Title: Large outbreak of urogenital schistosomiasis acquired in Southern Corsica: monitoring the early signs of the endemicization?

Supplementary material

1 Methods

1.1 Modelling the time series of weekly counts of Schistosoma Western blot positive cases before the intervention

In a first step we modeled the data collected before the start of the nationwide screening (the intervention), *i.e.* the first 181 weeks of the time series, from 1 January 2011 to 24 June 2014.

We analyzed this time series of counts with "integer-valued generalized autoregressive conditional heteroscedastic" (INGARCH) models [1] ,using the package `tscount` of the R software [2].

The number of WB-positive tests Y_t in each week t was supposed to follow either 1) a Poisson distribution, or 2) a negative binomial distribution of mean λ_t and dispersionϕ, conditionally on past observations:

$Y_t|F_{t-1} \sim P(\lambda_t)$ or $Y_t|F_{t-1} \sim NB(\lambda_t, \phi)$, where F_{t-1} is the σ-field generated by $\{Y_{t-1}, Y_{t-2}, \ldots\}$.

The mean λ_t was modelled as follows:

$$\log(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \log\left(Y_{t-i_k} + 1\right) + \sum_{\ell=1}^{q} \alpha_\ell \log\left(\lambda_{t-j_\ell}\right) + \eta^{\mathrm{T}} X_t, \qquad (1)$$

where β_0 is the intercept, the second and the third term enable to regress respectively on arbitrary lagged observations and lagged conditional means of the response, and the final term represent the linear effect of *n* time-varying covariates.

Model assessment.

For model selection, in addition to the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) we used several diagnostic plots included in the `tscount` package after model fitting:

- (i) the autocorrelation function of the response residuals. After fitting the time series models, we verified graphically the absence of residual autocorrelation by plotting the response residuals defined by: $r_t = Y_t - \lambda_t$ against its lagged values.
- (ii) the non-randomized probability integral transform (PIT, function `pit`), which will follow a uniform distribution if the predictive distribution is correct. We used the histogram plot developed by Czado et al [3]to identify underdispersion, showed by a U-shaped plot, or overdispersion, showed by an upside down U-shaped plot.
- (iii) and the marginal calibration (function `marcal`) defined as the difference of the average predictive cumulative distribution function (c.d.f.) and the empirical c.d.f. of the observations, that should be close to zero. As proposed by Christou and Fokianos, we compared the plot of the marginal calibration of the fitted models for values in the range of the original observations [4]. We compared model based deviations from zero.

1.2 Modelling the intervention(s)

We studied the effect of the national screening (intervention) on the weekly counts of WB-positive cases.

The model selected in step 1) is applied here to the whole time series and enhanced by incorporating terms that model the impact of *s* interventions.

The effect of the m^{th} intervention is modeled with the term $\omega_m {\delta_m}^{t-\tau_m} \mathbb{1}(t \geq \tau_m)$ [5], where ω_m is the size of the intervention, τ_m its occurrence time and δ_m its decay rate.

The resulting log-linear predictor is:

$$\log(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \log\left(Y_{t-i_k} + 1\right) + \sum_{\ell=1}^{q} \alpha_\ell \log\left(\lambda_{t-j_\ell}\right) + \eta^{\mathrm{T}} X_t + \sum_{m=1}^{s} \omega_m {\delta_m}^{t-\tau_m} \mathbb{1}(t \geq \tau_m), \quad (2)$$

While the parameters ω_m are estimated alongside the other parameters of the model (*e.g.* β_0, β_1, etc…), τ_m and δ_m are constants fixed before the fitting procedure.

We assumed that the interventions provoke a temporary increase in the weekly number of WB-positive cases, followed by a decrease. Thus, the constants δ_m were chosen in the range (0,1), to model an exponentially decaying intervention effect (transient shift). Different values of δ_m were tested, specifically 0.7,0.8 and 0.9, a strategy recommended earlier [6]. For this purpose, we employed the function `interv_multiple` that uses an iterative detection procedure for detection of the effect of multiple interventions described by Liboschik et al [5]. We tested for external interventions effects each time point after the launching of the nationwide screening$(181 \leq \tau_m \leq 262)$. For each value of$\delta_m$, the procedure computed a score statistic for intervention effects occurring at each time pointτ_m. The intervention effect with the lowest p-value was retained under the condition that this p-value was under 0.05. After removing the selected intervention effect, the procedure started over and the time series was tested for onward interventions until no significant intervention effect could be found.

1.3 Estimating the outbreak size

The outbreak size O was estimated as the cumulative difference, over the T weeks of the whole time series, of the means of the INGARCH models with and without the intervention covariates:

$$O = \sum_{t=1}^{T} \lambda_t - \sum_{t=1}^{T} (\lambda_t \,|\omega_m = 0, m = 1..s)$$

with

$$\lambda_t = \exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k \log\left(Y_{t-i_k} + 1\right) + \sum_{\ell=1}^{q} \alpha_\ell \log\left(\lambda_{t-j_\ell}\right) + \eta^{\mathrm{T}} X_t + \sum_{m=1}^{s} \omega_m {\delta_m}^{t-\tau_m} \mathbb{1}(t \geq \tau_m)\right)$$

and

$$\lambda_t|\omega_m = 0, m = 1..s = \exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k \log\left(Y_{t-i_k} + 1\right) + \sum_{\ell=1}^{q} \alpha_\ell \log\left(\lambda_{t-j_\ell}\right) + \eta^{\mathrm{T}} X_t\right)$$

We used the delta method [7] to approximate the variance of $\hat{O}$, using the estimated variance of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q, \eta^T, \ldots, \hat{\omega}_1, \ldots, \hat{\omega}_s$.

2 Results

Models were fitted to the data of the 181 first weeks of the time series using a Poisson and a negative binomial conditional distribution as shown in Table S1.

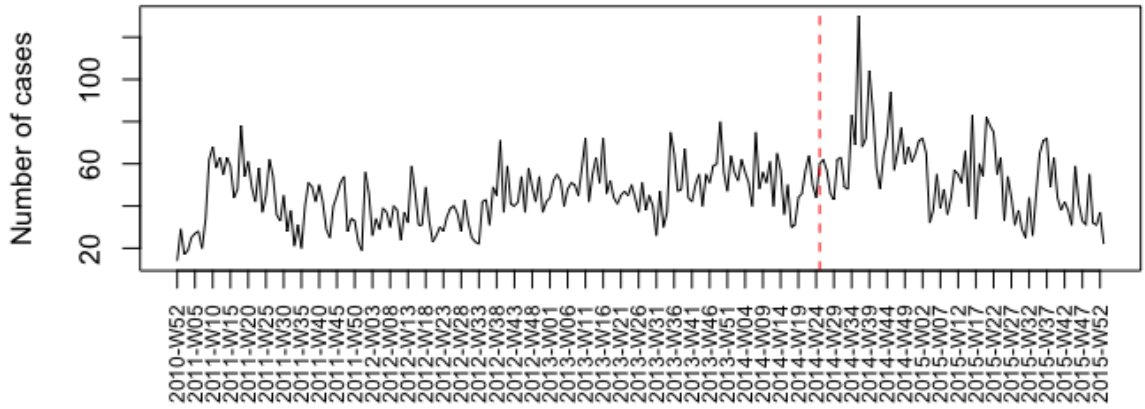

Figure S1. Weekly counts of Schistosoma Western Blot positive cases in France, 2011-2015

The vertical dashed red line represents the week of start of the nationwide screening (2014-06-24, week 2014-25).

Table S1. Summary of the results of model fitting on time series before the start of the nationwide screening for serological schistosomiasis in France, week 2010-52– week 2014-25.

Model characteristics		
Distribution family	Poisson	Negative binomial
Link function	logarithmic	logarithmic
Log-likelihood	-776·080	-697·1535
BIC	1567·754	1415·101
AIC	1558·159	1402·307
Regression parameter		
β_0: Intercept	2·075 [1·792-2·359]$	2·075[1·691-2·705]*
β_1: 1 week-lagged observation	0·429 [0·352-0·506]$	0·429 [0·268 0·530]*
η_1: linear trend	0·000991 [0·000548-0·00143]$	0·000991 [0·000307-0·00186]*
$\sigma^2 = \phi^{-1}$ overdispersion		0·0486 [0·0350-0·0648]*

$ Standard errors and 95% confidence intervals were calculated by normal approximation. *Standard errors and 95% confidence intervals were calculated by parametric bootstrap with 500 replications.

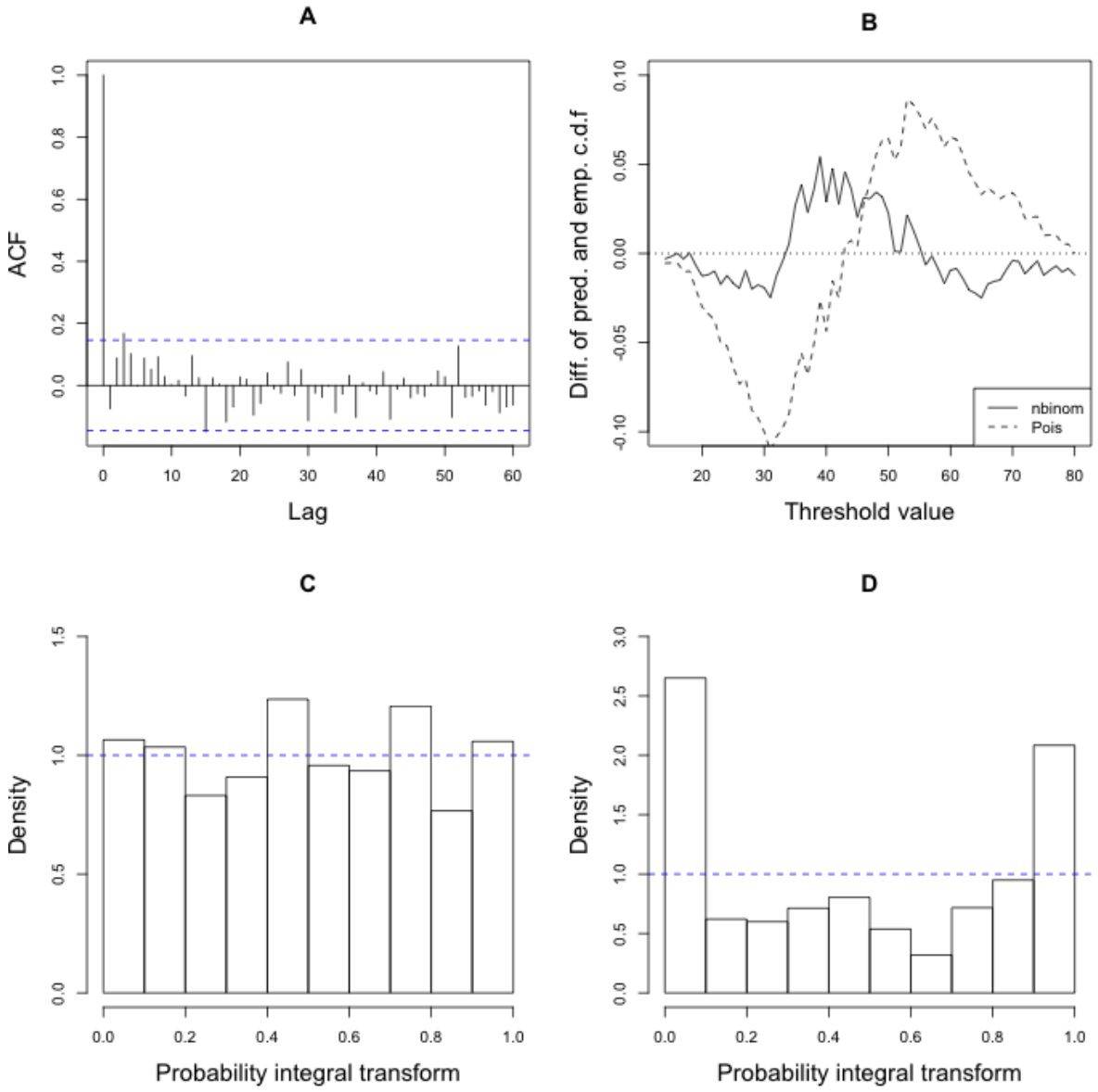

Figure S2. Diagnostic plots after model fitting to the weekly counts of WB-positive cases.

Both conditional distributions had strictly identical response residuals. As shown in the plot of the autocorrelation function (A), there was no maximum autocorrelation indicative of seasonal fluctuation or persisting serial correlation. The non-randomized probability integral transform of the Negative binomial model (C) shows the distribution approaches uniformity better. The marked U-shape of the PIT histogram of the Poisson model (D) is indicative of underdispersion. Finally the marginal calibration plot of the negative binomial model is close to zero while major deviations are observed for the Poisson model.

Table S2. Summary of the results of the intervention analysis, 1 January 2011-31 December 2015

Model characteristics

Distribution family	Negative binomial
Link function	logarithmic
Log-likelihood	-1031·334
BIC	2072·668
AIC	2090·51

Regression parameter

β_0: Intercept	2·223 [1·867-2·722]*
β_1: 1 week-lagged observation	0·399 [0·262- 0·493]*
η_1: linear trend	0·000558 [0·000114-0·00110]*
ω_1: intervention size	0·502 [0·285-0·767]*
τ_1 : time of occurrence of the intervention	191
δ_1 :decay rate	0.9
$\sigma^2 = \phi^{-1}$ overdispersion	0·0514 [0·0388-0·0638]*

*Standard errors and 95% confidence intervals were calculated by parametric bootstrap with 500 replications.

The intervention analysis identified a transient shift 11 weeks after the start of the nationwide screening for schistosomiasis (week 191). A covariate was included in the fitted negative binomial model to account for the intervention effect

$$Y_t|\mathcal{F}_{t-1} \sim \text{NegBinom}\ (\lambda_t, \phi = 19.47),$$

$$\lambda_t = 2.22 + 0.40\log(Y_{t-1}+1) + 0.00056\,t + 0.50\ \cdot 0.9^{t-191}\,\mathbb{1}\,(t \geq 191), \quad t = 1, \ldots, 262$$

We went on to estimate the outbreak size O based by computing the cumulative difference of the means of the model with and without the covariate accounting for the identified significant intervention effect.

$$0 = \sum_{t=1}^{T} \lambda_t - \sum_{t=1}^{T} (\lambda_t\,|\omega_1)$$

The estimated overall outbreak size was estimated at 338 cases [95% Confidence Interval: 166-510].

References

[1] T. Liboschik, R. Fried, K. Fokianos, P. Probst, tscount: Analysis of Count Time Series. R package version 1.3.3., 2016. http://tscount.r-forge.r-project.org.).
[2] R Core Team, R: A language and environment for statistical computing, 2016. http://www.R-project.org/.).
[3] C. Czado, T. Gneiting, L. Held,Predictive model assessment for count data, Biometrics 65(4) (2009) 1254-61.
[4] V. Christou, K. Fokianos,On count time series prediction, Journal of Statistical Computation and Simulation 85(2) (2015) 357-373.
[5] T. Liboschik, P. Kerschke, K. Fokianos, R. Fried,Modelling interventions in INGARCH processes, International Journal of Computer Mathematics 93(4) (2016) 17.
[6] K. Fokianos, R. Fried,Interventions in log-linear Poisson autoregression, Statistical Modelling 12(4) (2012) 33.
[7] G.W. Oehlert,A Note on the Delta Method, The American Statistician 46(1) (1992) 27-29.