Japanese Journal of Statistics and Data Science (2021) 4:1133–1155
https://doi.org/10.1007/s42081-021-00124-0

ORIGINAL PAPER

Recent Statistical Methods for Survival Analysis

Zero-inflated-censored Weibull and gamma regression models to estimate wild boar population dispersal distance

Eduardo de Freitas Costa[1] · **Silvana Schneider**[2] · **Giulia Bagatini Carlotto**[2] · **Tainá Cabalheiro**[2] · **Mauro Ribeiro de Oliveira Júnior**[3]

Received: 27 February 2021 / Revised: 25 April 2021 / Accepted: 29 April 2021 /
Published online: 15 May 2021

Abstract

The dynamics of the wild boar population has become a pressing issue not only for ecological purposes, but also for agricultural and livestock production. The data related to the wild boar dispersal distance can have a complex structure, including excess of zeros and right-censored observations, thus being challenging for modeling. In this sense, we propose two different zero-inflated-right-censored regression models, assuming Weibull and gamma distributions. First, we present the construction of the likelihood function, and then, we apply both models to simulated datasets, demonstrating that both regression models behave well. The simulation results point to the consistency and asymptotic unbiasedness of the developed methods. Afterwards, we adjusted both models to a simulated dataset of wild boar dispersal, including excess of zeros, right-censored observations, and two covariates: age and sex. We showed that the models were useful to extract inferences about the wild boar dispersal, correctly describing the data mimicking a situation where males disperse more than females, and age has a positive effect on the dispersal of the wild boars. These results are useful to overcome some limitations regarding inferences in zero-inflated-right-censored datasets, especially concerning the wild boar's population. Users will be provided with an R function to run the proposed models.

Keywords Dispersal · Distance · Wild boar · Zero-inflated data · Censored data

1 Introduction

The dynamics of wild animal populations is an important issue not only for ecological purposes, but also for agricultural and livestock production (Cumming et al., 2012). The wild boar (*Sus scrofa*) is considered to have one of largest geographical

✉ Eduardo de Freitas Costa
eduardo.costa@wur.nl

Extended author information available on the last page of the article

ranges of all terrestrial mammals (Lewis et al., 2017). It causes several types of losses related to the environment, agricultural activities, and animal production/trade by spreading diseases (Meng et al., 2009; Sánchez-Vizcaíno et al., 2019). With the re-emergence of African Swine Fever (ASF) as a threat to the global pig industry (Sánchez-Cordón et al., 2018), there is an ongoing need for better quantitative descriptions of the movement and land-occupation behaviors of wild boars (McClure et al., 2015; Morelle et al., 2014).

Animal dispersal occurs when individuals leave their social group or home range, redistributing the population (Breed & Moore, 2016). According to Casas-Díaz et al. (2013), dispersal is an important characteristic of the wild boar's ecology and should be taken into account in the design of disease monitoring programs as these animals can affect the spread of diseases and the probability of new outbreaks. Several techniques are used to measure animal dispersal [e.g., radio telemetry, GPS, ground surveys, or hunting statistics (Morelle et al., 2014)], and datasets produced by recapture with hunting bags are important sources of information for veterinary epidemiology (Keuling et al., 2018; Scillitani et al., 2010; Vicente et al., 2018).

Animal dispersal data may have a complex structure: it may be right-skewed given the short-distance dispersal; have excessive zeros, because many animals do not disperse at all or they disperse distances that are inside the error margin of telemetry (Keuling et al., 2008). Also, censored observations may occur, when an animal's track is lost (e.g., collars or ear tags are lost, or animals move away from the hunting perimeter) (Podgórski et al., 2014; Prévot & Licoppe, 2013; Truvé & Lemel, 2003). This data structure is challenging from an inferential perspective, creating issues for researchers attempting to formulate hypotheses or build models using dispersal distance information (Bowman et al., 2002). Thus, many authors discard the censored data and also avoid fitting generalized linear models. As a consequence, the datasets are not correctly used in making inferences and testing hypotheses about dispersal distances (Prévot & Licoppe, 2013; Truvé & Lemel, 2003). Evidently, improvements in quantitative analytical techniques on dispersal distance are still needed (Whitmee & Orme, 2013).

According to Lambert (1992), the zero values can have the following sources: they can come from a population where the value would always be zero, called "structural zeros"; alternatively, they can be "sampling zeros" from a population whose observations belong to some probability distribution. To analyze a zero-inflated non-negative outcome (semi-continuous outcome) data, Manning et al. (1981) proposed a two-part model (2PM), separating the zero and positive values explicitly in two submodels (parts). Liu et al. (2010) proposed an extension of the 2PM, assuming a generalized gamma distribution of the positive values. Lee et al. (2010) proposed a two-part multilevel modeling, in which the zero proportion is modeled by logistic regression and the continuous values by gamma regression. Gebregziabher et al. (2017) presented a family of models for zero-heavy continuous outcomes, with Weibull mixture regression as a special case, and with a marginal inference. Louzada et al. (2018) proposed a zero-inflated non-default rate regression, assuming that the positive values follow a Weibull distribution.

To consider the whole data structure of wild boar dispersal distances, we propose an extension of the zero-inflated gamma model (Lee et al., 2010),

incorporating the right-censored observations in the continuous values through the gamma regression. Furthermore, we consider the zero-inflated, right-censored Weibull regression (Louzada et al., 2018), discarding the cure rate of the model. In the models proposed in this paper, both techniques allow the incorporation of the whole data (without excluding zeros and the right-censored observations) to make inferences. This includes covariates in the model, as the survival models can also be used to assess distances, as seen in Reader (2000) and Chatwin et al. (2013). The gamma is a flexible distribution; it describes a different type of survival pattern according to the hazard rate: increasing, decreasing, or constant; it fits a variety of lifetime data adequately and has the exponential distribution as a particular case (Lawless, 2011). The Weibull distribution is also flexible and has a monotonically increasing, decreasing, or constant hazard rate (Klein & Moeschberger, 2003); is probably the most used parametric lifetime model (Wienke, 2011); it has been widely used to model right-skewed data and has also motivated proposals of various types of generalizations (Liao et al., 2020; Ramos et al., 2018; Shinohara et al., 2020).

Given the lack of a dataset fully describing zeros, dispersal distances, and censored data, we apply both techniques to an artificial dataset that mimics the wild boar population's dispersal. The simulated animal population includes three subpopulations: animals for which no dispersal was registered (zeros); animals that showed some measurable distance, and animals whose track has been lost (censored). Such data mimic situations of capture–recapture using hunting bags in which a proportion of animals moved away from the hunting perimeter. The manuscript is structured as follows.

First, we present the definitions and the construction of the zero-inflated-censored Weibull (ZIWeibull) and the zero-inflated-censored gamma (ZIGamma) models. Next, we show the properties of the models' estimators and compare the two models. Finally, we demonstrate their application to the artificial wild boar dispersal dataset. An R function was created to enable users to run the models with their own data. Instructions regarding this function and package installation details are available in Appendix (instructions for installing the R functions and run the models).

2 Formulation of the models

Let T denote the random variable of interest, and it is the observable time to event, called lifetime or failure time. Finally, consider $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ three vectors of the covariates, where $\mathbf{x}_1$ relates to the zero-inflation probability, $\mathbf{x}_2$ relates to the scale, and $\mathbf{x}_3$ relates to the shape of the lifetime distribution.

We propose an alteration of the zero-inflated non-default regression model (Louzada et al., 2018), using the zero-inflated and regression parts while removing the cure rate part of the model. Thus, the probability density function (PDF) and the cumulative density function (CDF) of the observable lifetime time T of the zero-inflated model are, respectively, given by

$$f_0(t) = \begin{cases} p_0, & \text{if } t = 0 \\ (1-p_0)f(t), & \text{if } t > 0 \end{cases}$$

and

$$F_0(t) = p_0 + (1-p_0)F(t), \quad t \geq 0, \tag{1}$$

where $p_0 \in (0,1)$ is the zero-inflation probability, $f(t)$ is the probability density function, and $F(t)$ is the cumulative density function of the observable lifetime time.

The observed data are denoted by $D_{\text{obs}} = \{(t_i, \delta_i), i = 1, \ldots, n\}$, where $\delta_i = 1$ if the event of interest occurs and $\delta_i = 0$ if it is right-censored, $i = 1, \ldots, n$. Assuming independent and non-informative censoring, the likelihood function $L(\boldsymbol{\theta};D_{\text{obs}})$ is given by

$$L(\boldsymbol{\theta};D_{\text{obs}}) = \prod_{i\,:\,t_i=0} p_{0i} \prod_{i\,:\,t_i>0} \left\{ \left[(1-p_{0i})f_i(t_i)\right]^{\delta_i} \left[(1-p_{0i})S_i(t_i)\right]^{1-\delta_i} \right\}, \tag{2}$$

where $\boldsymbol{\theta} = (p_{0i}, \boldsymbol{\vartheta})$, $\boldsymbol{\vartheta}$ denotes the parameter vector associated to the probability density function $f(t)$ and, consequently, to the survivor function $S(t)$; p_{0i} is the zero-inflate probability of the ith subject.

Considering the relation $f(t) = h(t)S(t)$ and $S(t) = \exp\{-H(t)\}$ in (2), where $h(t)$ is the hazard function and $H(t)$ is the cumulative hazard function, we obtain the following likelihood function:

$$L(\boldsymbol{\theta};D_{\text{obs}}) = \prod_{i\,:\,t_i=0} p_{0i} \prod_{i\,:\,t_i>0} \left\{ \left[h_i(t_i)\right]^{\delta_i} \left[(1-p_{0i}) \exp\left[-H_i(t_i)\right]\right] \right\}. \tag{3}$$

Therefore, the log-likelihood function for $\boldsymbol{\theta} = (p_{0i}, \boldsymbol{\vartheta})$, corresponding to the observed data and the likelihood function as in (3), is given by

$$\begin{aligned} \log\{L(\boldsymbol{\theta};D_{\text{obs}})\} &= \sum_{i\,:\,t_i=0} \log(p_{0i}) + \sum_{i\,:\,t_i>0} \delta_i \log\left[h_i(t_i)\right] \\ &\quad + \sum_{i\,:\,t_i>0} \log\left\{(1-p_{0i}) \exp\left[-H_i(t_i)\right]\right\} \\ &= \sum_{i\,:\,t_i=0} \log(p_{0i}) + \sum_{i\,:\,t_i>0} \delta_i \log\left[h_i(t_i)\right] \\ &\quad + \sum_{i\,:\,t_i>0} \log(1-p_{0i}) - \sum_{i\,:\,t_i>0} H_i(t_i). \end{aligned} \tag{4}$$

To obtain the maximum-likelihood estimates for the vector $\boldsymbol{\theta} = (p_{0i}, \boldsymbol{\vartheta})$, we must find the score functions $U(\boldsymbol{\theta}) = \frac{\partial \log\{L(\boldsymbol{\theta};D_{\text{obs}})\}}{\partial \boldsymbol{\theta}}$. To solve the non-linear system of equations $U(\boldsymbol{\theta}) = \frac{\partial \log\{L(\boldsymbol{\theta};D_{\text{obs}})\}}{\partial \boldsymbol{\theta}} = 0$, we chose the function "optim" in the R software (R Core Team, 2019) for numerical maximization, with the method "BFGS", according

to Louzada et al. (2018). Details on non-linear optimization can be found in Press et al. (2007).

The standard errors (SE) of the estimators are obtained through the diagonal of the inverted Fisher information matrix. The Fisher information matrix can be generically defined by

$$K(\boldsymbol{\theta}) = \begin{pmatrix} K_{pp} & 0 \\ 0 & K_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}} \end{pmatrix}, \tag{5}$$

where K_{pp} denotes second-order partial derivatives of $\boldsymbol{\theta}$ with respect to the covariates vector associated with p_0, $K_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}}$ denotes second-order partial derivatives of $\boldsymbol{\theta}$ with respect to the vector $\boldsymbol{\vartheta}$.

According to Ospina and Ferrari (2012), the confidence intervals of the components of the vector $\boldsymbol{\theta}$ can be obtained through $\hat{\boldsymbol{\theta}} \pm z_{1-\alpha/2}[K(\hat{\boldsymbol{\theta}})^{-1}]^{1/2}$, with asymptotic coverage $100(1-\alpha)\%$, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution. For more details on the asymptotic theory, see Ospina and Ferrari (2012) and Louzada et al. (2018).

The approach presented thus far can be used for any choice of hazard function $h(t)$ and, consequently, any cumulative hazard function $H(t)$. In the following section, we present the equations when the Weibull and gamma distributions are chosen to adjust the survival times.

2.1 Zero-inflated-censored Weibull model

If we consider the Weibull distribution to model the non-negative random variable $T|T>0$, in other words $T|T>0 \sim \text{Weibull}(\alpha_{\text{w}}, \lambda_{\text{w}})$, which here denotes the observable survival time, we obtain the following probability density function:

$$f(t) = \frac{\alpha_{\text{w}}}{\lambda_{\text{w}}}\left(\frac{t}{\lambda_{\text{w}}}\right)^{\alpha_{\text{w}}-1} \exp\left[-\left(\frac{t}{\lambda_{\text{w}}}\right)^{\alpha_{\text{w}}}\right], \quad t \geq 0, \tag{6}$$

where $\alpha_{\text{w}} > 0$ is the shape parameter and $\lambda_{\text{w}} > 0$ is the scale parameter of the distribution.

Replacing the hazard function $h(t)$ and the cumulative hazard function $H(t)$ from the log-likelihood function (4) with the hazard function and the cumulative hazard function from the Weibull distribution (6), we obtain the following log-likelihood function:

$$\begin{aligned} \log\{L(\boldsymbol{\theta};D_{\text{obs}})\} = & \sum_{i\,:\,t_i=0} \log(p_{0i}) + \sum_{i\,:\,t_i>0} \delta_i\left[\log\left(\frac{\alpha_{\text{w}}}{\lambda_{\text{w}}}\right) + (\alpha_{\text{w}}-1)\log\left(\frac{t_i}{\lambda_{\text{w}}}\right)\right] \\ & + \sum_{i\,:\,t_i>0} \log(1-p_{0i}) - \sum_{i\,:\,t_i>0}\left(\frac{t_i}{\lambda_{\text{w}}}\right)^{\alpha_{\text{w}}}. \end{aligned} \tag{7}$$

The inclusion of covariates determines the effect of covariates on the observable survival time and the probability of zeros. The parameters $(p_{0i}, \alpha_{\text{w}}, \lambda_{\text{w}})$ are related to

the proportion of zeros, the shape, and the scale parameters of the Weibull distribution, respectively. The systematic components of the regression version of the zero-inflated-censored Weibull model are given by

$$\begin{cases} G_1(p_{0i}) & = \xi_{0i} \\ G_2(\lambda_{\mathrm{w}i}) & = \eta_{1i} \\ G_3(\alpha_{\mathrm{w}i}) & = \eta_{2i}, \end{cases} \tag{8}$$

where $\xi_{0i} = \mathbf{x}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1$, $\eta_{1i} = \mathbf{x}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2$, and $\eta_{2i} = \mathbf{x}_{3i}^{\mathrm{T}}\boldsymbol{\beta}_3$ are the linear predictors and; $\boldsymbol{\beta}$'s are the unknown regression coefficient vectors. The link functions $G_1(.)$, $G_2(.)$, and $G_3(.)$ provide the connection between the linear predictor and the parameters of the probability density function. According to the fit by Louzada et al. (2018), $G_1(.)$ is set as the logistic regression $G_1(p_{0i}) = \log\left(\frac{p_{0i}}{1-p_{0i}}\right)$, with $p_{0i} = \frac{e^{\mathbf{x}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1}}{1+e^{\mathbf{x}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1}}$. Since the scale and shape parameters are defined on the positive real line, the $G_2(.)$ and $G_3(.)$ link functions are chosen as $G_2(\lambda_{\mathrm{w}i}) = \log(\lambda_{\mathrm{w}i})$ and $G_3(\alpha_{\mathrm{w}i}) = \log(\alpha_{\mathrm{w}i})$. Thus, $\lambda_{\mathrm{w}i} = e^{\mathbf{x}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2}$ and $\alpha_{\mathrm{w}i} = e^{\mathbf{x}_{3i}^{\mathrm{T}}\boldsymbol{\beta}_3}$.

Considering the observable time $T|T > 0 \sim \mathrm{Weibull}(\alpha_{\mathrm{w}}, \lambda_{\mathrm{w}})$ and the probability density function $f_0(t)$ (see (1)), the expected value of T is defined by

$$\mathbb{E}(T) = (1 - p_0)\lambda_{\mathrm{w}}\Gamma\left(1 + \frac{1}{\alpha_{\mathrm{w}}}\right), \tag{9}$$

where $\Gamma(.)$ is the gamma function.

Replacing the parameters in (9) by the relations in (8), we have the following expected value given covariates:

$$\mathbb{E}(T|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \left(1 - \frac{e^{\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}_1}}{1 + e^{\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}_1}}\right)e^{\mathbf{x}_2^{\mathrm{T}}\boldsymbol{\beta}_2}\Gamma\left(1 + \frac{1}{e^{\mathbf{x}_3^{\mathrm{T}}\boldsymbol{\beta}_3}}\right). \tag{10}$$

2.2 Zero-inflated-censored gamma model

The zero-inflated-censored gamma model is an extension of Lee's model (Lee et al., 2010). It is based on the concept of two-part or "hurdle" models, in which zeros and non-zeros are considered two independent processes (Mullahy, 1986; Nobre et al., 2017), and the right-censored data are considered in the continuous values. Here, the zeros will be modeled as "success events" using a logistic regression. The positive observable survival time is modeled by a gamma regression survival model.

Let Z be a binary variable, where $Z = 1$ when the observable time was zero in the database, and $Z = 0$ otherwise. Therefore, Z has Bernoulli distribution with p_0 parameter, $Z \sim \mathrm{Ber}(p_0)$. Since the zero probability, $P(Z = 1) = p_0$, is independent of the time distribution, the likelihood function (2) can be rewritten as follows:

$$L(\boldsymbol{\theta};D_{\text{obs}}) = \left\{ \prod_{i\,:\,Z_i=1} p_{0i} \prod_{i\,:\,Z_i=0} (1-p_{0i}) \right\} \left\{ \prod_{i\,:\,t_i>0} f_i(t_i)^{\delta_i} S_i(t_i)^{1-\delta_i} \right\}, \tag{11}$$

where $\boldsymbol{\theta} = (p_{0i}, \boldsymbol{\vartheta})$ and $\boldsymbol{\vartheta}$ denotes the parameter vector associated with the probability density function $f(t)$. Factoring the likelihood allows for the estimation of p_{0i} to be made independently of the fit for the time distribution. To model the zero probability, p_{0i}, we used a logistic regression with $G_4(p_{0i}) = \log\left(\frac{p_{0i}}{1-p_{0i}}\right)$.

Considering the observable time $T|T>0 \sim \text{Gamma}(\alpha_{\text{g}}, \lambda_{\text{g}})$, with a shape parameter α_{g} and a scale parameter λ_{g}, we obtain the following probability density function:

$$f(t) = \frac{1}{\lambda_{\text{g}}^{\alpha_{\text{g}}}\,\Gamma(\alpha_{\text{g}})} t^{\alpha_{\text{g}}-1} \exp\left(-\frac{t}{\lambda_{\text{g}}}\right), \quad t>0, \tag{12}$$

where $\alpha_{\text{g}} > 0$ is the shape parameter and $\lambda_{\text{g}} > 0$ is the scale parameter of the distribution.

Considering the likelihood function (11), the relations $f(t) = h(t)S(t)$ and $S(t) = \exp\{-H(t)\}$, assuming gamma distribution to the observable survival time, produce the following log-likelihood function:

$$\begin{aligned} \log\{L(\boldsymbol{\theta};D_{\text{obs}})\} = & \sum_{i\,:\,Z_i=1} \log(p_{0i}) + \sum_{i\,:\,Z_i=0} \log(1-p_{0i}) \\ & + \sum_{i\,:\,t_i>0} \delta_i \left[\alpha_{\text{g}} \log\left(\frac{t_i}{\lambda_{\text{g}}}\right) - \log\left[t_i \Gamma(\alpha_{\text{g}})\right] - \left(\frac{t_i}{\lambda_{\text{g}}}\right) \right] \\ & + \sum_{i\,:\,t_i>0} (1-\delta_i) \left[\log\left[S(t_i)\right]\right], \end{aligned} \tag{13}$$

where $S_i(t_i) = \int_{t_i}^{\infty} \frac{1}{\lambda_{\text{g}}^{\alpha_{\text{g}}}\,\Gamma(\alpha_{\text{g}})} u^{\alpha_{\text{g}}-1} \exp\left(-\frac{u}{\lambda_{\text{g}}}\right) \text{d}u$.

We obtain the maximum-likelihood estimates for $\boldsymbol{\theta} = (p_{0i}, \alpha_{\text{g}}, \lambda_{\text{g}})$ using iterative techniques from the "glm" function for estimate p_{0i} and from the package "flexsurv" (Jackson, 2016) to estimate α_{g} and λ_{g}. Both functions are implemented in the software R (R Core Team, 2019).

The inclusion of the covariates associated with the observable time was achieved through the scale and shape parameters of the gamma distribution. The systematic component of the regression version of the zero-inflated-censored gamma model is given by

$$\begin{cases} G_4(p_{0i}) &= \xi_{0i} \\ G_5(\lambda_{\text{g}i}) &= \eta_{1i} \\ G_6(\alpha_{\text{g}i}) &= \eta_{2i}, \end{cases} \tag{14}$$

where $\xi_{0i} = \mathbf{x}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1$, $\eta_{1i} = \mathbf{x}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2$, and $\eta_{2i} = \mathbf{x}_{3i}^{\mathrm{T}}\boldsymbol{\beta}_3$ are the linear predictors, and β's are the unknown regression coefficient vectors. The link functions $G_4(.)$, $G_5(.)$, and $G_6(.)$ provide the connection between the linear predictor and the parameters of the probability density function. $G_4(.)$ is a set of logistic regression $G_4(p_{0i}) = \log\left(\frac{p_{0i}}{1-p_{0i}}\right)$, with $p_{0i} = \frac{e^{\mathbf{x}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1}}{1+e^{\mathbf{x}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1}}$. According to Jackson (2016) $G_5(.)$ and $G_6(.)$, link functions are $G_5(\lambda_{gi}) = \log(\lambda_{gi})$ and $G_6(\alpha_{gi}) = \log(\alpha_{gi})$. Thus, $\lambda_{gi} = e^{\mathbf{x}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2}$ and $\alpha_{gi} = e^{\mathbf{x}_{3i}^{\mathrm{T}}\boldsymbol{\beta}_3}$.

Assuming the failure time $T|T > 0 \sim \text{Gamma}(\alpha_g, \lambda_g)$ and the probability density function (1), the expected value of T is defined by $\mathbb{E}(T) = (1 - p_0)\alpha_g\lambda_g$. Considering the inclusion of the covariates, we obtain the following expected value given covariates:

$$\mathbb{E}(T|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \left(1 - \frac{e^{\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta_1}}}{1 + e^{\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta_1}}}\right)\exp(\mathbf{x}_2^{\mathrm{T}}\boldsymbol{\beta_2})\exp(\mathbf{x}_3^{\mathrm{T}}\boldsymbol{\beta_3}). \tag{15}$$

3 Simulation studies

In this section, we present a simulation study to evaluate the models proposed in the previous section. The ZIWeibull model was fitted in a dataset generated from a Weibull distribution, and the ZIGamma model was fitted in a dataset generated from a gamma distribution. The models' parameters were estimated as described in Sect. 2. Computer codes are available online (see "Availability of data and codes" section for more information). To check the performance of the estimators, we examine the coverage probabilities of the 95% confidence intervals, the bias and root-mean-square errors, as well as the estimator's asymptotic consistency. We consider five sample sizes $n = (200, 400, 600, 800, 1000)$, and for each scenario, we generate 1000 datasets. In the process of generating each data set, we considered failure times to follow Weibull and gamma distributions with the following regression structure: $\alpha_i = e^{\beta_1 + \mathrm{x}_{3i}\beta_2}$, $\lambda_i = e^{\beta_3 + \mathrm{x}_{2i}\beta_4}$, and $p_{0i} = \frac{e^{\beta_5 + \mathrm{x}_{1i}\beta_6}}{1+e^{\beta_5 + \mathrm{x}_{1i}\beta_6}}$, where the covariate vectors are specified by $\mathrm{x}_i \equiv \mathrm{x}_{i1} = \mathrm{x}_{i2} = \mathrm{x}_{i3} \sim \text{Normal}(0, 1)$ and the regression coefficients by $\beta_1 = 0.5$, $\beta_2 = 0.5$, $\beta_3 = 1.5$, $\beta_4 = 2$, $\beta_5 = -3$, $\beta_6 = 1$.

To simulate random samples of size n, we suppose that the time until the occurrence of the event of interest has the cumulative distribution function $F(t)$ given by $F_0(t) = p_{0i} + (1 - p_{0i})F(t)$, $t \geq 0$. The algorithmic steps to generate the random samples are as follows:

1. Generate $\mathrm{x}_i \sim \text{Normal}(0, 1)$ and calculate $p_{0i}, \alpha_i, \lambda_i$;
2. Generate u_i from a uniform distribution $U(0, 1)$;
3. Generate v_i from a uniform distribution $U(p_{0i}, 1)$;
4. If $u_i \leq p_{0i}$, set $t_i = 0$;
5. If $u_i > p_{0i}$, set t_i equals to the inverse function $F_0^{-1}(v_i)$;
6. Generate the censoring times c_i from a gamma distribution $G(0.25, 128.2)$, specified to control the censoring percentage;
7. If $t_i < c_i$, set $\delta_i = 1$; otherwise, set $\delta_i = 0$;
8. If $t_i = 0$; set $Z_i = 1$, otherwise, set $Z_i = 0$.

For the zero-inflated-censored model, since failure times have a Weibull distribution, the average time until the event of interest is calculated by

$$\mathbb{E}(T|\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3)=\left(1-\frac{e^{\beta_5+\mathrm{x}_1\beta_6}}{1+e^{\beta_5+\mathrm{x}_1\beta_6}}\right)e^{\beta_3+\mathrm{x}_2\beta_4}\ \Gamma\left(1+\frac{1}{e^{\beta_1+\mathrm{x}_3\beta_2}}\right). \tag{16}$$

For the zero-inflated-censored model, since failure times have a gamma distribution, the average time until the event of interest is calculated by

$$\mathbb{E}(T|\mathrm{x}_1,\mathrm{x}_2,\mathrm{x}_3)=\left(1-\frac{e^{\beta_5+\mathrm{x}_1\beta_6}}{1+e^{\beta_5+\mathrm{x}_1\beta_6}}\right)e^{\beta_1+\mathrm{x}_3\beta_2}\cdot e^{\beta_3+\mathrm{x}_2\beta_4}. \tag{17}$$

For both models proposed, considering the 1000 replications of each scenario, Fig. 1 shows that as the sample size increases from 200 to 1000, the means of the six estimated parameters $(\hat{\beta}_1,\ldots,\hat{\beta}_6)$ converge on the true values of the parameters $(0.5, 0.5, 1.5, 2, -3, 1)$.

Figure 2 shows that for both models, the increase of the sample size from 200 to 1000 reduces asymptotically the variance, pointing to the minimal variance of the parameters estimated in these models. In Appendix (parameter estimates in the simulation studies), several summary statistics for the parameters are provided, such as mean, standard deviation (SD), confidence interval, bias, mean squared error, and coverage probability. Notably, for all sample size scenarios, the means of the standard errors (SE) are similar to the SD, indicating the good performance of the models. The coverage probability is remarkably close to the nominal value of 95%.

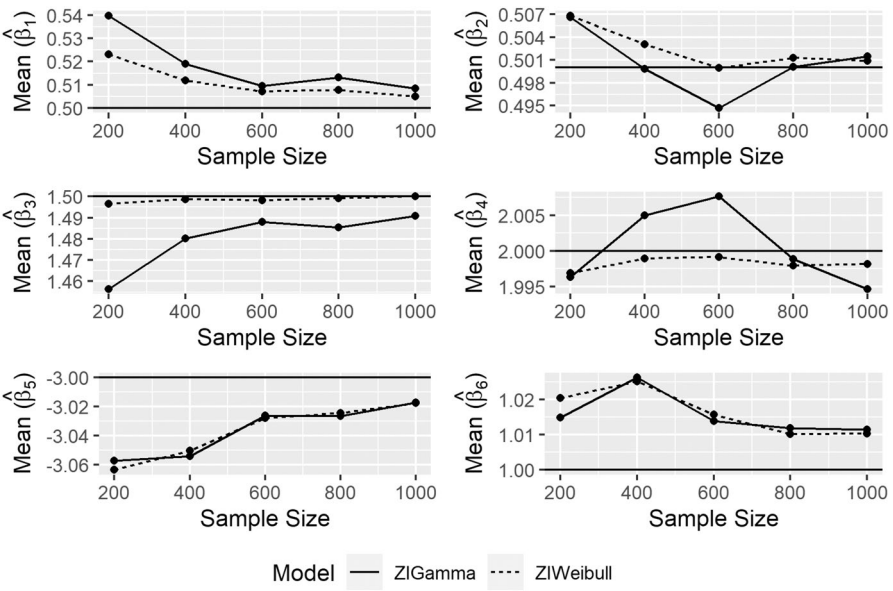

Fig. 1 Mean values of the $\hat{\beta}$ estimated by the ZIWeibull model (when dataset were generated by the ZIWeibull model) and ZIGamma model (when datasets were generated by the ZIGamma model) using different sample sizes. The solid horizontal line is the true value

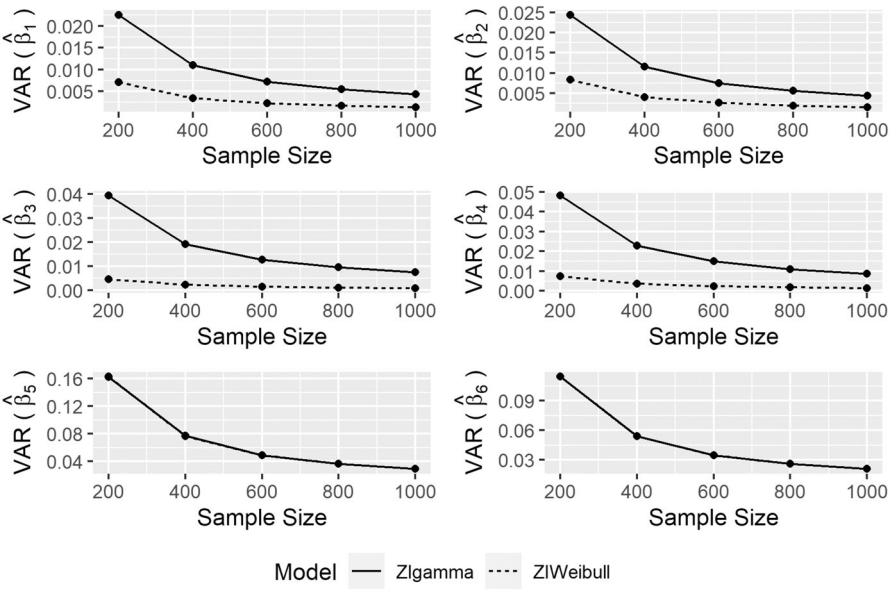

Fig. 2 Variance for the $\hat{\beta}$ estimated by the ZIWeibull model (when dataset was generated by the ZIWeibull model) and ZIGamma model (when dataset was generated by the ZIGamma model) using different sample sizes

For both ZIGamma and ZIWeibull models, the expected values of the proportion of zeros were similar, for all sample sizes, converging asymptotically to the true value (Table 1). The average times-to-event estimated by the ZIWeibull and ZIGamma models was close to the true value, regardless the scenario of gamma or Weibull distribution for both datasets generation (Table 1).

4 Wild boar dispersal application

In this section, we adjust the ZIWeibull and ZIGamma models to the wild boar natal dispersal study based on recapture by hunting bag (Truvé & Lemel, 2003). Dispersal occurs when wild boars leave their territory for new territories or home ranges (Breed & Moore, 2016). Dispersal accounts for three distinct subpopulations of animals: a segment that will not disperse, i.e., have zero distance; a segment that will disperse measurably (event), and animals whose dispersal is unknown (right-censored).

Examples of such a data structure in wild boar dispersal are available in Keuling et al. (2010), Prévot and Licoppe (2013), and Jerina et al. (2014). These authors report a dataset structure compatible with that of Koenig et al. (1996) and Paradis et al. (1998) who reported that recaptures often occur in the same location at which the animal was marked, creating zeros, or animals moving more than the limit set for recapture in generating right-censored data. Unfortunately, the full datasets for these papers are not publicly available. To address this limitation, we

Table 1 The simulation results of estimating the probability of zero and for the expected time, given these two values for the covariates x = −1 and x = 1 and gamma and Weibull scenarios for data generation

True values for Weibull dataset	Model fitted	Sample size				
		200	400	600	800	1000
$\mathbb{P}(p_{0i}\|\mathrm{x}=-1)=0.018$	ZIGamma	0.017	0.017	0.017	0.017	0.017
	ZIWeibull	0.016	0.017	0.017	0.018	0.018
$\mathbb{P}(p_{0i}\|\mathrm{x}=1)=0.119$	ZIGamma	0.115	0.116	0.118	0.118	0.119
	ZIWeibull	0.114	0.116	0.118	0.118	0.119
$\mathbb{E}(T\|\mathrm{x}=-1)=0.595$	ZIGamma	0.594	0.598	0.599	0.602	0.604
	ZIWeibull	0.592	0.591	0.593	0.593	0.594
$\mathbb{E}(T\|\mathrm{x}=1)=25.920$	ZIGamma	25.566	25.540	25.428	25.411	25.371
	ZIWeibull	25.856	25.961	26.001	25.919	25.943
True values for gamma dataset	Model fitted	Sample size				
		200	400	600	800	1000
$\mathbb{P}(p_{0i}\|\mathrm{x}=-1)=0.018$	ZIGamma	0.016	0.016	0.017	0.017	0.017
	ZIWeibull	0.016	0.016	0.017	0.017	0.017
$\mathbb{P}(p_{0i}\|\mathrm{x}=1)=0.119$	ZIGamma	0.114	0.116	0.118	0.117	0.118
	ZIWeibull	0.114	0.116	0.118	0.118	0.119
$\mathbb{E}(T\|\mathrm{x}=-1)=0.595$	ZIGamma	0.592	0.593	0.593	0.595	0.597
	ZIWeibull	0.592	0.591	0.593	0.593	0.594
$\mathbb{E}(T\|\mathrm{x}=1)=79.280$	ZIGamma	79.620	79.862	79.387	79.210	78.965
	ZIWeibull	81.554	81.152	81.850	81.711	81.596

$\mathbb{P}(p_{0i}|\cdot)$ denotes the probability of zero; $\mathbb{E}(T|\cdot)$ denotes the expected time

created a dataset that mimics a real situation. Details about the data generation are available in Appendix (Information used to generate the applied dataset). The computer code used to generate the data and the dataset spreadsheet are available as Electronic supplementary material.

Applied to wild boars, our interest is to adjust the dispersal distance, denoted by the random variable D, representing the observable distance. Therefore, the likelihood contribution of a dispersal distance D_i of an animal, or i, can assume two different values: p_{0i}, if animal i has a dispersal equal to zero, and $(1-p_{0i})f_0(d_i)$, if animal i has a dispersal larger than zero.

The models' parameters were tested under null hypothesis of equality to zero using a Wald test. For both ZIWeibull and ZIGamma models, the regression coefficients β_2 (age), β_3 (sex) are accounting for the shape parameter (α), and β_5 (age), β_6 (sex) are associated with the scale parameter (λ). The effects of the regression coefficients β_8 (age) and β_9 (sex) are associated with the proportion of zeros (p_0). Thus, regression coefficients have an indirect effect on the proportion of zeros estimated and on the expected dispersal distance, as we can see below for the ZIWeibull and ZIGamma models, respectively

$$\mathbb{E}(D|\text{age, sex}) = (1-p_0) \times \alpha_{\text{w}} \times \Gamma\left(1+\frac{1}{\lambda_{\text{w}}}\right)$$

and (18)

$$\mathbb{E}(D|\text{age, sex}) = (1-p_0) \times \alpha_{\text{g}} \times \lambda_{\text{g}},$$

where $\alpha_{\text{w}} = e^{\beta_1+\beta_2\text{age}+\beta_3\text{sex}}$, $\lambda_{\text{w}} = e^{\beta_4+\beta_5\text{age}+\beta_6\text{sex}}$, and $p_0 = \frac{e^{\beta_7+\beta_8\text{age}+\beta_9\text{sex}}}{1+(e^{\beta_7+\beta_8\text{age}+\beta_9\text{sex}})}$ obtained through the ZIWeibull model adjust, and $\alpha_{\text{g}} = e^{\beta_1+\beta_2\text{age}+\beta_3\text{sex}}$, $\lambda_{\text{g}} = e^{\beta_4+\beta_5\text{age}+\beta_6\text{sex}}$, and $p_0 = \frac{e^{\beta_7+\beta_8\text{age}+\beta_9\text{sex}}}{1+(e^{\beta_7+\beta_8\text{age}+\beta_9\text{sex}})}$ obtained through the ZIGamma model.

In the ZIGamma model, the proportion of zeros in the simulated wild boar dataset is not significantly dependent on age ($\beta_8 = -0.016$ P value $= 0.163$), or sex ($\beta_9 = 0.423$ P value $= 0.135$), as shown in Table 2. On the other hand, age ($\beta_5 = 0.04$ Pvalue < 0.001) and sex ($\beta_6 = -0.64$ P value $= 0.001$) are associated with the scale parameter of the curve. For the shape of the curve, there is association of age ($\beta_2 = -0.01$ P value $= 0.028$) but not sex ($\beta_3 = 0.027$ P value $= 0.98$). For the ZIWeibull model, the proportion of zeros in the simulated wild boar dataset is also not dependent on age (P value $= 0.162$) or on sex (P value $= 0.134$), see β_8 (age) and β_9 (sex) in Table 2. However, age and sex are associated with the scale of the curve ($\beta_5 = 0.032$ and $\beta_6 = -0.644$ Pvalue < 0.001 for both); and age was associated with the shape ($\beta_2 = -0.007$ P value $= 0.042$) (Table 2).

Because the generated dispersal distance is affected by sex and age, we have chosen to calculate the mean distance using the average age for both sexes. For males, the average age was 15 months, so the estimated mean distance was 13.9 km and 13.66 km with the ZIWeibull and ZIGamma models, respectively. For females, the mean age was 12 months, at which age the estimated mean distance was 6.1 km and 6.08 km with the ZIWeibull and ZIGamma models, respectively.

For the regression models proposed here, increase in the mean distance is associated with increased age and males dispersing more than females, as shown in Fig. 3. On average, males are known to disperse more than females. For both sexes, increased age increases the distances. These findings are expected, as we intentionally generated an applied dataset while attempting to account for the matrilineal, territorial societal structure (females and their offspring having very low dispersion), as well as for the solitary males that disperse more to seek territory and mating opportunities (Keuling et al., 2010).

To assess the models' goodness-of-fit, we evaluated the following information criteria: Akaike and Bayes information criteria, (AIC and BIC). These were calculated according to $qk - 2\log(\hat{L})$, where k is the number of parameters, $\hat{L}$ is the likelihood, and $q = 2$ for AIC, or $q = \ln(n)$, $n =$ sample size for BIC (Brewer et al., 2016). As the ZIGamma model is a hurdle model, the full log-likelihood was calculated adding the log-likelihood of the logistic regression and the log-likelihood of the gamma regression (McDowell, 2003).

AIC are similar for both proposed models, but larger difference is observed in BIC which is smaller for the ZIGamma model (see Table 2). However, our aim here

Table 2 Results of the ZIGamma and ZIWeibull models fitted for the simulated dataset of wild boar dispersal distance

	ZIGamma model			
	Estimate	SE	95% CI	*P* value
$\beta_1(\alpha_g)$	1.06	0.130	(0.832; 1.349)	–
β_2 (age)	− 0.01	0.005	(− 0.002; − 0.001)	0.028
β_3 (sex)	0.027	0.141	(− 0.277; 0.279)	0.98
$\beta_4(\lambda_g)$	8.27	1.49	(6.33; 12.3)	–
β_5 (age)	0.04	0.008	(0.028; 0.061)	< 0.001
β_6 (sex)	− 0.64	0.194	(− 1.021; − 0.258)	0.001
β_7 (Intercept)	− 1.726	0.258	(− 2.251; − 1.237)	–
β_8 (age)	− 0.016	0.011	(− 0.039; 0.005)	0.163
β_9 (sex)	0.423	0.282	(− 0.126; 0.985)	0.135
	ZIWeibull model			
	Estimate	SE	95% CI	*P* value
$\beta_1(\alpha_w)$	0.014	0.081	(− 0.145; 0.173)	–
$\beta_2(age)$	− 0.007	0.004	(− 0.014; − 0.0002)	0.042
$\beta_3(sex)$	0.029	0.092	(− 0.178; 0.184)	0.974
$\beta_4(\lambda_w)$	2.242	0.109	(2.027; 2.457)	–
β_5 (age)	0.032	0.005	(0.021; 0.043)	< 0.001
β_6 (sex)	− 0.644	0.127	(− 0.893; − 0.395)	< 0.001
β_7 (Intercept)	− 1.725	0.257	(− 2.231; − 1.219)	–
β_8 (age)	− 0.016	0.011	(− 0.037; 0.006)	0.162
β_9 (sex)	0.422	0.282	(− 0.131; 0.976)	0.134

SE is the standard error; CI is the confidence interval; for the hypothesis test, it was used a Wald test. α and λ are the shape an scale parameters, respectively. $\text{ZIWeibull}_{\text{AIC}} = 2324.28$, $\text{ZIGamma}_{\text{AIC}} = 2320.45$, $\text{ZIWeibull}_{\text{BIC}} = 2360.21$, $\text{ZIGamma}_{\text{BIC}} = 2348.39$

is not strictly to select the best model from between ZIGamma and ZIWeibull, but to present both as alternative methods in this applied context, allowing for the incorporation of all the information contained in the dataset to calculate the likelihood function.

5 Discussion

Although the lack of a real dataset hampers the attempt to draw a direct conclusion about an actual wild boar population, our aim was to highlight that both ZIWeibull and ZIGamma can be useful tools for statistical inferences in wild boar dispersal. According to Whitmee and Orme (2013), dispersal is one of the principal mechanisms underlying wild animals' land occupation, and there is an urgent need to obtain quantitative empirical data and methods to analyze dispersal distance. The

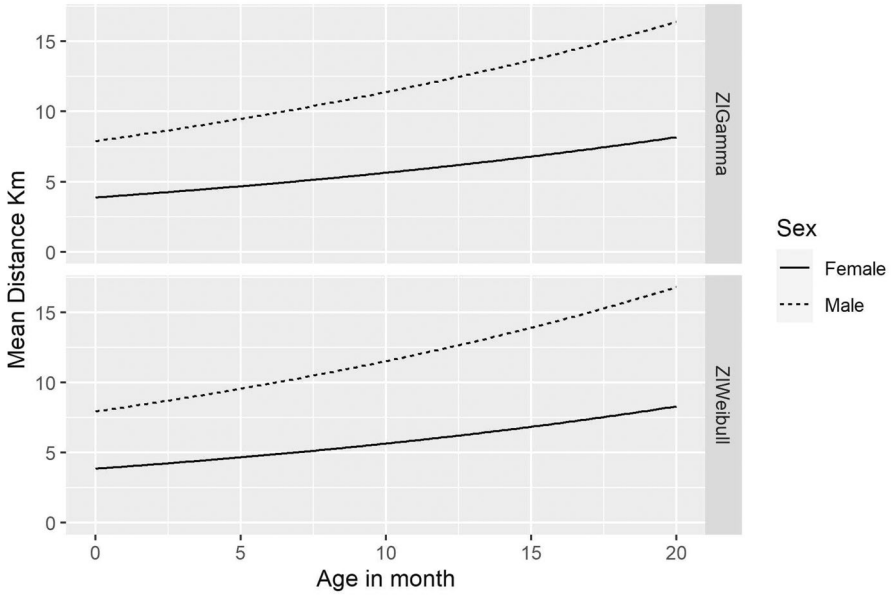

Fig. 3 Values for the mean dispersal distance (Km) estimated by the ZIWeibull and ZIGamma regression models

models presented here are designed to accommodate the proportions of individuals that will not disperse at all and zeros created by measurement error (accounting for the excess of zeros), the observable dispersal distance, and the cases in which the dispersal is greater than our capacity to follow it, or incomplete followings (right-censoring). In this sense, researchers could count with an inferential tool avoiding data exclusion using the full dataset in the same likelihood function. For instance, Truvé and Lemel (2003) excluded censored data and split the dataset to fit a non-parametric model to each covariate.

From an applied perspective, estimating the characteristics of wild boar populations [e.g., average dispersal across sexes and ages (Keuling et al., 2010)] is important in enabling animal health managers and decision-makers to make evidence-based decisions (EFSA, 2014; Morelle et al., 2014). For instance, risk assessment models for the introduction/spread of African Swine Fever are heavily dependent on parameters regarding the land occupation (De la Torre et al., 2015). Surveillance programs for commercial swine herds often consider the dynamics of the wild boar population in preventing diseases (Casas-Díaz et al., 2013). As such, the models proposed here are useful for generating inferences about dispersal distance and testing if covariates are associated with the distance, thus helping to overcome current gaps in our understanding of wild boar behavior (Guinat et al., 2016; EFSA, 2019).

For the zero-inflated-censored Weibull (ZIWeibull) model, the estimators were consistent and efficient, and both the standard deviation values and the estimated standard errors values are quite close, indicating a good adjustment. Likewise, for the ZIWeibull model, the coverage probabilities for the ZIGamma model were close

to the nominal level, and both the standard deviation values and the standard error values are remarkably close. The large sample theory for the models proposed in this paper is still an open question. However, our simulation studies show that the asymptotic validity seems to be satisfied. Finally, we note that the zero-inflated gamma model proposed in this paper can be extended to include the cure rate in the model. Furthermore, the zero-inflated-censored Weibull and gamma regression models can be extended to interval-censored survival data.

Appendix

Appendix for zero-inflated-censored Weibull and gamma regression models to estimate wild boar population dispersal distance by Eduardo Costa, Silvana Schneider, Giulia Carlotto, Tainá Cabalheiro, and Mauro Oliveira

Instructions for installing the R functions and run the models

The function is available as a package format. The user may face some issues: if your R version does not support the package "devtools", you will be not able to install "ZIdispersal" package. In this sense, the user should follow the steps:

1. Install and load the package devtools:

   ```
   install.packages ("devtools")  library(devtools)
   ```

   OBS: If the installation of "devtools" do not not occur, then the user will not succeed the "ZIdispersal" installation.

2. Install and load the package ZIdispersal:

   ```
   install_github("eduardodefreitascosta/Zidisper-
   sal/ZIdispersal",  force = T) library(ZIdispersal)
   ```

   This package has 1 function: `ZIreg`.
   For more details, please see the documentation `(?ZIreg)`.

3. Download the applied data (wild boar dispersal) and run the applied models: https://github.com/eduardodefreitascosta/ZIdispersal/blob/master/ZIdispersal/wild_boar.rds

   To run the ZIGamma

   ```
   ZIreg(dist age + sex, zero = zero, censor = delta,
   data = wild_boar, dist = "gamma")
   ```

   To run the ZIWeibull

   ```
   ZIreg(dist  age + sex, zero = zero, censor =
   delta, data =wild_boar, dist = "weibull"
   ```

Table 3 Descriptive statistics of the mean dispersal distance, percentage of zeros, and censored observations for males and females wild boars (below and above the mean age) in the applied dataset

Wild boar group	Mean distance (km)	Percentage of zero	Percentage of censored
Male (total)	11.111	12.500	15.000
Male < 15 month	9.106	12.500	7.812
Male > 15 month	14.675	12.500	27.789
Female (total)	6.174	18.500	4.500
Female < 12 month	4.305	20.968	0.000
Female > 12 month	9.224	14.474	11.850

Information used to generate the applied dataset

Dispersal distance was generated based on Truvé and Lemel (2003). The distances are dependent on sex and age (month). For females, the distance increases between seven and nine months, and the general average dispersal for females is 6 km. For males, the distance increases from 10 to 13 months, and the general dispersal distance for males is approximately 10 km. The proportion of zeros was 7% for males and 12% for females; 5% of censored observations were generated for females and 16% for males. To generate data as similar as possible to the real data found in Truvé and Lemel (2003), we consider a sample size $n = 400$ for animal dispersal consisting of 50% males and 50% females. We consider exponential distribution for age, with scale parameters of 12 for females and 15 for males. We also consider the zeros proportion of 15% and right-censoring when the distance is greater than 30 km.

The general average distance in the generated dataset was 8.68 km, 6.17 km for females and 11.11 km for males. The proportions of zeros were 12.5 % and 18.5 % for males and females, respectively. The censoring proportions were 15 % and 4.5 % for males and females, respectively. More details can be seen in Table 3.

For both sexes, the observed dispersal distance distribution is right-skewed, with more zeros for females and a larger tail on the right, and more right-censored data for males (Fig. 4).

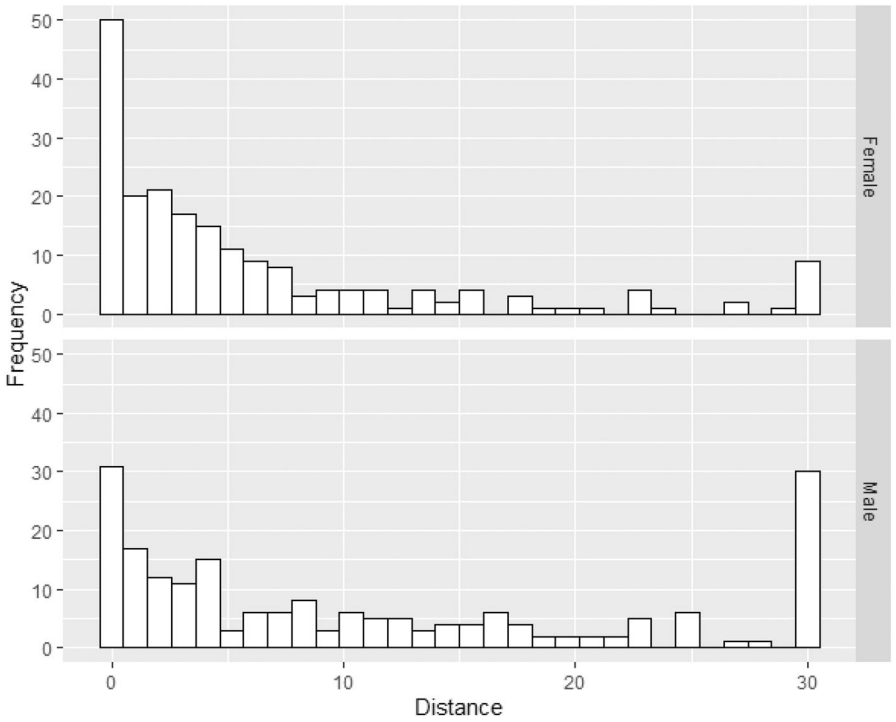

Fig. 4 Distribution of the dispersal distance in km for males and females wild boars

Parameter estimates in the simulation studies

See Tables 4 and 5.

Table 4 Results of the simulation study for the zero-inflated-censored Weibull model adjustment considering different sample sizes

	True	Mean (SD)	SE	CI	Bias	MSE	CP
Sample size n = 200							
β_1	0.5	0.522 (0.089)	0.084	(0.347; 0.696)	0.022	0.008	0.938
β_2	0.5	0.505 (0.096)	0.091	(0.316; 0.694)	0.005	0.009	0.945
β_3	1.5	1.494 (0.070)	0.067	(1.357; 1.631)	− 0.006	0.005	0.948
β_4	2	1.996 (0.088)	0.087	(1.823; 2.169)	− 0.004	0.008	0.948
β_5	− 3	− 3.076 (0.408)	0.405	(− 3.875; − 2.277)	− 0.076	0.172	0.946
β_6	1	1.029 (0.344)	0.339	(0.354; 1.704)	0.029	0.119	0.955
Sample size $n = 400$							
β_1	0.5	0.509 (0.057)	0.059	(0.398; 0.620)	0.009	0.003	0.951
β_2	0.5	0.504 (0.061)	0.062	(0.385; 0.623)	0.004	0.004	0.953
β_3	1.5	1.495 (0.049)	0.048	(1.399; 1.592)	− 0.005	0.002	0.949
β_4	2	2.002 (0.060)	0.060	(1.885; 2.119)	0.002	0.004	0.953
β_5	− 3	− 3.039 (0.267)	0.275	(− 3.562; − 2.516)	− 0.039	0.073	0.947
β_6	1	1.013 (0.230)	0.231	(0.563; 1.464)	0.013	0.053	0.947
Sample size $n = 600$							
β_1	0.5	0.509 (0.050)	0.048	(0.412; 0.607)	0.009	0.003	0.947
β_2	0.5	0.504 (0.051)	0.051	(0.403; 0.605)	0.004	0.003	0.946
β_3	1.5	1.499 (0.038)	0.039	(1.424; 1.573)	− 0.001	0.001	0.951
β_4	2	2.002 (0.050)	0.049	(1.904; 2.100)	0.002	0.003	0.954
β_5	− 3	− 3.027 (0.219)	0.219	(− 3.456; − 2.599)	− 0.027	0.049	0.944
β_6	1	1.017 (0.195)	0.187	(0.653; 1.399)	0.017	0.038	0.948
Sample size $n = 800$							
β_1	0.5	0.506 (0.042)	0.041	(0.424; 0.588)	0.006	0.002	0.948
β_2	0.5	0.501 (0.044)	0.043	(0.415; 0.587)	0.001	0.002	0.944
β_3	1.5	1.497 (0.034)	0.034	(1.431; 1.564)	− 0.003	0.001	0.948
β_4	2	2.000 (0.043)	0.042	(1.915; 2.085)	0.000	0.002	0.946
β_5	− 3	− 3.014 (0.187)	0.190	(− 3.381; − 2.647)	− 0.014	0.035	0.946
β_6	1	1.006 (0.155)	0.161	(0.702; 1.309)	0.006	0.024	0.952
Sample size $n = 1000$							
β_1	0.5	0.504 (0.038)	0.037	(0.430; 0.578)	0.004	0.001	0.944
β_2	0.5	0.500 (0.042)	0.039	(0.418; 0.583)	0.000	0.002	0.949
β_3	1.5	1.499 (0.031)	0.030	(1.438; 1.560)	− 0.001	0.001	0.943
β_4	2	2.000 (0.040)	0.038	(1.923; 2.078)	0.000	0.002	0.946
β_5	− 3	− 3.012 (0.168)	0.170	(− 3.342; − 2.682)	− 0.012	0.028	0.951
β_6	1	1.009 (0.143)	0.143	(0.730; 1.289)	0.009	0.020	0.948

SD is the standard deviation calculated from 1000 parameter estimates, SE is the mean of the 1000 estimated standard errors, CI is the mean of the confidence intervals, MSE is the mean of the 1000 estimated mean squared errors, and CP is the empirical coverage probability of the estimated 95% confidence intervals

Table 5 Results of the simulation study for the zero-inflated-censored gamma model adjustment considering different sample sizes

	True	Mean (SD)	SE	CI	Bias	MSE	CP
Sample size $n = 200$							
β_1	0.5	0.539 (0.160)	0.150	(0.224; 0.855)	0.040	0.027	0.940
β_2	0.5	0.506 (0.156)	0.156	(0.199; 0.813)	0.007	0.025	0.943
β_3	1.5	1.456 (0.210)	0.198	(1.043; 1.869)	− 0.044	0.046	0.954
β_4	2	1.996 (0.224)	0.219	(1.556; 2.435)	− 0.004	0.050	0.952
β_5	− 3	− 3.05 (0.420)	0.403	(− 3.882; − 2.232)	− 0.057	0.180	0.950
β_6	1	1.014 (0.348)	0.339	(0.331; 1.698)	0.015	0.122	0.948
Sample size $n = 400$							
β_1	0.5	0.518 (0.105)	0.105	(0.311; 0.726)	0.019	0.012	0.954
β_2	0.5	0.499 (0.105)	0.107	(0.292; 0.707)	0.000	0.011	0.950
β_3	1.5	1.480 (0.140)	0.139	(1.203; 1.756)	− 0.020	0.020	0.943
β_4	2	2.004 (0.148)	0.151	(1.714; 2.295)	0.005	0.022	0.955
β_5	− 3	− 3.05 (0.287)	0.278	(− 3.617; − 2.4905)	− 0.054	0.086	0.941
β_6	1	1.026 (0.235)	0.232	(0.565; 1.487)	0.026	0.056	0.944
Sample size $n = 600$							
β_1	0.5	0.509 (0.084)	0.085	(0.344; 0.674)	0.010	0.007	0.948
β_2	0.5	0.494 (0.088)	0.087	(0.321; 0.667)	− 0.005	0.008	0.941
β_3	1.5	1.487 (0.109)	0.113	(1.273; 1.702)	− 0.012	0.012	0.954
β_4	2	2.007 (0.122)	0.122	(1.767; 2.248)	0.008	0.015	0.952
β_5	− 3	− 3.02 (0.222)	0.222	(− 3.462; − 2.590)	− 0.026	0.050	0.956
β_6	1	1.013 (0.184)	0.186	(0.652; 1.375)	0.014	0.034	0.957
Sample size $n = 800$							
β_1	0.5	0.513 (0.073)	0.074	(0.369; 0.656)	0.013	0.006	0.950
β_2	0.5	0.500 (0.075)	0.075	(0.351; 0.648)	0.000	0.006	0.945
β_3	1.5	1.485 (0.095)	0.098	(1.297; 1.673)	− 0.015	0.009	0.954
β_4	2	1.998 (0.10)	0.105	(1.790; 2.207)	− 0.001	0.011	0.948
β_5	− 3	− 3.02 (0.195)	0.191	(− 3.410; − 2.642)	− 0.027	0.039	0.950
β_6	1	1.011 (0.164)	0.161	(0.689; 1.333)	0.012	0.027	0.950
Sample size $n = 1000$							
β_1	0.5	0.508 (0.066)	0.066	(0.375; 0.638)	0.008	0.004	0.950
β_2	0.5	0.501 (0.067)	0.066	(0.369; 0.633)	0.001	0.005	0.948
β_3	1.5	1.490 (0.086)	0.087	(1.320; 1.661)	− 0.009	0.008	0.948
β_4	2	1.994 (0.096)	0.094	(1.804; 2.184)	− 0.005	0.009	0.950
β_5	− 3	− 3.010 (0.177)	0.170	(− 3.364; − 2.670)	− 0.017	0.032	0.949
β_6	1	1.011 (0.147)	0.143	(0.722; 1.299)	0.011	0.022	0.949

SD is the standard deviation calculated from 1000 parameter estimates, SE is the mean of the 1000 estimated standard errors, CI is the mean of the confidence intervals, MSE is the mean of the 1000 estimated mean squared errors, and CP is the empirical coverage probability of the estimated 95% confidence intervals

Supplementary Information The online version supplementary material available at https://doi.org/10.1007/s42081-021-00124-0.

Funding Not applicable.

Data availability The dataset generated for the applied study is available online on: https://github.com/eduardodefreitascosta/ZIdispersal/blob/master/ZIdispersal/wild_boar.rds Also, it is available in a spreadsheet format in Electronic supplementary material. The codes used to generate the applied dataset are available in the Electronic supplementary material. For the datasets used in the simulated studies, please see code availability subsection.

Code availability All codes for datasets, analysis, and plots are available on GitHub repository: https://github.com/eduardodefreitascosta/ZIdispersal Please, read the README.md file for instructions.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

Bowman, J., Jaeger, J. A. G., & Fahrig, L. (2002). Dispersal distance of mammals is proportional to home range size. *Ecology, 83*(7), 2049–2055.

Breed, M. D., & Moore, J. (2016). Movement: Search, Navigation, Migration, and Dispersal. In: Breed, M. D., & Moore, J. (eds) Animal behavior, chap 8 (2nd ed., pp. 253–292). Elsevier, Amsterdam. https://doi.org/10.1016/B978-0-12-801532-2.00008-8

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution, 7*(6), 679–692.

Casas-Díaz, E., Closa-Sebastià, F., Peris, A., Miño, A., Torrentó, J., Casanovas, R., et al. (2013). Recorded dispersal of wild boar (*Sus scrofa*) in northeast Spain: Implications for disease-monitoring programs. *Wildlife Biology in Practice, 9*(3), 19–26.

Chatwin, T. A., Joy, R., & Burger, A. E. (2013). Set-back distances to protect nesting and roosting seabirds off Vancouver island from boat disturbance. *Waterbirds, 36*(1), 43–52.

Cumming, G. S., Gaidet, N., & Ndlovu, M. (2012). Towards a unification of movement ecology and biogeography: Conceptual framework and a case study on Afrotropical ducks. *Journal of Biogeography, 39*(8), 1401–1411.

De la Torre, A., Bosch, J., Iglesias, I., Muñoz, M. J., Mur, L., Martínez-López, B., et al. (2015). Assessing the risk of African swine fever introduction into the European union by wild boar. *Transboundary and Emerging Diseases, 62*(3), 272–279.

EFSA. (2014). Scientific opinion on African swine fever. EFSA panel on animal health and welfare (AHAW). *EFSA Journal, 12*(4), 3628.

EFSA. (2019). Research gap analysis on African swine fever. *EFSA Journal, 17*(8), e05811.

Gebregziabher, M., Voronca, D., Teklehaimanot, A., & Ana, E. J. S. (2017). Weibull mixture regression for marginal inference in zero-heavy continuous outcomes. *Statistical Methods in Medical Research, 26*(3), 1476–1499.

Guinat, C., Gogin, A., Blome, S., Keil, G., Pollin, R., Pfeiffer, D. U., & Dixon, L. (2016). Transmission routes of African swine fever virus to domestic pigs: Current knowledge and future research directions. *The Veterinary Record, 178*(11), 262–267.

Jackson, C. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software, 70*(8), 1–33.

Jerina, K., Pokorny, B., & Stergar, M. (2014). First evidence of long-distance dispersal of adult female wild boar (*Sus scrofa*) with piglets. *European Journal of Wildlife Research, 60*(2), 367–370.

Keuling, O., Lauterbach, K., Stier, N., & Roth, M. (2010). Hunter feedback of individually marked wild boar *Sus scrofa* L.: Dispersal and efficiency of hunting in northeastern Germany. *European Journal of Wildlife Research, 56*(2), 159–167.

Keuling, O., Sange, M., Acevedo, P., Podgorski, T., Smith, G., Scandura, M., et al. (2018). Guidance on estimation of wild boar population abundance and density: Methods, challenges, possibilities. *EFSA Supporting Publications, 15*(7), 1–48.

Keuling, O., Stier, N., & Roth, M. (2008). Annual and seasonal space use of different age classes of female wild boar *Sus scrofa* L. *European Journal of Wildlife Research, 54*(3), 403–412.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis. Techniques for censored and truncated data. Statistics for biology and health* (2nd ed.). Springer.

Koenig, W. D., Van Vuren, D., & Hooge, P. N. (1996). Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology & Evolution, 11*(12), 514–517.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1–14.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data. Wiley Series in Probability and Statistics*. Wiley.

Lee, A. H., Zhao, Y., Yau, K. K., & Xiang, L. (2010). How to analyze longitudinal multilevel physical activity data with many zeros? *Preventive Medicine, 51*(6), 476–481.

Lewis, J. S., Farnsworth, M. L., Burdett, C. L., Theobald, D. M., Gray, M., & Miller, R. S. (2017). Biotic and abiotic factors predicting the global distribution and population density of an invasive large mammal. *Scientific Reports, 7*(1), 44152–44164.

Liao, Q., Ahmad, Z., Mahmoudi, E., & Hamedani, G. G. (2020). A new flexible bathtub-shaped modification of the Weibull model: Properties and applications. *Mathematical Problems in Engineering, 29,* 110–123.

Liu, L., Strawderman, R. L., Cowen, M. E., & Shih, Y.-C.T. (2010). A flexible two-part random effects model for correlated medical costs. *Journal of Health Economics, 29*(1), 110–123.

Louzada, F., Moreira, F. F., & de Oliveira, M. R. (2018). A zero-inflated non default rate regression model for credit scoring data. *Communications in Statistics Theory and Methods, 47*(12), 3002–3021.

Manning, W. G., Morris, C. N., Newhouse, J. P., Orr, L. L., Duan, N., Keeler, E. B., et al. (1981). A two-part model of the demand for medical care: Preliminary results from the health insurance study. *Health, Economics, and Health Economics, 137,* 103–123.

McClure, M. L., Burdett, C. L., Farnsworth, M. L., Lutman, M. W., Theobald, D. M., Riggs, P. D., et al. (2015). Modeling and mapping the probability of occurrence of invasive wild pigs across the contiguous United States. *PLoS One, 10*(8), e0133771.

McDowell, A. (2003). From the help desk: Hurdle models. *The Stata Journal: Promoting Communications on Statistics and Stata, 3*(2), 178–184.

Meng, X., Lindsay, D., & Sriranganathan, N. (2009). Wild boars as sources for infectious diseases in livestock and humans. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1530), 2697–2707.

Morelle, K., Lehaire, F., & Lejeune, P. (2014). Is wild boar heading towards movement ecology? A review of trends and gaps. *Wildlife Biology, 20*(4), 196–205.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics, 33*(3), 341–365.

Nobre, A. A., Carvalho, M. S., Griep, R. H., Fonseca, M. D. J. M. D., Melo, E. C. P., Santos, I. D. S., & Chor, D. (2017). Multinomial model and zero-inflated gamma model to study time spent on leisure time physical activity: An example of ELSA-Brasil. *Revista de saude publica, 51,* 1–7.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis, 56*(6), 1609–1623.

Paradis, E., Baillie, S. R., Sutherland, W. J., & Gregory, R. D. (1998). Patterns of natal and breeding dispersal in birds. *Journal of Animal Ecology, 67*(4), 518–536.
Podgórski, T., Scandura, M., & Jedrzejewska, B. (2014). Next of kin next door: Philopatry and sociogenetic population structure in wild boar. *Journal of Zoology, 294*(3), 190–197.
Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (2007). *Numerical recipes. The art of scientific computing* (3rd ed.). Cambridge University Press.
Prévot, C., & Licoppe, A. (2013). Comparing red deer (*Cervus elaphus* L.) and wild boar (*Sus scrofa* L.) dispersal patterns in southern Belgium. *European Journal of Wildlife Research, 59*(6), 795–803.
R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
Ramos, P. L., Nascimento, D. C., Cocolo, C., Nicola, M. J., Alonso, C., Ribeiro, L. G., et al. (2018). Reliability-centered maintenance: Analyzing failure in harvest sugarcane machine using some generalizations of the Weibull distribution. *Modelling and Simulation in Engineering, 2018,* 1–12.
Reader, S. (2000). Using survival analysis to study spatial point patterns in geographical epidemiology. *Social Science & Medicine, 50*(7–8), 985–1000.
Sánchez-Cordón, P., Montoya, M., Reis, A., & Dixon, L. (2018). African swine fever: A re-emerging viral disease threatening the global pig industry. *The Veterinary Journal, 233,* 41–48.
Sánchez-Vizcaíno, J. M., Laddomada, A., & Arias, M. L. (2019). *African swine fever virus* (Vol. 25, pp. 443–452). New York: Wiley.
Scillitani, L., Monaco, A., & Toso, S. (2010). Do intensive drive hunts affect wild boar (*Sus scrofa*) spatial behaviour in Italy? Some evidences and management implications. *European Journal of Wildlife Research, 56*(3), 307–318.
Shinohara, S., Lin, Y.-H., Michimae, H., & Emura, T. (2020). Dynamic lifetime prediction using a Weibull-based bivariate failure time model: A meta-analysis of individual-patient data. *Communications in Statistics: Simulation and Computation*. https://doi.org/10.1080/03610918.2020.1855449.
Truvé, J., & Lemel, J. (2003). Timing and distance of natal dispersal for wild boar *Sus scrofa* in Sweden. *Wildlife Biology, 9*(4), 51–57.
Vicente, J., Plhal, R., Blanco-Aguiar, J. A., Sange, M., Podgórski, T., Petrovic, K., et al. (2018). Analysis of hunting statistics collection frameworks for wild boar across Europe and proposals for improving the harmonisation of data collection. *EFSA Supporting Publications, 15*(12), 1–33.
Whitmee, S., & Orme, C. D. L. (2013). Predicting dispersal distance in mammals: A trait-based approach. *Journal of Animal Ecology, 82*(1), 211–221.
Wienke, A. (2011). *Frailty models in survival analysis*. Chapman & Hall/CRC Biostatistics Series, CRC Press.

Authors and Affiliations

Eduardo de Freitas Costa[1] · Silvana Schneider[2] · Giulia Bagatini Carlotto[2] · Tainá Cabalheiro[2] · Mauro Ribeiro de Oliveira Júnior[3]

Silvana Schneider
silvana.schneider@ufrgs.br

Giulia Bagatini Carlotto
gbcarlotto@gmail.com

Tainá Cabalheiro
cabalheirot@gmail.com

Mauro Ribeiro de Oliveira Júnior
mauro.oliveira@unieuro.edu.br

[1] Department of Epidemiology, Bioinformatics and Animal Models, Wageningen Bioveterinary Research, Lelystad, The Netherlands

[2] Department of Statistics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

[3] Centro Universitário UNIEURO, Brasília, Distrito Federal, Brazil