Modelling and Forecasting Malaria Incidence using Generalized Time Series Models

Siuli Mukhopadhyay*

*Department of Mathematics
IIT Bombay,India

jointly with
Drs. Nithya Gogtay**, Urmila Thatte**, P. Shetty** and Rashmi Tiwari*
**KEM Hospital, Mumbai, India

Disease Models

- Disease modelling - time and space-time model
- Example: Monthly malaria counts recorded in Mumbai from 2015-2020
- Some Data complexities:
  1. Disease data is in form of counts and proportions
  2. too many zeroes
  3. non-gaussian behaviour
  4. correlation due to time and space
  5. disease spreads affected by demographics, socio-economic variables, weather etc.
- Popular method: Count time series models

What is the need for disease modelling?

- Do we need to model diseases?
- What do we gain from these models?
- In what ways does it help to make our health system better?

What are these models?

- Epidemiological/Statistical
- Are different models needed for different diseases?
- How do we model Infectious and non infectious diseases?

How it all started

- Visit to KEM hospital, largest tertiary care hospital in Mumbai
- Visit to Municipality authorities
- Data collection for dengue and malaria, (2016).
- Data Analysis

Working Team in 2015

- Data collectors: 4
- PhD student: 1
- Masters students: 4
- Post doctoral student: 1

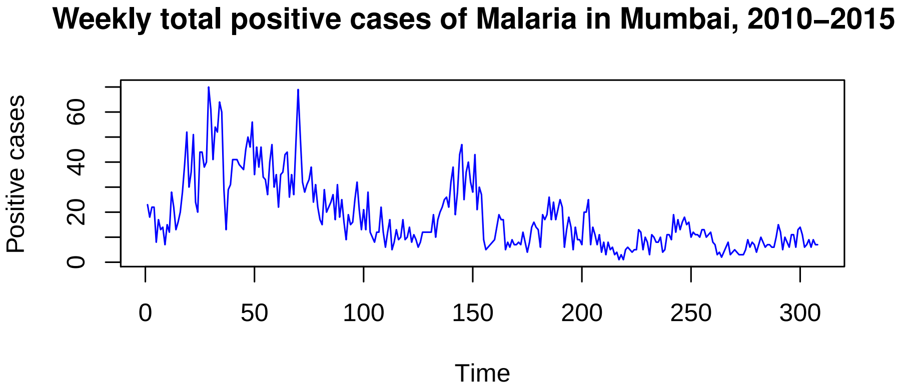
Data Description

- Source 1: Malaria data Collected from largest tertiary care hospital in Mumbai city.
- Source 2: Dengue data Collected from BMC.
- Details: Patients testing positive for malaria and dengue.
- Duration: January 2010 - November 2015.
- Format: Available in weekly format (308 weeks).

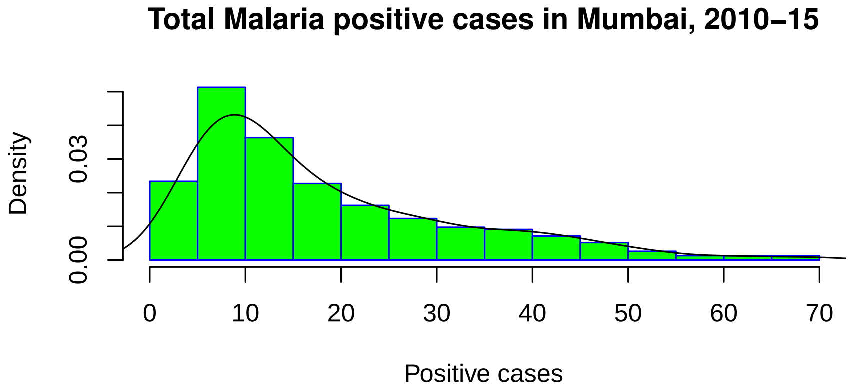
Other Data Collected

- Weather data collected from meteorology department.
- Variables: Rainfall, maximum & minimum temperature, humidity.
- 308 observations in weekly format.

Time Series Plot: Time in Weeks



Histogram of Weekly Positive Cases



- Note the lack of symmetry.
- Skewness on the right

Decomposition of Data

- Trend: Downward trend detected.
- Seasonality: Detected
- Time Series Model - AR, MA or ARMA?

- Shape of distribution: Non-normal time series
- Type of data: Positive counts of malaria

Proposed Type of Model

- Use Generalized Linear Model (GLM) theory for model fitting.

Why did we use GLM over Usual Linear Model (LM) theory?

- Can handle distributions other than normal
- Can handle positive count data

Covariate Analysis

- Which covariates should we include?

Coeff.	Estimate	S.Error	p-value
Rainfall	7.591e-05	1.637e-04	0.64289
Tmax	4.436e-02	8.632e-03	2.77e-07 ***
Tmin	-1.457e-03	1.209e-03	0.22834
Humidity	1.858e-02	1.881e-03	$<$ 2e-16 ***

- Rainfall & Tmax are significant at $\alpha = 5\%$.

Data sets

- Training data: January 2010 - November 2014, i.e. 256 weeks
- Test data: December 2014-November 2015, i.e. 52 weeks

GLM: Poisson Model

- Assume: Data generated by a Poisson distribution with mean λ.
- Distributional form: $f(y_t|pastinfo) = \frac{\exp^{-\lambda_t} \lambda_t^{y_t}}{y_t!}$, $y_t = 0, 1, \ldots.$
- Mean and Variance: λ_t.
- Mean Model: $\log(\lambda_t) = \eta_t$.
- η_t: known function of covariates, time and some unknown parameters.

Models for η_t: Poisson Distribution

$M1$:

$$\eta_t = x_t'\beta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + \phi_5 Y_{t-6} + \theta_1 e_{t-1} + \theta_2 e_{t-2}$$

$M2$:

$$\eta_t = x_t'\beta + \phi_1 \log(Y_{t-1}) + \phi_2 \log(Y_{t-2}) + \phi_3 \log(Y_{t-3}) + \phi_4 \log(Y_{t-4}) + \phi_5 \log(Y_{t-5}) + \phi_6 \log(Y_{t-6})$$

$M3$:

$$\eta_t = x_t'\beta + \theta_1(\log(Y_{t-1}) - x_{t-1}'\beta) + \theta_2(\log(Y_{t-2}) - x_{t-2}'\beta) + \theta_3(\log(Y_{t-3}) - x_{t-3}'\beta) + \theta_4(\log(Y_{t-4}) - x_{t-4}'\beta) + \theta_5(\log(Y_{t-5}) - x_{t-5}'\beta) + \theta_6(\log(Y_{t-6}) - x_{t-6}'\beta) + \phi_1 e_{t-2}$$

$M4$:

$$\eta_t = x_t'\beta + \phi_1 e_{t-1} + \phi_2 e_{t-2} + \phi_3 e_{t-3} + \phi_4 e_{t-4}.$$

where $x_t'\beta = \beta_0 + \beta_1 t + \beta_2 HMD_t + \beta_3 T_{max\,t} + \beta_4 cos(2\pi t/13)$

GLM: Negative Binomial Model

- Distributional form:
$f(y_t|pastinfo) = \frac{\Gamma(y_t+k)}{\Gamma(k)\Gamma(y_t+1)}(\frac{\lambda_t}{\lambda_t+k})^{y_t}(\frac{k}{\lambda_t+k})^k \, y_t = 0, 1, \ldots.$
- Mean λ_t.
- Variance: $\lambda_t + \frac{{\lambda_t}^2}{k}$
- Mean Model: $\log(\lambda_t) = \eta_t$.
- η_t: known function of covariates, time and some unknown parameters.

$M5:$

$$\eta_t = x_t'\beta + \phi_1 Y_{t-1} + \phi_2 Y_{t-4} + \phi_3 Y_{t-6} + \theta_1 e_{t-1} + \theta_2 e_{t-2}$$

$M6:$

$$\eta_t = x_t'\beta + \phi_1 \log(Y_{t-1}) + \phi_2 \log(Y_{t-2}) + \phi_3 \log(Y_{t-4}) + \phi_4 \log(Y_{t-6})$$

$M7:$

$$\eta_t = x_t'\beta + \theta_1(\log(Y_{t-1}) - x_{t-1}'\beta) + \theta_2(\log(Y_{t-2}) - x_{t-2}'\beta) + \theta_3(\log(Y_{t-4}) - x_{t-4}'\beta) + \theta_4(\log(Y_{t-6}) - x_{t-6}'\beta) + \phi_1 e_{t-2}$$

$M8:$

$$\eta_t = x_t'\beta + \phi_1 e_{t-1} + \phi_2 e_{t-2} + \phi_3 e_{t-3}.$$

Summary Table

M	p	WR	MSE_{WR}	$MSE_{Response}$	χ^2	D	df	AIC	BIC
1	11	30.94	0.12	57.81	566.17	580.84	236	1752.3	1794.49
2	10	31.73	0.13	54.63	552.99	569.59	239	1748.00	1786.72
3	7	34.45	0.14	56.62	575.87	596.55	242	1768.90	1797.12
4	8	35.51	0.14	66.13	670.77	683.62	243	1866.60	1898.40
5	9	32.84	0.13	64.24	238.23	252.08	240	1642.50	1681.22
6	8	31.64	0.13	55.51	237.57	253.20	241	1624.70	1659.87
7	5	33.85	0.14	58.05	238.11	255.14	244	1633.50	1658.17
8	7	35.55	0.14	67.39	239.06	254.13	245	1708.36	1708.36

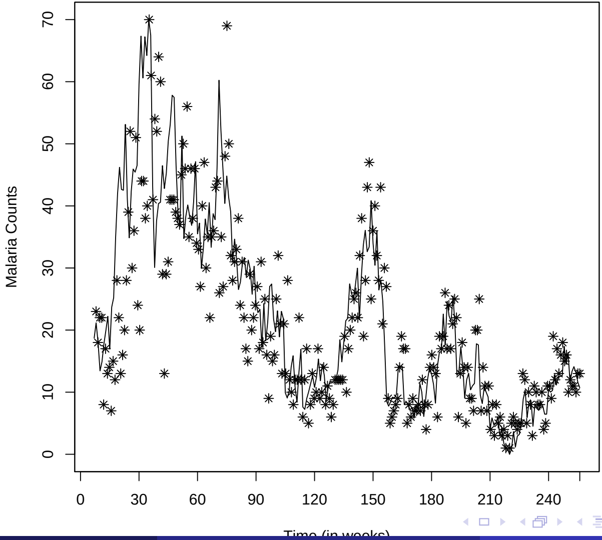
Best model?

- Best Model: M6
- Lowest values of AIC, χ^2, MSE.

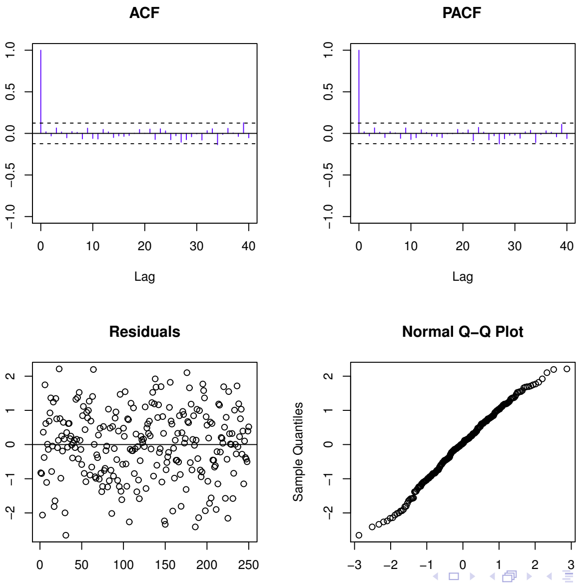
Parameter Estimates and Standard Errors, model M6

Parameter	Estimates	Std Error	P-values
(Intercept)	-1.063	0.613	0.083
t	-0.001	0.0004	0.003
$\cos(\frac{2\pi t}{13})$	-0.098	0.033	0.002
Y_{t-1}	0.386	0.057	< 0.001
Y_{t-2}	0.127	0.058	0.031
Y_{t-4}	0.119	0.054	0.028
Y_{t-6}	0.140	0.051	0.006
HMD	0.010	0.002	< 0.001
Tmax	0.036	0.014	0.009

Fitted versus Observed Data



Randomized quantile residual analysis for model M6



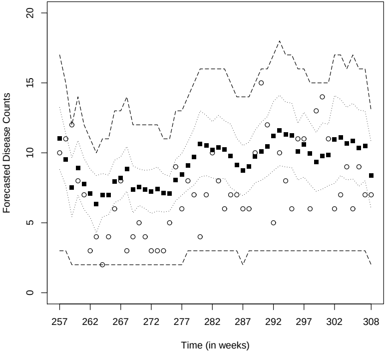
One Step at a time Forecast

- The forecasted estimate:

$$\begin{aligned}\hat{\mu}_{t+1|t+1} &= \exp[\hat{\beta}_0 + \hat{\beta}_1(t+1) + \hat{\beta}_2 HMD_{t+1} + \hat{\beta}_3 T_{max\,t+1} \\ &+ \hat{\beta}_4 cos(2\pi(t+1)/4)\hat{\phi}_1 \log(Y_t) \\ &+ \hat{\phi}_2 \log(Y_{t-1}) + \hat{\phi}_3 \log(Y_{t-3}) + \hat{\phi}_4 \log(Y_{t-5})], \\ & t = N, N+1, \ldots.\end{aligned}$$

- For model M6, $\hat{\sigma}^2_{t+1|t+1} = \hat{\mu}_{t+1} + \hat{\mu}^2_{t+1}$.
- $Q_{t+1|t+1}(p)$ is estimated by the pth quantile of the negative binomial distribution.

Actual counts (o), Forecasted Values (•), Interval of Estimated 50% Quantiles (_), Interval of 3 $\hat{\sigma}_{m|m}$ (...)



For our forecasts: MAD= 2.85, MSD= 3.31

Comparison with Other Malaria Models

Model	RMSE	MAD
Model M6	3.31	2.87
Linear Regression model with lagged weather covariates	5.58	4.87
Poisson Regression model with lagged weather covariates	3.55	3.12
ARMA $(4,0)+$ lagged weather covariates	4.57	3.73
ARIMA $(3,1,0)+$ lagged weather covariates	5.39	4.69
SARIMA $(3,1,2)(1,2,1)^{52}+$ lagged weather covariates	9.61	8.11