ps"	P-splines	Any combination of	Evenly spaced knots
		base and *order*	Not easily interpretable
			Non-optimal estimator

Bivariate spline



Changing the scale



Outline

The problem

How do these variables behave in relation to each other?

- Age $\rightarrow$ external causes deaths from 5 to 45 years
- Income $\rightarrow$ cardiovascular diseases
- Distance to health services $\rightarrow$ mammography
- Adherence to HIV treatment $\rightarrow$ development of virus resistance
- ...
- Time $\rightarrow$ transmissible diseases
- Space $\rightarrow$ vector-borne diseases

GAM – definition

- extension of GLM, where the linear predictor η is not limited to linear regression
- the model includes any function of the independent covariates (x_i):

$$\eta \ = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots$$

- $f(x) \rightarrow$ can be a non-parametric function such as lowess
- When to use? When the covariate effect changes depending upon its value

Why not to use

- Statistical models aim to explain the observed data, not to simply reproduce it – overfitting
- Parametric models in general are better to estimate standard errors or confidence intervals
- Parametric models are more efficient, if correctly specified (smaller number of observations)

Outline

The problem

- Going back to the leptospirosis example.
- To estimate the effect of rainfall, humidity and temperature on the number of cases of leptospirosis
- Why not just apply a regression model?
  - Trend
  - Seasonality
  - Autocorrelation

Autocorrelation

- Autocovariance is the covariance of the variable against a time-shifted version of itself

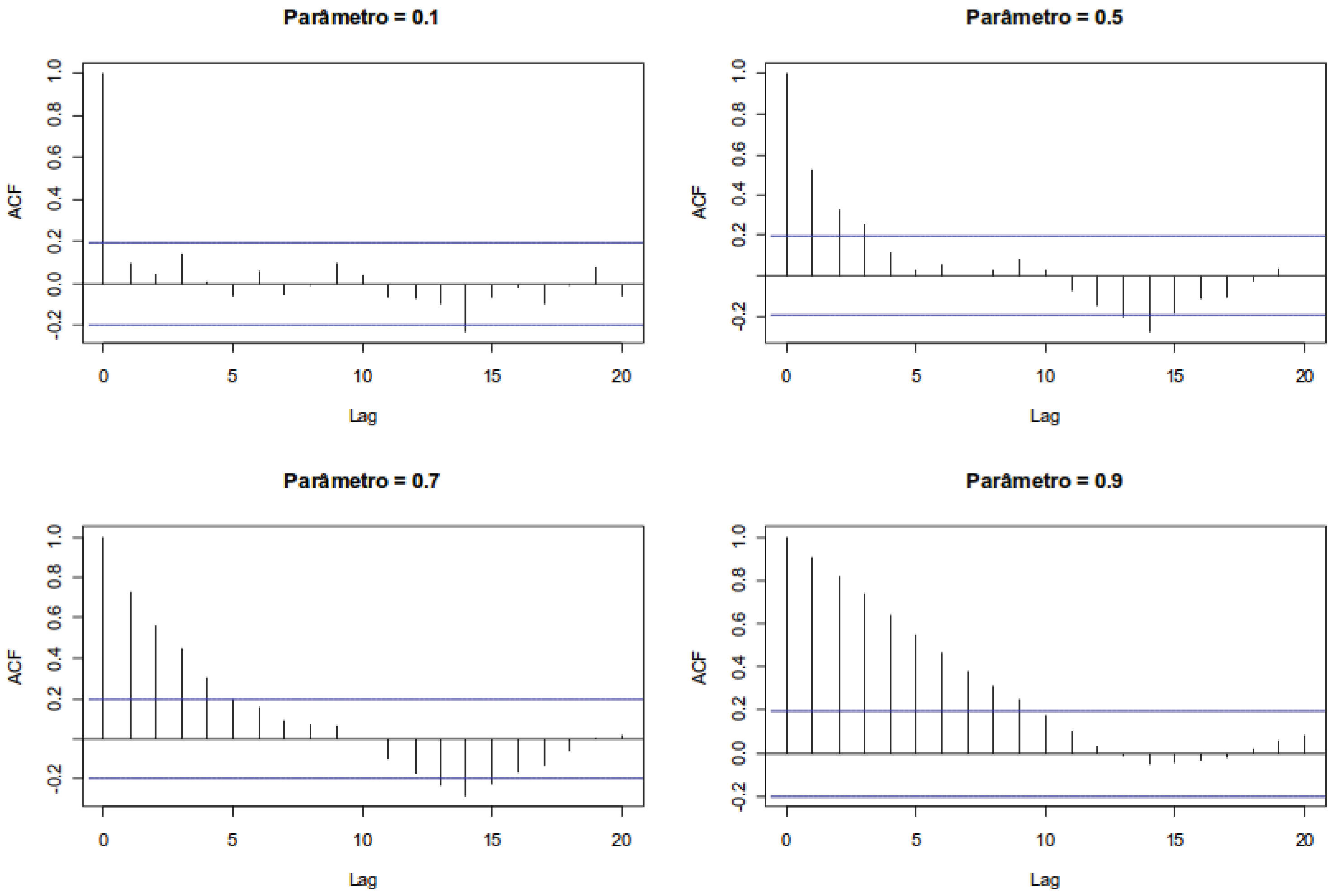
$$C_{xx}(t,s) = E[(X_t - \mu_t)(X_s - \mu_s)] - \mu_t \mu_s$$

- If $X(t)$ is stationary $\rightarrow \mu_t = \mu_s = \mu$ and

$$C_{xx}(t,s) = C_{xx}(t,s) = C_{xx}(\tau)$$

- Autocorrelation $c_{xx}(\tau) = C_{xx}(\tau)/\sigma^2$
  $\tau \rightarrow$ the lag
  $\sigma^2 \rightarrow$ the variance
- It is a measure of how similar a series is to a time-shifted version of itself
- Range: $[-1, 1]$

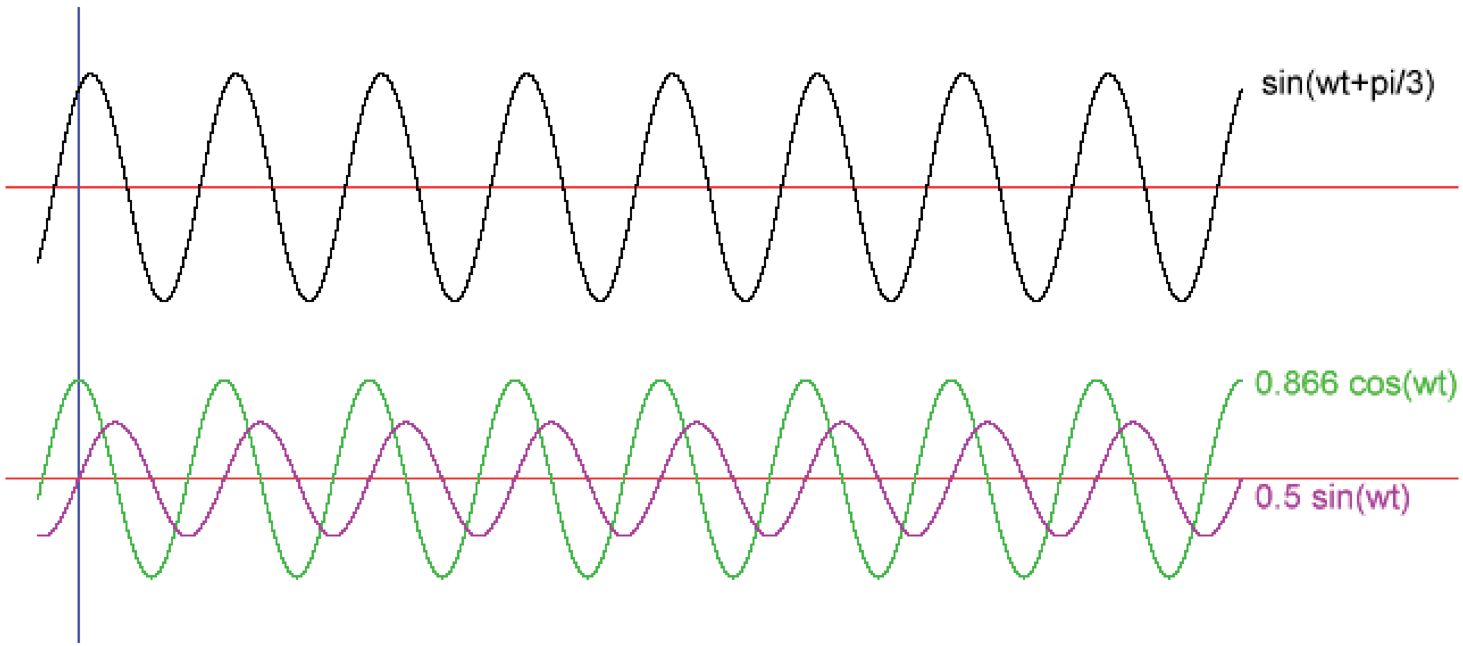
Autocorrelation



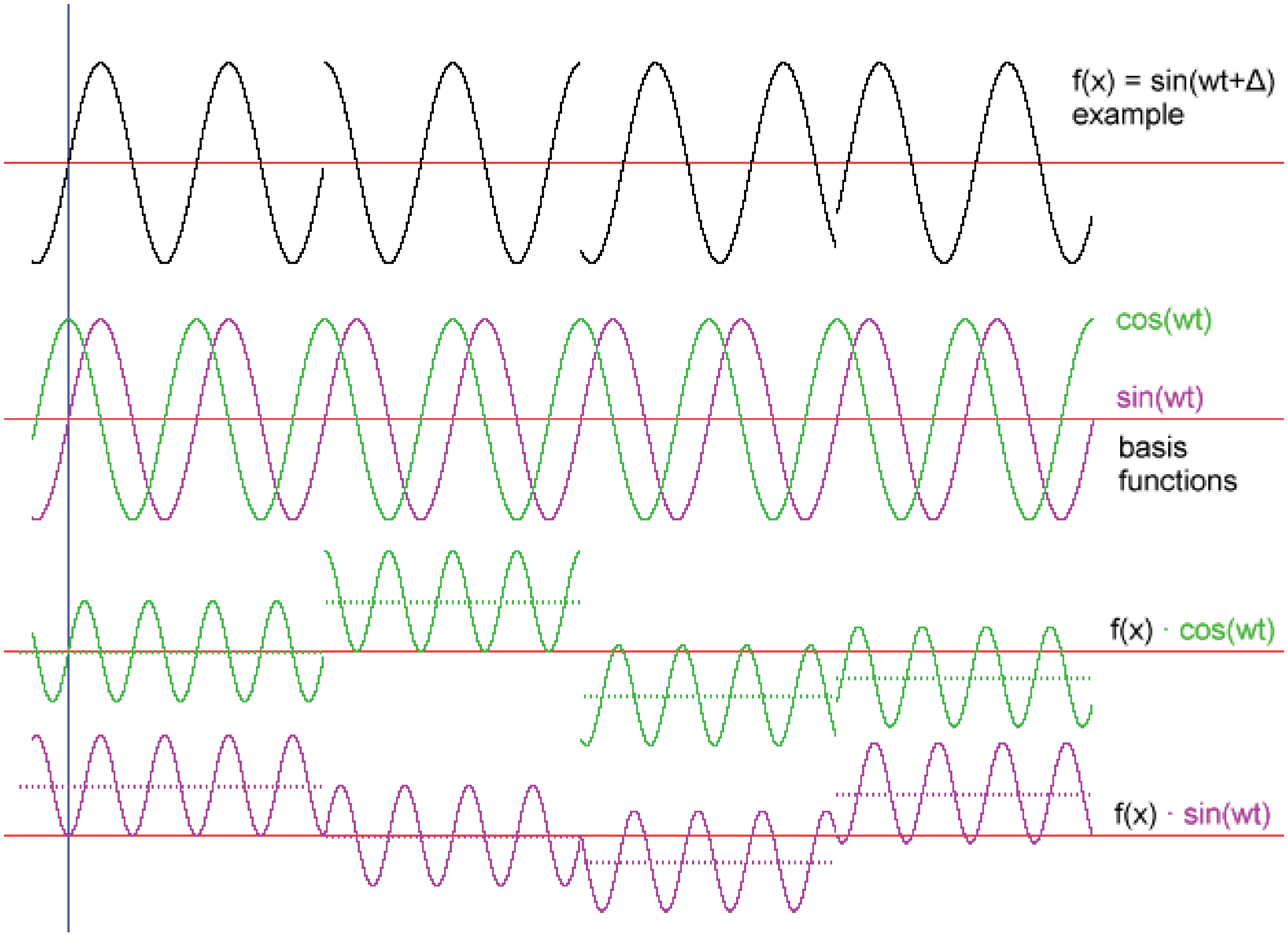
Seasonality

- Component of a time series which is defined as the repetitive and predictable movement around the trend line
- Not necessarily related to climate seasons
- Can be either removed or modelled:
    - sinusoid
    - including each month (or season) as a categorical variable

Seasonality: sinusoid



Seasonality: sinusoid



Trend

- A stationary process is a stochastic process whose joint probability distribution does not change when shifted in time (or space)
- Mean and variance, if they exist, are constant
- Trend model: linear (?!?), polynomial, splines
- Do we really want to remove the trend?

Modelling time series

- Time series books – ARIMA models
- Not much used in epidemiology:
  - Intervention
  - Explanation
  - “Causes”
- Regression models including (if needed) AR components
- Emphasis on covariates

GAM for Time Series

- The main idea is to model the effect of covariates on some health event over time
- Reasons:
  - allow the inclusion of time dependence
  - non-linear relationship
  - trend and seasonality can be easily incorporated

GAM for Time Series

- Considering the response variable a count, the best choices in GLMs are:
  - Poisson: $\lambda =$ expected values and $=$ variance $\rightarrow$ overdispersion
  - Quasipoisson – it is not a distribution, but a way to relax the previous assumption and allow for overdispersion. It does not present AIC.
- Other models, very often used:
  - Negative Binomial – has a mean μ, scale parameter θ and variance function $V(\mu) = \mu + \mu^2/\theta$.
  - Zero-inflated models – mixture models combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial – are available as well (package VGAM)

GAM for Time Series

$$\text{Lepto}(t) = \text{rain}(t-?) + \text{humidity}(t-?) + AR(t, t-1) + trend + seasonality + \varepsilon$$

- Trend and seasonality $\rightarrow$ smooth function
- Covariates – time lag
- It is possible to include the variation on the population at risk (offset)

Outline

Why Distributed Lags?

- When risk factors and health events are measured on populations:
  - asthma & air pollution
  - cold weather & heart attack
  - flooding & leptospirosis
- Between climate and health event $\rightarrow$ time interval – lag
- Questions:
  - How much time after?
  - How long does the effect last?
  - When does the effect disappear?
  - Is there a threshold?

Recommended reading

- Schwartz J. The distributed lag between air pollution and daily deaths.*Epidemiology*, 2000;11(3):320-326.
- Welty, LJ. & Zeger, SL. Are the Acute Effects of Particulate Matter on Mortality in the National Morbidity, Mortality, and Air Pollution Study the Result of Inadequate Control for Weather and Season? A Sensitivity Analysis using Flexible Distributed Lag Models. *American Journal of Epidemiology*, 2005;162:(1):80-88.
- Gasparrini A., Armstrong, B., Kenward M. G. Distributed lag non-linear models. *Statistics in Medicine*. 2010; 29(21):2224-2234.
- Armstrong B. Models for the relationship between ambient temperature and daily mortality. *Epidemiology*. 2010, 17(6):624-631.

Problems

- Effects change over time – increasing and decreasing
- Covariates – temperature, humidity, rainfall and pollution – highly correlated
- Possible non-linear structure

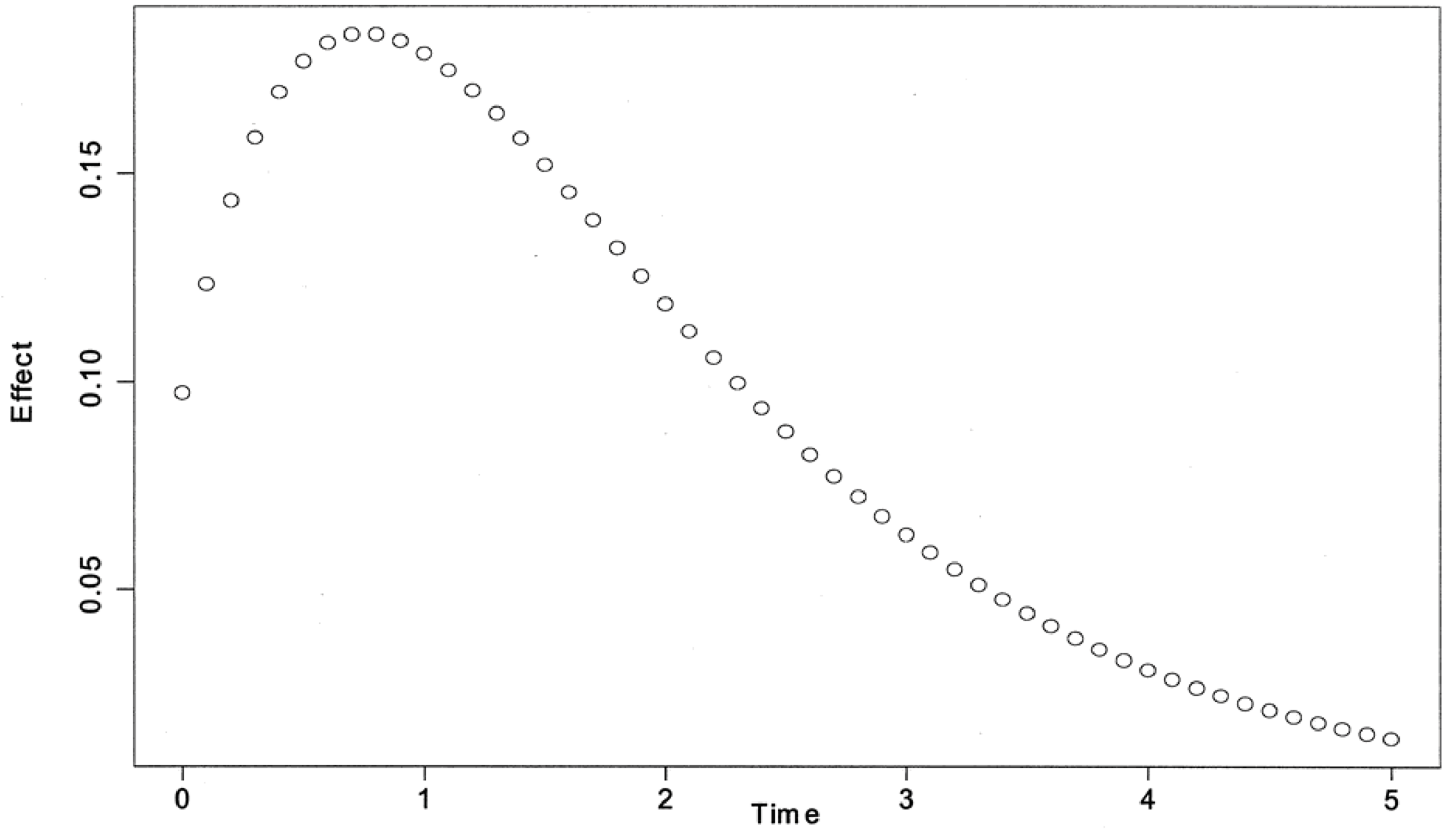
Effect throughout time

- In a linear model the sum of the effect of all independent variables, shifted by each time lag, is associated with the outcome

$$y(t) = \nu + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \ldots + \beta_k x_{t-k} + \varepsilon_t \qquad (1)$$

- Supposing that the number of events in a week follows the rainfall one week before
- This number increases up to two weeks after, and decreases smoothly up to the 5^{th} lag, the graphic of the β's of the model would present a curve such as:

Effect throughout time



A hypothesized curve showing the impact of an environmental toxin over time. The effect rises, and then falls, possibly with a long tail. The goal of this analysis is to determine what the actual shape of the curve representing the time course of deaths after exposure to PM10 is.

From: Schwartz: Epidemiology, Volume 11(3).May 2000.320-326

Alternative models

- Running average of the predictor → the shape of increase and decrease cannot be observed
- One parameter for each lag → no supposition about the shape of the curve
- To restrict the parameters to a specific shape → PDL (*Polynomial Distributed Lag*)
- To combine possible non linear effects with lag → DLNM (*Distributed lag non-linear models*)

Effect throughout time

- We use a transformation to represent the accumulated effect of X, weighted by a polynomial (2°degree)
- With this transformation of $X \Rightarrow Z$:
  - colinearity disappears
  - the shape induced on the relationship (in the example quadratic), imposes a restriction on the parameters
- After estimation of the parameters α of z, parameters β for X are obtained via back transformation
- The error of α goes back as well to β

When the effect is non-linear

- The solution is a combination of splines and lags
- cross-basis: a bidimensional space of functions describing simultaneously the shape and the effect distributed over time
- The idea is to specify two independent set of base functions
- PDL is a particular cases of DLNM, with a linear predictor

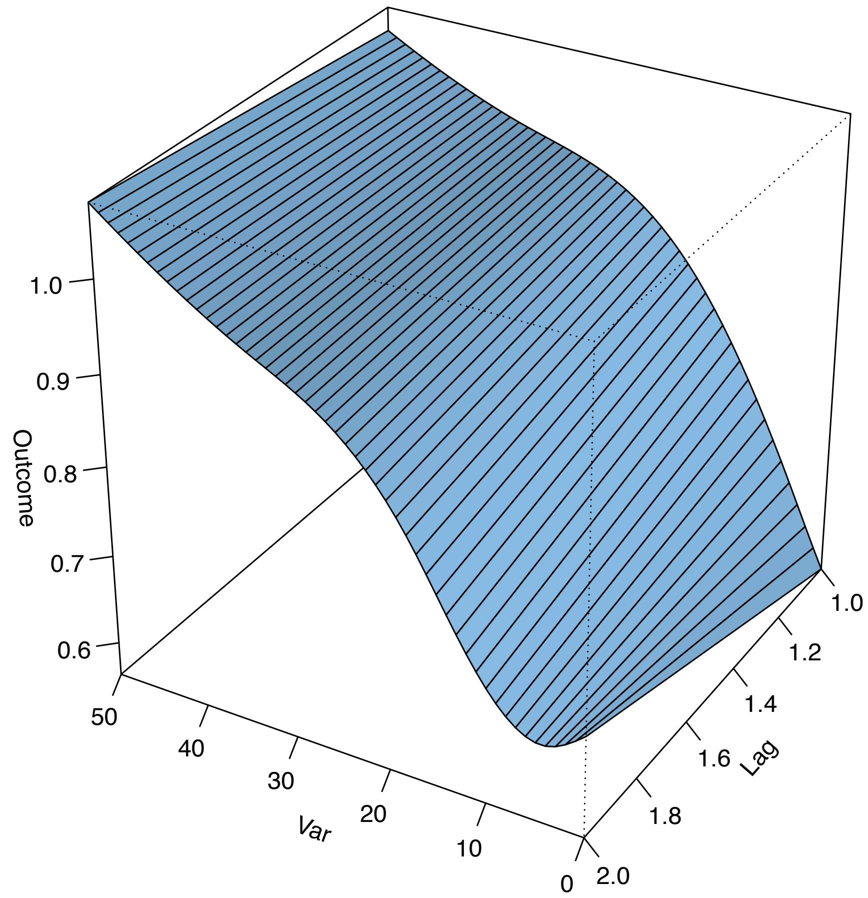
When the effect is non-linear

- The solution is a combination of splines e lags
- cross-basis: a bidimensional space of functions describing simultaneously the shape and the effect distributed over time
- The idea is to specify two independent set of base functions
- PDL is a particular cases of DLNM, with a linear predictor

How to interpret

- A grid is built on possible predicted values over time
- It is possible to evaluate the effect of a given value of the predictor over time $\rightarrow$ cut-points
- Or observe on each lag the shape of the relationship between predictor and outcome
- It is also possible to estimate the cumulative effect over time for values of the predictor

Rainfall & Leptospirosis



Outline

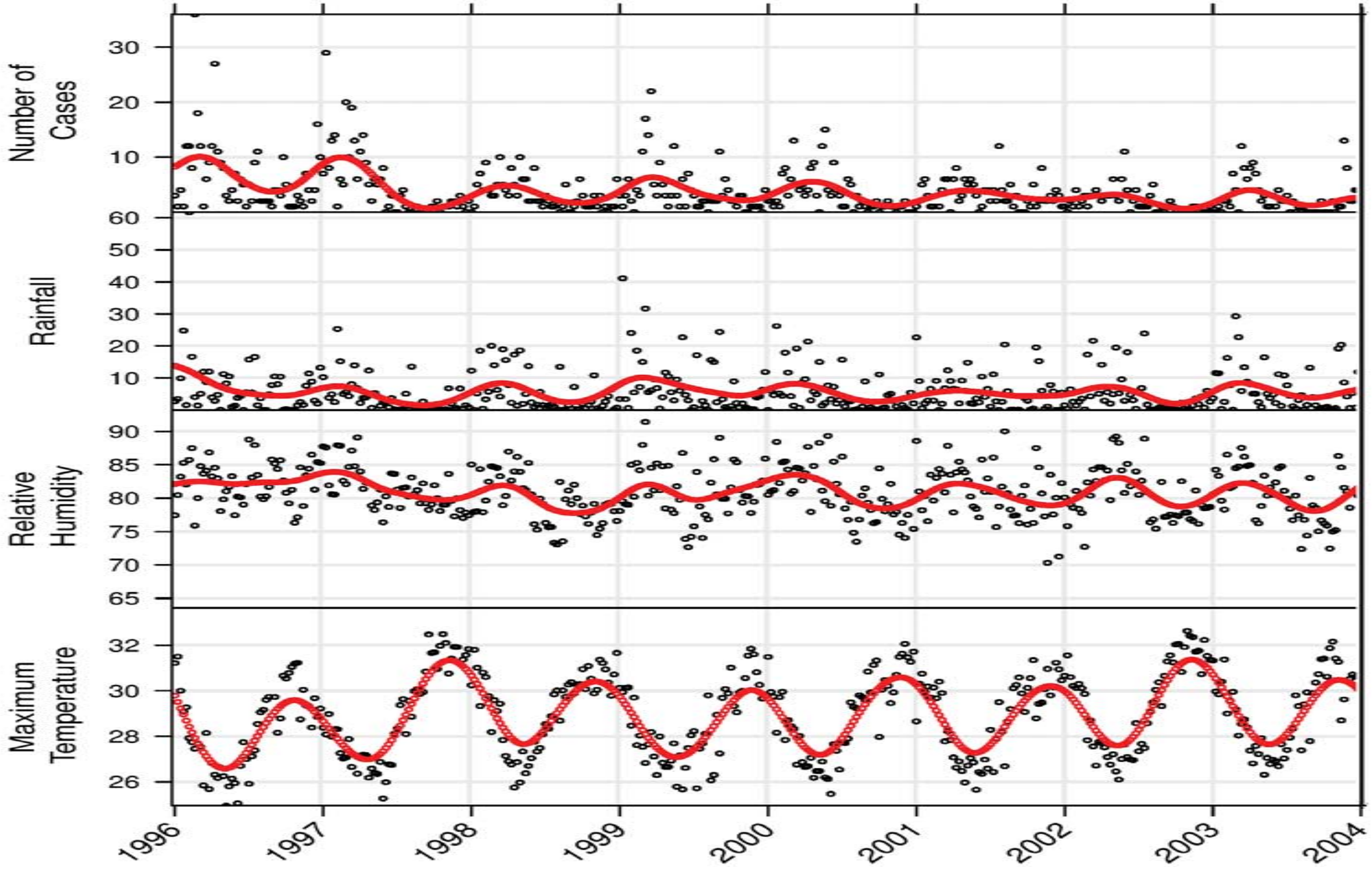
Models for Time Series

- ARIMA or SARIMA models: regression models where independent variables are just a shifted version of the dependent variable.
- Stationary time series:
  - stochastic process whose joint probability distribution does not change when shifted in time or space
  - mean and variance do not change over time or position
  - removing trend and seasonality (S and I terms)
- detection of order of autoregressive and moving average terms
- fit, evaluation, ...
- prediction
- Dynamic models!

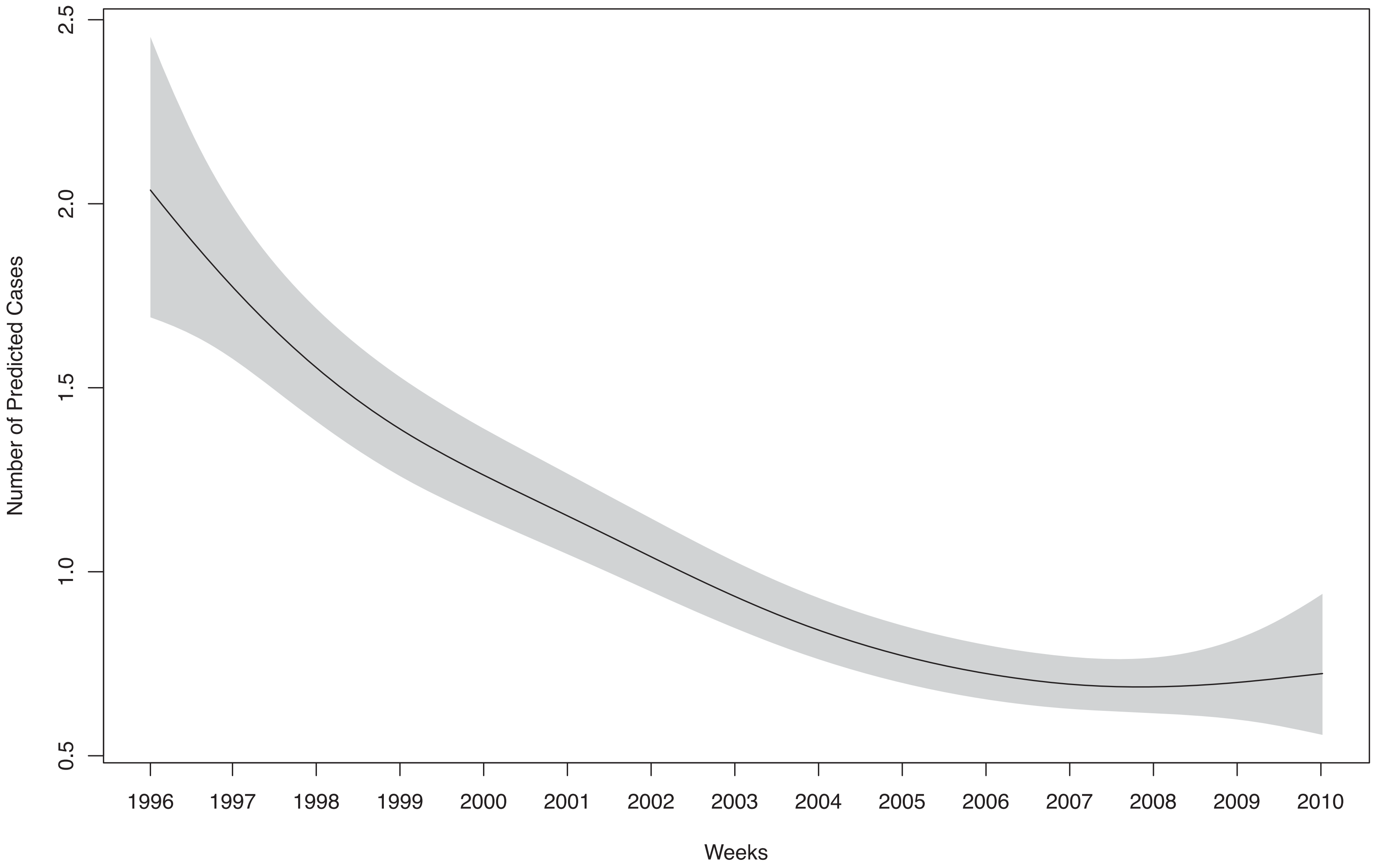
Modelling for:

- Explaining why events happen this way over time:
  - Independent variables are associated with events y in $t \rightarrow$ regression
  - Past events are “cause” of present events $t \rightarrow$ dynamic models
- How to predict $t + k$?

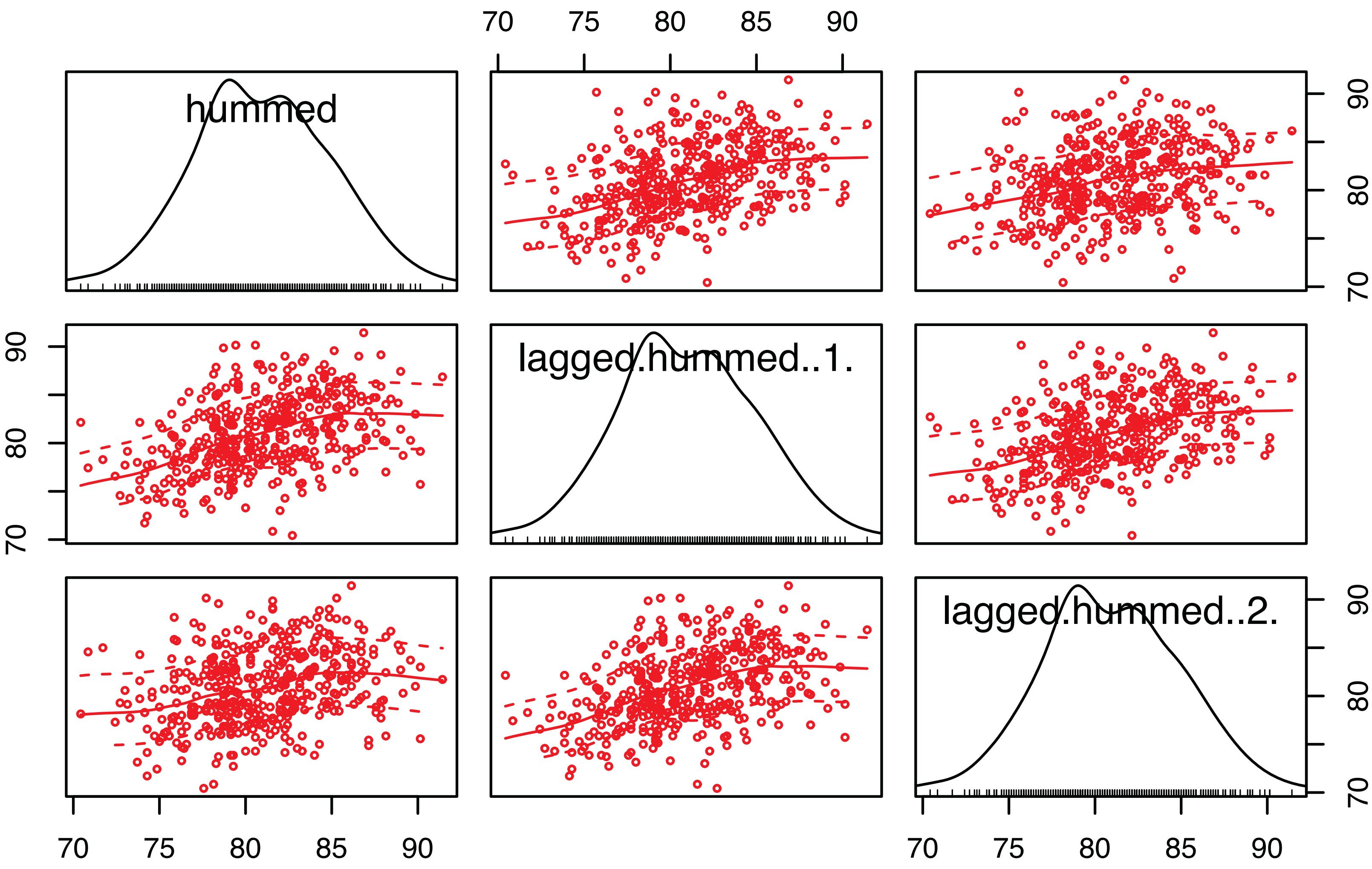
Exploratory analysis



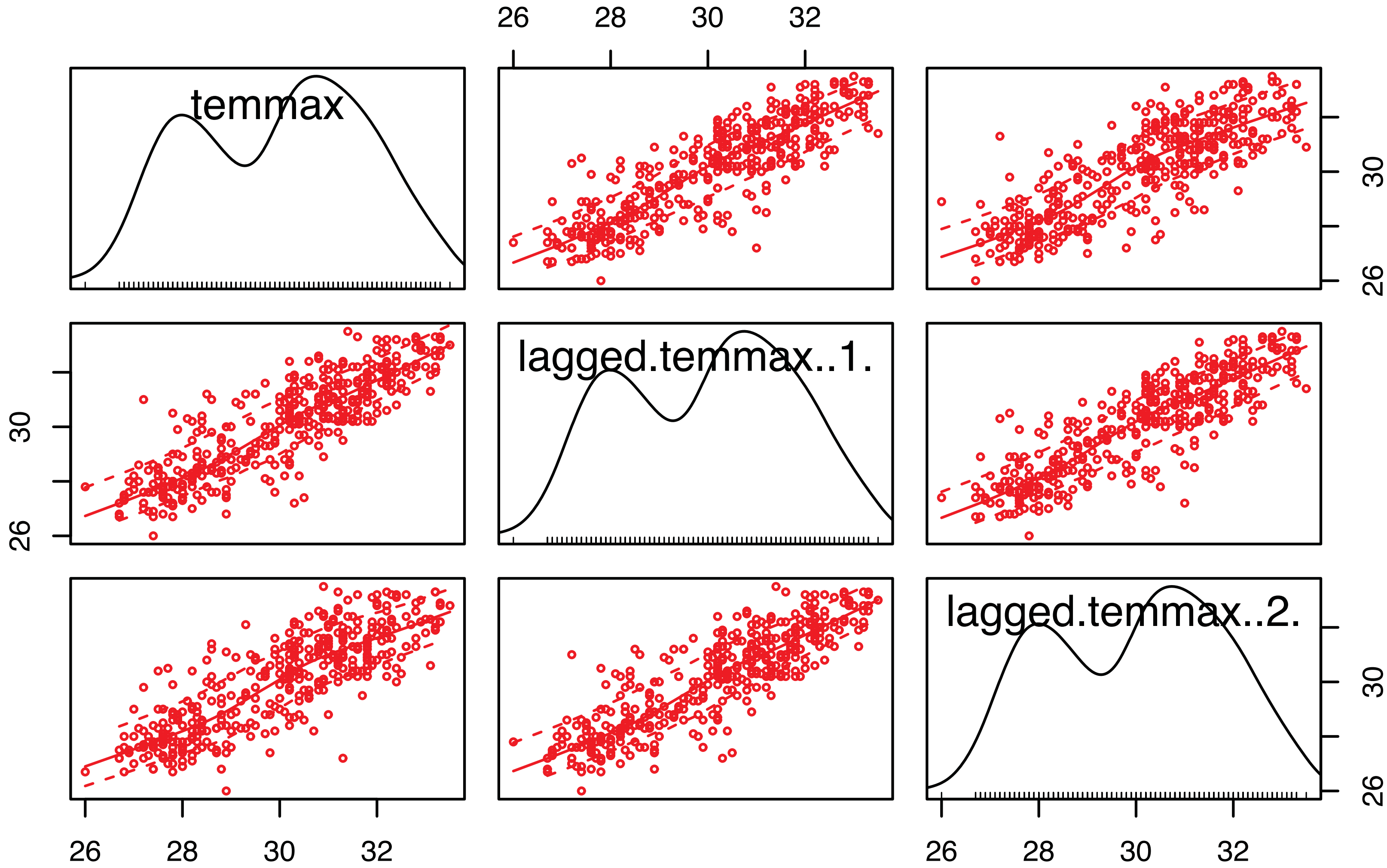
Exploratory analysis – trend



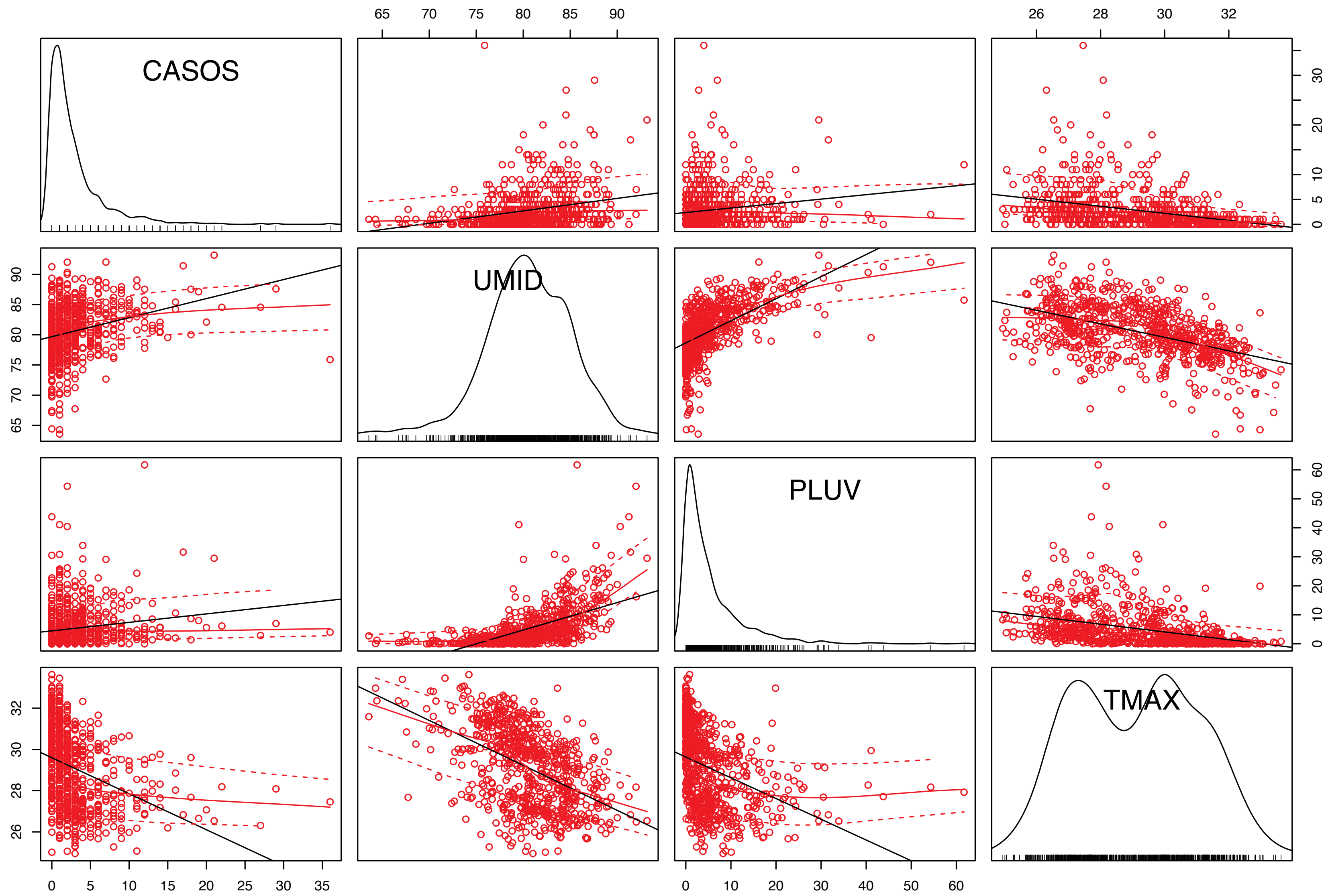
Colinearity



Colinearity



Colinearity



Structure

- Autocorrelation
- Components

TS Components

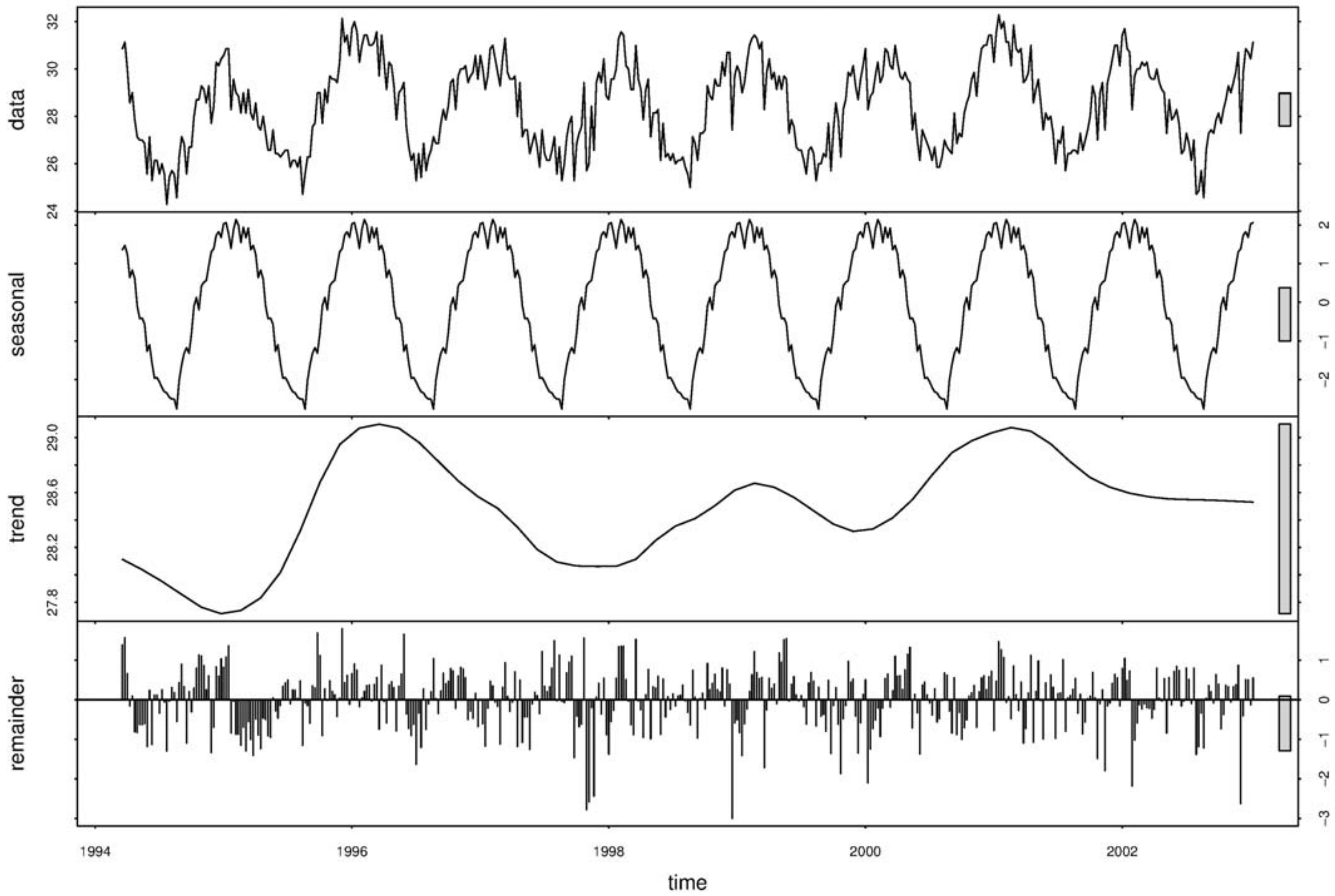
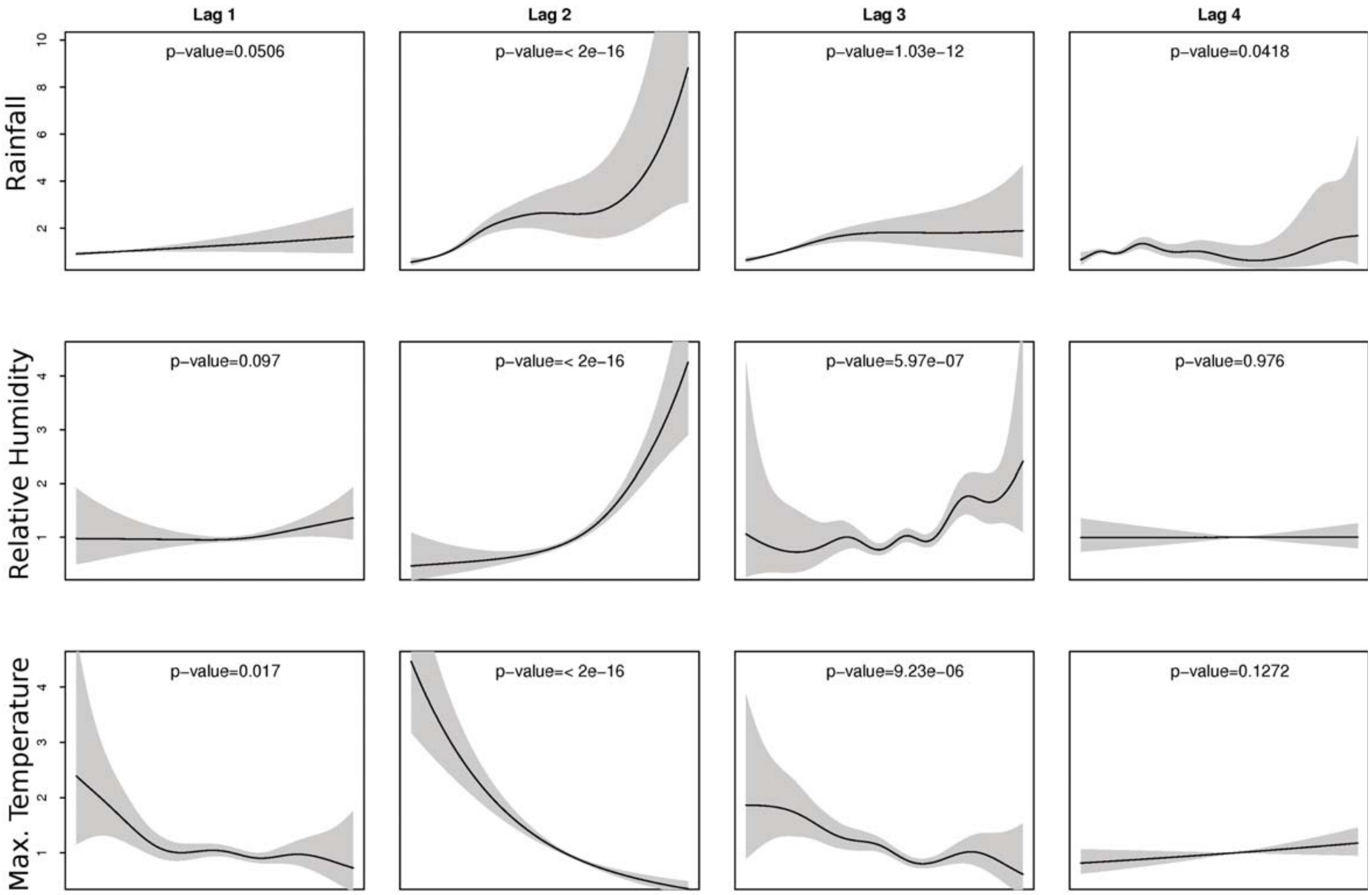


Fig.: Maximum temperature

Functional form



Seasonality

- Exclude the seasonality of the independent variables using sinusoid functions.
- Use the residuals of this seasonal model as independent variables
- Include a seasonal term in the complete model
- Interpretation is the same, as the residuals keep the same measure unit: the meaning of the parameter estimated is the same

Multiple model

- Test the significance of each time lag, respecting the functional form
- Join all lags and covariates
- When the functional form is not linear $\rightarrow$ categorise, segmented regression, CART model (Classification and regression trees)
- Splines & PDL

Residuals

- ACF of residuals again
- still trend?
- inclusion of AR term

Summary

- Counts: Poisson, Quasipoisson or Negative Binomial → overdispersion!
- Trend and seasonality → `s(tempo)` e `s(tempo, k=52)`
- Removal of seasonality of independent variables
- Regression model

```
gam(cases ~ offset(log(pop)) + s(time) +
           sin(2*pi*(1:\text{length(dataset)}/52.14) +
           covs + lag(cases, 1),
           family=negbin(c(1,10), data=dataset)
```