DALARNA UNIVERSITY

Working papers in transport, tourism, information technology and microdata analysis

Estimating zones of influence using threshold regression

Youngjo Lee

Moudud Alam

Per Sandström

Anna Skarin

Editor: Hasan Fleyeh

Nr: 2020:01

Working papers in transport, tourism, information technology and microdata analysis
ISSN: 1650-5581

Estimating zones of influence using threshold regression.

Youngjo Lee[1], Moudud Alam[2], Per Sandström[3], and Anna Skarin[4]

[1]Department of Statistics, Seoul National University, Seoul, Republic of Korea

[2]School of Technology and Business Studies/Statistics, Dalarna University, Falun, Sweden

[3]Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden.

[4]Department of Animal Nutrition and Management, Swedish University of Agricultural Science, Uppsala, Sweden

Date: 05-04-2020

Abstract

In environmental impact assessments, it is important to be able to estimate influence of anthropogenic activities on animal populations. To quantify the influence, it is common to estimate how far, in distance, from a given disturbance source there is an influence on the animals' habitat selection through estimating a zone of influence (ZOI). Usually, ZOI is estimated for one disturbance source at a time. In this work, we demonstrate how threshold regression modelling can be used for estimating ZOI from several possible sources of disturbances, simultaneously. Based on the theoretical properties of different estimation methods for the estimation of threshold regression we select a set of estimation methods and compare their merits through a simulation study and a real data example. The simulation results revealed that Adaptive Lasso, and Hierarchical likelihood (HL) methods, are two reasonable methods for dealing with the problem. HL performed better than Adaptive Lasso in that it had much higher success rate in identifying correct threshold with small sample size whereas Adaptive Lasso requires large sample to assure good performance. While Adaptive lasso needed to be aided with suitable weights, which are not easy to find, HL method did not require any prior weights. These two methods were applied to estimate the ZOI around 40 wind turbines and surrounding public roads on reindeer habitat selection in winter, by using GPS positioning data from 42 reindeer in north of Sweden in December to March (2012-2015). The results showed that both the disturbance sources have a negative effect on reindeer habitat selection in winter. The HL approach showed that the negative ZOI from the nearest wind turbine was 1.8 km (approx.), however the trend of higher selection of areas further away from the wind turbines was evident up to 4 km (approx.) from the active wind turbines.

Key words: Reindeer, Cumulative effect, effect threshold, zones of influence, threshold regression, penalized likelihood, model selection.

1. Introduction:

In environmental impact assessment of important animal habitats, it is often of interest to estimate the zones of influence (ZOI) caused by possible disturbances (e.g. establishment of wind farms, mining sites, etc.) on local animal populations (e.g. Boulanger et al. 2012; Ficetila and Daoel, 2009). In many situations the impact of an increasing number of, same or different, potentially disturbing objects are of interest to assess; often referred to as cumulative impact assessment (Johnson et al., 2005; Walker and Johnston, 2009, Gillingham et al. 2016).

This study is motivated by the problem of cumulative impact assessment of anthropogenic activities on semi-domesticated but free-ranging reindeer (*Rangifer tarandus tarandus*) in the Sami reindeer husbandry area in Sweden. Over the last century anthropogenic activities (e.g. road traffic, mines, roads, power grid, etc.) have developed over the time and contributed to the habitat loss for the reindeer population (Kivinen et al. 2015; Vistnes and Nellemann 2008; Skarin & Åhman 2014). A fairly new emerging source of disturbance are construction of wind farms to foresee a production of renewable energy, which have shown to have adverse effect on reindeer habitat selection (Skarin et al. 2018; Skarin and Alam 2017). To be able to manage the increased level of infrastructure in the landscape, it is vital to know at which distance does the combined effects (we call it "cumulative effect") of all the locally available infrastructures on reindeer habitat use vanishes.

Massive expansion of wind farms is a relative recent phenomenon (Sawyer, 2017), and there has been quite a few studies attempting to estimate their effects on large animals, e.g. reindeer. Recently, Tsegay et al. (2017) studied the effect of wind farm on reindeer habitat use at a Norwegian island, and did not find any negative effect. However, whether their findings (Tsegay, 2017) can be extended to other landscapes than the specific offshore setting is unclear. Skarin and Alam (2017) studied the effect of wind farm on reindeer habitat selection in calving season in an area with small forested mountains in northern Sweden and found significant negative effect, however, they did not give proper attention to precisely estimate the zones of influence. Research on developing and evaluating analytical methods for precise detection of ZOI is scarce in the literature, with Boulanger et al. (2012), and Ficetola and Danoel (2009) being a couple of rare examples.

In estimating zones of influence, the concept of ecological threshold (Holling, 1973; Hugget, 2005), and analytical procedures developed therein (see e.g. Ficetola and Danoel, 2009) are used (Boulanger et al., 2012). Under this framework, the estimation of the zones of influence is often carried out by repeatedly fitting some kind of piecewise regression model (Ficetola and Danoel, 2009). The disturbance effect is modelled as linear (possibly after some transformation) with respect to the distance from the source of disturbance, and is assumed to vanish abruptly at some point in distance (giving the boundary of ZOI). This requires a piece wise linear regression model formulation. The point of intersection of two pieces of a broken regression line, one piece of line from the source of disturbance up to the boundary of ZOI and a flat line afterwards giving no effect, is chosen on the basis of the maximum (log-) likelihood (see e.g. Boulanger et al. 2012; Hudson, 1966), or measure of sensitivity/specificity (see Ficetola and Danoel, 2009, and references cited therein) through a grid search, over the possible boundaries of ZOI. This requires repeated fitting of a series of models. So far, this approach is mainly used in the situation where only one source of disturbance is of concern (Boulanger et al. 2012), or the different sources are merged to create one factor (e.g. Polfus et al., 2011). The model computation can quickly get burdensome as number of influential factors increases, because the grid search of thresholds has to be carried out for different combinations of the threshold parameters. Drawing inference on the boundary of the influence zone (or intersection points) remains as a big challenge (Hinkley, 1969) and the problem gets even worse for correlated data, e.g. in presence of spatial correlation.

While estimating ZOI, it is often ignored that the impact of anthropogenic activity on surrounding environment may be non-linear, and may not disappear abruptly at certain distance. Therefore, a smoothed regression (such exponential or smoothed polynomial) is also used (Nielsen 2009), instead of a piecewise regression with a single cut point (as used in Boulanger et al. 2012). However, exponential decay rates are often subjectively selected (Nielsen et al., 2009), to get rid of non-linearity in the model parameters. In comparison between several models Ficetola and Danoel (2009) found, using a simulation study, that generalized additive model (GAM) and piecewise regression both works fine in detecting change point (threshold) when there is one abrupt change in response at a certain point in one covariate.

Even though, it is possible to estimate the ecological threshold parameter objectively (without requiring grid search; see e.g. Muggeo, 2003), it is very hard to motivate why the effect of disturbance should be linear and vanish abruptly (at a single threshold), or why it should decay according to a specified curvature, e.g. negative exponentially. To overcome these problems we suggest a multiple piecewise regression following the framework of threshold regression (see e.g. Ng et al. 2018, Hansen, 2017, Chan et al. 2017). Further, it is straightforward to add random effects to threshold regression while no previous study has considered random-effects model in ZOI estimation, even though random effects models are frequently used in resource selection model (Gilles et al. 2006; Skarin and Alam, 2017).

Our work has investigated the usefulness of regression threshold method in estimating cumulative ZOIs. Using the piecewise (or segmented) regression model we converted the problem of ecological threshold estimation into a variable selection problem. However, the identification of the effect threshold, from the data, is still a challenging task. Often, the shrinkage (or penalized) estimation methods (Zhou, 2006) are used for the threshold model estimation (Chan et al. 2017), but such techniques have to be applied with due caution because threshold regression model often produce a highly collinear model matrix. To overcome this challenge, we investigate two options. First, we explore the usefulness of the additional information that the effects of disturbance diminishes with distance from the disturbance, and eventually wiped off at some long distance, as the weights (Zhou, 2006) in the shrinkage estimation. Second, we use hierarchical likelihood (HL; Lee et al. 2014) approach that incorporates sharper penalty than the traditional shrinkage methods, e.g. Ridge regression, Lasso, and Adaptive Lasso (Zhou, 2006). A fully probabilistic model specification of HL approach makes it easy to incorporate any random effects in the model, and allows us to draw inference in a single phase estimation while the traditional shrinkage method require two phase estimation, i.e. the model subset selection, and drawing inference by refitting the selected model (often referred to as selective inference; Taylor and Tibshirani, 2015).

We compared the performance of HL, and other alternatives namely Lasso, elastic net, and Adaptive Lasso, using simulation study. It is well known that in presence of high correlation between covariates many of the model selection techniques do not work well (see e.g. Zhou and Hastie, 2005). We investigate this issue in the simulation study. We also present a real data example by estimating the ZOI of windfarm, and road on the reindeer winter habitat selection by apply the best performing model, as per the simulation study, to the real data on reindeer GPS positions collected from reindeer in a forested winter grazing range in northern Sweden. In modelling the reindeer habitat selection, using GPS-data, we utilize the framework of resource selection model (Johnson et al. 2005; Gilles et al. 2006).

The article is organized as follows. Section 2 presents the statistical models, and their estimation method, Section 3 presents a comparison of the estimation methods via a simulation study. Section 4 presents a real data application of the models and methods presented in Section 2. Section 4 offers a discussion of the results.

2. Estimation of the ZOI

We present a modelling framework of the multiple threshold model in the context of ecological threshold estimation. We motivate why certain estimation technique can consistently estimate the threshold, while many other cannot.

2.1 Random effects threshold regression model for estimating ZOI

Assume we want to estimate the effect of K (known) disturbance variables $Z_k; k = 1,2, \dots, K$ (e.g. distance from a wind farm, distance from a mine, distance from the nearest road, etc.) on an ecological outcome variable Y (e.g. habitat selection). For a resource selection model, Y is a binary response variable with 1 represents actual observed location (GPS-position), and 0 represent randomly selected location within the home range of the GPS-tagged animal. Further, we assume it necessary to control for the effects of other P confounding variables $X_p; p = 1,2, \dots, P$. Then, following the framework for generalized linear models (GLM; McCullagh and Nelder, 1989), we can formulate the effect of covariates and disturbances on Y as

$$g\big(E(Y_i)\big) = \boldsymbol{X}_i\boldsymbol{\beta}_1 + h(\boldsymbol{Z}_i, \boldsymbol{\psi})\boldsymbol{\beta}_2 \text{ ; i=1,2,…,n} \qquad (1)$$

where g is a link function (e.g. for a resource selection model, g is a logit function leading to a logistic regression mode for the binary outcome, Y_i), $\boldsymbol{X}_i = \big(X_{1,i}, \dots, X_{P,i}\big)$, $\boldsymbol{Z}_i = \big(Z_{1,i}, \dots, Z_{K,i}\big)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$ are regression parameters, $\boldsymbol{\psi} = (\psi_1, \dots, \psi_K)^T$ are the threshold parameters which are parameters of primary interest, and h is a function determining smooth or abrupt disappearance of the effects of the disturbances. Keeping in mind that the GPS-data will be analysed using a resource selection model (RSF, i.e. logistic regression), throughout this paper, we limit our focus on binary Y_i, and a logistic link function.

If we let $h(\boldsymbol{Z}_i, \boldsymbol{\psi}) = \Big(Z_{1,i}, \dots, Z_{K,i}, \big(Z_{1,i} - \psi_1\big)_+, \dots, \big(Z_{K,i} - \psi_K\big)_+\Big)$ where $\big(Z_{k,i} - \psi_k\big)_+ = \big(Z_{k,i} - \psi_k\big).I\big(Z_{k,i} > \psi_k\big)$, with $I(.)$ being an indicator function, and $\boldsymbol{\beta}_2 = \big(\beta_{2,1,1}, \beta_{2,1,2}, \dots, \beta_{2,1,K}, \beta_{2,2,1}, \beta_{2,2,2}, \dots, \beta_{2,2,K}\big)^T$ then model (1) reduces to usual threshold regression model (e.g. the ones used in Muggeo, 2003). Further, assuming $h(\boldsymbol{Z}_i, \boldsymbol{\psi}) = \Big(min\big(Z_{1,i}, \psi_1\big), \dots, min\big(Z_{K,i}, \psi_K\big)\Big)$ and $\boldsymbol{\beta}_2 = \big(\beta_{2,1}, \beta_{2,2}, \dots, \beta_{2,K}\big)^T$ we get usual ecological threshold model (e.g. the one used in Boulanger, 2012). Depending on the sampling scheme, it might be necessary to add additional terms (e.g. random effects to account for spatial correlation) to the right hand side of equation (1).

In application, if the threshold is estimated within a certain m units (e.g. $m = 50$ meter) of accuracy, this error is negligible. Further, we may limit our investigation within some "b" meter from the boundary of a study area. Under these considerations, we present a different representation of model (1) which does not assume any linear or smooth cancellation of the effect of $Z_k's$.

$$g\big(E(Y_i)\big) = \boldsymbol{X}_i\boldsymbol{\beta}_1 + \sum_{k=1}^{K} \widetilde{\boldsymbol{Z}}_{k,i}\boldsymbol{\beta}_{2,k} \qquad (2)$$

Where $\widetilde{\boldsymbol{Z}}_{k,i}$ is the k:th element of $\widetilde{\boldsymbol{Z}}_i = \Big(Z_{k,i}, \big(Z_{k,i} - m\big)_+, \big(Z_{k,i} - 2m\big) \dots, \big(Z_{k,i} - q_k m\big)\Big)$, and q_k is subjectively chosen so that any effect after $q_k m$ units of distance from the disturbance is uninteresting (if not impossible to exist, e.g. within 50 meter from the boundary of the study area). Associated parameter vector $\boldsymbol{\beta}_2$ is a $(q_k + 1) \times 1$ vector of parameters such that many of its elements are supposed

to be 0. In particular, after a certain threshold p_k^* $(1 < p_k^* < q_k + 1)$ we have $\beta_{2,k,p_{k+1}^*} = \beta_{2,k,p_{k+2}^*} = \cdots = \beta_{2,k,q_k} = 0$. With this model specification, the estimation of ecological threshold converts into a problem of model selection, which can be handled by using existing techniques for regression subset selection such as Adaptive Lasso (Zhou, 2006), and HL approach (Lee et al. 2015).

The idea of converting a regression threshold estimation problem into a model selection problem is not new (see e.g. Chan et al., 2015, and Ng and Lee 2018). Though the threshold regression models are widely used in time series analysis (see articles featured in a special issue of the Journal of Business and Economic Statistics, 2017, Vol. 35, Issue 2), this approach has never been explored in the area of ecological threshold estimation. Unlike time series analysis, in ecological threshold estimation, we have additional (prior) information that the disturbance-effect eventually vanishes at some long distances from the source of disturbance (e.g. close to the boundary of the study area). In the following section, we explore the usefulness of the additional information, naturally available in studies on estimation of the zones of influence.

2.2 Model fitting and inference

From model (2) we see that the successive columns of $\widetilde{\boldsymbol{Z}} = \{\widetilde{\boldsymbol{Z}}_{k,i}\}$ are almost identical to each other except that a certain number of elements in the receding column are restricted to 0. Therefore, the estimation of model (2) is a challenging task because the columns of $\widetilde{\boldsymbol{Z}} = \{\widetilde{\boldsymbol{Z}}_{k,i}\}$ becomes highly correlated. In a similar problem, Chan et al. (2015) used grouped Lasso (same as Adaptive Lasso), and Ng and Lee (2018) used modified smoothly clipped absolute deviation (SCAD) approach (which is a special case of HGLM). The Adaptive Lasso estimates the model parameter by optimizing the following objective function

$$L_{AL} = l(\boldsymbol{\beta} \,,, \phi; y) + \lambda_n \sum_{j=1}^{P+K} w_j |\beta_j| \qquad (3)$$

where l is the likelihood function of model (2), is w_j is some known weight, and λ_n is a penalty parameter. Zhou (2006) suggested w_j to be replaced by $w_j = 1/|\tilde{\beta}_j|^\gamma$ where $\tilde{\beta}_j$ is any root-n-consistent estimate of β_j, such as ordinary maximum likelihood estimator, or ridge estimator, and γ is a constant. With $w_j = 1 \forall j$, Adaptiv Lasso reduces to ordinary Lasso. Adaptive Lasso approach may not be applicable for model (2) because it would be a highly challenging task to find a numerically stable estimate of β_j when some columns in the design matrix are highly correlated.

In model (2), parameter selection in $\boldsymbol{\beta}_1$ part is uninteresting. Therefore, we propose only the parameter $\boldsymbol{\beta}_2$ to be penalized, for which we may choose $w_j \propto q_j m$. Intuitively, with $q > l$, the parameters associated q^{th} column of $\widetilde{\boldsymbol{Z}}_i$ receives higher penalty than the one associated with l^{th} column of $\widetilde{\boldsymbol{Z}}_i$, to ensure that the effects have higher chance of being wiped off at longer distance from the source of the disturbance.

The Adaptive Lasso method can be motivated from a Bayesian perspective in that we are using Laplace prior for β_j's. Still, the parameters γ and w_j have to be chosen subjectively, and λ_n has to be estimated in some ad hoc method, e.g. via cross validation. Further, if the weights, w_j's, are not chosen in a data driven way, Zhou's (2006) argument about the oracle property of Adaptive Lasso, does not readily apply (see, Leng et al., 2006). We try to overcome these limitations, by using HL approach, as follows.

HL approach for variable selection is proposed by Lee and Oh (2014), and Lee and Lee et al. (2015), and applied successfully for threshold regression model (in time series context) by Ng et al. (2018). In

HL approach, we consider $\boldsymbol{\beta}_2$ as random parameter (similar to Bayesian approach), but we set up the distribution of each element of $\boldsymbol{\beta}_2$ as follows.

$$\beta_j \sim N(0, \theta \quad \frac{u_j}{w_j}) \text{ and } u_j \sim Gamma(\omega) \tag{4}$$

where θ and ω are parameters and w_j is a known (prior) weight. Unlike Adaptive Lasso, HL approach is developed on the basis of a fully specified probabilistic model. A logistic regression model with the random effects, $\boldsymbol{\beta}_2$, becomes a hierarchical generalized linear model (HGLM; Lee and Nelder, 1996). Parameter ω may be selected by some ad hoc method (e.g. through cross validation) but any large value e.g. $\omega = 30$ works fine (Lee and Oh, 2014). The remaining parameters in (2) and (4) can be estimated by profile likelihood method (Lee et al, 2015). However, it is also possible to include additional weights w_j to disproportionately penalize $\beta_{2,j}$ parameters. We may use the same weights, w_j, as proposed for Adaptive Lasso but there is a problem that the scale of w_j can be consumed by the free parameter θ . Therefore, we propose a scale free weights such as $w_j^* = \frac{w_j}{\sum w_j}$ for HL approach. Further, we can include an independent Gaussian random effects, α_i $(i = 1,2, \dots, a)$, in model (2) with view to modelling within subjects (intra-class) correlation in response y, as well as adjusting the inference for unequal observations on different subjects (Skarin and Alam 2017). The random effects extend model (2), for i:th subject as

$$g\big(E(\boldsymbol{y}_i)\big) = \boldsymbol{X}_i\boldsymbol{\beta}_1 + \widetilde{\boldsymbol{Z}}_{1,i}\boldsymbol{\beta}_2^* \quad + \boldsymbol{Z}_{2,i}\boldsymbol{\alpha} \tag{5}$$

where $\boldsymbol{y}_i$ is an $n_i \times 1$ vector of responses correspond to subject i, $\boldsymbol{X_i}$, $\widetilde{\boldsymbol{Z}}_{1,i}$ $\boldsymbol{Z}_{2,i}$ are design matrices of conformable dimensions, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_a)^T$ is a vector of random effects, and we assume $\boldsymbol{\alpha} \sim \boldsymbol{N}(0, \tau\mathbf{I})$. Combining (4) and (5) we have the following joint log-likelihood (we call it hierarchical likelihood, or h-likelihood) function

$$h = \sum_{i=1}^{n} log\big(f(\boldsymbol{y}_i|\boldsymbol{u}, \boldsymbol{\alpha}; \boldsymbol{\beta}, \phi, \theta)\big) + \sum_{j=1}^{K} log\left(f\big(\beta_{2,j}|u_j; w_j^*, \theta\big)\right) + log\big(f(\boldsymbol{\alpha}; \tau)\big) + \sum_{j=1}^{K} log\left(f\big(u_j; \omega\big)\right) \tag{6}$$

From (6) we can estimate u_j by solving $\frac{\partial h}{\partial u_j} = 0$ which gives

$$u_j = \omega \frac{\left(\frac{2}{\omega} - 1\right) + \sqrt{\frac{8\beta_{2,j}^2 w_j^*}{\omega\theta} + \left(\frac{2}{\omega} - 1\right)^2}}{4} \tag{7}$$

Substituting u_j, in (6) we get a profile likelihood which is, with a known ω, a h-likelihood function of a binomial HGLM with Gaussian random effects $\boldsymbol{\beta}_2$, and $\boldsymbol{\alpha}$. Therefore, the IWLS algorithm of Lee and Nelder (1996) can be used for estimating the remaining parameters, for fixed u_j (above) and ω, leading to a two-step procedure: i) Solve equation (7) for u_j, and ii) use an IWLS for the other parameter estimation. It is necessary to iterate between these two steps, until convergence (R codes implementing this procedure is accompanied with this article, as a supplementary material). To assure numerical stability of the estimation procedure, any small u_j (we used the threshold $u_j < 1e - 10$) are replaced with a very small but non-zero number $(1e - 10)$, as suggested by Lee and Oh (2014).

Lee and Oh (2014) and Lee at al. (2015) showed a penalized likelihood representation of HL approach for model selection, and established its oracle property which readily applies here. Adaptive Lasso also enjoys the oracle property (Zhou, 2006) but in absence of root-n-consistent estimator of the model parameters, oracle property for Adaptive Lasso may not hold.

3. Simulation

The performance of the HL, Adaptive Lasso, Lasso, and Elastic net were assessed and compared in a simulation study, using a binomial family generalized linear model (GLM; McCullagh, and Nelder, 1989). We use a logistic link function and the following linear predictor (equation 8).

$$\eta_i^* = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 (d_i - 1.5)_+ + \beta_4 (d_i - 2.5)_+ \quad (8)$$

where d_i represents the distance from the disturbance (e.g. distance from wind turbine) in a range between 0 and 6 (equally spaced), $i = 1, \dots, n\ (= 500, 1000, \text{and } 5000)$, $\beta_0 = -1.5$, $\beta_1 = 0.8$, $\beta_2 = 1.8$, $\beta_3 = -1$, $\beta_4 = -0.8$. We simulate x_i independently from the standard normal distribution, which mimics a covariate (e.g. an environmental variable in the applied study). In a typical study on cumulative ZOI, there will be several covariates and more than one source of disturbance. However, for simplicity, we begin with two covariates, one non-distance covariate, x, and one representing distance from the source of disturbance, d. Then, we fit a logistic model with the simulated data with the following link function

$$\eta_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 (d_i - 1.5)_+ + \beta_4 (d_i - 2.5)_+ + \beta_5 (d_i - 3.5)_+ + \beta_6 (d_i - 5)_+ \quad (9)$$

If the model fitting method consistently select a correct threshold then it should give $\beta_5 = \beta_6 = 0$. We compared the performances of the different methods in term of success ratio in catching correct threshold, i.e. estimate $\beta_4 > 0$ but $\beta_5 = \beta_6 = 0$, out of 2000 Monte Carlo replication.

We performed the simulation study in **R** (R Core Team, 2017), using "**glmnet**" (Friedman et al., 2010) library for fitting GLM with Lasso, Adaptive Lasso, and Elastic net methods. For HL approach, we used our own programme written in **R**. In Adaptive Lasso we did not penalize β_0, and β_1, but used weight factors w = (0.75,1.5,2.5,3.5,5) for the parameters $(\beta_2, \dots, \beta_6)$. The simulation results are presented in Table 1.

Table 1. Success rates of Adaptive Lasso and Elastic net in catching the correct threshold

Method	**Success rate in catching correct threshold**		
	Sample size (n=500)	n=1000	n=5000
H-likelihood (w=30)	90%	98%	98%
Adaptive Lasso	51%	71%	91%
Elastic net (α =0.95)	40%	40%	38%

Note: We treated |estimated coefficient| <1e-5 as 0.

The simulation results revealed that only HL, and Adaptive Lasso were able to catch the correct threshold, consistently (as sample size grows). Elastic net performed its best with $\alpha = 0.95$, yet worse than adaptive Lasso. In the HL approach, we tried both fixed weight ($w_j = 1$, and $\omega = 30$) and varying weights, w = (0.05,1.5,2.5,3.5,5)/12.55, both gave numerically indistinguishable results. Overall, the HL

approach was found to be the best, outperforming Adaptive Lasso by a big margin, especially for small sample sizes (Table 1).

We also studied the bias in estimating model parameters, even though the effect threshold is the main concern in ZOI estimation. First, we considered bias in $\sum_{j=2}^{6}\beta_j$, because $\sum_{j=2}^{6}\beta_j = 0$ tells us that the effect vanish within the boundary of the study area. In this perspective, HL showed the lowest, bias followed by Adaptive Lasso. We also checked unbiasedness of each of the threshold parameter $(\beta_2, \dots, \beta_6)$ estimator and we found that all the methods were biased, especially with small sample sizes (n=500), with Adaptive Lasso showing the lowest bias, followed by HL, showing a downward bias. Elastic net and HL also underestimated these parameters. With increasing sample size, the bias of HL diminished and at n=5000, HL showed lower bias than Adaptive Lasso.

For HGLM, we carried out the same simulation study by adding a random effect in the linear predictor, modifying the linear predictor (equation 8) as follows:

$$\eta_{i,j} = -1.5 + 0.8x_{i,j} + 1.8z_{i,j} - 1.0\left(z_{i,j} - 1.5\right)_+ - 0.8\left(z_{i,j} - 2.5\right)_+ + \alpha_i \qquad (9)$$

where $i = 1,2,,\dots,30$, $j = 1,2,\dots,n\ (= 30 \text{ and } 120)$, $x_{i,j} \sim N(0,1)$, $\alpha_i \sim N(0,1.2)$, and $z_{i,j}$ was simulated from $N(5.1,1.9)$ but any $z_{i,j} < 0$ was recoded as 0 and $z_{i,j} > 10$ as 10. Left and right censoring of $z_{i,j}$ was done by keeping the structure of the real data (see Section 4) in mind, in particular that the distance cannot be negative and large $z_{i,j}$ value might create numerical problem in the logistic HGLM.

The success rate for the fitted HGLM, with a linear predictor containing all the terms in equation (9) plus two additional terms $\left(z_{i,j} - 3.5\right)_+$, and $\left(z_{i,j} - 5\right)_+$ in identifying correct threshold, was 86.95% for $n = 30$, and 96.65% for $n = 120$, respectively, which was slightly worse compared to the GLM (Table 1). The same simulation study was not possible to carry out for the Adaptive Lasso and Elastic Net because **glmnet** package does not allow random effects in the model specification, nor does any other **R** package. Therefore, detailed results from this simulation study are not reported.

4. A real data example: Effects of wind turbines and public roads on reindeer habitat selection in a winter grazing area.

We analysed real location data on reindeer, collected through 61 reindeer, equipped with GPS-collars, using a coastal-mountain-forest area in northern Sweden (Gabrielsberget, Nordmalings municipality, Västerbotten county). The area is located within Vilhelmina Norra and Vapsten reindeer herding communities and is used in winter by Byrijke, a Norwegian reindeer-herding district, as an agreement of exchange of pastures between the three parties (Skarin et al. 2016). Forty wind turbines were erected during 2011-2012. Reindeer GPS-data were collected after construction during three winter seasons, December to March in 2012/2013, 2013/2014, and 2014/2015.

We developed RSFs (Manly et al., 2002; Gillies et al. 2006) models with a use-availability design, using binomial family generalized linear mixed models, evaluating whether the wind turbines in the wind farm and the public roads affected reindeer habitat selection within the grazing area. During part of the study period reindeer were fed with supplementary feeding in the wind farm and as the reindeer were likely to move differently during feeding, positions collected during this time period were omitted in further analysis. The GPS-data were collected with different time intervals (as the main purpose of the collars was to support reindeer herding actions in the area). Reindeer with very few observations (<20 positions) were also removed and after only allowing for a regular time interval (one position/day) between

positions we used 1325 positions from 17, 15, and 10 individual reindeer from each winter, respectively. To estimate the RSF, we generated available points using a 1:1 ratio of used to available locations, the reindeer available area defined by the reindeer herder surveillance border of the animals. Resulting in 2648 observations, including both the observed and random locations. The habitat variables joined to each location included: slope in degrees, ruggedness index, elevation (m), the nearest distance (m) from a wind turbine (DW), the nearest distance (m) to public roads (DR) (width > 5m), vegetation type, predicted lichen cover, and whether the turbines where visible or not in relation to the topography. Two observations were deleted due to missing value in vegetation type. The wind turbines were not running between December 14, 2012 and January 22, 2013, we thus split the locations based on the wind farm running (WR) or not in interaction with distance to the wind turbines.

Initially, we fitted a logistic mixed model, without applying any threshold on the distance variables. In fitting the logistic regression model, we standardize the elevation, slope, and ruggedness index, to remove any numerical problem due to heterogeneous scale of the covariates. We also used square root transformation of the distances (measured in 100 m, to be consistent with Skarin and Alam, 2017). In the mixed model, we used animal effect as the random effects to account for correlation within animals and to adjust the inference for the unequal observations on each animal. However, the variance of the random effect estimate (by glmer function of lme4 library) turned out to be 0 which indicated no need for the random effects. The model was therefore refitted with a simple logistic model (ignoring random effects) (Table 2). The exact estimate of the effects of all covariates are not of primary interest for this study, except that we want to control for their effects on reindeer winter habitat selection, while estimating any effect of the wind farm and any other human disturbances of major interest.

Table 2. Results from the logistic regression model fitted without putting any threshold on distance variables

Coefficients/statistics	Estimate	Standard error	p-value
Intercept	3.26	0.60	<0.01
Standardized elevation	0.33	0.07	<0.01
Standardized slope	0.11	0.06	0.10
Standardized ruggedness index	0.07	0.06	0.30
Standardized vegetation index (LAV)	0.50	0.05	<0.01
Distance from nearest wind turbine (DW)[1]	-0.35	0.09	<0.01
Dummy for wind turbines being active (WR)	-0.72	0.28	0.01
Distance from road (DR)[1]	-0.33	0.05	<0.01
Dummy variable for wind turbine visibility			
No view	-	-	-
View open (VO)	-1.58	0.61	0.02
View cover (VC)	-1.40	0.49	<0.01
Interaction between DW and WR	0.17	0.06	0.01
Interaction between DW and VO	0.25	0.11	0.02
Interaction between DW and VC	0.30	0.08	<0.01
Residual deviance	3323.0		
Residual degrees of freedom	2635		

[1] Note: All the distance measures (in 100 m) were square root transformed.

Except for the effect of DR, the signs of all the significant coefficients in Table 2, agree with previous findings, obtained with data from other areas and with different data collection methods (see e.g. Skarin and Alam, 2017). The reindeer preference for areas close to the roads could be due to possible non-linearity in the effect, which can be better investigated using a piece-wise regression (Table 3). The positive (and significant) interaction between distance to wind farm and whether it was running

(DW:WR) indicates that the reindeer preferred to stay away from wind farms when the wind turbines were active. A third order interaction (DW:WR:Visibility dummies) were tried but the coefficient estimates turned out to be insignificant (p-value > 0.8), and dropped from the final model (Table 2). Subsequently, we estimated the effect threshold for wind farm and road, using HL, and Adaptive Lasso method (Table 3).

For the two distance variables, DW and DR, we considered effects threshold estimated with a precision of about 100 m around the source of disturbance as acceptable. Therefore, we set m = 1 in equation (2) for square root distance. However, because we used square root transformation of the distance variables in model fitting (Table 3), the second thresholds (2m = 2) correspond to 400 m from the source of disturbance, third threshold corresponds to 900 m, and so on. In fitting Adaptive Lasso, we took the estimate of the coefficient from the ordinary logistic model (Table 2) as the penalty factor for coefficients associated with non-distance variables, β_1. For β_2 parameters, we consider the effects thresholds as the penalty factors (as was done in the simulation study). Hyper parameter in Adaptive Lasso is selected by using 10-fold cross validation. In HL method, we did not penalize the β_1 parameters but we penalized β_2 with fixed hyper-parameter $\omega = 30$.

In presence of interaction terms (Table 3), the interpretation of the model parameters were complicated. However, comparing the estimates between Adaptive Lasso and HL we see that HL estimated more threshold parameters as 0. Possibly explained by HL setting a sharper penalty than Adaptive Lasso (Lee and Oh, 2014). The deviance and the Pearson's Chi-squared statistics of the threshold logistic model evaluated at the estimates from the HL method (3253.54, and 2663.01, respectively) were found smaller than the same statistics (3323.00, and 2697.15) from the ordinary logistic regression (Table 2), which indicated that the threshold regression model fitted better with the data than the ordinary logistic model. The deviance of the logistic model evaluated at the estimates from the Adaptive Lasso method was found to be 3273.04 which is larger than the same from the HL method, but the Pearon's Chi-squared statistics of the Adaptive Lasso (2620.46) was smaller than the HL model.

The many 0 estimates of the threshold parameters given by the HL approach, make it easy to compute a ZOI. To identify whether a running wind farm has any effect on reindeer resource selection, compared to when it is not running, the parameters associated with the dummy variable WR and its interaction with the distance from wind farm (DW) and its segments, $(DW\text{-}1)_+$.... $(DW\text{-}9)_+$ (Table 3). In particular, the solution of the equation $-1.3039 + 0.3024\sqrt{\frac{x}{100}} = 0$ giving $x = 1858.7m$ is the estimated effect threshold of the running wind farm, comparing to when it is not running and given everything else remains constant. In other words, given everything else is fixed, the negative effect of running windfarm on reindeer habitat selection wipes off at about 1.9 km from the wind farm. To visualize this fact, we plot the logarithm of odds-ratio (OR) of the distance to wind farm, as found from the HL method, for both the periods when the wind farm was running and when the wind farm were not running, assuming everything else is fixed (Figure 1).

Only one coefficient related to distance to road ("$(DR\text{-}3)_+$", in Table 3) was found no-zero, and negative. This implies that reindeer habitat selection within 900 m from the road were not associated with the distance from road, given that everything else remains constant, and the reindeer's habitat selection was negatively associated with the distance to roads beyond this point.

Table 3. Results from a threshold logistic model fitted using HL and Adaptive Lasso method

Covariates	Estimate (Std. err. in parenthesis)	
	Adaptive Lasso	**HL**
Intercept	-1.37	0.80 (0.170)
Standardized elevation	0.90	0.08 (0.080)
Standardized slope	0.13	0.10 (0.064)
Standardized ruggedness index	0.09	0.10 (0.065)
Standardized vegetation index (LAV)	0.42	0.45 (0.051)
Dummy for wind turbines being active (WR)	-0.43	-1.30 (0.221)
Dummy variable for wind turbine visibility (V)		
No view	-	-
View open	-0.04	0.07 (0.204)
View cover	1.41	0.41 (0.125)
Distance to wind farm (DW)[1]	0.35	.
$(DW-1)_+$	-0.25	.
$(DW-2)_+$	0.59	.
$(DW-3)_+$	0.07	.
$(DW-4)_+$	-1.02	.
$(DW-5)_+$	.	-0.31 (<0.001)
$(DW-6)_+$	-1.28	-0.81 (0.162)
$(DW-7)_+$	.	.
$(DW-8)_+$	-2.00	.
$(DW-9)_+$	-0.49	.
Distance to road (DR)[1]	0.76	.
$(DR-1)_+$	.	.
$(DR-2)_+$	-1.15	.
$(DR-3)_+$	-0.55	-0.58(0.055)
$(DR-4)_+$	0.35	.
$(DR-5)_+$	-0.14	.
$(DR-6)_+$	0.20	.
$(DR-7)_+$	0.29	.
Interaction terms[2]		
DW:WR	0.00	0.30 (0.043)
$(DW-1)_+$	-0.11	.
$(DW-2)_+$	.	.
$(DW-3)_+$	0.53	.
$(DW-4)_+$	.	.
$(DW-5)_+$	-0.48	.
$(DW-6)_+$	1.68	.
$(DW-7)_+$	.	.
$(DW-8)_+$	.	.
$(DW-9)_+$	.	.
DW (including its segments):V	Yes	Yes but all <1e-4

Note: Standard error (Std. err.) for Adaptive Lasso is not computed, and not reported for HL estimates which are virtually 0. A dot (".") indicates 0 coefficient estimate and dash ("-") indicates reference category.

[1]All the distance measures (in 100 m) were square root transformed.

[2]The estimates of the last interaction term, is not very interesting, and it is omitted to be able to fit the table in a single page.

The habitat (resource) selection (measured at the scale of log-odds-ratio) of the wind farm was affected by the distance from the windfarm (Figure 1). The flat part of the broken line in Figure 1, between distance 0 and 3.6 km, indicates that the reindeers' resource selection within 3.6 km was not affected, when the wind farms were not running. The intersection point (at Distance = 1.86, approx.) of the two lines (Figure 1) shows the threshold at which the negative effect of running wind farm vanishes.

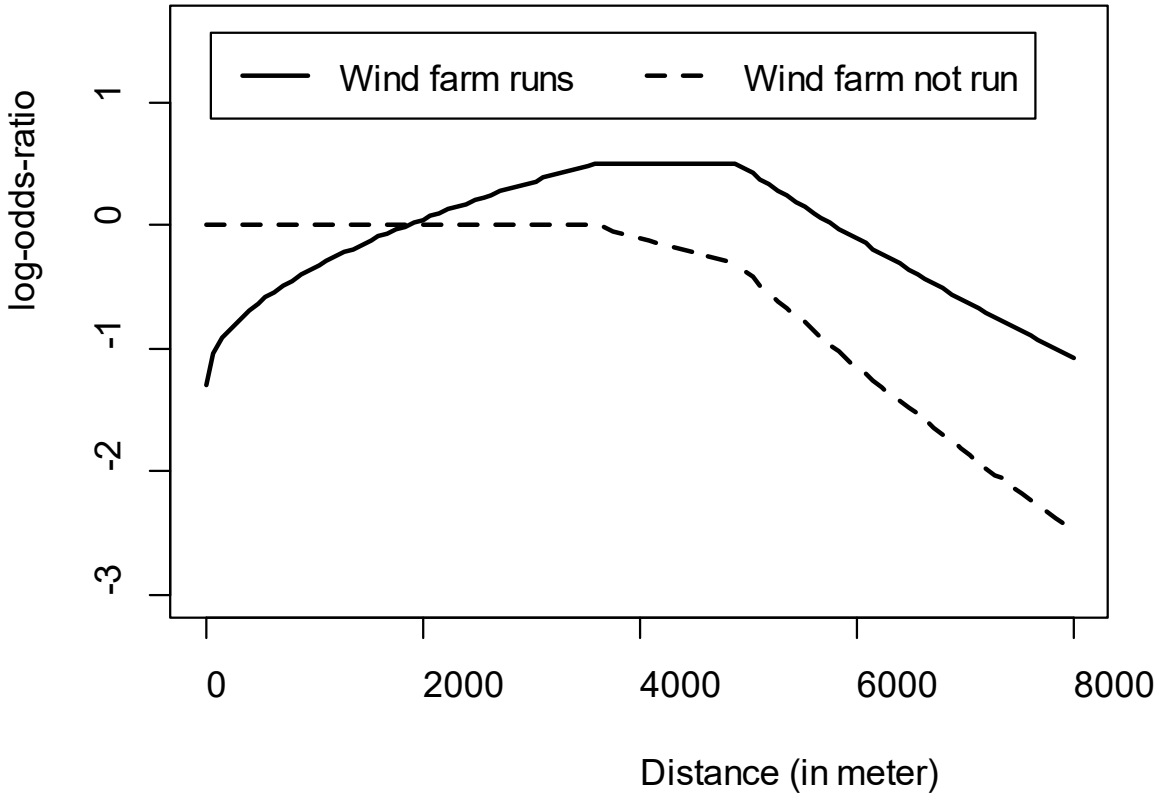

Figure 1: Plot of the HL method implied comparative reindeer resource selection, measured by log-odds-ratio, between the periods when wind turbines were running (solid line: $-0.31 * (DW-5)_+ - 0.81 * (DW-6)_+$**), and when they were not running (broken line:** $-1.30 - 0.31 * (DW-5)_+ - 0.81 * (DW-6)_+ + 0.30 * DW{:}WR$**), at different distance from the wind turbine, keeping everything else fixed.**

5. Discussion

We present threshold regression method for estimating zones of influence in environmental impact assessment. The core idea is already explored in econometrics, and a simplified form have been used in estimating ZOI (Boulanger et al. 2012). A limitation of for example Boulanger et al. (2012) is that the disturbance effects were assumed to be linear and to disappear abruptly at a single threshold. Moreover, the effect threshold was estimated through *ad hoc* method, e.g. via grid search. This article presents a way to identifying and estimate the effect thresholds of ZOI directly estimated from the data, without

requiring grid search. We also investigated the merits of different estimation techniques for ZOI estimation.

Based, on the simulation study, and considering theoretical properties of the estimator, we suggest HL method for model fitting. However, Adaptive Lasso method was also found useful, for large sample size. An advantage of HL over Lasso it that the HL method readily provide the standard error estimates (Lee and Oh, 2014), and random effects can be included in the model. A disadvantage of the HL method was the estimates' sensitivity to the starting value. Therefore, in real applications, we suggest HL model be estimated with different starting values (e.g. GLM, or ridge estimate), and use deviance or log-likelihood to assess the best fitting model. R codes for fitting threshold regression via HL method is available as supplementary materials (from the authors, e-mail: maa@du.se).

We apply Adaptive Lasso, and HL method to estimate effects of a wind farm and public road on reindeer habitat selection in winter, in a forested mountain area in the north of Sweden. However, in estimating ZOI, the results from the two methods were different. HL estimated more threshold parameters to be zero (or near zero), compared to Adaptive Lasso. Using the results from the HL method we conclude that, if everything else remains constant, the negative effect of the running wind farm on reindeer habitat selection vanished at 1.9 km from the wind farm.

Our results from real data analysis agrees partly with the existing knowledge in that reindeer response to human disturbances such as roads and power lines are often found to vanish at around 1-2 km from the source of the disturbance (Lundqvist 2007; Anttonen, Kumpula & Colpaert 2011; Panzacchi et al. 2012). However, our results contradicted previous results of Tsegaye et al. (2017) where no significant effect of a wind farm was found. However, their selection of study area (in an island) makes it difficult to generalize their results to any other environment (Panzacchi et al. 2015).

The HL estimation in this study has been carried out by using extended quasi likelihood method (Lee and Nelder, 1996). Further, in actual computation, any numerically zero value of the random effects in the dispersion term of the HL were replaced with a small non-zero number. Numerically better estimation can be obtained by using higher order correction, and using numerically more robust optimization techniques, such as concave-convex procedure (see e.g. Lee et al., 2015). In estimating cumulative ZOI we considered only two human disturbances, wind farm and public roads. Adding more sources of interest might also be interesting. We leave this issues for possible future work.

References:

Anttonen, M., Kumpula, J. & Colpaert, A. (2011), Range selection by semi-domesticated reindeer (Rangifer tarandus tarandus) in relation to infrastructure and human activity in the boreal forest environment, northern Finland, *Arctic* 64, 1–14.

Bergström, L., Kautsky, L.; Malm, T., Rosenberg, R., Wahlberg, M., Capetillo, N. Å., and Wilhelmsson, D. (2014), Effects of offshore wind farms on marine wildlife—a generalized impact assessment, *Environmental Research Letters* 9, 034012.

Boulanger, J., Poole, K. G., Gunn, A., and Wierzchowski, J. (2012), Estimating the zones of influence of industrial developments on wildlife: a migratory caribou rangifer trandus groenlandicus and diamond mine case study, *Wildlife Biology* 18, 164—179.

Calenge, C. (2006), The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. *Ecological Modelling* 197, 516-519

Chana, N. H., Yaub, C. Y., and Zhang, R. (2015), LASSO estimation of threshold autoregressive models, *Journal of Econometrics* 189, 285–296.

Ficetola, G. F., and Danoel, M. (2009), Ecological thresholds: an assessment of methods to identify abrupt changes in species—habitat relationships, *Ecography* 32, 1075-1084.

Friedman, J., Hastie, T., Tibshirani, R. (2010), Regularization paths for Generalized Linear Models via coordinate descent, *Journal of Statistical Software* 33(1), 1-22.

Gillies, C. S., Hebblewhite, M., Nielsen, S. E., Krawchuk, M. A., CAMERON L. Aldridge, C. L., Frair, J. L., Saher, D. J., Stevens, C. E., Jerde, C. L. (2006), Application of random effects to the study of resource selection by animals, *Journal of Animal Ecology* 75: 887-898.

Holling, C. S. (1973), Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics* 4, 1-23.

Hudson, D. J. (1966), Fitting segmented curves whose join points have to be estimated, *Journal of American Statistical Association* 61: 1097-1129.

Huggett, A. J. (2005), The concept and utility of ‘‘ecological thresholds’’ in biodiversity conservation. *Biol. Conserv.* 124, 301-310.

Hansen, B. E. (2017), Regression kink with an unknown threshold, *Journal of Business & Economic Statistics* 35(2), 228-240

Hinkley, D. V. (1969), Inference about the intersection in two-phase regression, *Biometrika* 56, 495-504.

Kivinen, S. (2015), Many a little makes a mickle: Cumulative land cover changes and traditional land use in the Kyrö reindeer herding district, northern Finland, *Applied Geography* (63), 204-211.

Johnson , C. M., Boyce, M. S., Case, R. L., Cluff, H. D., Gau , R. J., Gunn, A., and Mulders, R. (2005), Cumulative Effects of Human Developments on Arctic Wildlife, *Wildlife Monographs* 160, 1-36.

Leng, C., Lin, Y., and Wahba, G. (2006), A note on the lasso and related procedures in model selection, *Statistica Sinica* 16(4), 1273-1284.

Lee, Y., and Oh, H. (2014), A new sparse variable selection via random-effect model, *Journal of Multivariate Analysis* 125, 89-99.

Lee, S., Pawitan, Y., and Lee, Y. (2015), A random-effect model approach for group variable selection, *Computational Statistics and Data Analysis* 89, 147-157.

Lee, Y., and Nelder, J.A. (1996), Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society (B)* 58, 619-656.

Lundqvist, H. (2007), Ecological cost-benefit modelling of herbivore habitat quality degradation due to range fragmentation, *Transactions in GIS* 11, 745-763.

Manly, B.F.J., McDonald, L.L., McDonald, T.L., and Erickson, W.P. (2002), *Resource Selection by Animals*, 2nd edn. Kluwer Academic Publishers, Dordrecht.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall, London.

Muggeo, V. M. R. (2003), Estimating regression models with unknown break-points, *Statistics in Medicine* 22, 3055-3071.

Muggeo, V.M. R. (2008), segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News, 8(1), 20-25. URL https://cran.r-project.org/doc/Rnews/.

Nielsen, S. E., Cranston, J., and Stenhouse, G. B. (2009), Identification of priority areas for grizzly bear conservation and recovery in Alberta, Canada, *Journal of Conservation Planning* 5, 38-60.

Ng, C.T., Lee, W., and Lee, Y. (2018), Change-point estimators with true identification property, *Bernoulli* 24(1), 616-660.

Panzacchi, M., Van Moorter, B. & Strand, O. (2013), A road in the middle of one of the last wild reindeer migration routes in Norway: crossing behaviour and threats to conservation. *Rangifer* 33, 15-26.

Panzacchi, M., Van Moorter, B., Strand, O., Loe, L.E. and Reimers, E. (2015), Searching for the fundamental niche using individual-based habitat selection modelling across populations. *Ecography*, 38: 659-669. doi:10.1111/ecog.01075

Polfus, J.L., Hebblewhite, M., and Heinemeyer, K. (2014), Identifying indirect habitat loss and avoidance of human infrastructure by northern mountain woodland caribou, *Biological Conservation* 144, 2637-2646.

R Core Team (2017), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Skarin A., and Alam, M. (2017), Reindeer habitat use in relation to two small wind farms, during preconstruction, construction, and operation. *Ecology and Evolution* 7, 3870-3882.

Skarin, A., Sandström, P., Buhot, Y., and Nelleman, C. (2016), Renar och vindkraft II - Vindkraft i drift och effekter på renar och renskötsel (in English: Impacts of wind power infrastructure development on semi-domesticated reindeer and reindeer husbandry), Report 294, Department of

Animal Nutrition and Management, Swedish University of Agricultural Sciences, Uppsala, URL: https://pub.epsilon.slu.se/13562/7/skarin_a_et_al_160818.pdf (last accessed, Jan. 03, 2020).

Skarin, A. and Åhman, B. (2014), Do human activity and infrastructure disturb domesticated reindeer? The need for the reindeer's perspective. *Polar Biology*, 1-14.

Sawyer, S. (2017), GWEC annual wind power update, short term forecast more than 800 GW globally by 2021, RenewableEnergyWorld.com, URL http://www.renewableenergyworld.com/ugc/articles/2017/04/26/gwec-annual-market-update-and-short-term-forecast--more-than-800-gw-globally-by-2021.html (last accessed Jan. 03, 2020).

Taylor, J., and Tibhsirani, R.J. (2015), Statistical learning and selective inference, *Proceedings of the National Academy of Science of the United State of America* 112, 7629-7634.

Thaxter, C.B., Buchanan, G.M., Carr, J., Butchart, S.H.M., Newbold, T., Green, R.E., Tobias, J.A., Foden, W.B., O'Brien, S., and Pearce-Higgins, J.W. (2017), Bird and bat species' global vulnerability to collision mortality at wind farms revealed through a trait-based assessment. *Proceedings of the Royal Society B: Biological Sciences*, 284.

Tsegaye, D., Colman, J.E., Eftestøl, S., Flydal, K., Røthe, G., and Rapp, K. (2017), Reindeer spatial use before, during and after construction of a wind farm. *Applied Animal Behaviour Science*, 195, 103-111.

Tong, H. (2011), Threshold models in Time Series Analysis—30 Years On, *Statistics and its Interface* 4, 107-118.

Walker, L.J., and Johnston, J. (1999), Guidelines for the Assessment of Indirect and Cumulative Impacts as well as Impact Interactions, European Communities, Luxemburg.

Weber, M., N. Krogman, and T. Antoniuk. (2012), Cumulative effects assessment: linking social, ecological, and governance dimensions. *Ecology and Society* 17(2), 22.

Zhou, H. (2006), The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association* 101, 1418-1429.

Zhou, H., and Hastie, T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society (B)* 67, 301-320.