Bayesian models for misreported counts data: theoretical and applied issues

S. Arima
University of Salento

SISBAYES 2025, Padova 4-5 September 2025

Outline

Introduction

Data quality is emerging as an essential characteristic of all data driven processes.

- Data has been at the core of the COVID-19 response: daily case numbers, deaths, testing capacity and percentage of vaccinated persons;
- Data analysts play a significant role and they dominated the public scene (major policy decisions, such as lock downs, school closures and quarantine restrictions).

However, the reliability of statistical analysis strongly depends on the quality of the collected data.

Data sources: registry vs survey

Two main data sources:

- **survey data:**
  - + Accurately planned and designed;
  - + Data collected by experts (or at least, trained);
- - Expensive in terms of costs and time.
- **registry data:**
  - + Rich and based on the whole population;
  - + Continuously updated;
  - + Very low cost;
  - - Data collection is not accurate $\rightsquigarrow$ misreporting events

Data sources: registry vs survey

Two main data sources:

- **survey data:**
  - \+ Accurately planned and designed;
  - \+ Data collected by experts (or at least, trained);
- \- Expensive in terms of costs and time.
- **registry data:**
  - \+ Rich and based on the whole population;
  - \+ Continuously updated;
  - \+ Very low cost;
  - \- Data collection is not accurate $\rightsquigarrow$ misreporting events

Registry data

Data coming from official collection systems usually experience considerable **underreporting of events.**

Among the others:

- Death-birth registry (developing countries): it is common to miss the report of infants who die shortly after birth leading to underestimation of vital statistics, compromising the definition of appropriate government intervention policies and distribution of financial resources;
- Work and tax: It is very likely that the data about undeclared work are misreported with a substantial underestimation of the importance of the problem in the real life;
- Violence Against Women: police reports substantially underestimate the problem. WHO declares that 1 in 3 women have experienced a form of violence at least once in their lifetime.

Registry data

Data coming from official collection systems usually experience considerable **underreporting of events.**

Among the others:

- Death-birth registry (developing countries): it is common to miss the report of infants who die shortly after birth leading to underestimation of vital statistics, compromising the definition of appropriate government intervention policies and distribution of financial resources;
- Work and tax: It is very likely that the data about undeclared work are misreported with a substantial underestimation of the importance of the problem in the real life;
- Violence Against Women: police reports substantially underestimate the problem. WHO declares that 1 in 3 women have experienced a form of violence at least once in their lifetime.

Registry data

Data coming from official collection systems usually experience considerable **underreporting of events.**

Among the others:

- Death-birth registry (developing countries): it is common to miss the report of infants who die shortly after birth leading to underestimation of vital statistics, compromising the definition of appropriate government intervention policies and distribution of financial resources;
- Work and tax: It is very likely that the data about undeclared work are misreported with a substantial underestimation of the importance of the problem in the real life;
- Violence Against Women: police reports substantially underestimate the problem. WHO declares that 1 in 3 women have experienced a form of violence at least once in their lifetime.

Talk sources

- A Bayesian nonparametric approach to correct for underreporting in count data, Biostatistics, 2023, Vol. 25 (3) (joint work with S. Polettini, G. Pasculli, L. Gesualdo, F. Pesce, D.Procaccini);
- A zero-inflated Poisson spatial model with misreporting for wildfire occurrences in southern Italian municipalities,2025, Evironmetrics, Vol. 36(1) (joint work with A. Pollice and C. Calculli)

Talk sources

- A Bayesian nonparametric approach to correct for underreporting in count data, Biostatistics, 2023, Vol. 25 (3) (joint work with S. Polettini, G. Pasculli, L. Gesualdo, F. Pesce, D.Procaccini);
- A zero-inflated Poisson spatial model with misreporting for wildfire occurrences in southern Italian municipalities,2025, Evironmetrics, Vol. 36(1) (joint work with A. Pollice and C. Calculli)

A Bayesian nonparametric approach to correct for underreporting in count data

Biostatistics, 2023, **00**, 1–15
https://doi.org/10.1093/biostatistics/kxad027
Article

OXFORD

A Bayesian nonparametric approach to correct for underreporting in count data

Serena Arima[1,*], Silvia Polettini[2], Giuseppe Pasculli[3], Loreto Gesualdo[4], Francesco Pesce[5], Deni-Aldo Procaccini[6]

[1]Department of Human and Social Sciences, University of Salento, Via di Valesio, 73100, LECCE, Italy
[2]Department of Social and Economic Sciences, Sapienza University, P.le Aldo Moro, 5, 00185 ROMA, Italy
[3]Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University, Via Ariosto, 25, 00185 Roma RM, Italy
[4]Section of Nephrology, Department of Precision and Regenerative Medicine and Ionian Area (DiMePre-J), Azienda Ospedaliero Universitaria Consorziale Policlinico di Bari, Piazza Giulio Cesare, 11 - 70124 Bari, Italy
[5]Division of Renal Medicine, "Fatebenefratelli Isola Tiberina–Gemelli Isola", 00186 Rome, Italy
[6]Section of Nephrology, Department of Precision and Regenerative Medicine and Ionian Area (DiMePre-J), Azienda Ospedaliero Universitaria Consorziale Policlinico di Bari, Piazza Giulio Cesare, 11 - 70124 Bari, Italy

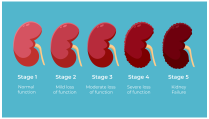
*To whom correspondence should be addressed: Serena Arima. Email: serena.arima@unisalento.it

SUMMARY

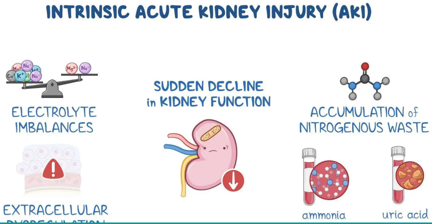
We propose a nonparametric compound Poisson model for underreported count data that introduces a latent clustering structure for the reporting probabilities. The latter are estimated with the model's parameters based on experts' opinion and exploiting a proxy for the reporting process. The proposed model is used to estimate the prevalence of chronic kidney disease in Apulia, Italy, based on a unique statistical database covering information on $m = 258$ municipalities obtained by integrating multisource register information. Accurate prevalence estimates are needed for monitoring, surveillance, and management

Motivating application

- CKD is the gradual loss of kidney function;
- The decrease in renal functionality is measured according to the glomerular filtration rate (GFR) (moderate, severe, or end-stage renal disease). Very low GFR is an indication for renal replacement therapies.



- It is a life-long disease: patients need continuous therapies (i.e. dialysis) and in the later stages transplantation is indicated;



Motivating application

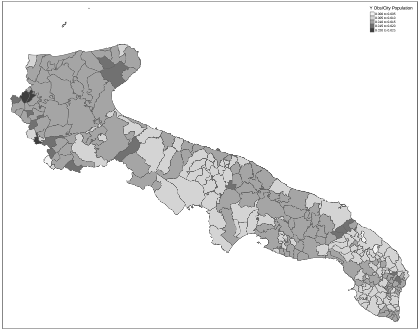
CKD is **chronic disease** $\rightsquigarrow$ strong social impact.

- Costs for the society are particularly severe: 1.8% of the total budget for health care in Italy was spent on CKD patients (2.5 billion euros);
- Interesting relationships with socio-economic factors:
  - the earlier a patient's access to nephrological care, the better the clinical, economic, and psychological outcome;
  - socio-economic status impacts on the access to appropriate medical care (e.g., renal/peritoneal dialysis, renal transplants) as well as distance from the closest health facility;
  - caregivers are heavily involved in terms of time, quality of life, and economic support.

Smart, N. A., Dieberg, G., Ladhani, M. and Titus, T. (2014). Early referral to specialist nephrology services for preventing the progression to end-stage kidney disease. Cochrane Database of Systematic Reviews 18, 1-92.

The motivating application: data souces

- Prevalence estimates come from ARES-Puglia (Regional Health Agency in Puglia): counts of patients living in Puglia affected by CKD for $m = 258$ Municipalities of Apulia region;
- counts come from registries:
  - patients enrolled in hospital or regularly registered in the system;
  - patients buying medicine with some doctor prescription (ticket and or discount for declaration of the disease);



The motivating application: data souces

However:

- experts suspect that crude counts are not trustworthy:
  - some patients (living far away, with other disease, economic/social issues) are not recorded;
  - some patients move to other regions for health care (health-care migration).

A model accounting for underreporting might improve estimates of CKD rates, thus allowing for an appropriate allocation of funds and healthcare facilities for CKD patients.

Use ad-hoc survey data/ auxiliary variables to **correct** underreported registry data thorugh auxiliary variables (proxy of accurateness of data recording)

The motivating application: data souces

However:

- experts suspect that crude counts are not trustworthy:
  - some patients (living far away, with other disease, economic/social issues) are not recorded;
  - some patients move to other regions for health care (health-care migration).

A model accounting for underreporting might improve estimates of CKD rates, thus allowing for an appropriate allocation of funds and healthcare facilities for CKD patients.

Use ad-hoc survey data/ auxiliary variables to **correct** underreported registry data thorugh auxiliary variables (proxy of accurateness of data recording)

Underreporting: notation and literature

Let the number of events of interest over a set of m areas, denoted by T_i.

$$\begin{aligned} T_i|\theta_i &\sim Poisson(E_i\theta_i) \\ \theta_i &\sim f(X_{i1}, X_{i2}, ..., X_{ip}) \end{aligned}$$

where for $i = 1, ..., m$

- θ_i is the event rate; E_i is the known number of exposed;
- $X_1, ..., X_p$ is a set of covariates.

However, in some or potentially all cases, we have not observed T_i but we observe

$$Y_i \leq T_i$$

We aim at estimating the event intensities θ_i, predict the true counts T_i based on underreported observations $Y_i < T_i$ for all $i = 1, ..., m$.

Underreporting: notation and literature

Let the number of events of interest over a set of m areas, denoted by T_i.

$$\begin{aligned} T_i|\theta_i &\sim Poisson(E_i\theta_i) \\ \theta_i &\sim f(X_{i1}, X_{i2}, ..., X_{ip}) \end{aligned}$$

where for $i = 1, ..., m$

- θ_i is the event rate; E_i is the known number of exposed;
- $X_1, ..., X_p$ is a set of covariates.

However, in some or potentially all cases, we have not observed T_i but we observe

$$Y_i \leq T_i$$

We aim at estimating the event intensities θ_i, predict the true counts T_i based on underreported observations $Y_i < T_i$ for all $i = 1, ..., m$.

Underreporting: Bailey et al. (2005)

The Poisson model has been extended to account for underreporting in various ways.

- Bailey and others (2005) extended the censored Poisson regression model in Caudill and Mixon (1995) assuming the counts of underreported areas are the lower bound for the true nonobserved counts (leprosy cases in the Brazilian region of Olinda)
  - Censored likelihoods $\rightsquigarrow$ does not account for the severity of the underreporting;
  - underlying counts at least some of the units must be completely observed;
  - a-priori knowledge of the censored areas (prior knowledge on the relationship between leprosy occurrence rate and a measure of social deprivation)

Bailey et al.(2005) Modeling of underdetection of cases in disease surveillance. Annals of Epidemiology 15, 335-343.

Caudill, S. B. and Mixon, F. G. (1995). Modeling household fertility decisions: estimation and testing of censored regression models for count data. Empirical Economics 20,183-196

Underreporting: Bailey et al. (2005)

The Poisson model has been extended to account for underreporting in various ways.

- Bailey and others (2005) extended the censored Poisson regression model in Caudill and Mixon (1995) assuming the counts of underreported areas are the lower bound for the true nonobserved counts (leprosy cases in the Brazilian region of Olinda)
  - Censored likelihoods $\rightsquigarrow$ does not account for the severity of the underreporting;
  - underlying counts at least some of the units must be completely observed;
  - a-priori knowledge of the censored areas (prior knowledge on the relationship between leprosy occurrence rate and a measure of social deprivation)

Bailey et al.(2005) Modeling of underdetection of cases in disease surveillance. Annals of Epidemiology 15, 335-343.

Caudill, S. B. and Mixon, F. G. (1995). Modeling household fertility decisions: estimation and testing of censored regression models for count data. Empirical Economics 20,183-196

Underreporting: CP model

Stoner et al. (2019) extended the model introduced by Winkelmann (1996) and proposed a Bayesian hierarchical Compound Poisson Model (CPM) to account for the underreporting in tuberculosis counts in Brazil:

- $T_i|\lambda_i \sim Poisson(\lambda_i)$ and the relative risk is $\theta_i = \lambda_i/E_i$;
- T_i is not fully observed:
  - $W_t \sim Bernoulli(\epsilon_i)$ $t = 1, ..., T_i$ where ϵ_i is the reporting probability;
  - $Y_i = \sum_{t=1}^{T_i} W_t$ has a CP distribution since

$$Y_i|T_i, \epsilon_i \sim Binomial(T_i, \epsilon_i) \qquad T_i|\theta_i \sim Poisson(E_i\theta_i)$$

and, marginalizing,

$$Y_i|\theta_i, \epsilon_i \sim Poisson(E_i\theta_i\epsilon_i)$$

O. Stoner et al. (2019) A hierarchical framework for correcting under-reporting in count data. JASA,114(528), 1481-1492.

Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. Empirical Economics 21, 575-587.

Underreporting: CP model

Stoner et al. (2019) extended the model introduced by Winkelmann (1996) and proposed a Bayesian hierarchical Compound Poisson Model (CPM) to account for the underreporting in tuberculosis counts in Brazil:

- $T_i|\lambda_i \sim Poisson(\lambda_i)$ and the relative risk is $\theta_i = \lambda_i/E_i$;
- T_i is not fully observed:
  - $W_t \sim Bernoulli(\epsilon_i)$ $t = 1, ..., T_i$ where ϵ_i is the reporting probability;
  - $Y_i = \sum_{t=1}^{T_i} W_t$ has a CP distribution since

$$Y_i|T_i, \epsilon_i \sim Binomial(T_i, \epsilon_i) \qquad T_i|\theta_i \sim Poisson(E_i\theta_i)$$

  and, marginalizing,

$$Y_i|\theta_i, \epsilon_i \sim Poisson(E_i\theta_i\epsilon_i)$$

O. Stoner et al. (2019) A hierarchical framework for correcting under-reporting in count data. JASA,114(528), 1481-1492.

Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. Empirical Economics 21, 575-587.

Underreporting: Stoner et al. (2019)

Stoner et al. (2019) extedend the model by

$$log(\theta_i) = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + v_i$$

$$logit(\epsilon_i) = \gamma_0 + \sum_{k=1}^{K} \gamma_j z_{ij} + u_i$$

where random effects may be added to account for spatial as well as residual variability.

CPM: identifiability issue

- If no further information is available, only the parameter $\eta_i = \theta_i \epsilon_i$ is identifiable.
- To guarantee the CPM identifiability, it is necessary to introduce external information on the reporting process.

Different solutions:

- Many papers (e.g. Whittemore and Gong (1991), Stamey, Young, and Boese (2006) and Dvorzak and Wagner (2015)) resort to a validation dataset on the reporting process.
- Moreno and Giron (1998) : directly model the uncertainty about ϵ_i using informative beta prior distributions
- Stoner et al. (2019): assure identifiability by
    i. X and Z different (not the orthogonality not necessary);
    ii. informative prior distribution for the mean reporting rate (e.g. β_0)

Moreno, E. and Giron J. (1998). Estimating with incomplete count data: A Bayesian approach. JSPI, 66(1), 147-159.

CPM: identifiability issue

- If no further information is available, only the parameter $\eta_i = \theta_i \epsilon_i$ is identifiable.
- To guarantee the CPM identifiability, it is necessary to introduce external information on the reporting process.

Different solutions:

- Many papers (e.g. Whittemore and Gong (1991), Stamey, Young, and Boese (2006) and Dvorzak and Wagner (2015)) resort to a validation dataset on the reporting process.
- Moreno and Giron (1998) : directly model the uncertainty about ϵ_i using informative beta prior distributions
- Stoner et al. (2019): assure identifiability by
    i. X and Z different (not the orthogonality not necessary);
    ii. informative prior distribution for the mean reporting rate (e.g. β_0)

Moreno, E. and Giron J. (1998). Estimating with incomplete count data: A Bayesian approach. JSPI, 66(1), 147-159.

CPM: identifiability issue

- If no further information is available, only the parameter $\eta_i = \theta_i \epsilon_i$ is identifiable.
- To guarantee the CPM identifiability, it is necessary to introduce external information on the reporting process.

Different solutions:

- Many papers (e.g. Whittemore and Gong (1991), Stamey, Young, and Boese (2006) and Dvorzak and Wagner (2015)) resort to a validation dataset on the reporting process.
- Moreno and Giron (1998) : directly model the uncertainty about ϵ_i using informative beta prior distributions
- Stoner et al. (2019): assure identifiability by
  1. X and Z different (not the orthogonality not necessary);
  2. informative prior distribution for the mean reporting rate (e.g. β_0)

Moreno, E. and Giron J. (1998). Estimating with incomplete count data: A Bayesian approach. JSPI, 66(1), 147-159.

CPM: identifiability issue

- If no further information is available, only the parameter $\eta_i = \theta_i \epsilon_i$ is identifiable.
- To guarantee the CPM identifiability, it is necessary to introduce external information on the reporting process.

Different solutions:

- Many papers (e.g. Whittemore and Gong (1991), Stamey, Young, and Boese (2006) and Dvorzak and Wagner (2015)) resort to a validation dataset on the reporting process.
- Moreno and Giron (1998) : directly model the uncertainty about ϵ_i using informative beta prior distributions
- Stoner et al. (2019): assure identifiability by
    i. X and Z different (not the orthogonality not necessary);
    ii. informative prior distribution for the mean reporting rate (e.g. β_0)

Moreno, E. and Giron J. (1998). Estimating with incomplete count data: A Bayesian approach. JSPI, 66(1), 147-159.

Underreporting: De Oliveira et al. (2022)

- de Oliveira et al.(2022) also refer to the compound Poisson model and propose to cluster areas according to ϵ:
  - m areas are grouped into K known data quality clusters, where $K \leq m$;
  - they define the clustering indicator $\mathrm{h_i} = (h_{1i}, ..., h_{Ki})^T$;
  - $\epsilon_i = (1 - h_i^T \gamma)$ and $\gamma = (\gamma_1, ..., \gamma_K) \in [0, 1)$ and $\sum_{j=1}^{K} \gamma_j \leq 1$.
- For each area, $\mathrm{h_i}$ must be a-priori specified and reflects the clustering structure.
  - $\mathrm{h_i} = (1, 0, 0, ..., 0)^T$ then the $i-$th area has the highest level of data quality. In such area the reporting probability is $\epsilon_i = 1 - \gamma_1$, larger than the other areas;
  - $\mathrm{h_i} = (1, 1, 1, ..., 1)^T$ then the $i-$th area lies in the worst data quality category an data in this region are recorded with a lower probability $\epsilon_i = 1 - \gamma_1 - \gamma_2 - ... - \gamma_K$.

de Oliveira, G.L. et al. (2022) "Bias Correction in Clustered Underreported Data." Bayesian Anal. 17 (1) 95 - 126, March 2022

Underreporting: De Oliveira et al. (2022)

- To be identifiable, it only requires information about the reporting probabilities in the best areas;
- ϵ_i is represented in terms of interpretable parameters (as in the previous slide);
- they derive Jeffreys prior for γ.

Underreporting: Our proposal

We propose to extend the model in de Olivera et al. (2022) in the following aspects:

- the number of clusters is not a-priori defined;
- areas are clustered (through a probabilistic mechanism) according to one or more proxy data-quality variable.

We propose a Bayesian nonparametric model based on a Dirichlet Process (DP) model on the underreporting probabilities.

Underreporting: Our proposal

$$Y_i|\theta_i, \epsilon_i \sim Poisson(E_i \theta_i \epsilon_i), \quad i = 1, \dots, m \tag{1a}$$

$$log(\theta_i) = \sum_{j=1}^{J} \beta_j x_{ij} + u_i + s_i \tag{1b}$$

$$u_i|\sigma_u^2 \sim N(0, \sigma_u^2) \tag{1c}$$

$$s_i|\sigma_s^2 \sim ICAR(\sigma_s^2) \tag{1d}$$

$$\epsilon_i|z, G_z \sim G_z, \tag{1e}$$

where $\epsilon_i|z$ are clustered according to a non parametric process involving covariates Z, considered as proxies of the data quality.

Underreporting: Our proposal

- Within the wide class of predictor-dependent stick-breaking priors, we rely on the probit stickbreaking process (PSBP) under which the mixing weights arise through a probit model on the covariate Z:

$$G_z(\cdot) = \sum_{j=1}^{\infty} \left(\Phi(\eta_j(z)) \prod_{l<j} (1 - \Phi(\eta_l(z))) \right) \delta_{\psi_j}(\cdot), \tag{2}$$

  where the weights are obtained through normally distributed random variables $\eta_j(z) = z'\gamma_j$ transformed to the unit interval via the standard normal cdf and $\delta_{\psi_j}(\cdot)$ denotes the Dirac measure at location ψ_j, and ψ_j $(j = 1, 2, \ldots)$ are independent draws from G_0, that do not depend on the covariates.

- In our approach variables Z are proxies for data quality.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. JASA 104, 1646-1660

Underreporting: Our proposal

- Within the wide class of predictor-dependent stick-breaking priors, we rely on the probit stickbreaking process (PSBP) under which the mixing weights arise through a probit model on the covariate Z:

$$G_z(\cdot) = \sum_{j=1}^{\infty} \left(\Phi(\eta_j(z)) \prod_{l<j} (1 - \Phi(\eta_l(z))) \right) \delta_{\psi_j}(\cdot), \qquad (2)$$

  where the weights are obtained through normally distributed random variables $\eta_j(z) = z'\gamma_j$ transformed to the unit interval via the standard normal cdf and $\delta_{\psi_j}(\cdot)$ denotes the Dirac measure at location ψ_j, and ψ_j $(j = 1, 2, \dots)$ are independent draws from G_0, that do not depend on the covariates.
- In our approach variables Z are proxies for data quality.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. JASA 104, 1646-1660

Our proposal: some details

- The sequence of atoms ψ_j is constructed by introducing a partial ordering on a sample from a baseline measure G_0 using the following mechanism: ψ_1 is drawn from $G_0^* = U[a_0, b_0]$ and $\psi_j \sim G_{0j}^*$, $j = 2, \dots$ with $G_{0j}^* = U[a_1, b_1]$ with $a_1 < b_1 < a_0$ to ensure $\psi_j < \psi_1$ for $j = 2, \dots$.
- The choice of a_0 and b_0 should reflect the prior information about the best reporting probability.
- De Oliveira et al. (2020) show that under a clusterization of the areas, eliciting a prior on the reporting probability in the best reported areas using experts' knowledge is sufficient for identification.

Other priors (standard approach):

- $u_i \sim N(0, \sigma_u^2)$; $\sigma_u^{-2} \sim Gamma(a_u, b_u)$;
- $\sigma_s^{-2} \sim Gamma(a_s, b_s)$; $\beta_j \sim N(0, \sigma_\beta)$ $j = 1, ..., p$; $\gamma_k \sim N(0, \sigma_\gamma)$ $k = 1, ..., K$;

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. Bayesian Analysis 6, 145-177.

Simulation study: take home message

We consider very different simulation settings:

- large proportion of areas have high reporting probabilities: de Oliveira (2022) and the proposed model perform similarly;
- small number of observation in the best quality: the proposed approach is more robust;
- weak the proxy: our proposal has a certain robustness, provided the proxy is at least moderately correlated with the original one;
- decreasing reporting probabilities decrease all models severely deviate from the reference model

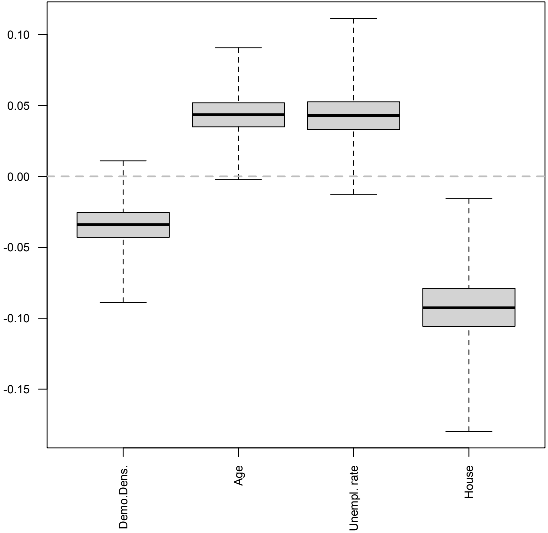
CKD data: results

For each municipality, we collect information auxiliary information:

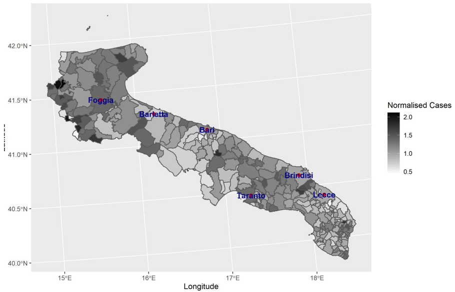
- Demographic density;
- Aging Index;
- Unemployment rate;
- % of house ownership.

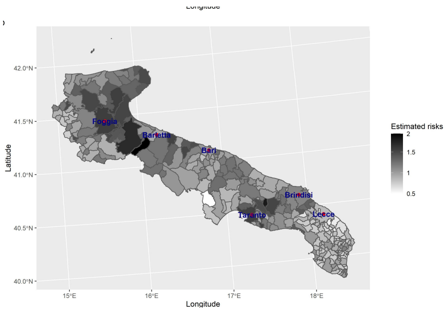
Building on epidemiologial literature, difficulties to access healthcare facilities are considered as barriers to detect and notify events to the surveillance, we computed a proxy (Z) measuring the minimum travel distance from the center of each area to the closest specialized health facility using the GoogleMaps distance calculator tool.

CKD data: posterior estimates of β



CKD data: Observed vs Predicted risks





Conclusion

- Registry data: precious source of information;
- Correct underrreporting of registry data with ad-hoc quality proxies;
- Compound Poisson model in a Bayesian non-parametric framework;
- We believe our proposal can be considered a flexible alternative, in that it supplies the least amount of prior information about the reporting process.

A zero-inflated Poisson spatial model with misreporting for wildfire occurrences in southern Italian municipalities, 2025, Evironmetrics,

Environmetrics
The official journal of The International Environmetrics Society
An Association of The International Statistical Institute

RESEARCH ARTICLE

A zero-inflated Poisson spatial model with misreporting for wildfire occurrences in southern Italian municipalities

Serena Arima, Crescenza Calculli, Alessio Pollice
First published: 02 May 2024

https://doi.org/10.1002/env.2853

Abstract

We propose a Poisson model for zero-inflated spatial counts contaminated by measurement error: we accommodate the excess of zeroes in the counts, consider the possible under/over reporting of the response and account for the neighboring structure of spatial areal units. Bayesian inferences are provided by MCMC implementation through the R package NIMBLE. To evaluate the model performance, a simulation study is carried out under configurations that allow for structured and unstructured spatial random effects. The proposed model is applied to investigate the distribution of the counts of wildfire occurrences in the municipal areas of two neighboring Italian regions for the summer season 2021. Fire counts are obtained by processing MODIS satellite data, while several socio-economic and environmental-driven potential risk factors are also considered in the model formulation. Data from multiple sources with different spatial support are processed in order to comply with the municipal units. Results suggest the appropriateness of the approach and provide some insights on the features of wildfire occurrences.

Wildfires in Southern Italy

INCENDI IN PUGLIA

Dalla Capitanata al Salento, continua l'emergenza incendi in Puglia: decine di ettari distrutti

Secondo Coldiretti Puglia nell'estate di quest'anno gli incendi nella regione sono triplicati con danni, per milioni di euro, all'ambiente, all'economia, al lavoro e al turismo

6 Agosto 2021 10:22

INCENDI

Incendi, dopo 4 giorni ancora fiamme nel bosco di Gravina in Puglia: il fuoco riprende vigore, oltre 1.000 ettari in fumo – FOTO

L'incendio che da qualche giorno interessa il bosco Difesa Grande a Gravina in Puglia "ha ripreso vigore, spinto dal vento": oltre 1.000 ettari distrutti

1 Agosto 2021 19:48

INCENDI IN PUGLIA

Incendi in Puglia: fiamme in provincia Lecce, distrutti molti ulivi

Incendi in Puglia: centinaia di alberi d'ulivo avvolti dalle fiamme in provincia di Lecce

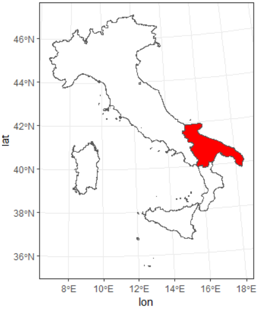
2 Luglio 2021 09:56

INCENDI IN BASILICATA

Incendi in Basilicata: fiamme a Maratea nella zona del Cristo Redentore, case minacciate

Vasto incendio a Maratea, in prossimità del Cristo Redentore: il fuoco ha lambito diverse abitazioni

30 Giugno 2021 14:50



⇈ occurrence and magnitude of wildfire

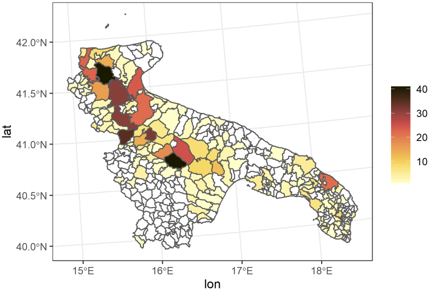
- fire-prone regions
- raising temperatures and prolonged drought periods (global trend)
- agricultural practices and land use
- social and cultural heritage (population characteristics, deprivation)

Active fires from remote sensing

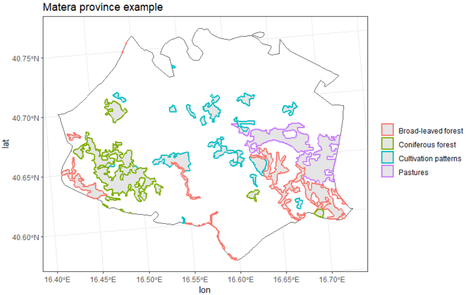
- Moderate Resolution Imaging Spectroradiometer (**MODIS** Collection 6 land product)
- 1 km spatial resolution

Fire counts in the study area

- municipality level (388 administrative units)
- summer season (june-august 2021)



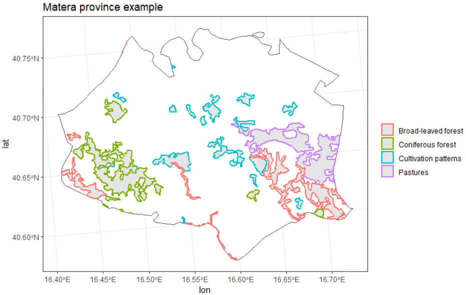
- Land use
  - **CORINE** Land Cover inventory (Copernicus Land Monitoring Service)
  - satellite images into **classes**
  - **% of covered land** by classes for **each municipality**
  - 100 m spatial resolution



- Rainfall data
  - meteorological monitoring networks Apulia and Basilicata Civil Protection Departments
  - number of days with rain → **proportion of summer days without rain** for each municipality

- Territorial indicators (ISTAT)
  - population Censuses
  - social and material vulnerability
  - number of cattle units (adults)

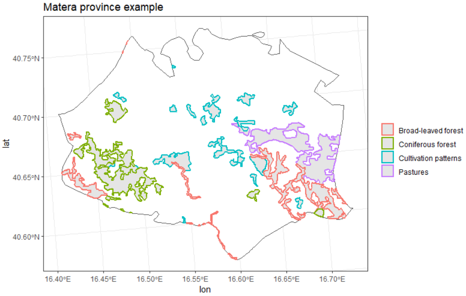
- Land use
  - **CORINE** Land Cover inventory (Copernicus Land Monitoring Service)
  - satellite images into **classes**
  - **% of covered land** by classes for **each municipality**
  - 100 m spatial resolution



- Rainfall data
  - meteorological monitoring networks Apulia and Basilicata Civil Protection Departments
  - number of days with rain → **proportion of summer days without rain** for each municipality

- Territorial indicators (ISTAT)
  - population Censuses
  - social and material vulnerability
  - number of cattle units (adults)

- Land use
  - **CORINE** Land Cover inventory (Copernicus Land Monitoring Service)
  - satellite images into **classes**
  - **% of covered land** by classes for **each municipality**
  - 100 m spatial resolution



- Rainfall data
  - meteorological monitoring networks Apulia and Basilicata Civil Protection Departments
  - number of days with rain → **proportion of summer days without rain** for each municipality

- Territorial indicators (ISTAT)
  - population Censuses
  - social and material vulnerability
  - number of cattle units (adults)

Proposed model

- Zero-inflated model: more than 60% of the counts are equal to 0;
- Measurement errors in count responses:
  - positive inflation (e.g. high temperature zones, small and localized fires,etc.)
  - negative inflation (e.g. rain, clouds may invalidate the measurements)

$\rightsquigarrow$ Ignoring measurement error may lead to biased estimates and invalid inference results

Following Zhang et al.(2022)[a], we propose a Bayesian Zero-inflated Poisson Model with Measurement Error in the response

[a] Zhang, Q. and Yi, G. Y. (2022) Zero-Inflated Poisson Models with Measurement Error in the Response, Biometrics 64, 1-25

The model (1): ZIP

For $i = 1, \ldots, n$ let we denote:

- Y_i^*: observed counts;
- Y_i: true (unobserved) counts;
- X_i: potential covariates

Relying on the Bayesian approach, given two possibly different vectors of covariates in $X_{i\mu}$ and $X_{i\phi}$, the conditional distribution of Y_i can be written as a hierarchical model:

$$
\begin{aligned}
P(Y_i = 0 | \mu_i, \phi_i, X_{i\mu}) &= (1 - \phi_i) + \phi_i e^{-\mu_i} \\
P(Y_i = y_i | \mu_i, \phi_i, X_{i\phi}) &= \phi_i \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}
\end{aligned}
$$

The model (1): ZIP

To facilitate the dependence of ϕ_i and μ_i on covariates X_i, we consider a complementary log-log regression model for ϕ_i and a log-linear model for μ_i :

$$
\begin{aligned}
P(Y_i = 0|\mu_i, \phi_i, X_{i\mu}) &= (1-\phi_i) + \phi_i e^{-\mu_i} \\
P(Y_i = y_i|\mu_i, \phi_i, X_{i\phi}) &= \phi_i \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \\
\log(\mu_i) &= \beta_{0\mu} + \beta_{\mu}^T X_{i\mu} + u_i \\
\text{cloglog}(\phi_i) &= \beta_{0\phi} + \beta_{\phi}^T X_{i\phi}
\end{aligned}
$$

The model (1): ZIP

- Random effects u_i's are assigned a proper Gaussian conditional autoregressive, CAR:

$$\mathrm{u}|C, M, \tau, \gamma \sim \mathrm{MVN}\left(0, \frac{1}{\tau}\left(I - \gamma C\right)^{-1} M\right)$$

where

- C are weights associated with each pair of neighboring; areas;
- M is a diagonal matrix of conditional variances;
- γ measures the overall degree of spatial dependence;
- τ is the precision of the Gaussian CAR prior.

We specify a Uniform prior distribution in the interval $[-1, 1]$ for γ and a Gamma prior with both parameters equal to 0.01 for τ.

The model (2): ME

Let

- Z_{i+} denote the count due to the *add-in* error
- Z_{i-} denote the count due to the *leave-out* error

The **ME model** is specified as

$$Y_i^* = Y_i + Z_{i+} - Z_{i-}$$

where

$$\begin{aligned}
Z_{i+}|\lambda_i, W_{i+} &\sim \text{Poisson}(\lambda_i) \\
Z_{i-}|\pi_i, Y_i, W_{i-} &\sim \text{Binomial}(Y_i, \pi_i) \\
\log(\lambda_i) &= \alpha_{0+} + \alpha_+^T W_{i+} \\
\text{logit}(\pi_i) &= \alpha_{0-} + \alpha_-^T W_{i-}
\end{aligned}$$

The model

Write $\alpha = (\alpha_{0+}, \alpha_{w+}^T, \alpha_{0-}, \alpha_{w-}^T)^T$. Let $\beta = (\beta_{\phi 0}, \beta_{\phi x}^T, \beta_{\mu 0}, \beta_{\mu x}^T)$ and $\theta = (\beta^T, \alpha^T)^T$. We are interested in conducting inference about β, accounting for the nuisance parameter α.

The distribution of the surrogate variable Y_i^* is given by

$$P(Y_i^* = y_i^* | X_i) = \underbrace{(1-\phi_i)\frac{\lambda_i^{y_i^*} e^{-\lambda_i}}{y_i^*!}}_{\text{mix 1}} + \underbrace{\phi_i \frac{\mu_i^* y_i^* e^{-\mu_i^*}}{y_i^*!}}_{\text{mix 2}}$$

where $\mu^* = (1-\pi_i)\mu_i + \lambda_i$

- *Identifiability:* $\pi(\theta)$ is proper, then the posterior $\pi(\theta | y_i^*, x_i)$ is proper.

Fire data: model specification

$$
\begin{array}{ll}
Y_i \sim P_0, & \text{with probability } 1-\phi_i \\
Y_i \sim Poisson(\mu_i) & \text{with probability } \phi_i \\
cloglog(\phi_i) = \beta_{\phi 0} + \beta_{\phi x}^T S_i &
\end{array}
$$

where

$$
\begin{aligned}
\log(\mu_i) &= \beta_{0\mu} + \beta_{\mu[1]}\mathrm{DIndx}_i + \beta_{\mu[2]}\mathrm{ALT}_i + u_i \\
cloglog(\phi_i) &= \beta_{0\phi} + \beta_{\phi[1]}\mathrm{ARTIF\ SURF}_i + \beta_{\phi[2]}\mathrm{HA}_i
\end{aligned}
$$

- DIndx: deprivation index; ALT: altitude;
- ARTIF SURF: artificial surface; HA: Heterogeneous agricultural areas.

Fire data: model specification

$$Y_i^* = Y_i + Z_{i+} - Z_{i-}$$
$$Z_{i+} \sim Poisson(\lambda_i)$$
$$log(\lambda_i) = \alpha_{0+} + \alpha_{w+}^T W_{i+}$$

where

$$\log(\lambda_i) = \alpha_{0+} + \alpha_+ \mathrm{RAIN}_i$$

$$Y_i^* = Y_i + Z_{i+} - Z_{i-}$$
$$Z_{i-} \sim Binomial(Y_i, \pi_i)$$
$$logit(\pi_i) = \alpha_{0-} + \alpha_{w-}^T W_{i-}$$

where

$$logit(\pi_i) = \alpha_{0-} + \alpha_- \mathrm{FOR}_i$$

- RAIN: % of days without rain;
- FOR: Forest and semi natural areas.

Fire data: model specification

$$
\begin{aligned}
&Y_i^* = Y_i + Z_{i+} - Z_{i-}\\
&Z_{i+} \sim Poisson(\lambda_i)\\
&log(\lambda_i) = \alpha_{0+} + \alpha_{w+}^T W_{i+}
\end{aligned}
$$

where

$$\log(\lambda_i) = \alpha_{0+} + \alpha_{+}\mathrm{RAIN}_i$$

$$
\begin{aligned}
&Y_i^* = Y_i + Z_{i+} - Z_{i-}\\
&Z_{i-} \sim Binomial(Y_i, \pi_i)\\
&logit(\pi_i) = \alpha_{0-} + \alpha_{w-}^T W_{i-}
\end{aligned}
$$

where

$$logit(\pi_i) = \alpha_{0-} + \alpha_{-}\mathrm{FOR}_i$$

- RAIN: % of days without rain;
- FOR: Forest and semi natural areas.

Results

- Negative impact of the percentage of heterogeneous agricultural land use on the probability of fire presence: agricolture areas are characterized by a mosaic of small parcels of annual crops, pastures, and permanent crops, under strict control of the owner, preventing fire propagation and damages;
- slightly negative effect of the number of rainy days on the add-in model component suggesting that the over-reporting is associated with lower values of the RAIN covariate;
- the effect of the percentage of forest covariate positively affects the leave-out component of the ME model. This result can be explained by considering the limited satellite capacity in detecting and reporting small or large fires that are consistent with burning forests engulfed in thick;
- significant spatial effect.

Model comparison

Model	Specification	**WAIC**
M1	ZIP + ME + CAR	940.030
M2	ZIP + CAR	956.949
M3	ZIP + ME	1254.495
M4	ZIP	1088.767

Table: WAIC goodness-of-fit measures for four different models fitted to the data on wildfires occurences.

Conclusions

- **ME** can be framed as a missing data problem and it has to be addressed with **proper data modelling**;
- Identifiability
- Computationally **feasible approach**.

Conclusions

- **ME** can be framed as a missing data problem and it has to be addressed with **proper data modelling**;
- Identifiability
- Computationally **feasible approach**.

Conclusions

- **ME** can be framed as a missing data problem and it has to be addressed with **proper data modelling**;
- Identifiability
- Computationally **feasible approach**.

Thanks for you attention!

Email: serena.arima@unisalento.it