Statistical modelling of malaria in Botswana using the R language for statistical computing

Rachel Lowe

Institut Català de Ciències del Clima (IC3)
rachel.lowe@ic3.cat

International Conference Climate Services 3, 4-6 December 2013, Montego Bay, Jamaica.
Malaria in Africa

Malaria life-threatening disease caused by parasites transmitted to people through the bites of infected Anopheles mosquitoes.

Each year, malaria causes more than 1 million deaths, mostly in sub-Saharan Africa, among children.

Malaria is a major public health problem in Africa and its control is critical to achieving the MDGs.

Malaria is preventable and curable.
Malaria in Botswana

- Botswana: semi-arid country, malaria restricted by lack of rainfall.

- Good surveillance (malaria notifiable disease).

- Laboratory confirmed data routinely available.

- >20 years of data.

- Incorporate climate information into malaria control planning.

- Monitor routinely confounding factors such as drug and insecticide resistance.
Long-term trends

Trends in malaria incidence may result from trends in climate but mostly indicate changes in vulnerability, e.g. drug or insecticide resistance, declining control services, etc.

The long term increasing trend to 1996 ends when revisions to national control policy and practice occurred in 1997 (new drugs, new insecticide, revitalised programme).
Malaria and climate

Climate drivers:

- Temperature – various stages malaria life cycle
- Humidity – activity and survival of mosquitoes
- Rainfall – availability of breeding sites

In Botswana, the disease is highly seasonal (peak April-May) and follows the rainy season with a lag of about 2–3 months.
Malaria incidence in Botswana strongly related to rainfall during the peak rainfall season December – February (Thomson et al., 2005).

The relationship is non-linear. Possible reasons:

- Rainfall provides breeding sites, increasing risk.
- Intense rainfall and/or flooding can flush out larvae, reducing risk.
Seasonal climate forecasts

Higher than expected malaria years associated with above average precipitation.

Lower than expected malaria years associated with below average precipitation.

Forecasts able to predict below and above average rainfall during these years.

Using seasonal climate forecast information, National Malaria Control Programmes are able to:

- strengthen vector control measures
- prepare emergency containers with mobile treatment centres
- mobilise localised response
Objectives

Learn how to conduct a simple temporal climate and malaria analysis using the statistical software R.

- Analyse the relationship between malaria incidence and rainfall in Botswana.

- Investigate the long term trends in disease and vulnerability changes.

- Discuss the application of forecasts in disease prediction and control.
Data for Botswana 1982-2005

- Annual confirmed malaria cases
- Annual unconfirmed malaria cases
- Population estimates
- December-February (DJF) mean rainfall estimates
Exploratory statistics

- **Categorical variable** factor with two or more levels

- **Boxplot** useful for displaying differences
  - middle bar: median
  - box: $Q_3 - Q_1 = \text{IQR} \rightarrow 50\%$ data, whiskers: min and max

- **T-test** hypothesis test with $H_0$: means of two samples the same
Data distribution

Skewness measures extent to which distribution has long tails on one side
- Normal distribution skew=0
- Positive skew to right
- Negative skew to left

Transformations to normalise data: square roots, logs, reciprocals.
Linear regression

Simple linear regression

\[ y_t \sim N(\beta_0 + \beta_1 x_t, \sigma^2) \]

Polynomial regression

Fit higher powers of \( x \) to explain curvature in the relationship between \( y \) and \( x \)

\[ y_t \sim N(\beta_0 + \beta_1 x_t + \beta_2 x_t^2, \sigma^2) \]
Goodness of fit

\( R^2 \) proportion of variability in data accounted for by model. \( R^2 = 1 \), regression line fits data perfectly.

F statistic ratio of explained (\( R^2 \)) and unexplained (1- \( R^2 \)) variability divided by corresponding degrees of freedom. The larger the F statistic, the more useful the model.

F test (ANOVA) statistical test with null hypothesis that model 2 does not provide a significantly better fit than model 1.

\[
R^2 = 1 - \frac{RSS}{TSS}
\]

\[
RSS = \sum_i (y_i - \hat{y}_i)^2
\]

\[
TSS = \sum_i (y_i - \bar{y})^2
\]

\[
F = \left( \frac{RSS_1 - RSS_2}{p_2 - p_1} \right) \left( \frac{RSS_2}{n - p_2} \right)
\]
Model checking

Plot residuals = observed-fitted values, against:

- Fitted values (non-constant variance: heteroscedasticity)
- Explanatory variables (evidence of curvature)
- Standard normal deviates (non-normal errors)
Influential observations

- If removing an observation results in a substantial modification of the parameter estimates, the observation is said to be influential.

- Leverage indicates how heavily an observation contributes to fitted values.

- Rule of thumb: point highly influential if leverage $> 3(p-1)/n$.

- Refit model removing influential points.
Out-of-sample predictions

Observed and predicted malaria incidence
Botswana 1982-2007

- observed
- predicted
- 95% prediction interval

log(incidence)

Year

Discussion

Discuss how useful this model would be as an operational malaria early warning system.

What are the limitations of this approach?
References

