Zero-inflated negative binomial mixed models

for predicting number of wildfires

M. Bugallo[1], M.D. Esteban[1], D. Morales[1], M.F. Marey-Pérez[2]

[1]Center of Operations Research, Miguel Hernández University of Elche, Spain
[2]Higher Polytechnic School of Engineering, Santiago de Compostela University, Spain

Stochastic and Physical Monitoring Systems (SPMS 2024)
Group of Applied Mathematics and Stochastics (GAMS)
Department of Mathematics, Czech Technical University in Prague
Sokol Dobrichovice, Czech Republic, June 20-24, 2024

UNIVERSITAS Miguel Hernández RESEARCH INSTITUTE

UNIVERSITAS Miguel Hernández

Overview

Data and problem of interest

Wildfires

- Figure 1 shows the distribution and extent of burnt areas in Europe and the Mediterranean in 2021.
- Red circles represent fire events.

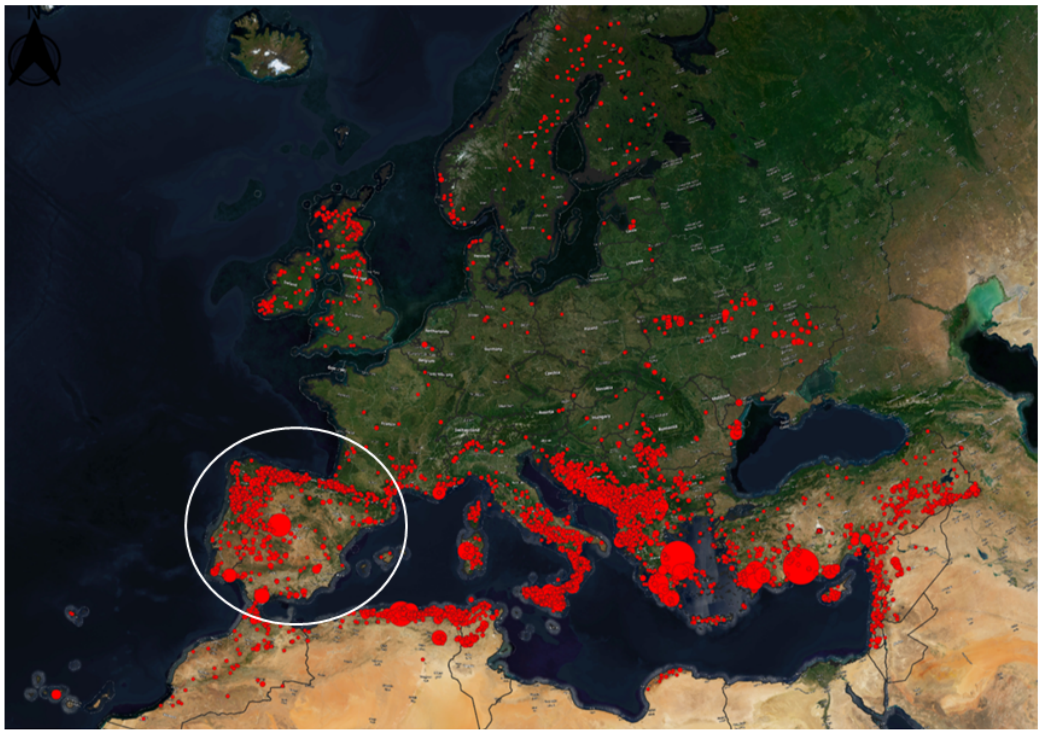


Figure 1: Wildfires in 2021.

Target variable

- We collect data on Spanish forest fires, at the provincial level for all months 2002-2015, with a total of 216,538 events.
- The dependent variable of interests, y_{ijk}, counts the number of wildfires in year i, month j and province k.
- We have $D = IJK = 8400$ domains, defined by the crossings of years ($I = 14$), months ($J = 12$) and provinces ($K = 50$).
- There are 951 domains where no forest fire was recorded.

2002	2003	2004	2005	2006	2007	2008
107	112	63	51	90	52	60
2009	2010	2011	2012	2013	2014	2015
47	89	44	55	60	77	44

Table 1: Number of zeros per year.

Target variable

Jan.	Feb.	March	April	May	June
177	112	52	40	22	20
July	Aug.	Sept.	Oct.	Nov.	Dec.
5	6	9	59	206	243

Table 2: Number of zeros per month.

Jan.	Feb.	March	April	May	June
6372	19270	30432	16701	11844	18494
July	Aug.	Sept.	Oct.	Nov.	Dec.
28912	37742	26124	12033	4343	4271

Table 3: Number of wildfires per month.

Auxiliary variables

- *year3* is a explanatory variable, with three categories indicating groups of years.
- *year3.1* corresponds to the years 2002-2006. It contains 3000 domains and 423 zeros (14.1%).
  - The average $(50 \times 3 \times 5)$ of total burned areas in July-September was 668.51 Ha, with a total of 48,518 active fires.
- *year3.2* is for the years 2007-2012. It contains 3600 domains and 347 zeros were observed (9.64%).
  - The average of total burned areas in July-September was 333.56 Ha, with a total of 30,020 active fires.
- *year3.3* is for years 2013-2015. It contains 1800 domains, such that 181 zeros were observed (10.06%)
  - The average of total burned areas in July-September was 265.48 Ha, with a total of 14,240 active fires.

Auxiliary variables

Variable	Description	Units
e	average vapor pressure	tenths of hPa
hr	average relative humidity	tenths of mm
n.fog	foggy days	% days
n.gra	hail days	% days
n.llu	rainy days	% days
n.nie	snowy days	% days
np.001	precipitation $\geq$0.1mm	% days
np.010	precipitation $\geq$1mm	% days
np.300	precipitation $\geq$30mm	% days
nt.00	min. temperature $\leq$0°C	% days
nt.30	max. temperature $\geq$30°C	% days
n.tor	storm days	% days
nw.55	wind speed $\geq$55km/h	% days
nw.91	wind speed $\geq$91km/h	% days
p.max	max, daily rainfall	mm
p.mes	total precipitation	mm

Table: Description of the auxiliary variables.

Auxiliary variables

Variable	Description	Units
q.mar	mean sea-level pressure	KPa
q.max	max. absolute pressure	KPa
q.med	average pressure	KPa
q.min	max. min. pressure	KPa
ta.max	absolute max. temperature	°C
ta.min	absolute min. temperature	°C
ti.max	lowest max. temperature	°C
tm.max	average max. temperature	°C
tm.mes	average temperature	°C
tm.min	average min. temperature	°C
ts.min	highest min. temperature	°C
w.med	average speed elaborated from 07, 13, 18 UTC	km/h
unemp	unemployment rate	%
year3	year group variable	–

Table: Description of the auxiliary variables.

Variables

Target variables

- y_{ijk} is a count variable taking values on $\mathbb{N} \cup \{0\}$, $i \in \mathbb{I} = \{1, \ldots, I\}$, $j \in \mathbb{J} = \{1, \ldots, J\}$, $k \in \mathbb{K} = \{1, \ldots, K\}$.
- y_{ijk} is the number of wildfires in year i, month j and province k.
- z_{ijk} is a zero-inflation latent (non observable) variable.
- $D = IJK$ is the total number of domains.

Explanatory variables

- $\boldsymbol{x}_{1,ijk} = (x_{1,ijk1}, \ldots, x_{1,ijkq_1})$ and $\boldsymbol{x}_{2,ijk} = (x_{2,ijk1}, \ldots, x_{2,ijkq_2})$ are $1 \times q_1$ and $1 \times q_2$ row vectors with explanatory variables.

$$\boldsymbol{y}_{jk} = \underset{1 \leq i \leq I}{\text{col}}(y_{ijk}), \boldsymbol{z}_{jk} = \underset{1 \leq i \leq I}{\text{col}}(z_{ijk}), \boldsymbol{y} = \underset{1 \leq j \leq J}{\text{col}}(\underset{1 \leq k \leq K}{\text{col}}(\boldsymbol{y}_{jk})), \boldsymbol{z} = \underset{1 \leq j \leq J}{\text{col}}(\underset{1 \leq k \leq K}{\text{col}}(\boldsymbol{z}_{jk}))$$

$$\boldsymbol{X}_{a,jk} = \underset{1 \leq k \leq K}{\text{col}}(\boldsymbol{x}_{a,ijk}), \quad \boldsymbol{X}_a = \underset{1 \leq j \leq J}{\text{col}}(\underset{1 \leq k \leq K}{\text{col}}(\boldsymbol{X}_{a,jk})), \quad a = 1, 2.$$

Random effects

- $u_{1,j},\ u_{1,k},\ u_{2,j},\ u_{2,k},\ j \in \mathbb{J},\ k \in \mathbb{K}$, independent and $N(0,1)$.
- Define the vector $\boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_2')'$, where

$$\begin{aligned}
\boldsymbol{u}_{1,jk} &= (u_{1,j}, u_{1,k})',\ \boldsymbol{u}_{2,jk} = (u_{2,j}, u_{2,k})',\ \boldsymbol{u}_{jk} = (\boldsymbol{u}_{1,jk}', \boldsymbol{u}_{2,kk}')', \\
\boldsymbol{u}_1 &= \underset{1\le j\le J}{\text{col}}\left(\underset{1\le k\le K}{\text{col}}(\boldsymbol{u}_{1,jk})\right) \sim N_{2JK}(0, \boldsymbol{I}), \\
\boldsymbol{u}_2 &= \underset{1\le j\le J}{\text{col}}\left(\underset{1\le k\le K}{\text{col}}(\boldsymbol{u}_{2,jk})\right) \sim N_{2JK}(0, \boldsymbol{I}),\ \boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_2')'.
\end{aligned}$$

The AZINB11 model (NB2 parameterization)

- $(z_{ijk}, y_{ijk}) \sim AZINB11,\ i \in \mathbb{I},\ j \in \mathbb{J},\ k \in \mathbb{K}$, if

$$z_{ijk} \sim \text{Bern}(p_{ijk}),\ \ P(y_{ijk} = 0 / z_{ijk} = 1) = 1,\ \ y_{ijk}|_{z_{ijk}=0} \sim NB(r, \mu_{ijk}),\ \text{i.e.}$$

$$P(y_{ijk} = t / z_{ijk} = 0) = \frac{\Gamma(t+r)}{\Gamma(t+1)\Gamma(r)} \left(\frac{\mu_{ijk}}{r+\mu_{ijk}}\right)^t \left(\frac{r}{r+\mu_{ijk}}\right)^r, t \in \mathbb{N}\cup\{0\},$$

where $p_{ijk} \in (0,1),\ r > 0,\ \mu_{ijk} > 0$.

The model

To complete the definition of the AZINB11 model, we assume

- The link functions are

$$\begin{aligned} \text{logit}(p_{ijk}) &= \log \frac{p_{ijk}}{1-p_{ijk}} = \boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}, \\ \log(\mu_{ijk}) &= \boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_{21}u_{2,j} + \phi_{22}u_{2,k}. \end{aligned}$$

- Conversely, for $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, we have

$$\begin{aligned} p_{ijk} &= \frac{\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}\}}, \\ \mu_{ijk} &= \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_{21}u_{2,j} + \phi_{22}u_{2,k}\}. \end{aligned}$$

- The vectors $(y_{ijk}, z_{ijk})'$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, are independent conditioned to $\boldsymbol{u}$.

The likelihood

Model parameters

- $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1', \phi_{11}, \phi_{12})'$, $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2', \phi_{21}, \phi_{22})'$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$,

Conditioned likelihood factors

$$
\begin{aligned}
P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) \quad = \quad & \left(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1+\phi_{11}u_{1,j}+\phi_{12}u_{1,k}\}\right)^{-1} \\
\cdot \quad & \left\{\xi_{ijk}\left[\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1+\phi_{11}u_{1,j}+\phi_{12}u_{1,k}\}\right.\right. \\
+ \quad & \left.\left.\exp\left\{r\log r - r\log\left(r+\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2+\phi_{21}u_{2,j}+\phi_{22}u_{2,k}\}\right)\right\}\right] \\
+ \quad & (1-\xi_{ijk})\exp\left\{y_{ijk}(\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2+\phi_{21}u_{2,j}+\phi_{22}u_{2,k})\right. \\
- \quad & (y_{ijk}+r)\log\left(r+\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2+\phi_{21}u_{2,j}+\phi_{22}u_{2,k}\}\right) \\
+ \quad & \left.r\log r+\log\frac{\Gamma(y_{ijk}+r)}{\Gamma(y_{ijk}+1)\Gamma(r)}\right\}.
\end{aligned}
$$

The likelihood

Conditioned likelihood

$$P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) = \prod_{j=1}^{J}\prod_{k=1}^{K} P(\boldsymbol{y}_{jk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}), \quad P(\boldsymbol{y}_{jk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) = \prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}).$$

Marginal likelihood

$$P(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{j=1}^{J}\prod_{k=1}^{K} \int_{\mathbb{R}^4} \prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) f_{N_4(0,I)}(\boldsymbol{u}_{jk})\, d\boldsymbol{u}_{jk}.$$

Log-likelihood function

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{j=1}^{J}\sum_{k=1}^{K} \log \int_{\mathbb{R}^4} \prod_{i=1}^{I} P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) f_{N_4(0,I)}(\boldsymbol{u}_{jk})\, d\boldsymbol{u}_{jk}.$$

ML estimators

- Maximum likelihood (ML) estimators

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\, \ell(\boldsymbol{\theta}; \boldsymbol{y}), \quad \Theta = \mathbb{R}^{q_1+q_2} \times \mathbb{R}_+^4, \quad \mathbb{R}_+ = (0, \infty).$$

- As $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ contains integrals in $\mathbb{R}^4$, we apply a Laplace approximation algorithm to
  - maximize $\ell(\boldsymbol{\theta}; \boldsymbol{y})$
  - calculate the ML estimators of the model parameters: $\hat{\boldsymbol{\theta}}$,
  - obtain mode predictors of random effects: $\hat{\boldsymbol{u}}$.

Predictors

Prediction

- We are interested in predicting expected counts

$$\begin{aligned}
\mu_{yijk} &= E[y_{ijk}|\boldsymbol{u}_{jk}] = (1 - p_{ijk}(\boldsymbol{\theta}_1, \boldsymbol{u}_{1,jk}))\mu_{ijk}(\boldsymbol{\theta}_2, \boldsymbol{u}_{2,jk}) \\
p_{ijk}(\boldsymbol{\theta}_1, \boldsymbol{u}_{1,jk})) &= \frac{\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}\}}{1 + \exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_{11}u_{1,j} + \phi_{12}u_{1,k}\}}, \\
\mu_{ijk}(\boldsymbol{\theta}_2, \boldsymbol{u}_{2,jk}) &= \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_{21}u_{2,j} + \phi_{22}u_{2,k}\}.
\end{aligned}$$

- The plug-in predictor of μ_{yijk} is

$$\hat{\mu}^{in}_{yijk} = (1 - p_{ijk}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{u}}_{1,jk}))\mu_{ijk}(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{u}}_{2,jk})$$

- We are not interested in predicting expected NB counts μ_{ijk} by using the EBP $E\left[\mu_{ijk}|\boldsymbol{y}\right]$.

1. Calculate the ML estimate $\hat{\boldsymbol{\theta}}$. Run **Algorithm B1**.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) Generate $u_{1,j}^{*(b)} \sim N(0,1)$, $u_{1,k}^{*(b)} \sim N(0,1)$, $u_{2,j}^{*(b)} \sim N(0,1)$ and $u_{2,k}^{*(b)} \sim N(0,1)$. Calculate

$$p_{ijk}^{*(b)} = \frac{\exp\left\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_{11}u_{1,j}^{*(b)} + \hat{\phi}_{12}u_{1,k}^{*(b)}\right\}}{\left(1 + \exp\left\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_{11}u_{1,j}^{*(b)} + \hat{\phi}_{12}u_{1,k}^{*(b)}\right\}\right)},$$

$$\mu_{ijk}^{*(b)} = \exp\left\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_{21}u_{2,j}^{*(b)} + \hat{\phi}_{22}u_{2,k}^{*(b)}\right\}.$$

   (b) Generate $z_{ijk}^{*(b)} \sim \text{Bern}(p_{ijk}^{*(b)})$. If $z_{ijk}^{*(b)} = 1$, do $y_{ijk}^{*(b)} = 0$. If $z_{ijk}^{*(b)} = 0$, generate $y_{ijk}^{*(b)} \sim NB(r, \mu_{ijk}^{*(b)})$.
   (c) Use $(y_{ijk}^{*(b)}, \boldsymbol{x}_{ijk})$ to calculate the ML estimate $\hat{\theta}_\ell^{*(b)}$.
3. Sort the values $\hat{R}_\ell^{*(b)} = D^{1/2}(\hat{\theta}_\ell^{*(b)} - \hat{\theta}_\ell)$, $b = 1, \ldots, B$, from smallest to largest: $\hat{R}_{\ell(1)}^* \leq \ldots \leq \hat{R}_{\ell(B)}^*$. A $(1-\alpha)\%$ basic percentile bootstrap confidence interval for θ_ℓ is

$$\left(\hat{\theta}_\ell - D^{-1/2}\hat{R}_{\ell(\lfloor(\alpha/2)B\rfloor)}^*,\ \hat{\theta}_\ell + D^{-1/2}\hat{R}_{\ell(\lfloor(1-\alpha/2)B\rfloor)}^*\right).$$

MSE estimators and prediction intervals

1. Calculate the ML estimate $\hat{\boldsymbol{\theta}}$. Run **Algorithm B2**.
2. Repeat B times $(b = 1, \ldots, B)$:
   1. Run steps (a) and (b) of Algorithm B1.
   2. For $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, calculate $\mu_{yijk}^{*(b)} = (1 - p_{ijk}^{*(b)})\mu_{ijk}^{*(b)}$.
   3. Use the bootstrap sample $(y_{ijk}^{*(b)}, \boldsymbol{x}_{ijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, to calculate the ML estimate $\hat{\boldsymbol{\theta}}^{*(b)}$ and the predictor $\hat{\mu}_{yijk}^{*(b)}$.
3. For $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, calculate

$$mse^*(\hat{\mu}_{yijk}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\mu}_{yijk}^{*(b)} - \mu_{yijk}^{*(b)})^2, rmse^*(\hat{\mu}_{yijk}) = \left(mse^*(\hat{\mu}_{yijk})\right)^{1/2},$$

$$rrmse^*(\hat{\mu}_{yijk}) = \frac{rmse^*(\hat{\mu}_{yijk})}{\hat{\mu}_{yijk}}, \bar{y}_{ijk}^* = \sum_{b=1}^{B}\frac{y_{ijk}^{*(b)}}{B}, \text{var}^*(y_{ijk}) = \frac{1}{B-1}\sum_{b=1}^{B}(y_{ijk}^{*(b)} - \bar{y}_{ijk}^*)^2,$$

$$PI_{ijk}^{\alpha} = \left(\hat{\mu}_{yijk} - z_{1-\alpha/2}\left(\text{var}^*(y_{ijk})\right)^{1/2}, \hat{\mu}_{yijk} + z_{1-\alpha/2}\left(\text{var}^*(y_{ijk})\right)^{1/2}\right).$$

Bernoulli regression parameters

Selected model (2002-2014)

- The Bernoulli (BE) mixed model has $q_1 = 6$ covariables.
  - $x_{1,1} =$ intercept, $x_{1,2} = hr$, $x_{1,3} = np.300$,
  - $x_{1,4} = ta.max$, $x_{1,5} = year3.2$, $x_{1,6} = year3.3$.

β	$\hat{\beta}$	SE	z-value	$P(>\|z\|)$	CI 95% (asymp.)	CI 95% (boot)
$\beta_{1,1}$	-14.353	2.716	-5.28	0.000	(-19.67, -9.03)	(-17.73, -11.21)
$\beta_{1,2}^{hr}$	0.192	0.028	6.93	0.000	(0.138, 0.246)	(0.158, 0.232)
$\beta_{1,3}^{np.300}$	0.143	0.048	3.00	0.003	(0.049, 0.236)	(0.070, 0.213)
$\beta_{1,4}^{ta.max}$	-0.132	0.040	-3.39	0.001	(-0.208, -0.056)	(-0.175, -0.094)
$\beta_{1,5}^{year3.2}$	-1.048	0.219	-4.79	0.000	(-1.477, -0.619)	(-1.367, -0.768)
$\beta_{1,6}^{year3.3}$	-0.754	0.271	-2.78	0.005	(-1.285, -0.223)	(-1.253, -0.311)

Table: ML regression parameters of the Bernoulli mixed model.

Bernoulli regression parameters

If the remaining variables are fixed, the negative signs of the ML estimates of the regression parameters show that an increase of

- maximum temperature recorded (*ta.max*),

contribute to reduce the number of absolute zeros.

On the other hand, an increase of

- humidity (*hr*),
- percentage of days with precipitations greater than 30mm (*np.300*).

increases the number of absolute zeros.

- The negative signs of *year3.2* and *year3.3* show that the number of absolute zeros has decreased from 2002-2006 to 2007-2012 and 2013-2016.

Bernoulli variance parameters

The estimates of the standard deviation parameters are

- $\hat{\phi}_{11} = 0.1035$ for the month random effects $u_{1,j}$,
- $\hat{\phi}_{12} = 1.8717$for the province random effects $u_{1,k}$.

The asymptotic and bootstrap 95% CIs are

- ϕ_{11}: (0.0037, 2.9343) and $(1.0621 \cdot 10^{-9}, 0.2972)$.
- ϕ_{12}: (1.4243, 2.4595) and (1.3193, 2.3050).
- No interval contains the zero: $u_{1,j}$ and $u_{1,k}$ are needed for modelling the mixture between 0 and the NB counts.

NB regression parameters

Selected model (2002-2014)

- The NB mixed model has $q_2 = 18$ covariables.
  - $x_{2,1} =$ intercept,
  - $x_{2,2} = e$, $x_{2,3} = hr$,
  - $x_{2,4} = n.llu$, $x_{2,5} = n.nie$,
  - $x_{2,6} = np.300$, $x_{2,7} = nt.00$,
  - $x_{2,8} = nw.55$, $x_{2,9} = nw.91$,
  - $x_{2,10} = q.mar$, $x_{2,11} = q.max$,
  - $x_{2,12} = q.min$, $x_{2,13} = ta.max$,
  - $x_{2,14} = ta.min$,
  - $x_{2,15} = tm.mes$,
  - $x_{2,16} = tm.min$,
  - $x_{2,17} = year3.2$,
  - $x_{2,18} = year3.3$.

β	$\hat{\beta}$	SE	z-value	$P(>\|z\|)$	CI 95% (asymp.)	CI 95% (boot)
$\beta_{2,1}$	-10.51	3.801	-2.764	0.0057	(-17.960, -3.058)	(-12.409, -8.227)
$\beta^{e}_{2,2}$	0.009	0.001	7.245	0.0000	(0.007, 0.012)	(0.009, 0.010)
$\beta^{hr}_{2,3}$	-0.045	0.003	-13.366	0.0000	(-0.051, -0.038)	(-0.046, -0.043)
$\beta^{n.llu}_{2,4}$	-0.014	0.001	-14.018	0.0000	(-0.016, -0.012)	(-0.015, -0.014)
$\beta^{n.nie}_{2,5}$	-0.032	0.003	-9.295	0.0000	(-0.039, -0.025)	(-0.034, -0.030)
$\beta^{np.300}_{2,6}$	-0.029	0.007	-4.070	0.0000	(-0.043, -0.015)	(-0.034 -0.024)
$\beta^{nt.00}_{2,7}$	0.016	0.001	14.686	0.0000	(0.014, 0.018)	(0.016, 0.017)
$\beta^{nw.55}_{2,8}$	0.018	0.002	10.821	0.0000	(0.015, 0.021)	(0.017, 0.019)
$\beta^{nw.91}_{2,9}$	-0.025	0.008	-2.986	0.0028	(-0.041, -0.008)	(-0.030, -0.020)
$\beta^{q.mar}_{2,10}$	0.146	0.038	3.873	0.0001	(0.072, 0.220)	(0.125, 0.165)
$\beta^{q.max}_{2,11}$	-0.057	0.019	-2.973	0.0029	(-0.095, -0.019)	(-0.068, -0.046)
$\beta^{q.min}_{2,12}$	0.049	0.019	2.601	0.0093	(0.012, 0.086)	(0.039, 0.060)
$\beta^{ta.max}_{2,13}$	0.053	0.006	8.933	0.0000	(0.041, 0.064)	(0.050, 0.056)
$\beta^{ta.min}_{2,14}$	0.025	0.007	3.407	0.0007	(0.011, 0.040)	(0.021, 0.030)
$\beta^{tm.mes}_{2,15}$	0.064	0.023	2.823	0.0048	(0.020, 0.109)	(0.055, 0.074)
$\beta^{tm.min}_{2,16}$	-0.150	0.021	-7.215	0.0000	(-0.191, -0.109)	(-0.158, -0.140)
$\beta^{year3.2}_{2,17}$	-0.180	0.021	-8.549	0.0000	(-0.222, -0.139)	(-0.189, -0.171)
$\beta^{year3.3}_{2,18}$	-0.272	0.030	-9.016	0.0000	(-0.331, -0.212)	(-0.285, -0.257

Table: ML regression parameters of the NB mixed model.

NB regression parameters

If the remaining variables are fixed, the negative signs of the ML estimates of the regression parameters show that an increase of

- humidity (*hr*),
- number of general rainy days (*n.llu*),
- number of on snowy days (*n.nie*)
- number of days whose precipitation is higher than 300mm (*np.300*),
- number of days with wind speed greater than or equal to 91km/h (*nw.91*),
- maximum absolute pressure (*q.max*),
- average minimum temperature (*tm.min*)

contribute to reduce the number of wildfires.

- The negative signs of *year3.2* and *year3.3* show that the count of fires has decreased from 2002-2006 to 2007-2012 and 2013-2016.

NB regression parameters

On the other hand, an increase of

- the average vapor pressure (e),
- days with minimum temperatures below 0°C ($nt.00$),
- days with wind speed greater than or equal to 55km/h ($nw.55$),
- mean pressure at sea level ($q.mar$),
- maximum minimum pressure ($q.min$),
- absolute maximum temperature ($ta.max$),
- average temperature($tm.mes$),
- absolute minimum temperatures ($ta.min$),

increases the number of forest fires.

NB variance and dispersion parameters

The estimates of the standard deviation parameters are

- $\hat{\phi}_{21} = 0.3426$ for the month random effects $u_{2,j}$,
- $\hat{\phi}_{22} = 1.1555$ for the province random effects $u_{2,k}$.

The asymptotic and bootstrap 95% CIs are

- ϕ_{21}: (0.2264, 0.5187) and (0.1765, 0.4511).
- ϕ_{22}: (0.9485, 1.4078) and (0.9140, 1.3703).
- No interval contains the zero: $u_{2,j}$ and $u_{2,k}$ are needed.

The estimate of the dispersion parameter $\gamma = 1/r$ is $\hat{\gamma} = 2.1514$, with an SE of 0.0214.
The asymptotic and bootstrap 95% CIs for γ are

- (2.0630, 2.2437) and (2.0069, 2.7547), respectively.

Residual analysis

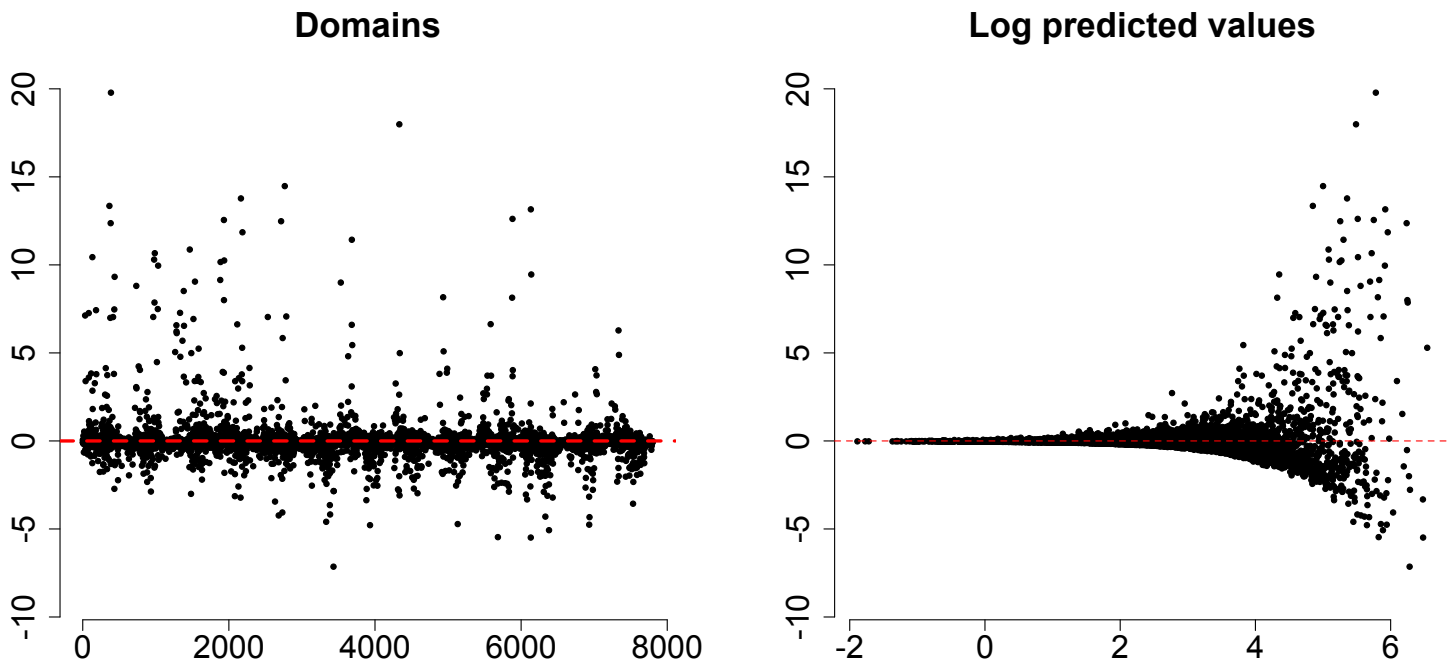


Figure: Standardized residuals vs domain indexes and log-predictions.

Residual analysis

Comments

- St.residuals fluctuate around zero, but there are more positive high residuals than negative ones.
- The asymmetry around zero is due to underprediction in provinces where the number of observed wildfires in summer is extremely high.
- The second plot also shows the mentioned asymmetry, which increases as the log-predictions grows.
- As the log-predictions increases, so does the variability of the residuals. This phenomenon is in agreement with the theoretical overdispersion of the BN distribution.

Residual analysis

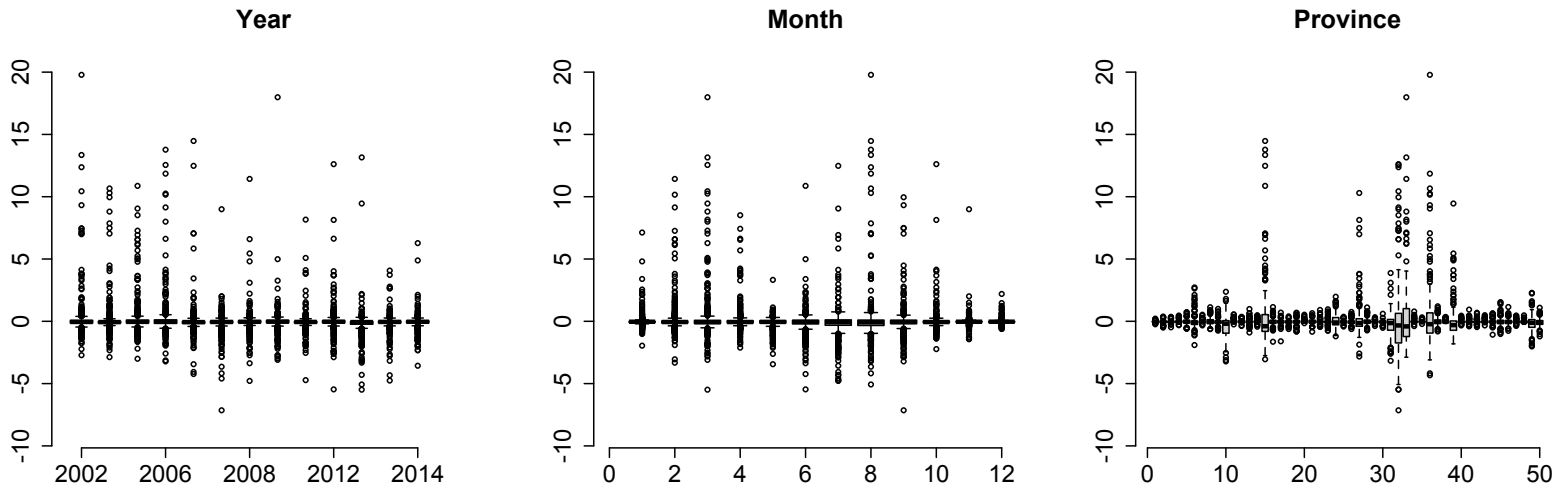


Figure: Boxplot of standardized residuals vs year, month and province.

Residual analysis

Comments

- Year does not seem to be discriminatory in the outliers's detection because no pattern is observed.
- Month seems to have more importance, with more atypical data at the end of winter and in summer.
- Province is a crucial factor in the outliers's detection.
- There are six provinces with absolute st.residuals greater than 3, with 82 observations (0.976% atypical data).
- The provinces are in the northwest of Spain: La Coruña (18), Lugo (5), Orense (22), Pontevedra (17), Asturias (14) and Cantabria (6).

Forecasting

- This section deals with the prediction of the number of forest fires in 2015 (near future).
- As prediction scenario, we assume the 2015 data.
  - $\boldsymbol{x}_{1,ijk}$ and $\boldsymbol{x}_{2,ijk}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, $i = 2015$.
  - $(\boldsymbol{\beta}_1', \phi_{11}, \phi_{12})'$, $(\boldsymbol{\beta}_2', \phi_{21}, \phi_{22})'$, $\hat{\boldsymbol{u}}_{1,jk}$, $\hat{\boldsymbol{u}}_{2,jk}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$.
- Based on the fitted AZINB11 model, we give predictions $\hat{\mu}_{yljk}$ for a short future horizon (one year).
- As the observed counts of wildfires y_{ljk} for 2015 are available, we can also check how good the predictions are.

Forecasting

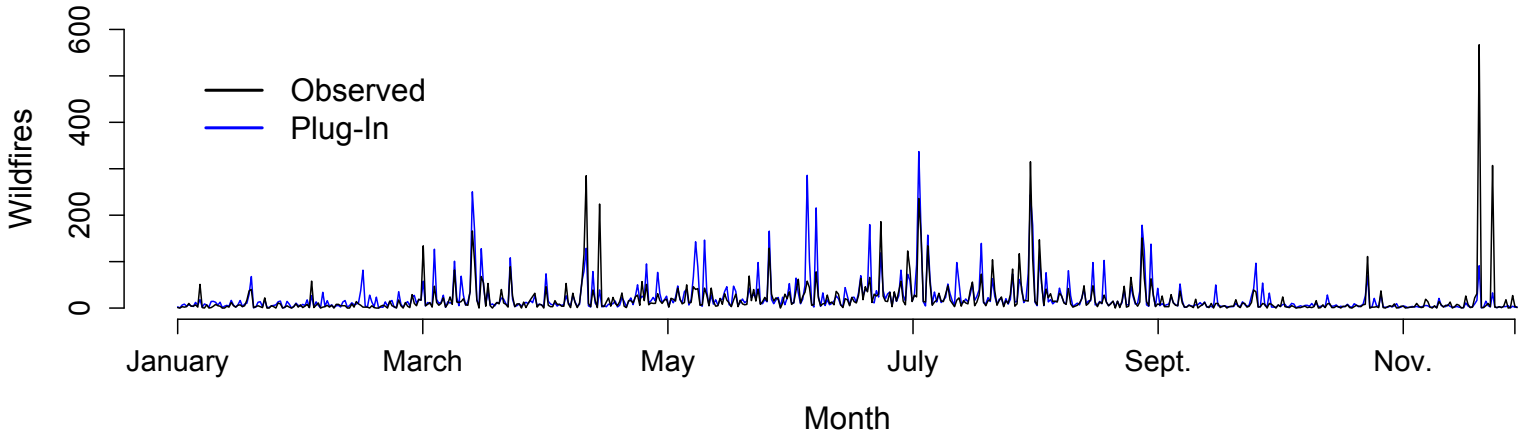
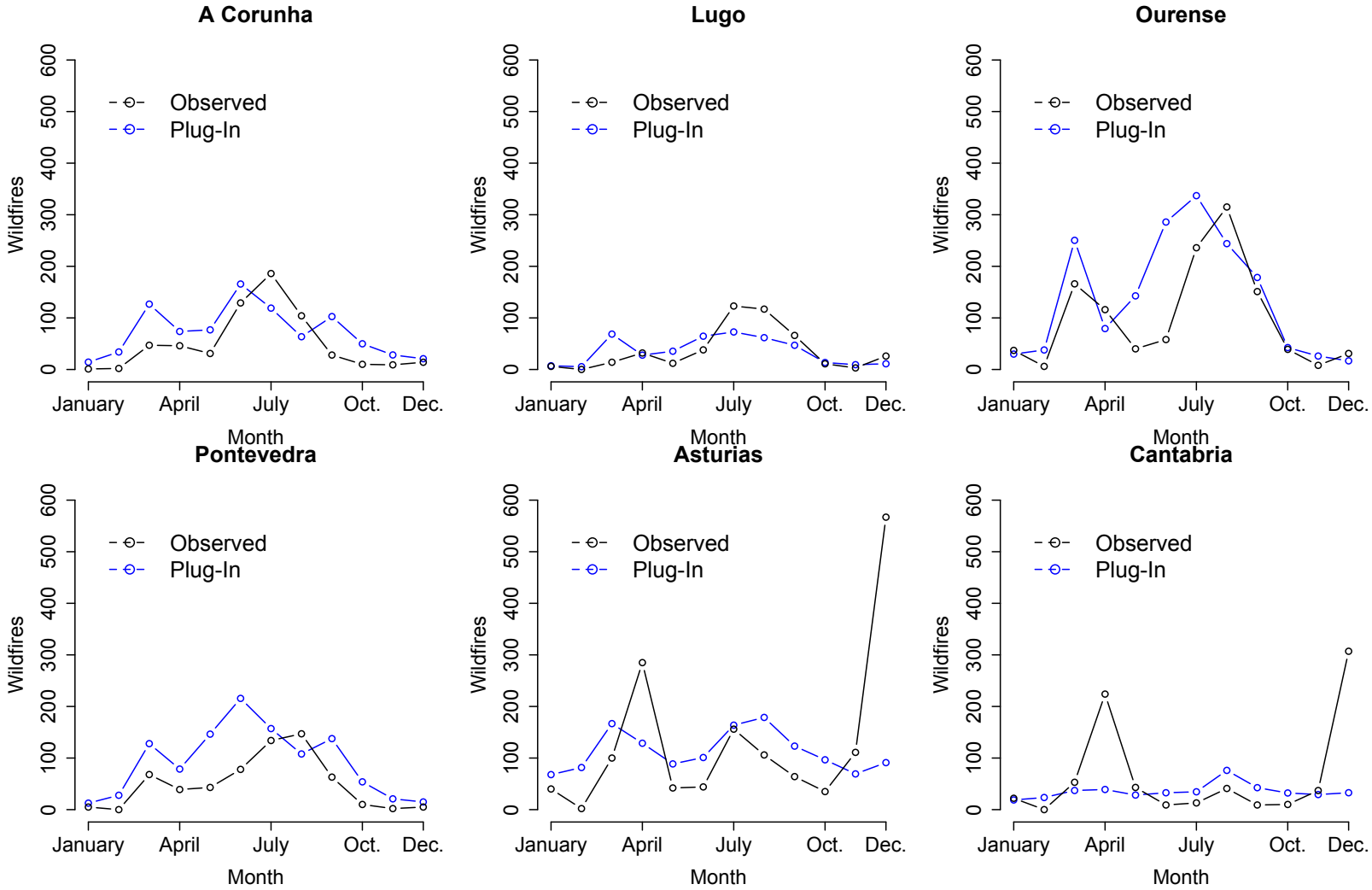


Figure: Wildfire forecasting for 2015 sorted by months and provinces (50 provinces per month).

- In general, predictions follows the trend of observations.
- Some predictions are far from observations in conflictive provinces (Asturias, Cantabria) during the summer. See next Figure.

Forecasting



Forecasting

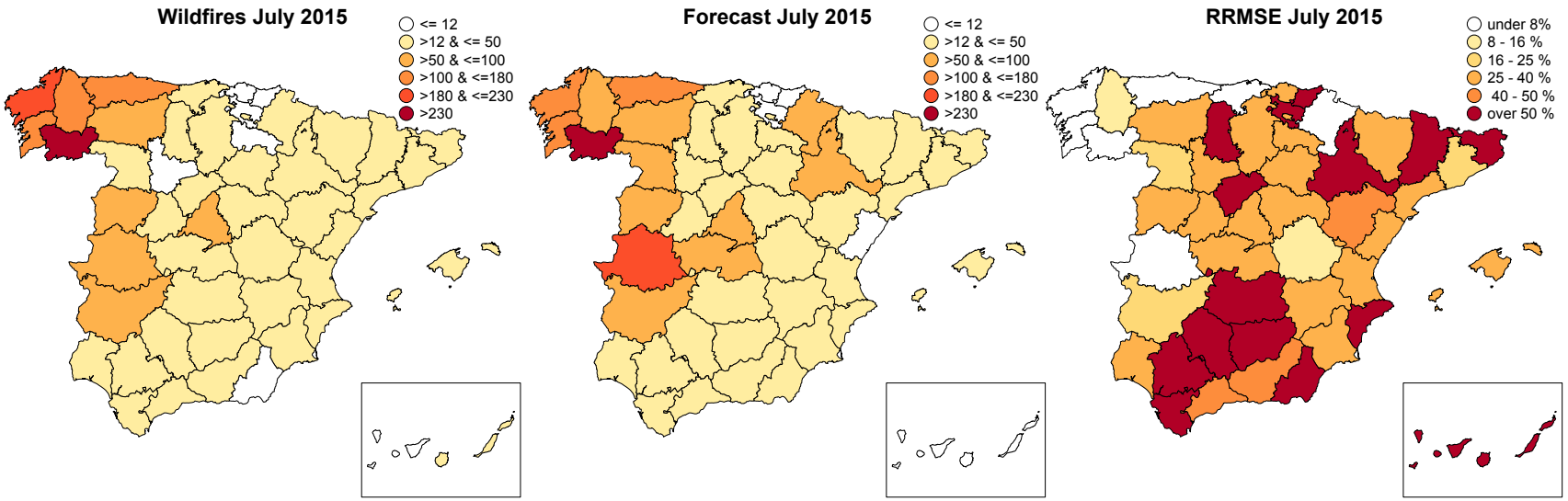


Figure: Observed (*left*) and predicted (*center*) wildfires and RRMSE (*right*) in July 2015.

Forecasting

Relative squared prediction errors (RSPE) for provinces in 2015

$$RSPE_{I.k} = \frac{\sqrt{\sum_{j=1}^{12}(y_{Ijk} - \hat{\mu}_{Ijk}^{in})^2}}{\sum_{j=1}^{12} y_{Ijk}}, \quad I = 14$$

Coverage probabilities for provinces in 2015 ($I = 14$)

$$C_{I.k}^{\alpha} = \frac{1}{12}\sum_{j=1}^{12} C_{Ijk}^{\alpha}, \;\; C_{Ijk}^{\alpha} = I\left(y_{ijk} \in PI_{Ijk}^{\alpha}\right), \;\; k = 1, \ldots, K.$$

For months of 2015, the RSPEs and the coverage probabilities are

$$RSPE_{Ij.} = \frac{\sqrt{\sum_{k=1}^{50}(y_{Ijk} - \hat{\mu}_{Ijk}^{in})^2}}{\sum_{k=1}^{50} y_{Ijk}}, \;\; C_{Ij.}^{\alpha} = \frac{1}{50}\sum_{k=1}^{K} C_{Ijk}^{\alpha}, \;\; C_{Ijk}^{\alpha} = I\left(y_{Ijk} \in PI_{Ijk}^{\alpha}\right),$$

Forecasting

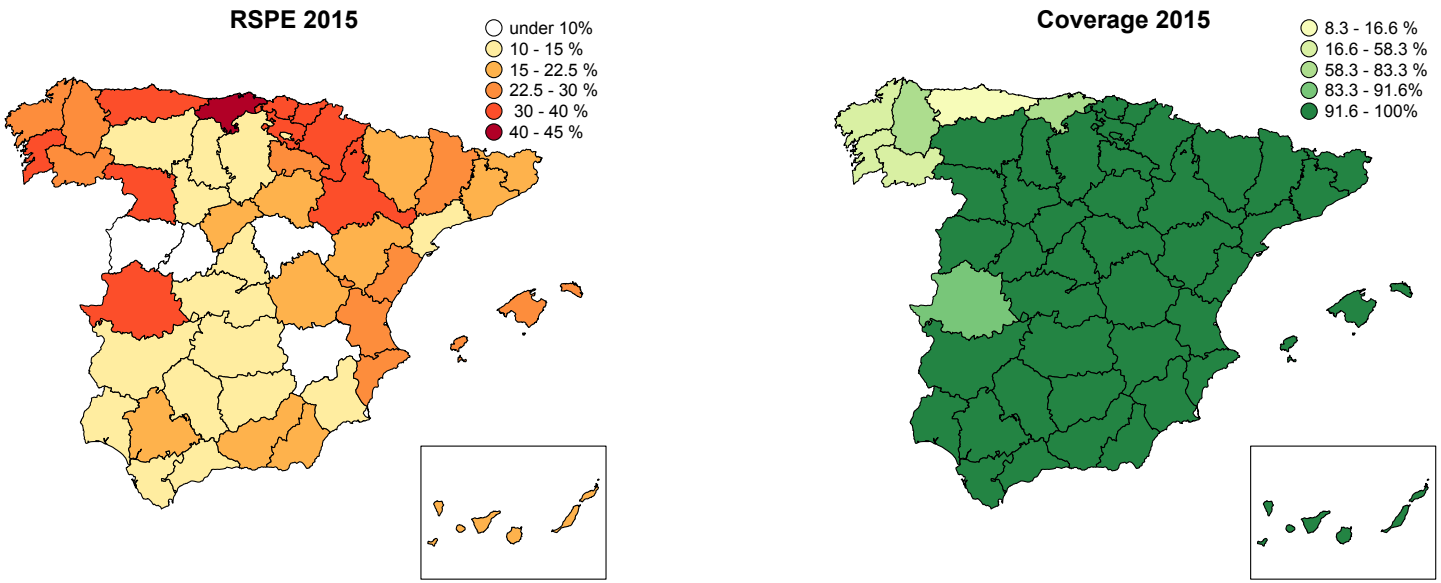


Figure: RSPE and coverage probabilities for each province in 2015, both in %.

Forecasting

	Jan.	Feb.	March	April	May	June
$RSPE_{lj.}$	16.96	20.30	14.41	20.57	18.43	24.32
$C^{\alpha}_{lj.}$	94	88	88	90	92	94
	July	Aug.	Sept.	Oct.	Nov.	Dec.
$RSPE_{lj.}$	10.96	9.69	14.94	27.61	19.03	46.86
$C^{\alpha}_{lj.}$	98	94	92	94	92	94

Table: Monthly RSPEs and coverage probabilities for 2015, both in %.

Conclusions

- Zero-inflated negative binomial mixed models are flexible tools to
  - describe the behavior of the number of fires in Spanish provinces and months during the period 2002-2014.
  - to predict the number of fires in 2015.
- We calculated the ML estimators of the model parameters and the mode predictors of the random effects by applying the Laplace approximation algorithm.
- We introduced plug-in predictors of number of fires and parametric bootstrap estimators of their MSEs.
- For the out-of-sample data (2015), we also gave prediction intervals.

Data and R codes

- Fire data: General Forest Fire Statistics (EGIF) of the Spanish Government.

  https://www.miteco.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/informacion-disponible/incendios-forestales.html
- Meteorological data: Spanish Meteorological Agency (AEMET)

  https://www.aemet.es/es/datos_abiertos
- Labour market data: Spanish Statistical Office (INE).

  https://www.ine.es/prodyser/microdatos.htm
- R codes: Downloadable from the repository

  https://github.com/mbugallo/aZINB11Fires
- R package glmmTMB: To fit the aZINB11 model.

Thank you

Thank you

for your attention

M. Bugallo, M.D. Esteban, M.F. Marey-Pérez, D. Morales (2023).
Wild fire prediction using zero-inflated negative binomial mixed models: Application to Spain.
Journal of Environmental Management 328, 116788, 1-16.

Appendix

- $h : \mathbb{R}^m \mapsto \mathbb{R}$ is a twice continuously differentiable function with a global maximum at the column vector $\boldsymbol{x}_0$, i.e.
- $\dot{h}(\boldsymbol{x}_0) = \left.\frac{dh}{d\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} = 0$ and $\ddot{h}(\boldsymbol{x}_0) = \left.\frac{d^2h}{d\boldsymbol{x}^2}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}$ is negative definite.
- A Taylor series expansion of $h(\boldsymbol{x})$ around $\boldsymbol{x}_0$ yields to

$$\begin{aligned} h(\boldsymbol{x}) &\approx h(\boldsymbol{x}_0) + \dot{h}'(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)'\ddot{h}(x_0)(\boldsymbol{x}-\boldsymbol{x}_0) \\ &= h(\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)'\ddot{h}(x_0)(\boldsymbol{x}-\boldsymbol{x}_0). \end{aligned}$$

- The multivariate Laplace approximation is

$$\begin{aligned} \int_{\mathbb{R}^m} e^{h(\boldsymbol{x})}\, d\boldsymbol{x} &\approx \int_{\mathbb{R}^m} e^{h(\boldsymbol{x_0})} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)'\left(-\ddot{h}(\boldsymbol{x}_0)\right)(\boldsymbol{x}-\boldsymbol{x}_0)\right\} d\boldsymbol{x} \\ &= (2\pi)^{m/2}\left|-\ddot{h}(\boldsymbol{x}_0)\right|^{-1/2} e^{h(\boldsymbol{x_0})}, \end{aligned} \tag{1}$$

Appendix

The likelihood of the AZINB11 model is

$$P(\boldsymbol{y};\boldsymbol{\theta}) = \int_{\mathbb{R}^{4JK}} P(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} = \int_{\mathbb{R}^{4JK}} \exp\left\{h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta})\right\} d\boldsymbol{u}, \tag{2}$$

where

$$\begin{aligned} h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \log P(y_{ijk}|\boldsymbol{u}_{jk};\boldsymbol{\theta}) \\ &- \frac{4JK}{2}\log 2\pi - \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K}(u_{1,j}^2 + u_{1,k}^2 + u_{2,j}^2 + u_{2,k}^2). \end{aligned}$$

- To apply the Laplace approximation to the integral in (2), we maximize $h(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta})$ in $\boldsymbol{u}$, given $\boldsymbol{y}$ and $\boldsymbol{\theta}$.
- For simplicity, we write $h(\boldsymbol{u})$.

Appendix

- The Newton-Raphson updating equation is

$$\boldsymbol{u}^{(i+1)} = \boldsymbol{u}^{(i)} - \ddot{h}^{-1}(\boldsymbol{u}^{(i)})\, \dot{h}(\boldsymbol{u}^{(i)}). \tag{3}$$

- Let $\boldsymbol{u}^{\circ}$ be the argument of maxima of the function $h(\boldsymbol{u})$.
- It holds $\dot{\boldsymbol{h}}(\boldsymbol{u}^{\circ}) = 0$ and the matrix $\ddot{h}(\boldsymbol{u}^{\circ})$ is negative definite.
- The loglikelihood of the AZINB11 model can be approximated by

$$\log P(\boldsymbol{y}; \boldsymbol{\theta},) \approx 2JK \log 2\pi + h(\boldsymbol{u}^{\circ}) - \frac{1}{2} \log |-\ddot{h}(\boldsymbol{u}^{\circ})| \triangleq g(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}^{\circ}).$$

Appendix

- The following step is to maximize $g(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}^{\circ})$ in $\boldsymbol{\theta} \in \Theta$.
- For simplicity, we write $g(\boldsymbol{\theta})$.
- Let us define $M = \dim(\Theta) = q_1 + q_2 + 4$.
- Let $\dot{g}$ and $\ddot{g}$ denote the $M \times 1$ vector and the $M \times M$ matrix of first and second order partial derivatives of $g(\boldsymbol{\theta})$.
- The Newton-Raphson updating equation is

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \ddot{g}^{-1}(\boldsymbol{\theta}^{(i)})\, \dot{g}(\boldsymbol{\theta}^{(i)}). \tag{4}$$

- The final ML-Laplace approximation algorithm combines (3) and (4).

Appendix

1. Set the initial values $i = 0$, $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, $\varepsilon_3 > 0$, $\varepsilon_4 > 0$, $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(-1)} = \boldsymbol{\theta}^{(0)} + 1$, $\boldsymbol{u}^{(0)} = 0$, $\boldsymbol{u}^{(-1)} = 1$, where 0 and 1 are column vectors of zeros and ones respectively.
2. Until $\|\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{(i-1)}\|_2 < \varepsilon_1$, $\|\boldsymbol{u}^{(i)} - \boldsymbol{u}^{(i-1)}\|_2 < \varepsilon_2$, do
    1. Apply algorithm (3) with seeds $\boldsymbol{u}^{(i)}$, convergence tolerance ε_3 and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(i)}$ fixed. Output: $\boldsymbol{u}^{(i+1)}$.
    2. Apply algorithm (4) with seed $\boldsymbol{\theta}^{(i)}$, convergence tolerance ε_4 and $\boldsymbol{u} = \boldsymbol{u}^{(i+1)}$ fixed. Output: $\boldsymbol{\theta}^{(i+1)}$.
    3. $i \leftarrow i + 1$.
3. Output: $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(i)}$ and $\hat{\boldsymbol{u}} = \boldsymbol{u}^{(i)}$.