2453-16

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

15 - 26 April 2013

Beyond linearity: Using GAMs for modelling dengue and malaria data

COLON GONZALEZ Felipe de Jesus

Abdus Salam International Centre for Theoretical Physics
Earth System Physics Section
Strada Costiera 11, 34151 Trieste
ITALY

Beyond linearity: Using GAMs for modelling dengue and malaria data

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

Trieste, Italy

19 March 2013

ICTP
The Abdus Salam
International Centre
for Theoretical Physics

What we wanted to know

- Are there significant relationships between occurrence and weather?
- What is the functional form of those relationships?
- Are there any thresholds?
- Other determinants showing a statistically significant effect?

Theoretical relationships

- Temperature and precipitation regulate many aspects of the ecology of vector-borne diseases
  - Temperature: Extrinsic incubation period (EIP), vector survival, biting frequency, mosquito size, changes in host behaviour
    - Very low or very high temperatures may hamper the ecology of the disease
  - Rainfall: creation or wash out of breeding sites
    - Dengue: preference for man-made containers
    - Malaria: sunlit shallow pools, edges of rivers, swamps, lake edges

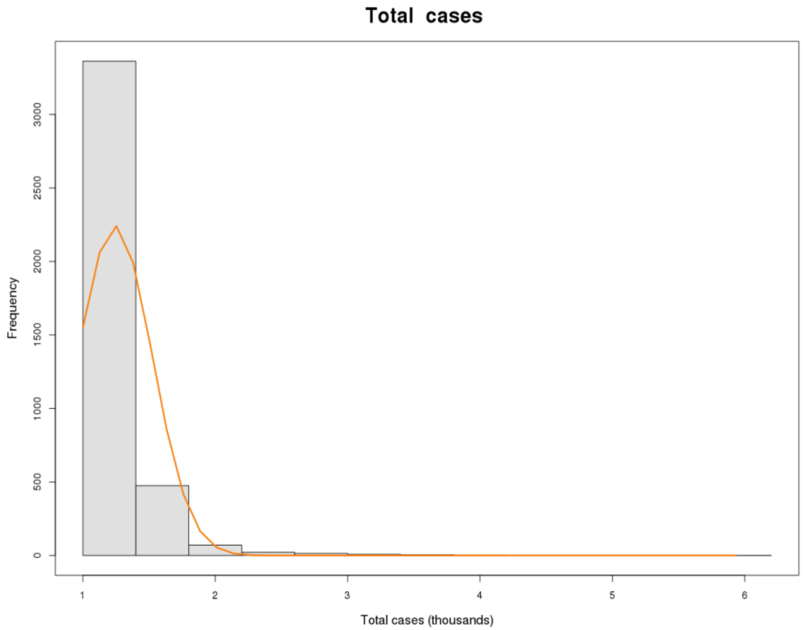
But....

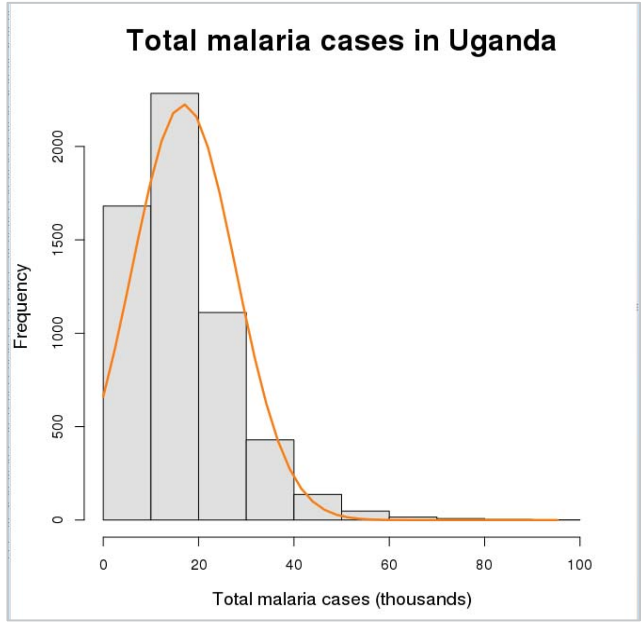
- These relationships are largely confounded by other factors:
  - Access to protective measures
  - Sheltering in sealed buildings
  - Control measures
  - Immunity
  - Urbanisation
  - Increased travelling and transportation

The need for a semi-parametric approach

- Epidemiological data
  - Dengue: 276 months of province-specific lab confirmed dengue reports (*n* = 32)
  - Malaria: 108 months of district-specific suspected cases (*n* = 56)
- Weather-related variables
  - Mexico: Tmin, Tmax, rainfall (station data)
  - Uganda: Air temperature, water fraction, soil moisture, mean temperature, rainfall (satellite and reanalyses data)
- Socioeconomic data
  - Usually yearly or decadal information

Data distribution





Quasi-Poisson models do not assume a particular probability distribution and relax the assumption of *variance* (σ^2) = *mean* (μ)

The Abdus Salam International Centre for Theoretical Physics

Model

- Poisson GAM:

$$g(\mu_{ti}) = s(t') + \sum_{j=1}^{J} s(X_{jti}) + \sum_{k=1}^{K} \beta_k (Z_{kti}) + \log(\xi_{ti}) + d_i + \alpha$$

g(.) is a log-link function that ensures that the fitted values are **always** non-negative

Model

- Poisson GAM:

$$g(\mu_{ti}) = s(t') + \sum_{j=1}^{J} s(X_{jti}) + \sum_{k=1}^{K} \beta_k (Z_{kti}) + \log(\xi_{ti}) + d_i + \alpha$$

μ is the ***rate parameter*** (expected number of new disease cases at time *t* and administrative unit *i*)

ICTP The Abdus Salam International Centre for Theoretical Physics

Model

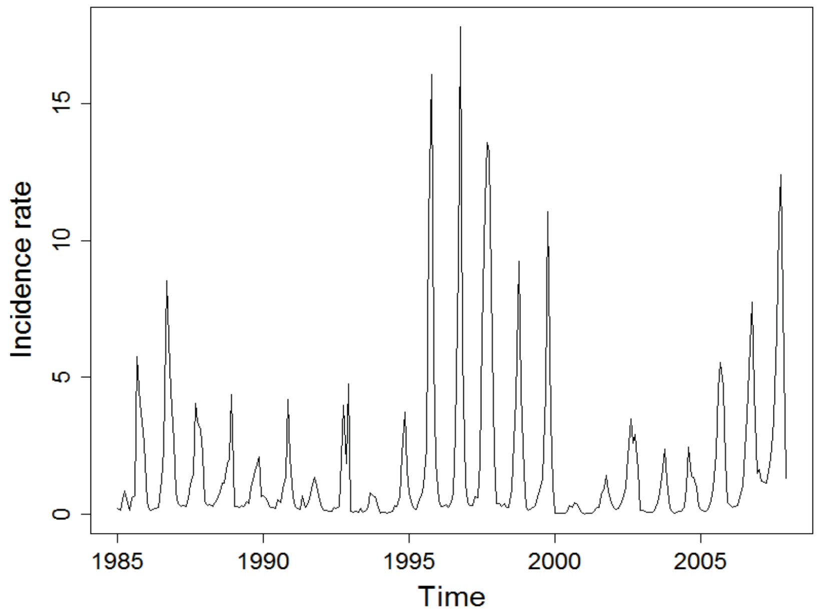
- Poisson GAM:

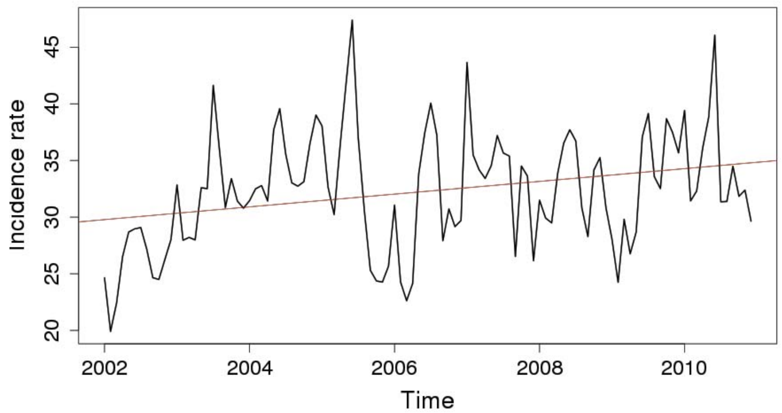
$$g(\mu_{ti}) = s(t') + \sum_{j=1}^{J} s(X_{jti}) + \sum_{k=1}^{K} \beta_k (Z_{kti}) + \log(\xi_{ti}) + d_i + \alpha$$

$s(.)$ indicates a **smooth function** (spline) for the time variable 1,...,n (months) with q degrees of freedom per year

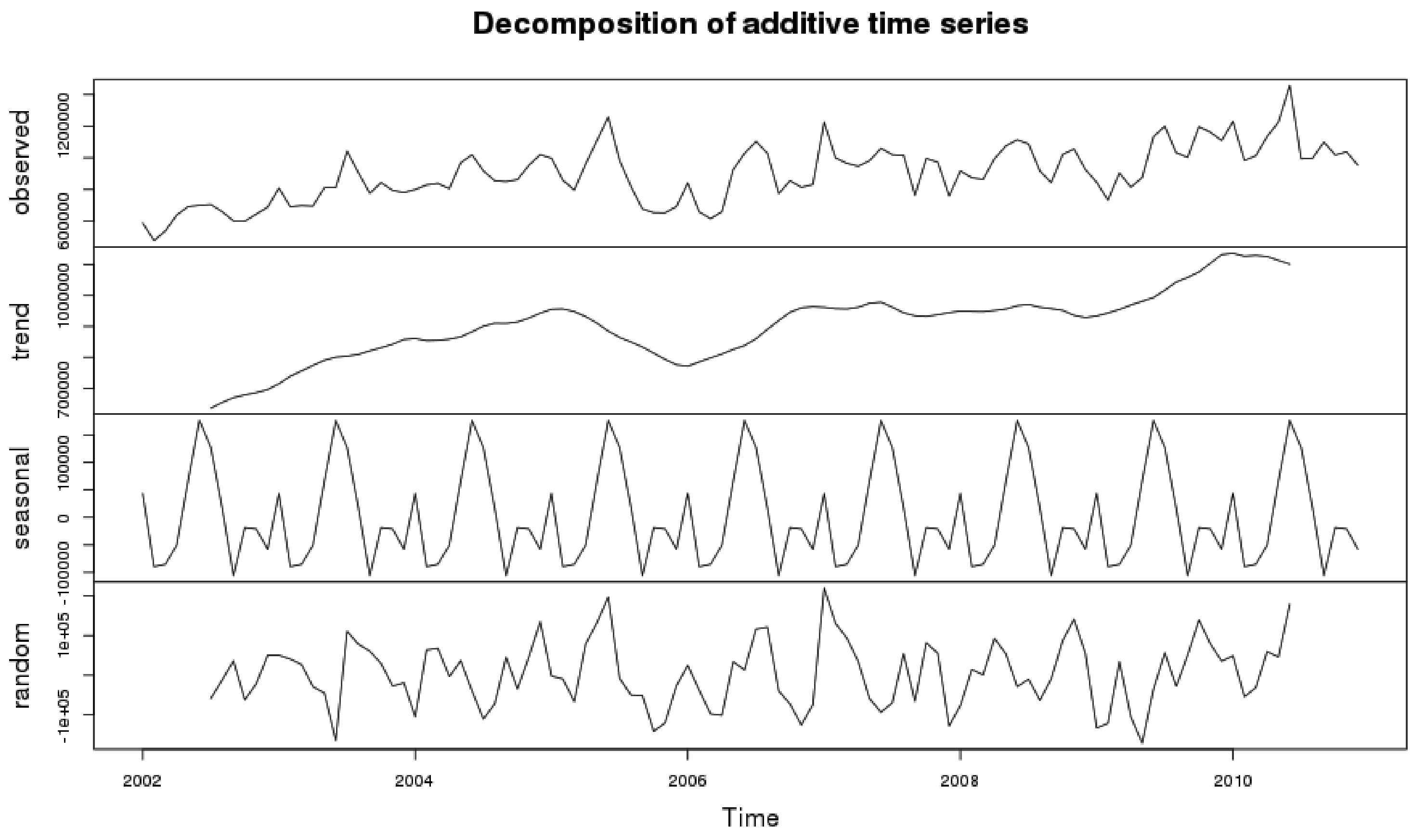
Annual cycle and inter-annual variability

- Holidays, seasonal water storage, seasonal migration
- New diagnostic techniques, micro-evolutionary changes, policy implementation





Malaria time series decomposition



The Abdus Salam International Centre for Theoretical Physics

Model

- Poisson GAM:

$$g(\mu_{ti}) = s(t') + \sum_{j=1}^{J} s(X_{jti}) + \sum_{k=1}^{K} \beta_k (Z_{kti}) + \log(\xi_{ti}) + d_i + \alpha$$

Nonlinear effect of the climate variables *(X)* at time *t* and administrative unit *i*

In the model

- Univariate smooth functions for *j* climate variables

$$\sum_{j=1}^{J} s(X_{jti})$$

$$s_1(Tmin_{ti}) + s_2(Tmax_{ti}) + s_3(Rain_{ti})$$

ICTP The Abdus Salam International Centre for Theoretical Physics

In the model

- Univariate smooth functions for *j* climate variables

$$\sum_{j=1}^{J} s(X_{jti})$$

$$s_1(Tmin_{ti}) + s_2(Tmax_{ti}) + s_3(Rain_{ti})$$

In the model

•Univariate smooth functions for *j* climate variables

$$\sum_{j=1}^{J} s(X_{j,ti})$$

$$s_1(Tmin_{ti}) + s_2(Tmax_{ti}) + s_3(Rain_{ti})$$

Model

- Poisson GAM:

$$g(\mu_{ti}) = s(t') + \sum_{j=1}^{J} s(X_{jti}) + \sum_{k=1}^{K} \beta_k (Z_{kti}) + \log(\xi_{ti}) + d_i + \alpha$$

Effect of non-climatic predictors in a linear fashion $\beta_1 Z_1 + \beta_2 Z_2 + \ldots \beta_k Z_k$

Model

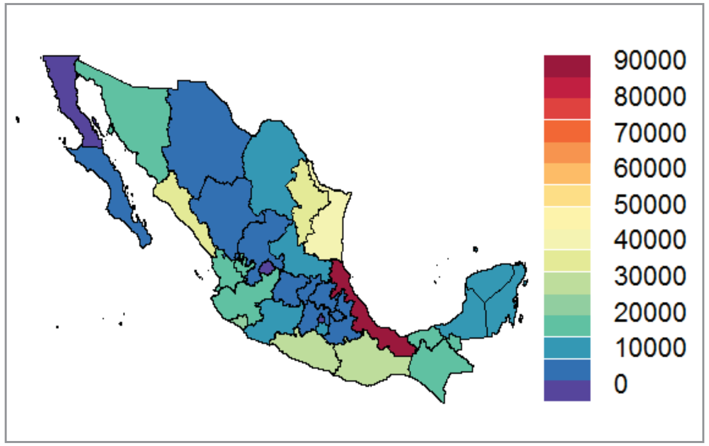
- Poisson GAM:

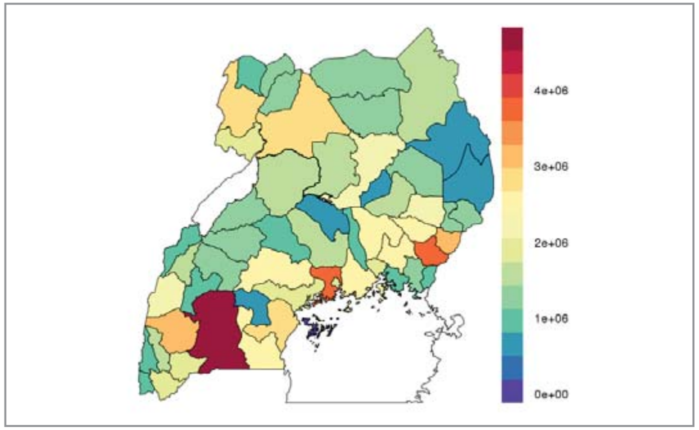
$$g(\mu_{ti}) = s(t') + \sum_{j=1}^{J} s(X_{jti}) + \sum_{k=1}^{K} \beta_k (Z_{kti}) + \log(\xi_{ti}) + d_i + \alpha$$

Log of the **population** (ξ) at time *t* and administrative unit *i* as an **offset**. No regression parameter estimated for it.

ICTP The Abdus Salam International Centre for Theoretical Physics

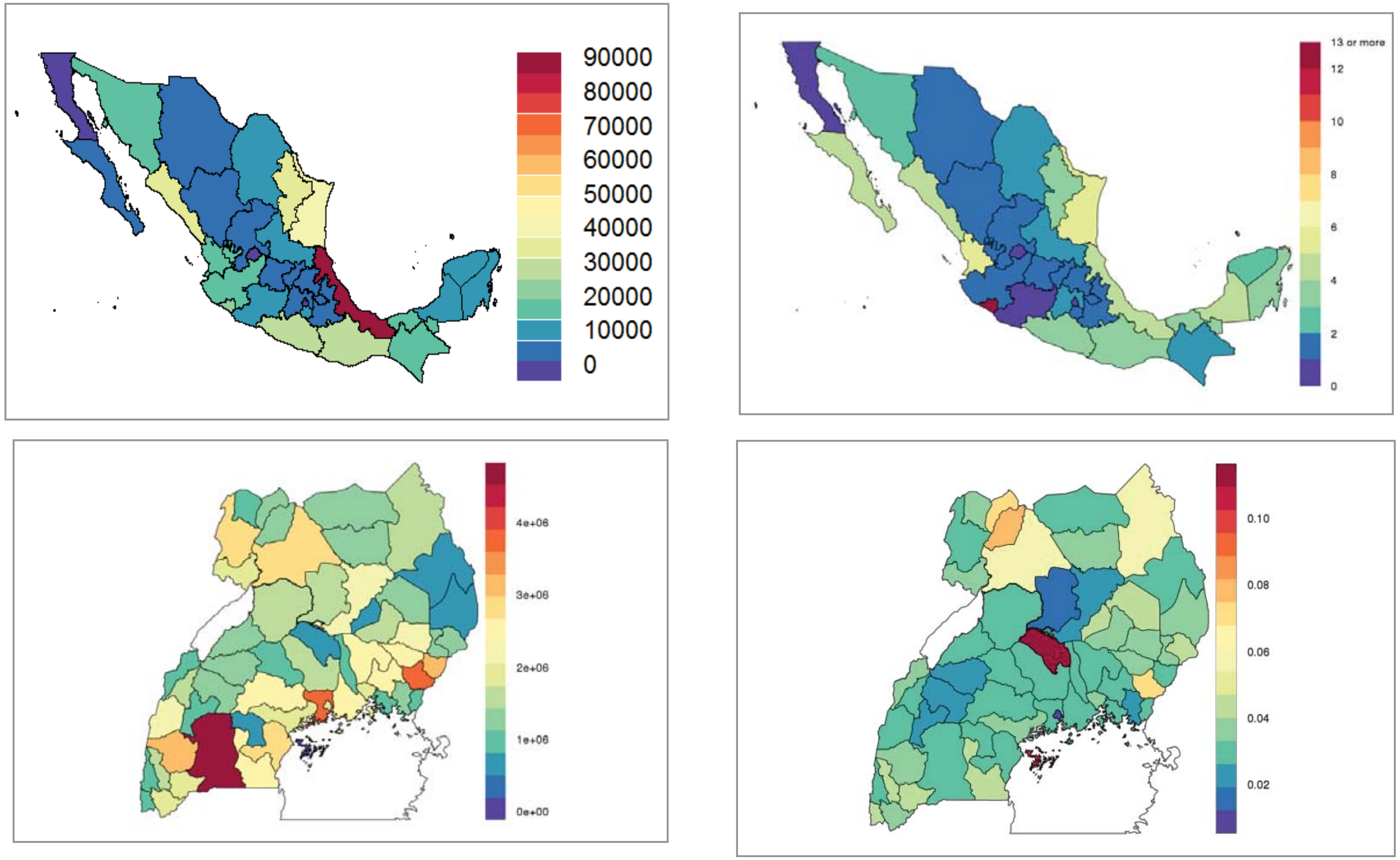
Why the offset?





You may see more cases in a region just because they have a larger population!

Why the offset?



Mean monthly disease occurrence per province (left panel)
vs. mean monthly incidence rate (right panel)

Model

- Poisson GAM:

$$g(\mu_{ti}) = s(t') + \sum_{j=1}^{J} s(X_{jti}) + \sum_{k=1}^{K} \beta_k (Z_{kti}) + \log(\xi_{ti}) + d_i + \alpha$$

Fixed-effects to control for possible omitted variable bias in each administrative unit

The Abdus Salam International Centre for Theoretical Physics

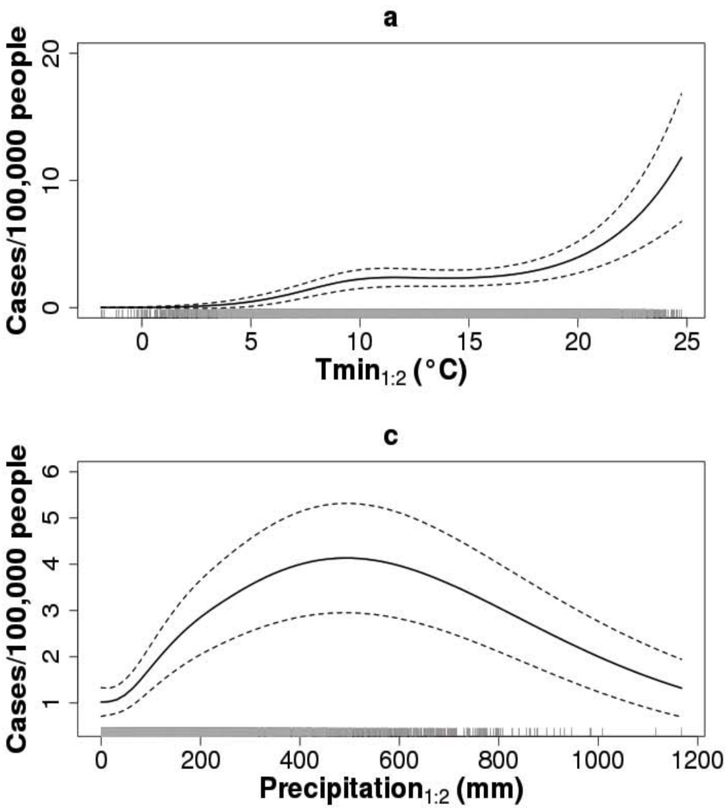
Model

- Poisson GAM:

$$g(\mu_{ti}) = s(t') + \sum_{j=1}^{J} s(X_{jti}) + \sum_{k=1}^{K} \beta_k (Z_{kti}) + \log(\xi_{ti}) + d_i + \alpha$$

Intercept (the log of the expected incidence rate when all variables in the model are evaluated at zero).

Dengue-weather



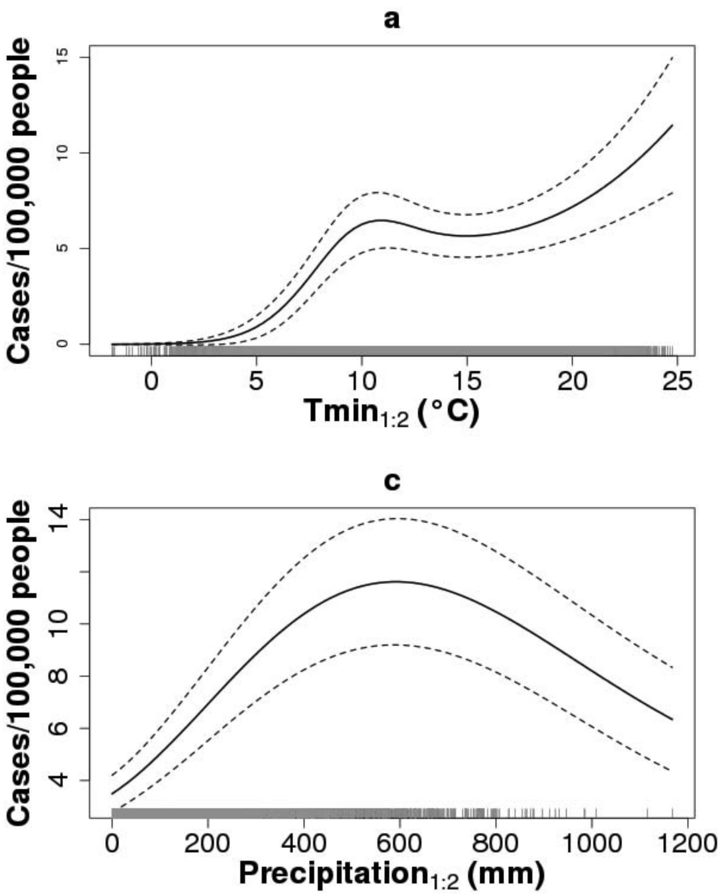
The Abdus Salam International Centre for Theoretical Physics

Autocorrelation?

- Dengue transmission implies one person leading to one or more secondary infections
  - Violation of independence
- Include log-transformed cases in the previous month (plus 1!)

The Abdus Salam International Centre for Theoretical Physics

AR1 adjusted

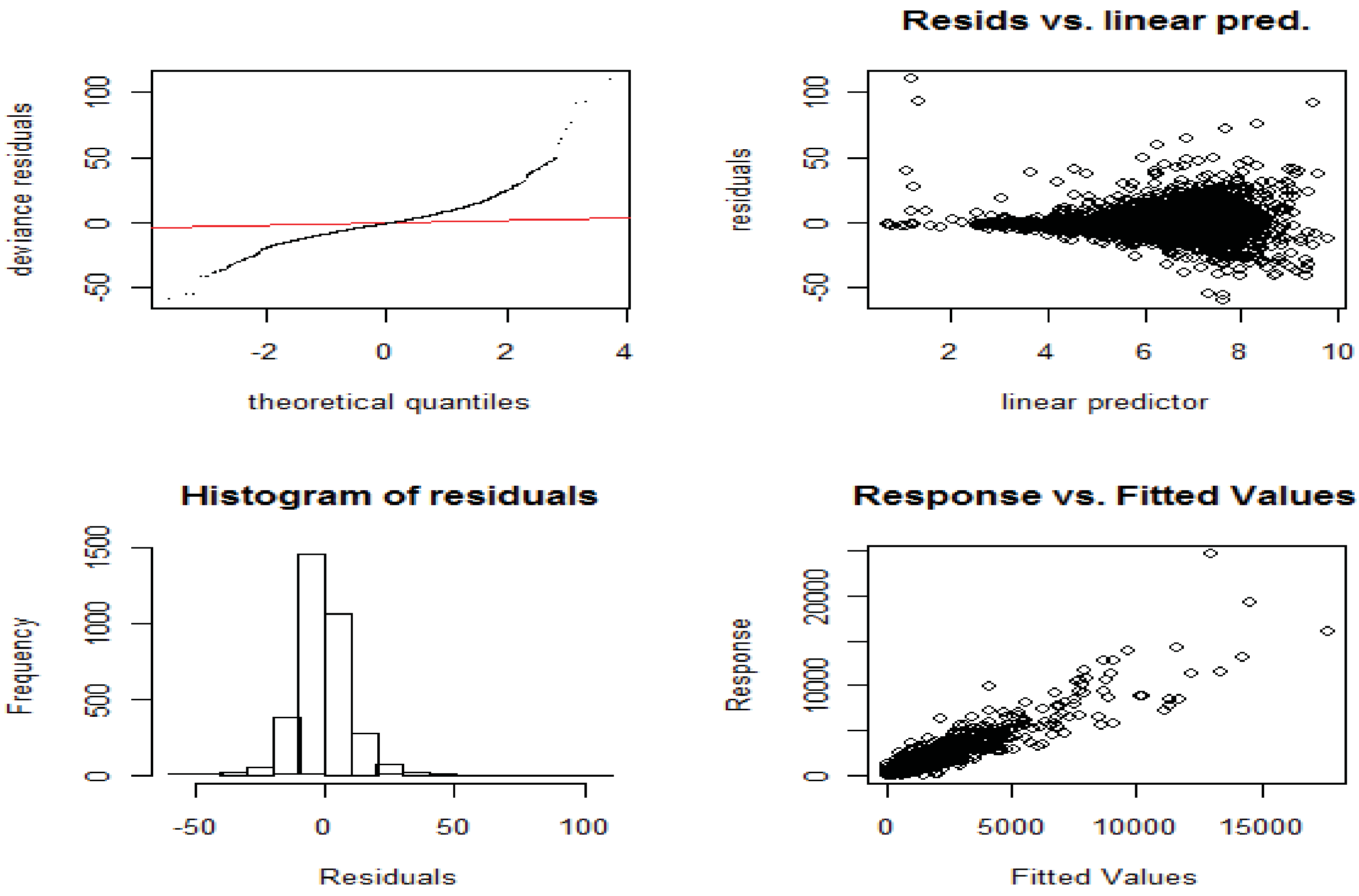


The Abdus Salam International Centre for Theoretical Physics

Model comparison

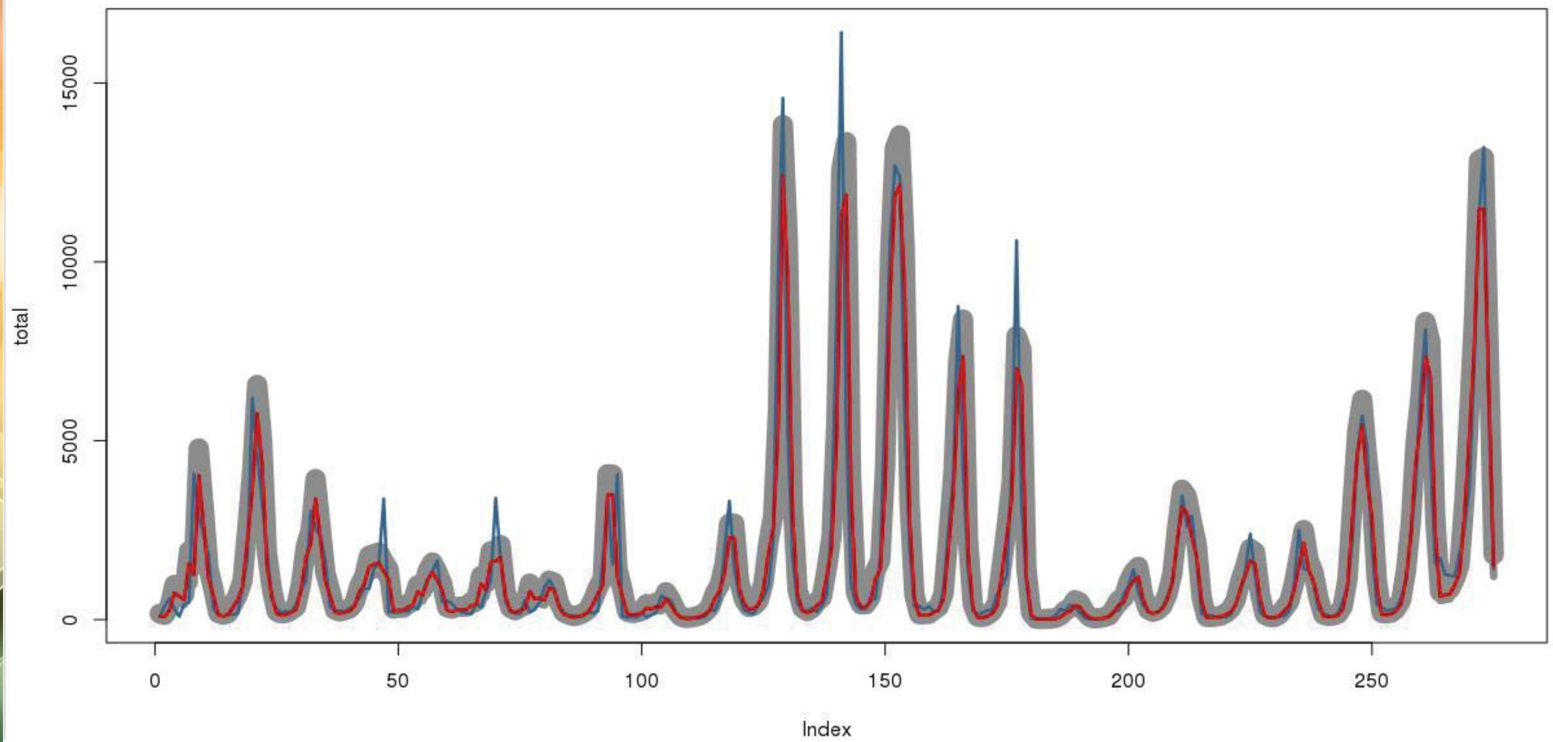
Model	AIC	GCV	PACF	Adj. R2	Deviance explained
Unadjusted	767573	88.1	0.63	0.38	62%
Autocorrelation adjusted	373465	41.8	0.06	0.68	82%

gam.check()



The Abdus Salam International Centre for Theoretical Physics

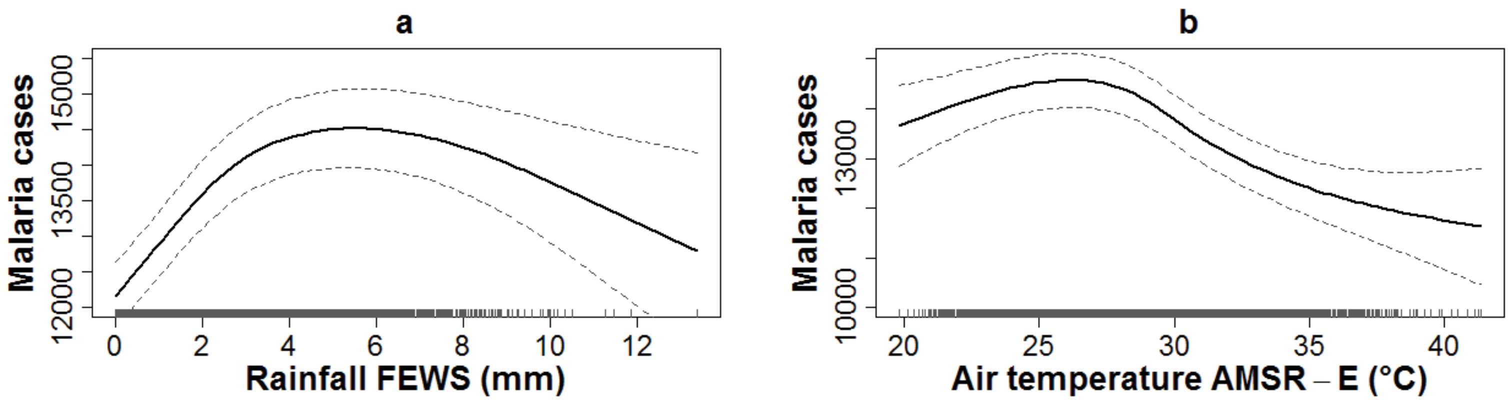
Fitted values

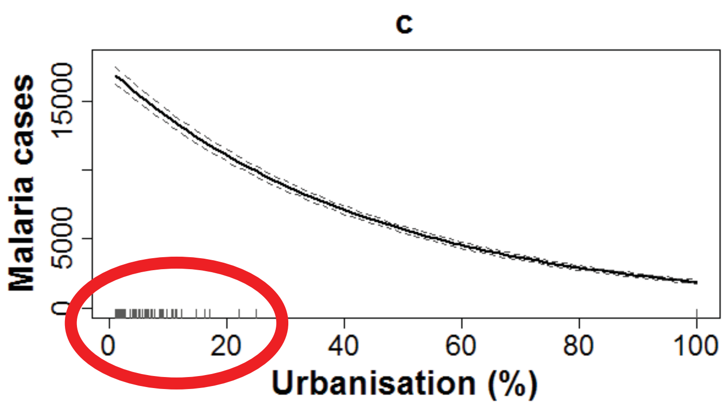


Total number of dengue cases in the country (blue line) and GAM-estimated number of cases (red line). The gray shadow indicates the 95% confidence interval.

ICTP The Abdus Salam International Centre for Theoretical Physics

Malaria results

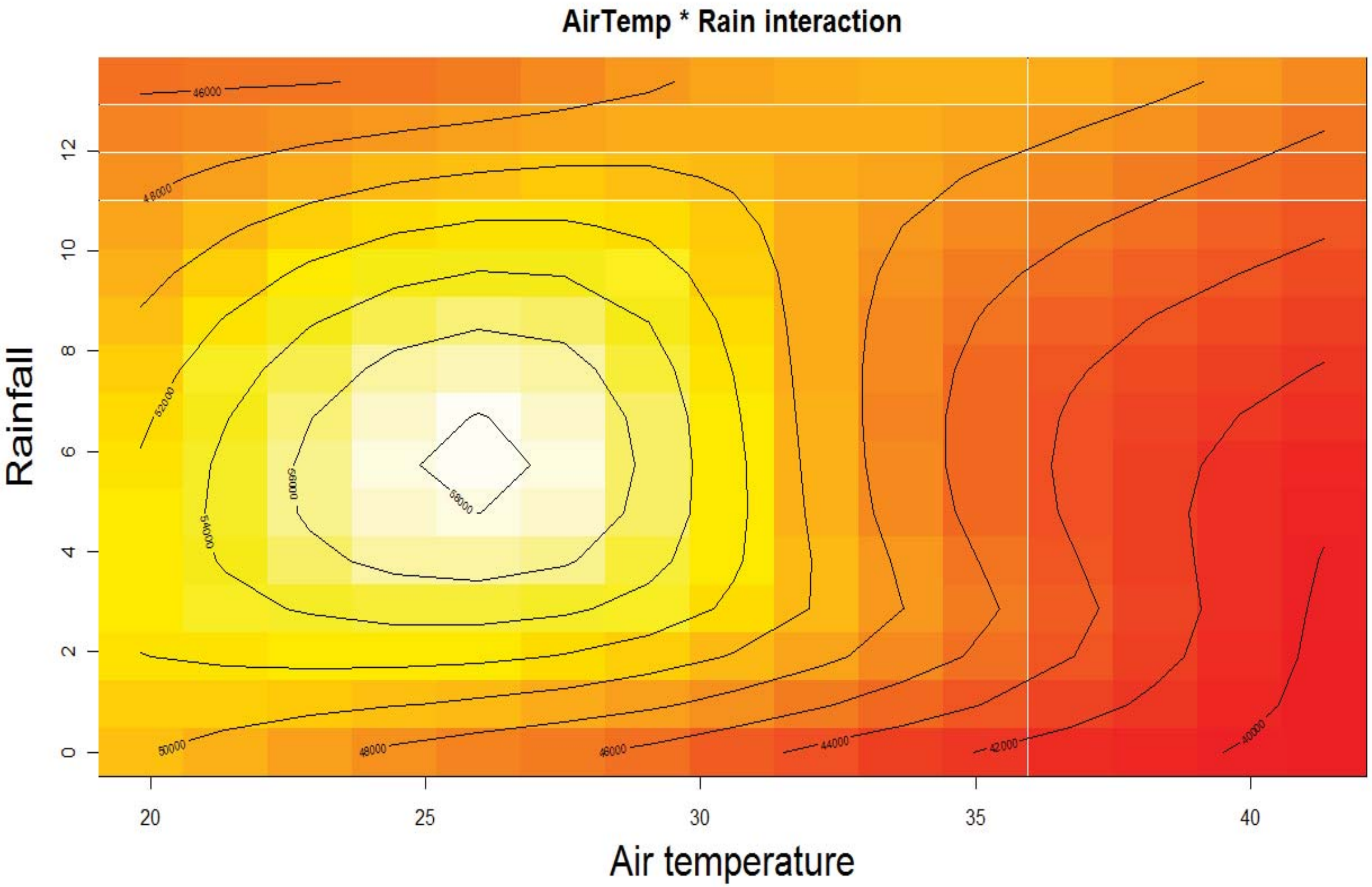




What about

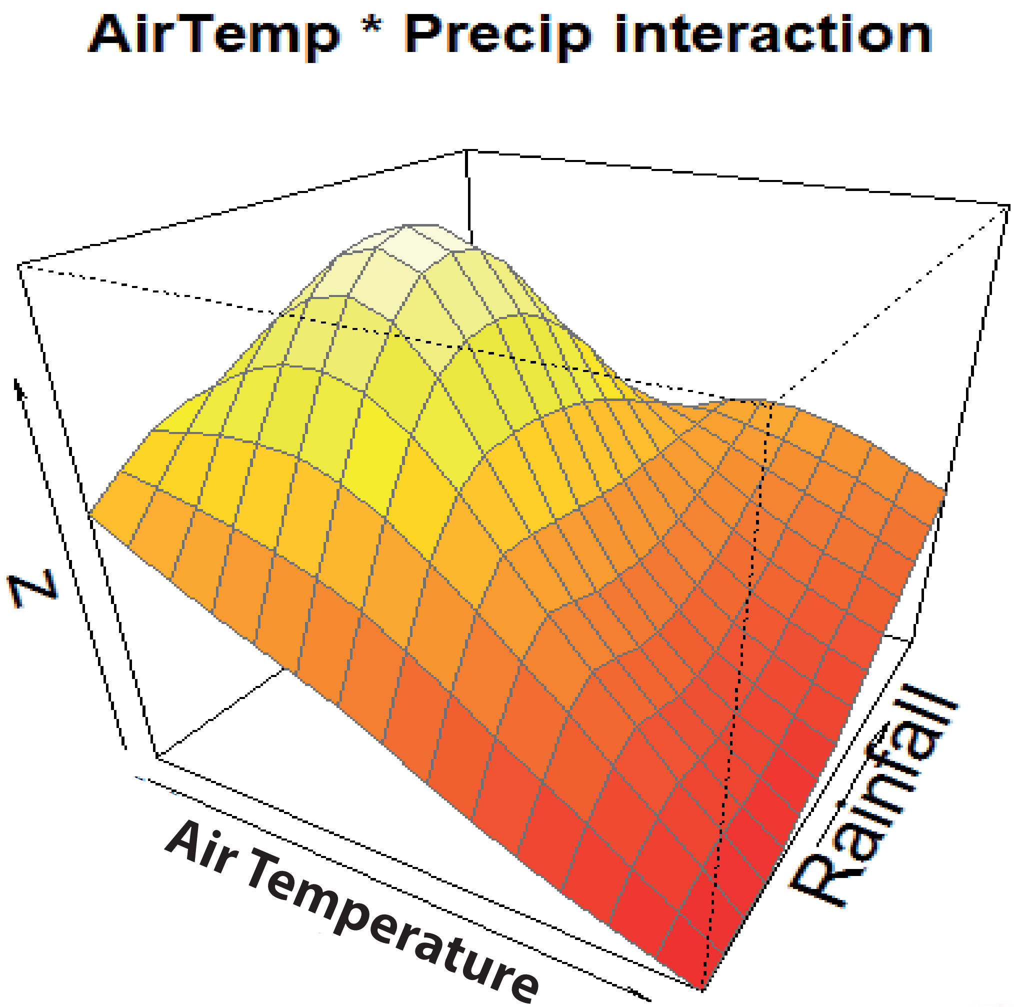
interactions?

Interaction



The Abdus Salam International Centre for Theoretical Physics

Interaction (perspective)



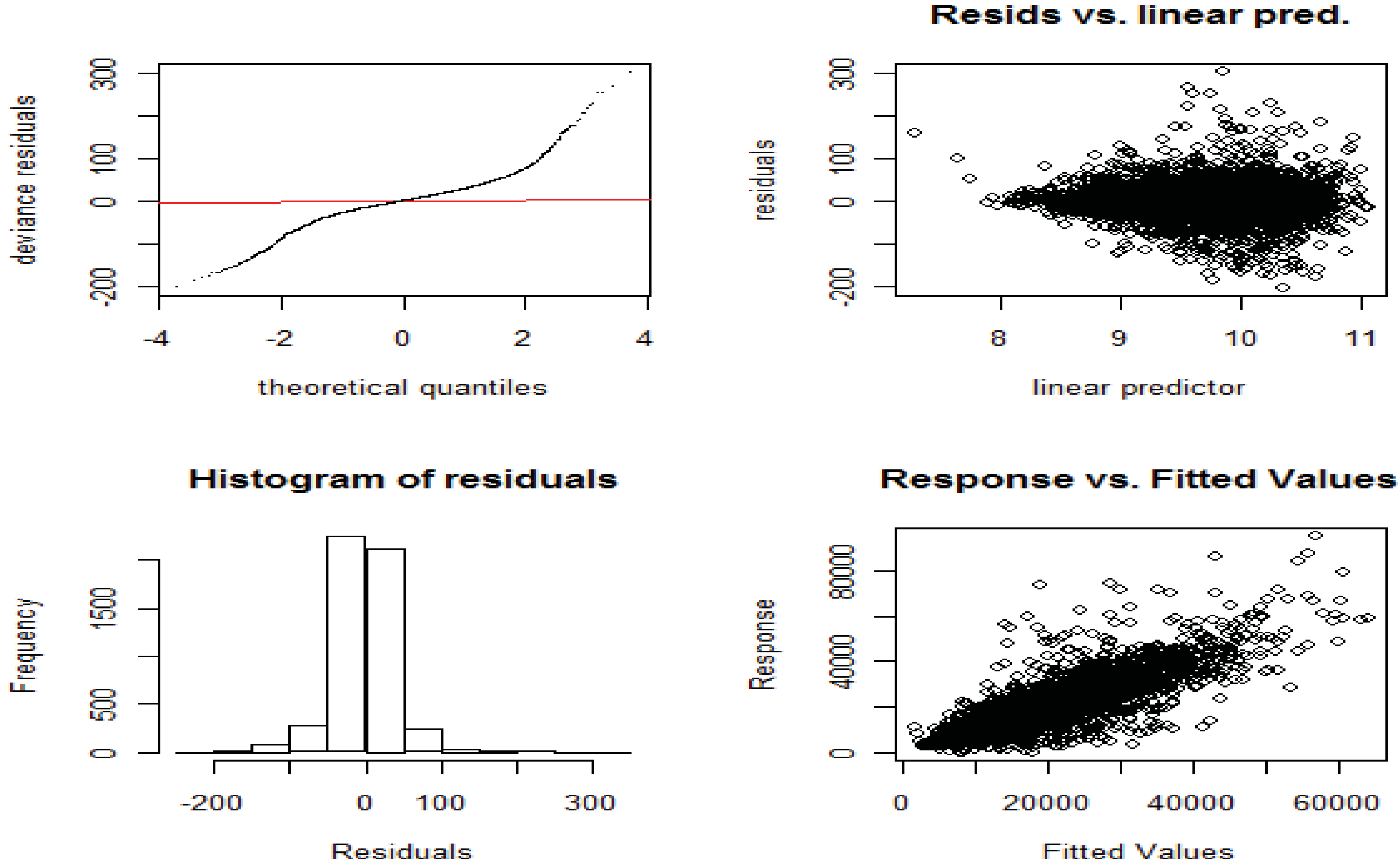
Model comparison

Model	AIC	GCV	PACF	Adj. R2	Deviance explained
No interaction	7632333	1592	-0.02	0.73	64%
With interaction	7621546	1592	-0.02	0.73	64%

```
 Resid. Df Resid. Dev    Df Deviance       F  Pr(>F)
1    4932.6    7574649
2    4930.3    7563858 2.271    10792 3.0974 0.03882 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```


The Abdus Salam International Centre for Theoretical Physics

gam.check()



The Abdus Salam International Centre for Theoretical Physics

Some limitations

- Underreporting
- Spatial autocorrelation
- Spatial aggregation
- Different seasonal trends between districts or regions
- Serotype fluctuations

Some coding tips!

```
# 1. Include a smooth for time with 2 df per year
# when the number of years = 6
# 2. Enter climate variables as smooth functions
# 3. Add other predictors in a linear fashion
# 4. Fit Poisson model with quasi-likelihood
# 5. Offset by population

mymodel <- gam(Totalcases ~
                + s(time, bs = 'cr', k = 12)
                + s(temperature, bs = 'cr')
                + s(rainfall, bs = 'cr')
                + urbanisation
                + offset(log(population))
               , # Do not delete this comma!
                data = mydata,
                family = poisson(link = 'log'),
                scale = -1) # for quasi-likelihood
summary(mymodel)
pacf(mymodel)
gam.check(mymodel)
?gam.check # To look for help
```

How to add an autocorrelation term

```
# 6. Adjust by autocorrelation

mymodel2 <- gam(Totalcases ~
                + s(time, bs = 'cr', k = 12)
                + s(temperature, bs = 'cr')
                + s(rainfall, bs = 'cr')
                + urbanisation
                + offset(log(population))
                + log(LagTotalcases+1)
               , # Do not delete this comma!
                data = mydata,
                family = poisson(link = 'log'),
                scale = -1) # for quasi-likelihood
summary(mymodel2)

# 7. compare the 2 models
anova(mymodel, mymodel2, test = 'F')
pacf(mymodel2)
gam.check(mymodel2)
```

Plotting a gam

```
# 8. plot your results

plot.gam(mymodel2,
         all.terms = F,
         scale = 0,
         pages = 1,
         rug = T,
         shade = T,
         se = T,
         lwd = 1.75,
         col = '#FF8000',
         shade.col = "gray88",
         trans = exp,
         seWithMean = F,
         main = "My model 2",
         cex.main = 1.2,
         cex.lab = 1.5)
```

Plotting a smooth interaction

```
# 9. Create a 3D plot for the interaction term
par(mar = c(5,5,3,3))
vis.gam(mymodel2, view = c('temperature', 'rainfall'),
        plot.type = 'persp', color = 'heat', n.grid = 15,
        border = 'gray45', type = 'response',
        main = 'Temperature * Rainfall interaction',
        phi = 30, theta = 30, cex.main = 1.5,
        cex.lab = 2, ylab = 'Rainfall',
        xlab = 'Temperature')

# 10. Plot a contour of the interaction
par(mar = c(5,5,3,3))
vis.gam(mymodel2, view = c('temperature', 'rainfall'),
        plot.type = 'contour', color = 'heat',
        n.grid = 15, contour.col = 'black',
        type = 'response',
        main = 'Temperature * Rainfall interaction',
        cex.main = 1.5, cex.lab = 2,
        ylab = 'Rainfall', xlab = 'Temperature')
```

Recommended reading

- Kirkwood, B & Sterne, J. *Essential medical statistics* (2008) Blackwell Publishing.
- Cohen, J. et al., *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2003) Lawrence Erlbaum
- Keele, L. *Semiparametric regression for the social sciences* 2008). Blackwell
- Shumway, & Stoffer, D. *Time Series Analysis and Its Applications With R Examples* (2006) Springer
- Gelman, A & , J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (2007) Cambridge University Press

Beyond linearity: Using GAMs for modelling dengue and malaria data

Felipe de Jesús Colón-González
fcolon_g@ictp.it

The Abdus Salam International Centre for Theoretical Physics