Crabtree RL, Conner MC, Treves A. Misleading overestimation bias in methods to estimate wolf abundance that use spatial models. Academia Biology. 2025;in press. https://doi.org/10.31220/agriRxiv.2023.00215.

Misleading biases in methods for estimating wolf abundance using spatial models

Robert Crabtree[1,*], Mary Conner[2], Adrian Treves[3]

Academic Editors: Eric Lewallen, Andre J. van Wijnen

Abstract: Abundance, the primary criterion used for management and conservation decision-making, is very difficult to estimate for carnivores, especially highly mobile, wide-ranging wolves. Due to their economic and ecological importance to game management, society, and ecosystem health, reliable methods are necessary for wolves. The recently developed integrated patch occupancy model (iPOM) claims to provide improved accuracy of carnivore abundance at finer scales while reducing cost. We evaluate its input data models, potential biases, precision, sampling design, model coherence, validation, and reproducibility. Testing their method's sensitive and vital assumptions to estimate the area occupied by wolf territories revealed that three sources of bias: false positive errors, closure violations, and resolution, combined to cause a substantial overestimation of abundance. The crucial confirmation step to independently determine individual wolf territorial centroids was eliminated, leading to a severe overestimation due to three sources of double-counting errors and mortality that occurred just before and during the survey period. iPOM lacks an inferential sampling design and relies upon a flawed survey of hunters' inadvertent recollections of wolf sightings. It also includes a static and deficient covariate model, which limits the ability to correct for the sources of overestimation bias. The variance method used to report a confidence interval is incorrect and omits multiple components of variation, resulting in substantial underreporting of uncertainty. With many recommendations for improvement, we conclude that iPOM produces unreliable predictions, is irreproducible, and is misleading by reporting accurate and precise abundance predictions when abundance is severely biased (overestimation) and imprecise.

Keywords: *Canis lupus*, carnivore, endangered species, abundance estimation, bias, variance, uncertainty, coherence, occupancy, closure violation

Citation: Crabtree R, Conner M, Treves A. Misleading biases in methods for estimating wolf abundance using spatial models. *Academia Biology* 2025;Volume. https://doi.org/10.20935/xxx

1. Introduction

Population size (abundance) is the leading criterion agencies use to manage and restore species populations and make critical decisions such as listing or delisting species. A reliable estimate, unbiased and precise over time, provides an efficient metric to account for the actions of decision-makers who can then adaptively respond and form sound policy [1,2]. This is important for large mammal populations, especially of carnivores, which humans have severely reduced through a loss of habitat and government-led eradication campaigns that began in the 18th century. Their subsequent population declines have negatively impacted ecosystems through density-mediated trophic interactions (see [3,4] for a review). When reliable carnivore abundance methods are embedded within an annual learning cycle, agencies can make adaptive decisions that effectively respond to changing conditions [2]. These iterations will have numerous and diverse benefits for game management and humans through the socioeconomic and ecosystem services carnivore populations provide [5–8], mainly through cumulative trophic pathways [9].

Unfortunately, large carnivore abundance is notoriously difficult to determine. Because they exist at low densities, display high mobility over large landscapes, are cryptic in their behavior, and exhibit aversive responses to people and to capture, they rarely lend themselves to complete enumeration or design-based approaches like survey sampling methods [10] that rely upon a probability structure inherited from randomized data collection. Instead, a model-based approach is used that depends upon assumptions of a stochastic model of the sample data and the sampling process [11], which can include marked or unmarked individuals. Individual marking of individuals or groups, like territorial packs, is highly preferred because raw unmarked counts are often misleading due to missing individuals that are not seen or duplicate observations of the same animal or group. Furthermore, such marking, whether physical, like using radio-collars and ear tags, or based on natural unique features, enables researchers to construct encounter histories, estimate detection probabilities, and account for detection errors by using distinguishing marks as the primary or secondary method (e.g., [12]). Collectively, this reduces biases caused by heterogeneous detection rates and imperfect detectability present in carnivore populations. As a result, nearly all carnivore studies and most mammal studies use mark–recapture and/or mark–resight

[1]Yellowstone Ecological Research Center, Bozeman, MT, USA.
[2]Department of Wildland Resources, Utah State University, Logan, UT, USA.
[3]Nelson Institute for Environmental Studies, University of Wisconsin at Madison, Madison, WI, USA.
*email: crabtree@yellowstoneresearch.org

([13,14] for carnivores) sampling methods in a model-based inference framework. An updated review of recent and related studies on wildlife population estimation methods can be found in Mills et al. ([15], chapter 4).

Wolves further exacerbate the challenges of estimating abundance because statistical models must be structured to account for high levels of heterogeneity in space and time owing to their complex socio-spatial behavior and seasonal response to ungulates, their primary prey. Furthermore, wolves do not lend themselves to the traditional, well-developed estimation methods like distance sampling, aerial surveys, or direct capture–recapture commonly used with vertebrate species. As a result, few reliable, transparent, and repeatable methods are commensurate with large carnivores' high economic, social, and ecological importance. Thus, much more attention is needed to develop, test, and evaluate new statistical models for wolf population abundance with a keen focus on their inherent biases, measures of precision (variance), field sampling designs, and other model properties that affect the ability to make valid inferences, which, in turn, will provide the information required to sustain wise decision-making and avoid costly mistakes. Reproducible methods, data sharing, and open validation of model assumptions and output—all open to independent review and public debate—will enhance the governance process and benefit public trust resources.

Here, we evaluate a newly developed method [16] called the integrated patch occupancy model (iPOM), which is representative of several new methods that non-traditionally apply spatial models to estimating abundance from observations of unmarked animals. We identify potential biases in iPOM's numerous components by testing assumptions—*biologically and statistically*—and, where possible, determine the direction and magnitude of these biases. Because spatial models are particularly sensitive to assumptions regarding animal space use, we focus on iPOM's use of occupancy modeling—its input data model, sampling design, and model biases. We also evaluate iPOM's second spatial model, an agent-based simulator that predicts territory size. These two spatial models determine the number of wolf packs which is the effective population unit for policy and management and also provide an understanding of ecological interactions and their functional value [17–19]. We then evaluate and test their variance estimates used to report uncertainty with their calculated confidence intervals for the statewide wolf population abundance. Next, because of iPOM's complexity, we assess iPOM's alignment with model coherency and reproducibility. Finally, we provide constructive criticism on improvement and a framework for biologists, the broader conservation community of decision-makers, and the public to assess any statistical model to determine the abundance of a challenging species.

2. Evaluation criteria

Bias and variance are the two fundamental criteria used to evaluate the quality of any statistical method that attempts to estimate or predict abundance and if, when, and how it should be used. Their minimization is the bedrock of valid inference, as both contribute to model error (uncertainty). Although managers primarily make decisions using the numerical value of abundance (the point estimate), its variance is also used to reliably track and detect changes over time and communicate ranges of possibility when thresholds are set by policy or law. We define the term reliability as the ability of a method to provide unbiased and precise measures of abundance over time. High variance adds uncertainty to the point that inference cannot be made for scientifically valid decision-making on quotas and minimum populations.

As independent attributes, these criteria differ because variance can be directly estimated from the sample data, whereas bias cannot. Thus, the only means to evaluate the bias of an estimation method is to state and test the validity of the model assumptions, which places full responsibility on developers and users to conduct such testing. Biological and statistical testing determines the likelihood of a bias and whether it is a negative or positive bias (or under- or overestimates the true value, respectively). Testing can also include simulation testing of potential model biases to determine the direction and magnitude of bias.

If a prediction of abundance is biased or has a high variance, then three of four possible outcomes (**Figure 1**) are deemed unreliable. If bias is ignored, decision-makers (managers, biologists, judges, conservationists, and policymakers) will likely face two outcomes (**Figure 1C,D**). Predictions become more uncertain and unreliable if the variance is high (**Figure 1B,D**). Even worse is a precarious, misleading situation when one claims the point estimate is unbiased and precise (**Figure 1A**) when, in fact, there is overestimation bias, and the variance provides false (overestimated) confidence (**Figure 1D**).

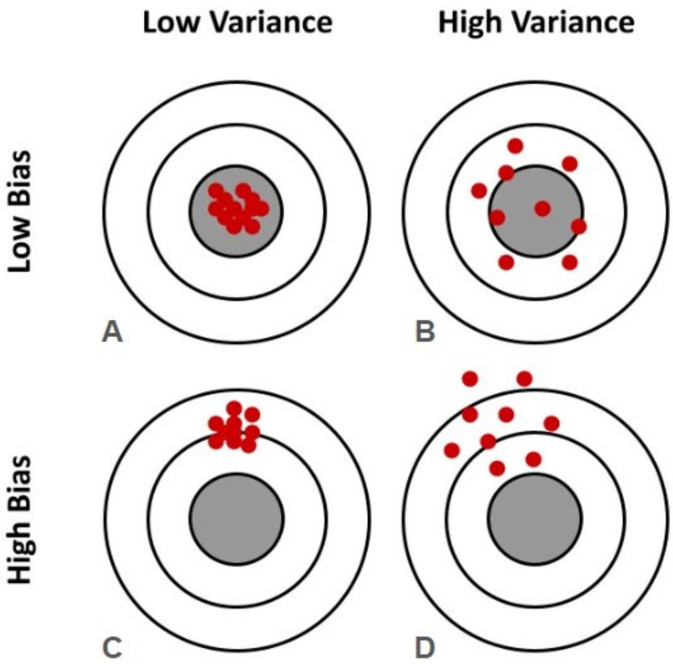

Figure 1. Four scenario outcomes for the two independent attributes used to assess the quality of an estimation or prediction method. Bias is the distance between the red dots and the target's center; precision, measured by the variance, is the spread of the red dots. (**A**) is unbiased and precise; (**B**) is unbiased and imprecise; (**C**) is biased and precise; and (**D**) is biased and imprecise. They are graphically analogous to shooting a rifle at a target, where the shot taken is the value of abundance (the point estimate). If the shooter's grip is shaky, then the variance (spread of shots) is higher (**B**,**D**) than with a steady hand (**A**,**C**). Bias is where the sighting of the rifle (curved barrel, bad scope) is "off" and shots are consistently off-center (**C**,**D**) versus on-center or unbiased (**A**,**B**).

Another fundamental concept for evaluating a statistical model is the tradeoff between bias and variance (**Figure 2**). When bias is detected due to failed tests of assumptions, investigators appropriately respond by adding variables (called covariates) to decrease these biases. In turn, this increases variance and often increases model complexity.

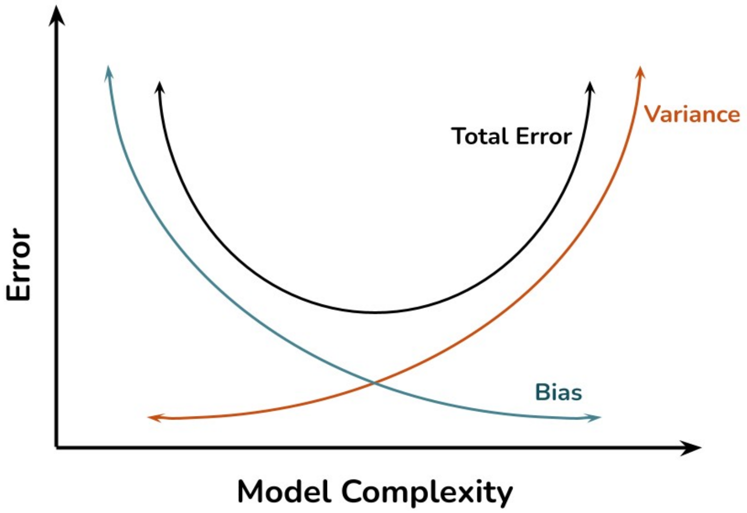

Figure 2. The tradeoff between precision and bias. Model complexity increases when investigators select parameters in the model (covariates) to decrease bias, which in turn increases variance given the sample data. Variance can be calculated from the sample data, whereas bias can only be assessed through biological and statistical testing of the model's assumptions. Thus, testing is required to design a statistical model that optimizes these tradeoffs by minimizing model error (variance + bias^2) with the efficient use of covariates—using the fewest minimally sufficient set of causal covariates that substantially reduces or eliminates biases. As the error of a statistical estimator increases, its ability to make inferences and successfully predict declines.

Investigators attempt to optimize reliability by minimizing model error by carefully adding covariates assumed to be causal, independent, and not confounding. However, serious problems arise when added covariates do not adequately capture and account for heterogeneity in space, time, and behavior. We highlight in particular the problem of including static covariates that are insensitive to changes in conditions under the control of investigators or policymakers. Conversely, and more problematic, is a deficient model where the analysis proceeds knowing causal covariates are missing [20]. Because iPOM combines submodels within multiple models and obtains direct and indirect samples from different populations at different times and spatial scales, we necessarily assess model coherence in making valid inferences. Finally, it is vitally important to ensure that a method or experiment, especially a new one, is reproducible—a cornerstone of science. This requires a transparent and adequate description of all methods and the sharing of data and analysis results from model components.

We acknowledge that analyzing messy ecological field data reveals numerous issues, potential biases, and problems. We also realize that compromise is nearly always a part of this process. With recommendations, we specifically focus on essential issues that can be improved upon with testing, verification, revision, and retesting—the iterative cycle of gaining reliable knowledge. We aim to lay out a path forward that eliminates or minimizes the problems that cripple inference.

3. Background on iPOM

With the recent reintroduction and expansion of wolves in the lower 48 states, some state agencies have abandoned field efforts to individually mark wolves to estimate abundance and other important demographic parameters (litter size, pack size, survival, dispersal) in favor of using new lower-cost methods to infer their abundance [21–23] using spatial models that reduce or replace empirical field sampling. This should raise concerns about uncertainty and reliability because nearly every field study of wolves, including those in Montana, Idaho, and Wyoming, has investigated their demography and complex social and spatial system using a sufficiently large sample of distinguishable marks on individuals (natural and radio- and GPS-marking) to account for the numerous and complex sources of heterogeneity causing bias. As a result, sampling of marked carnivores has become a widely accepted consensus method for estimating demographic parameters, especially abundance [24–26].

Recently, Idaho and Montana eliminated a radio-collaring program for wolf population monitoring statewide and instead transitioned to indirect observations (hunter surveys, cameras) of unmarked animals. For example, iPOM's goal is to integrate basic and applied research to provide accurate estimates of wolf abundance to inform decision-making and meet management needs while reducing reliance on intensive field monitoring efforts [16]. It combines three main models, area occupied (AO), territory size (TS), and pack size (PS), as well as additional correction factors, indices, and constants. Although [16] did not provide a coherent mathematical equation for wolf abundance specific to their sampling design, it did present submodel equations and an overall graphic illustration ([16], Figure 1) to map out how iPOM arrives at abundance. For clarity, we provide generalized equations (1 and 2) to help explain iPOM's deviation from past and current methods.

For decades, agencies have estimated abundance through field enumeration or estimating the number of packs (NP) yearly [17–19,24]. This sampling is verified by tracking radio-collared wolves to eliminate, or correct for, the persistent problem of double-counting packs, which is a common problem when relying on indirect signs (sightings, howling, scat, tracks, etc.). NP is then multiplied by empirical estimates of pack size (Equation (1)). Although these approaches have recently been discontinued in Idaho and Montana, they are currently being used to monitor wolves in Wyoming, Colorado, Oregon, Washington, and California.

iPOM and similar models called the patch occupancy model (POM [21]) and the scaled occupancy model (SOM [22]) first deviated from the normal annual field sampling of wolf packs (Equation (1)) by substituting a spatial modeling approach called occupancy modeling to first estimate the AO by territorial wolf packs and then dividing by the mean or median territory size to calculate NP. The POM, the SOM, and iPOM each use different versions of occupancy modeling to estimate the AO by stable territorial wolf packs. iPOM goes further and drops the normal empirical sampling approach to predicting TS and substitutes a second spatial model [27] that simulates TS. Thus, in an apparent attempt to reduce field costs and improve accuracy [16], iPOM substitutes two spatial models (AO and TS) for the normal empirical field sampling for NP (Equation (1)). Finally, iPOM further departs from annual empirical sampling to predict pack size (PS) using a regression model [16] to determine the total number of individual wolves belonging to the modeled territorial packs in Equation (2).

$$N_i = NP_i \times PS_i \left(\text{demographic without spatial modeling}\right) \tag{1}$$

$$N_i = \left(\sum AO_j\right)_i \times \left(1/TS\right)_i \times PS_i \left(\text{spatial models included}\right) \tag{2}$$

where N is the abundance of territorial pack members in the year i, and j designates each 600 km^2 grid cell that is summed across Montana. To account for the assumed small proportion of lone wolves (LW) not belonging to a pack, iPOM calculates a proportion, converts it into a multiplicative rate (1 + LW), and then includes it in the multiplication sequence in Equation (2). It is important to note that random variable N is a function of four random variables, AO, TS, PS, and LW, that all contain numerous components of variation to be included in the variance in N used to report uncertainty.

While iPOM should be a valid theoretical approach (Equation (2)), it is not an integrated model. Instead, it is a component model that combines three main models (assumed independent) in a sequential multiplicative series. Integration refers to integrated population models (IPMs) used extensively to estimate the parameters of wildlife populations ([28,29] for wolves in Idaho). IPMs are coherent and thus have numerous advantages over iPOM, including demographic mechanisms, reconciling spatial and temporal mismatch, estimating unmeasured demographic rates, density dependence, and assimilating overlapping data inputs from multiple sources [30].

Because wolves are exceptionally mobile, express complex social and territorial behaviors that vary seasonally, and range widely across heterogeneous landscapes, we examine the sensitivity of iPOM's spatial models that determine the territorial area a wolf pack observation represents. This directly affects the first two most critical of the four basic assumptions of this AO occupancy model: (1) the occupancy state of a grid cell remains constant between sampling times (the closure assumption); (2) no false detections of wolf packs—detections are representative of the wolf territories; (3) grid cells are independent spatially and temporally—the outcome of one survey does not influence the outcome of another survey site during the same season; and (4) a constant detection probability—the likelihood of detecting a wolf pack is consistent across all survey sites for a given sampling occasion. Recent improvements [31] have relaxed these assumptions by incorporating covariate submodels in an attempt to reduce bias. However, among the four basic assumptions, the closure assumption is arguably the most commonly violated [32,33]. It leads to overestimation bias for wide-ranging, highly mobile species like wolves. Closure can be violated in two ways: geographic changes in movement (immigration, dispersal, extra-territorial forays) and demographic changes (deaths) during the surveys.

To estimate the AO, iPOM uses a dynamic occupancy model [12,34] designed to correct detection errors, especially false positives (e.g., a detection is recorded but the wolf territory is absent). Dynamic occupancy models, first developed by MacKenzie et al. [35] with developments by Royle and Link [36] and Miller et al. ([34] used in iPOM), offer an initial framework for designing and analyzing large-scale animal distribution data. Such occupancy models reduce detection errors using covariate models that have additional assumptions, including those governing their selection in generalized linear models (see [37] for a review). These models allow for a change in occupancy between years. Still, occupancy must stay closed during the annual surveys, i.e., the territory and its spatial extent that observed wolf groups belong to must stay intact and remain stable, respectively, during iPOM's late fall survey period.

iPOM's occupancy modeling aims to separate the underlying biological processes driving distributional changes in wolf packs from the observational process, which begins with surveys of hunters' recollections of passive wolf pack observations from an unspecified portion of the survey units they are hunting. In iPOM, these units, called patches, are mutually exclusive, non-overlapping 600 km^2 grid cells and are pseudo-surveyed by subdividing post hoc hunter survey information into five one-week periods during the late fall hunting season. Wolf specialists sort through information and other indirect wolf signs to create two submodels: a critical "certain" data model built with unambiguous spatial observations of confirmed wolf pack territories (no false positives allowed) and an "uncertain" model from hunter surveys. The "certain" data model (the centroid model, hereafter) requires a confirmation step from a second method (e.g., tracking collared pack members) to confirm that the demarcated territory centroids are singular and unique from adjacent territorial packs before it is applied to estimating occupancy in uncertain and no-detection grid cells across Montana. Further description of iPOM's dynamic occupancy model can be found in Miller et al. [12].

Territory size (TS) in iPOM is determined by a simulation model developed using NetLogo, a software environment for basic agent-based simulations [38]. It uses numerous cost/benefit rules (each is an assumption) for patch ownership (occupancy), and territorial ownership dynamics are simulated for 127 packs on a grid with a cell area of 1 km^2 for Montana. It then maps spatially explicit territories and generates TS and the number of territorial packs (NP), which is not used in iPOM. It is important to note that iPOM's territory simulation model does not use any annual inputs of empirical sample data from Montana's wolf population. To predict PS, iPOM uses a Poisson regression of past empirical data on wolf pack sizes [16,39] and environmental covariates.

4. Input data and sampling design

We found the creation of the AO data model and its sampling design to be highly problematic with numerous sources of error. Such errors can cause bias in all three of iPOM's models because all use centroid information (violation of the independence assumption). These include assumption violations, the post hoc hunter surveys, the spatial sampling design, and the quality and timing (design) of the hunter surveys. iPOM's AO data model parallels the assumptions in mark–recapture methods: detections are correctly classified (marks are not lost) and are correctly recorded. iPOM's methods are far removed from the simple process in mark–recapture models, where these assumptions are seldom violated because the same trained individual(s) administers the distinguishable, unique mark and then carefully records it in near-real time (several days). In contrast, iPOM has a lengthy, arduous process, from hunter surveys to wolf specialists determining the quality and type (certain or uncertain) of observations used to hand-mark territorial centroids of wolf packs on a map. Staff and volunteers from Montana Fish, Wildlife and Parks (MFWP) primarily conduct phone interviews to determine a hunter's recollection of (1) the observation type (sightings, howl, scat, carcass, tracks), (2) the correct grid cell location, and (3) the correct time (week). We do not see any data verification, standardized protocols, or definitions used to sort through hunter survey data, nor through the wolf observations and indirect signs used to create the sensitive centroid model [16], which will directly inflate the AO and abundance (overestimation), even with small false-positive errors [40].

4.1. Quality of sampled hunter survey observations

Although there is increasing reliance on public members to contribute to data collection [41], there is a long list of potential biases in unstructured surveys by public members, especially when they are untrained or asked to recall observations afterwards, as in iPOM [16]. They plague monitoring efforts [42] and include cognitive bias leading to over-detecting a popular, charismatic, or controversial species (false-positive error). Other sources of bias stem from changes in field efforts over time (observation bias), incomplete and selective recording by observers (reporting bias), and geographical bias and detection bias [43]. Anderson [44] questioned why one would survey deer hunters as a basis for inference about humans. Similarly, why would one sample deer hunters' inadvertent recollections to make inferences about wolf populations, instead of sampling wolves directly?

In addition, observations of species not traceable to tangible material, such as museum specimens or digital recording devices (e.g., camera traps), include biases that may render them inefficient in representing occurrence and distribution [45]. Instead, Greenwood [41] recommends trained, recruited, retained, and experienced observers who use standardized protocols, transparent QA/QC, and validation. He further recommends independence from government and conservation organizations when political or personal motivations are present. Altwegg and Nichols [46] state that observations by members of the public can be valuable but present analytical challenges due to the observation process and thus advocate strengthening inference by designing surveys with specific questions, analysis strategies, and data characteristics in mind.

Unfortunately, iPOM does not recognize or adhere to the many criteria, recommendations, and concerns listed above. However, Miller et al. [12] did verify their initial centroid model using the rich information obtained from radio-collared wolves to ensure that each centroid designated a singular, unique wolf territory. Since then, this verification step has been eliminated from iPOM. Hunter's ability to distinguish coyotes from wolves under field conditions using indirect methods in MFWP hunter surveys is unknown, as is their ability to distinguish lone wolves from pack-living wolves. Even trained biologists make errors in field estimations of wolves from the size of carcasses [47], let alone distant animals, which are probably sometimes seen fleetingly.

4.2. Hunter survey sampling design and closure violations

iPOM essentially lacks a sampling design. Sells et al. [16] did not include the criteria commonly used to develop and plan a proper sampling design. These criteria include the selection of appropriate spatial and temporal scales, the incorporation of information appropriate to the time and place (e.g., empirical pack sizes), alignment of sampling units and efforts with observational and ecological processes, sample size consideration, and inherent environmental heterogeneity. And in their context, we would also like to know more about the observers, i.e., hunter behavior, independence, and validation. We found numerous problems and assumption violations in the decisions they made.

The fall timing exacerbates false-positive errors, closure violations, and errors in predicting PS. Closure was likely violated geographically and demographically in late fall (1) because many young adult wolves disperse from the pack then [48–50], (2) because non-dispersing pack members can split into smaller groups and make extraterritorial movements [48,51] prior to the formation of cohesive groups in the winter mating season, (3) due to mortality during the wolf hunting season, which starts in early September and continues through the late fall iPOM surveys, and (4) due to pack dissolution, which amplifies closure violations (see Section 5.1). By definition, pack size is estimated in the winter when social groups are most cohesive [52] and when 10-month-old pups are considered pack members [51,52]. Additionally, this adds another source of double-counting error because of the difficulty in distinguishing between adults and the nearly adult-sized 7- to 8-month-old pups that split off and travel together during the fall survey period. Finally, iPOM's five consecutive one-week periods of observation during the fall hunting season further violate assumptions because they are not temporally independent of each other nor randomized, which enhances closure violations.

iPOM used an inadvisable 'rule of thumb' to choose the resolution of its grid cell size to match the reported territory size of 600 km^2 [16,21]. MacKenzie [53] recommends explicitly selecting a grid cell smaller than the average home range size to ensure detection if the species of interest is wide-ranging and highly mobile because it could be in another portion of its home range during observations, especially if one does not survey the entire grid cell, as in iPOM. Furthermore, our simulations and those of Stauffer [22] indicate that the selection of grid cell size needs to be smaller than the average wolf territory size. Neglect of these fundamental findings inevitably leads to overestimation bias.

4.3. Centroid model determination

We found numerous problems and a lack of description in iPOM's methods for determining wolf territory centroids, making it impossible to replicate. First, Sells et al. [16] asserted a "fairly certain" wolf pack centroid quality code [54]. That method was neither described objectively nor on a case-by-case basis. In addition, there is little description of how the eight categories in their pack information table [54] were used to scientifically verify or justify wolf specialists' demarcation of pack centroids. Second, it is unclear whether they verified a "represented" wolf territory centroid, which must be singular and known with certainty [16]. We also find no description of (1) how they used seven types of information [16] to determine the spatial coordinates of a wolf territory or a centroid without a sufficient sample of radio-marked pack members and (2) how and if an assumed centroid was used to verify hunters' observations of wolf packs. Creel [55] also recognized the logic flaw in that the AO model was developed using centroids from the centers of territories determined by movements of GPS-collared wolves, which are unavailable under the current version of iPOM. Finally, Montana Fish, Wildlife, and Parks (MFWP) admitted they did not record or archive hunter reports, making it impossible to verify wolf observations, validate the AO model, and reproduce iPOM. The lack of archived data by itself makes Sells et al. [16] subject to retraction by the policies of the journal in which it was published because effectively the source data do not exist.

We also identified an additional source of double-counting of wolf pack observations (the demarcation of two centroids for one territory), which would cause a severe overestimation bias (see Section 5.5). Without time recording and locating distinguishing marks from hunter reports, the authors are unaware of how many of the same highly mobile and wide-ranging wolves were detected in different grid cells during the 5-week survey period. As such, subgroups from a single wolf pack are recorded as separate (two or more) wolf packs. Such double-counting errors are widely recognized in field surveys because animals can move between neighboring sites on the same day [56]. The rate of double- or triple-counting the same wolves would logically increase with the time between surveys (up to 34 days apart in iPOM). It can be corrected using locational marking devices like radio-collaring [57], which iPOM lacks. For birds [58], with a similar territorial spatial system to wolves, these false-positive error rates were 49% for new (naive) observers when movements were eliminated in a controlled setting. With double-observers, the rate was still high at 39%, and the expert, experienced observer rate was 9.5%. One can only imagine the false-positive error rates reported by untrained hunters who inadvertently recollect their passive observations of highly mobile, wide-ranging wolves, which, moreover, can resemble sympatric coyotes and some domestic dogs under field conditions. Despite their claim [16], we fail to see how their AO model can correct for the misclassification of coyote pairs because they are sympatric with wolf packs and select for many, if not all, of their selected covariates.

5. AO model

Biologically and statistically, AO is the most critical submodel. Hunter survey observations and centroids constitute the sample data ingested annually into iPOM through the AO model, which determines the area occupied by an assumed unique set of individual wolf territories. With assumption testing, we identified three major sources of model bias, which are different but related to the aforementioned data model errors. For iPOM's AO model, they are (1) two types of false-positive errors: misclassifications of non-wolf canids and non-pack wolves and double-counting in their sensitive centroid model; (2) three types of closure violations during the 5-week late fall survey; and (3) resolution bias due to the large grid cell size. Also, AO is the primary variable determining NP (AO/TS) because TS is an agent-based simulator that is not independent of AO (assumption violation), does not ingest empirical data annually, and has as many untested assumptions as it has decision rules. As such, we only evaluated critical AO model assumptions that were suspected to be sensitive (non-robust to failure) and are known sources that might cause bias in AO, NP, and abundance. We also evaluated assumptions peculiar to the AO covariate submodel, which is required to correct for false-positive detection errors.

To assess resolution bias, we conducted a simulation experiment to assess the potential magnitude of the well-known bias due to a large grid cell size (resolution). We could not complete our analysis to evaluate the size (magnitude) of the other two major sources of model bias, nor the overall bias in wolf abundance, because, after request, we could not obtain the needed data and analysis results from Sells et al. [16]. Nonetheless, we provide a qualitative assessment of each source to disentangle its relative contributions to bias. The extent to which their central AO model can correct for all three model biases is contingent upon assumption violations, the data model errors described above, and, most importantly, a proper covariate submodel, which we evaluate below.

5.1. Additional assumption violations at the territory scale

Here, we examine additional factors that cause bias in wolf abundance at the spatial scale of a wolf territory. They are (1) an inappropriate spatial sampling method for estimating the spatial extent of wolf territories and (2) pack dissolution [19,59,60], which amplifies the three sources of closure violations previously described (Section 4.2) and detection errors (Section 4.1) that result in overestimation of abundance. We demonstrate why these factors are critically important when converting annual estimates of the territory area (a) for each grid cell and (b) summed across the state of Montana (e.g., see **Figure 3**), which, when combined with the average TS, determines the number of wolf packs (NP in Equation (1)) each year. A corollary to closure violation for individual wolves is an unstated assumption that territories must be stable (closure) during the late fall survey period.

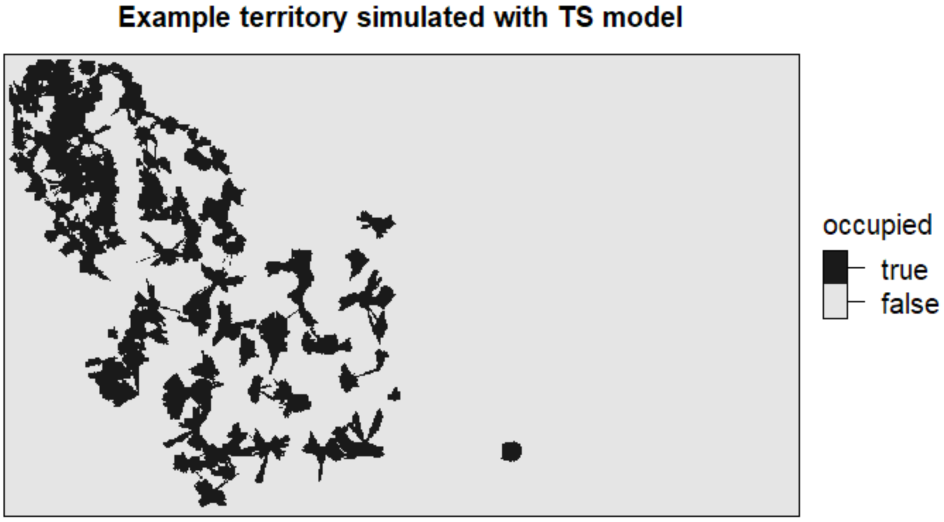

Figure 3. The simulation model of Sells and Mitchell [27] produced a simulated value of the total area occupied by wolf territories (black region).

The fundamental assumption underlying the first factor is that hunter surveys appropriately sample grid cells to “estimate the proportion of grid cells used by wolf packs during the late fall” [16]. This assumption is captured in the ergodic theorem as applied to animal movement, which states that the average time an animal spends moving across its home range is equal to the average over the entire space [61]. If the MFWP’s hunter survey method were valid, this would mean wolf pack observations from hunter surveys constitute an unbiased, representative sample that represents the average movement process or time spent in their territories. This seems impossible because iPOM transfers unrecorded locations of indirect information into a single centroid rather than apportioning use by the location of samples and the effort invested into obtaining those observations. But even if that step could be argued to represent wolf territorial space use without bias, several aspects of the average hunter’s observations make it highly improbable for them to have reported a representative sample of wolf pack observations in a stable territory, as described in Section 4.1. This demonstrates the violation of fundamental rules of unbiased scientific measurement because it is a convenience sample from untrained observers [44]; hunters have possible motivations to exaggerate reports so that a wolf hunting season will be inaugurated with a higher quota; data recording errors occur, not least of which is the missing data itself; and hunters themselves affect the behavior of wolves, which avoid hunting activities but may be attracted later to wounded ungulate prey and gut pile odors [62]. iPOM’s flawed sampling process is in stark contrast to pre-designed ergodic sampling where trained biologists systematically track the movements of a representative sample of radio-collared wolves, which can also suffer from sampling bias when determining space use [63,64].

Pack dissolution is the second territorial factor that causes instability over large areas and violates closure assumptions both demographically and geographically. Typically, one-third of wolf packs dissolve after a breeding adult is killed [59,60], which causes abandonment of the territory (the extinction of a territory centroid) as remaining pack members disperse. Dissolution usually happens during the initial months following breeder loss [60], with extraterritorial movements occurring earlier. Thus, the timing, rate, and speed of dissolution mostly coincide with iPOM’s late fall surveys because the rate and speed of dissolution increase with a smaller pack size [60] due to human-caused mortality. Overall, this effect causes increased movement across the adjacent eight grid cells (closure violation) surrounding the dissolved pack’s cell(s). It also increases double-counting because hunters may observe surviving wolves dispersing or traveling in search of other wolves and mistakenly classify multiple wolves as the presence of a pack.

A simple calculation illustrates the amplifying effect of pack dissolution on closure violation. Assume (1) an average dissolution rate of 40% when a human kills a breeder in smaller packs [59,60]; (2) a mortality rate of 20% for breeders ([54], Figure 9) during the fall hunting season (early September to December), which is before and during the 5-week late fall survey period; and (3) a population of 1000 wolves arranged into 125 packs with a PS = 8. These starting parameters yield ~10 dissolved packs, which produce 70 non-pack, non-breeding wolves that disperse as solo individuals, groups of 2+ wolves, and other temporary interactions with other wolves. Although this is only an 8% dissolution rate, it causes a variety of movement types across grid cells (closure violation) by (1) adjacent pack members, (2) scattered survivors of other dissolutions, and (3) other dispersers—all searching for an opportunity to breed in vacated territories. Unless MFWP seeks absence data from hunter reports and compares those to prior maps of known packs, the agency will not detect true, newly vacated territories.

Although our hypothetical example likely overestimates the double-counting during the fall survey, it does not include the sources of double-counting identified in previous sections. Furthermore, given the highly mobile, wide-ranging behavior of wolves, we suspect that an 8% dissolution rate affects an area as large as half the wolf range in Montana, as adjacent and more distant packs shift and disperse to compete for the vacant territory. It is also important to realize that the effect of killing breeders likely decreases the long-range dispersal (and gene flow) across populations because it provides otherwise long-range dispersers in saturated habitats, a nearby vacancy for a potential breeding opportunity.

5.2. Simulation testing of resolution bias

We simulated the effect of grid cell size resolution for multiple reasons. First, and foremost, was to isolate this known bias from the closure bias by using stable territories from iPOM's TS model within a naive model where the occupancy probability (ψ) = 1.0 for any detection. Second, it gave us insight into how well their AO model might correct for resolution bias due to the presence of a covariate submodel that downsizes a grid cell area's contribution to the AO when summed across Montana ([16], equation 6). Third was to explore the possible nonlinear pattern of sensitivity between resolution and overestimation bias. Fourth, we sought to understand how their TS simulator works, as we used it to represent the distributional dynamics of wolf territories and strengthen our evaluation. Last, we sought to understand the possible advantages of using the scaled occupancy argued for wolves in Wisconsin [22] compared to iPOM.

The premise of Sells et al.'s [16] analysis is that their simulations generate plausible territory layouts for determining the average TS used in iPOM (Equation (2)). Thus, we repurposed their TS simulator to generate wolf territory maps using their code [16], repeating 10 times to generate 10 unique sets of wolf territory maps. These served as example territories in our simulation experiment to assess the direction and magnitude of the potential bias. While the simulated maps do not correspond to actual wolf territories in Montana, they do represent realistic territorial layout patterns and suffice for our experiment. **Figure 3** presents 1 of the 10 simulated territory datasets from the TS simulator. In this simulation, the total area of wolf territories (the black region) is 56,713 km^2. Dividing this area by the 127 simulated packs produces an average territory size of 447 km^2, which is consistent with, but smaller than, the value of 484 km^2 for the period 2014–2019 in Montana, reported in Sells et al. [16]. Oddly, Sells et al. [16] cited work by Rich et al. [65] in the same study area that reported an empirical mean of 600 km^2, ranging from 188 to 2207 km^2. Therefore, we question why they used the smaller, non-empirical territory size produced by a simulator. Directly related, Creel [55] exposed an error in Sells et al.'s [16,39] analysis of TS used in iPOM. This leads to the unjustified use of smaller TS values, which directly causes an overestimate of wolf abundance.

5.3. Direction of resolution bias and closure bias in wolf occupancy modeling

We approximate the total area occupied (AO) by wolf territories (the black area in **Figure 3**). We start with a grid cell overlay at the coarse resolution of iPOM's 600 km^2 onto the wolf territory map. Then, all cells that contain any detected wolf pack observations are assumed to be occupied (ψ = 1.0) and summed to produce a naive AO. Logically, this will overestimate wolf territory size because when a cell partially overlaps with a territory, the non-overlapping areas within that cell are mislabeled as occupied. **Figure 4** illustrates this "edge effect" at three different grid cell sizes. As grid cell size increases, so does the area that is mislabeled as occupied. However, it is unclear from **Figure 4** alone exactly how much bias is introduced. We also wanted to describe the relationship between grid cell size and overestimation bias. So, for each of the 10 simulated territory maps described in the previous section, we tested sample grid cell sizes ranging from 1 km^2 to around 800 km^2. Then, we measured areal bias by computing the resulting AO values and comparing them to the actual area of wolf territories in the 10 simulations.

The resulting increase in overestimation bias for AO is plotted as a function of grid cell size in **Figure 5**. Resolution bias is highly nonlinear and increases rapidly with cell size. The resolution used in iPOM (600 km^2) would overestimate AO and abundance by around 150% (2.5×) if the dynamic model with covariates were not used, which attempts to correct for overestimation biases (resolution and false positives). The average proportion that a corrected grid cell contributed to AO in iPOM ~0.20, which indicated an 80% reduction (1 – ψ) in the AO across Montana. This also means that slight differences in the average occupancy can amount to significant changes in AO and wolf abundance. Even at a grid cell size of 120 km^2 or 20% of the grid cell size in Sells et al. [16], the resultant resolution bias would cause a 50% overestimation of abundance. Due to this, deficiencies in their covariates, and a flawed centroid model, we doubt their 80% reduction (1 – ψ) adequately compensates for the overestimation biases we identified, and we encourage validation by Sells et al. [16].

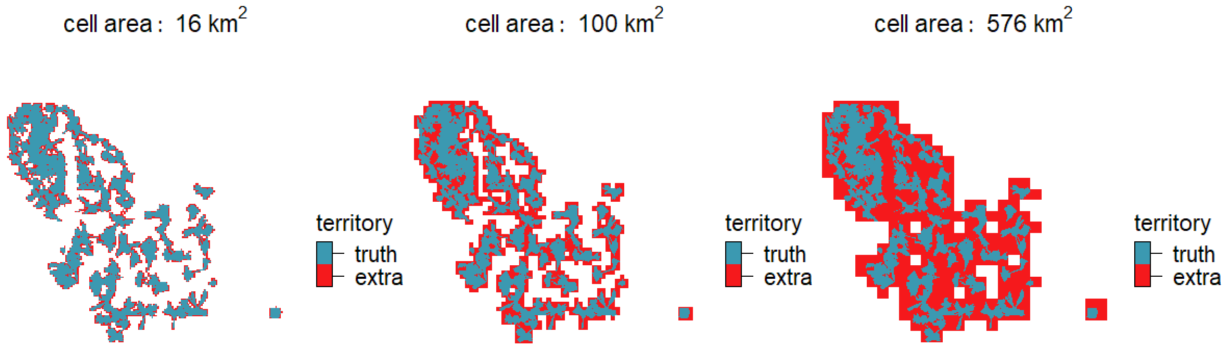

Figure 4. Simulation results that demonstrate the potential overestimation bias in AO at three different grid cell sizes. Wolf territory, in blue, is copied from the example in **Figure 3**. The areal component in red designates unoccupied areas erroneously included in the estimate of area occupied (AO). This would be the amount added to the AO if they applied a naive occupancy model (no covariate model).

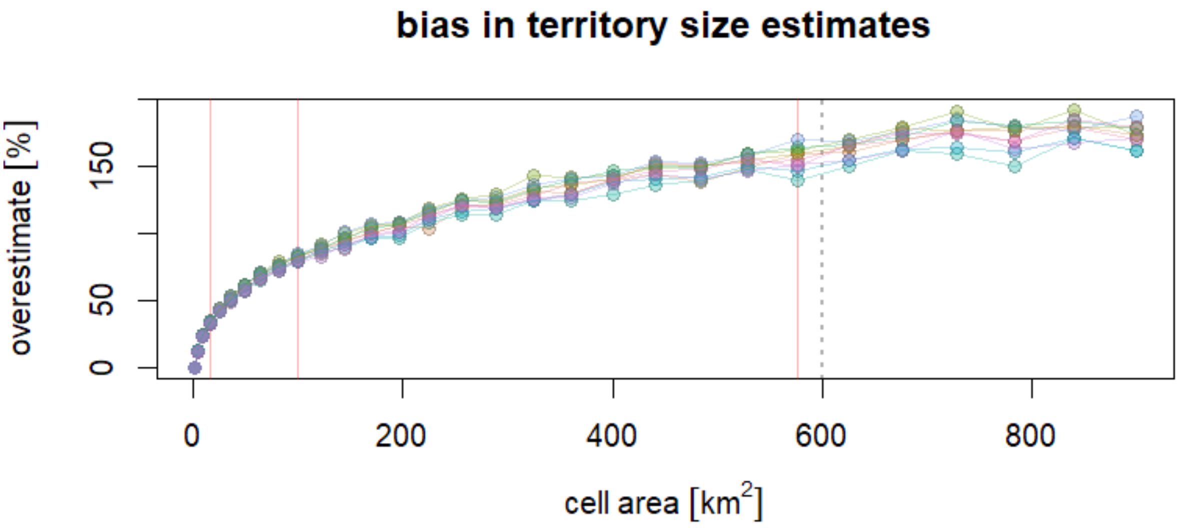

Figure 5. Simulations indicate that larger sampling grid cells (resolution) lead to overestimating the area occupied (AO). Each sample cell size range was tested with ten different (simulated) wolf territory maps. Like-colored points denote results from the same simulations. Solid red vertical lines indicate the three example cell sizes shown above, and the dotted vertical line indicates the resolution of the grid used in iPOM.

Because our simulations, as well as those of Stauffer et al. [22], demonstrate the positive bias caused by larger grid cell sizes, we strongly suggest that iPOM incorporate a scaled occupancy model that directly corrects for resolution bias by subsampling at, say, 20 km² and 100 km². The proper adoption of the SOM should avoid the many errors identified by [22]. There are also other options, such as dynamic scaled occupancy models [31] and structural equation models (SEMs) [66], which are inherently compatible because they provide means to make inferences on latent or unobserved quantities based on observed data and provide an excellent approach to building a non-deficient covariate model. In addition to serious problems in the AO covariate model, we further question whether iPOM's AO model is conditioned properly due to a flawed centroid model. Miller et al. [12] indicated lower sample sizes in the low-quality wolf habitat which makes up the large majority of the habitat in Montana. Because of this and Creel [55] noting that subtle 'tuning' by Sells et al. [16,39] clearly shows that large expanses of areas that were not known to be used by wolves were included in the simulated territories, we assessed whether minor effects in a model could cause significant biases. To illustrate, take the eastern two-thirds of Montana, which is not considered part of the current wolf range [67], and assume ψ at only 0.025 across some 400 low-habitat-quality grid cells, as partially depicted in Figure 1a,c of Miller et al. [12]. This alone would add 10 packs (400 × 0.025) that do not exist. Based on this, we recommend right-truncating the frequency distribution of the grid cell occupancy probabilities. If wolf specialists are able to verify a cell has a confirmed territory centroid for the "certain" model, then one can use similar biological knowledge to assign a zero probability.

5.4. Covariate submodels

The design and parameterization of the covariate submodels dictate how any model corrects, under-corrects, or over-corrects for bias. As such, we found two critical problems in iPOM's covariate submodels within their AO, TS, and PS component models: (1) deficient parameterization (excluding known, important covariates) and (2) the inclusion of mainly static covariates, which renders the AO model incapable of properly correcting for the overestimation biases we have identified. The most important and explicit assumption for covariate modeling is to select a minimally sufficient set of covariates (**Figure 2**) that independently causes the response variable to change. This overriding assumption was violated in all three of iPOM's covariate submodels, which casts serious doubt on the ability to correct for overestimation. Especially with a new method, it is incumbent upon Sells et al. [16] to conduct simulations or other biological testing on whether ψ estimates the proportion of a stable territory in a grid cell.

It is striking that iPOM's a priori selection of covariates is misaligned with substantial research and common biological knowledge about wolves' distributional and abundance dynamics [68]. The AO model excluded obvious causal covariates like snowpack, vegetation, and especially ungulate prey availability, which are well-known predictors of wolf space use (AO and TS), especially in late fall. Ungulate biomass is accepted as the primary predictor of the abundance of wolves and obligate ungulate predators [69], and access to forage predicts ungulate prey distributions. These covariates, including annual ungulate surveys, are available from MFWP; so, rather than providing a review of the problems with iPOM's covariate models, as in [55], we give an example illustrating these problems.

An example of the importance of using covariates that reflect biological mechanisms in predictive models of abundance, that includes numerous lessons for iPOM, is that of Moorcroft and Lewis [70], who developed mechanistic home range models using extensive wolf and coyote datasets, especially those that included mechanical path movements (iPOM's TS model does not). From these models, Moorcroft et al. [71] specifically included prey species densities by vegetation type and were able to accurately predict the shift of one pack into the dissolved territory of a neighboring pack. In a blind test (independent validation), the resulting new territorial distribution matched that from intensive radio-tracking. Nearly the same results would have been achieved by including the ranked importance of the six habitat types from other studies of habitat–prey associations, for example, by using selection models for primary prey (e.g., elk and deer for wolves).

A substantial amount of research conducted regionally and continentally demonstrates the tradeoff between wolf selection for landscape conditions that increase access to available prey, low slopes, more open habitat, and roads and avoidance of humans [72–74]. These tradeoffs can be captured along with climate, prey availability, and vegetation for the fall survey period. Instead, Sells et al. [16] reduced the dimensionality of five correlated covariates to the first principal component that explains 53% of the variation. Hence, this deficient parameterization of the AO covariate model captured mainly the effect of accessibility to wolves by hunters. Additionally problematic was the static nature of those covariates (forest cover, slope, elevation, low-use forested and unforested roads). This severely limits the ability of the AO model to respond each year to expected changes in climate, prey availability, and vegetation, resulting in relatively constant model output values year after year, especially if crucial dynamic input parameters are excluded, as in the TS and PS models.

5.5. Untangling primary sources of overestimation bias

Each of the three major sources of biases we identified in their AO model—false-positive errors, closure violation, and resolution—causes a positive (overestimation) bias when assumptions are violated. In order to understand and improve the false positives in the AO model, we seek to detangle them and identify their biological causation. As such, we identify three culprits that drive these biases to inflate the AO and subsequently overestimate wolf abundance: (1) substantial mortality before and during the late fall survey season leading to delineating wolf territories where none exist, (2) three sources of double-counting wolf territories when subjectively determining centroids, and (3) deficient and static covariate models.

The first culprit (#1) is insidious because it severely violates closure during the surveys and adds to double-counting errors that are already occurring due to uncorrectable misclassifications (#2). Both types of false-positive detection errors that we identified lead to severe overestimation bias ([12,75] for iPOM) due to the sensitivity of the critically important assumption that iPOM's AO centroid model contains only the "certain" centroids of singular wolf territories. The missing but essential confirmation step [76] should come from an independent source of observations similar in place and time. For example, Miller et al. [12] used data from tracking a sufficient sample of radio-collared wolves as confirmation in 2007–2010. Unfortunately, iPOM eliminated the confirmation step and used circular logic that the centroids of documented wolf packs were used in their "certain" centroid model. Wolf specialists using cameras, sightings, tracks, howls, verified wolf depredation sites, and "public tips" do not confirm that their hand demarcation of territory centroids on a map comes from "singularly" unique wolf packs because these observations suffer from the same false-positive biases caused by culprits #1 and #2 above.

We cannot emphasize enough how critically sensitive AO estimates are to even small amounts of false-positive errors, especially double-counting, directly leading to substantial overestimation of wolf abundance. McClintock et al. [77] found that false-positive error rates as small as 1% of all detections caused severe overestimation of occupancy (AO), colonization, and local extinction probabilities, even with experienced observers [78]. In a recent study of gray fox [79], a false-positive error rate of 40% yielded an abundance that was 4x greater than that when accounting for false positives. The same authors conducted a simulation using intensive sampling within spatially separated grids and confirmed this high sensitivity: AO was also severely overestimated with only a 10% false-positive rate.

Although we could not obtain supporting data for the crucial territorial centroid data model and hunter surveys to assess the false-positive rates in iPOM, we examined Parks et al. [54] and found evidence supporting double-counting errors. In the data from 2023, for example ([54], Figure 8), we noticed an unexpected clustering of centroid locations typical of those from 2021 to 2023, a period of liberal wolf hunting with increased mortality and no confirmation step for the centroids. We visually compared these centroid locations with those from an earlier period (2007–2010) when radio-collared wolves were used in a confirmation step to correct and determine singular territory centroids at a time at which wolf mortality was lower [16]. In the earlier period, centroids were evenly spaced out, which is what you would expect from nearly exclusive wolf territories averaging 582 km^2 in 600 km^2 grid cells. This pattern of even spacing even held for high-density areas indicated as saturated [12,16]. To validate this suspected clustering, we tallied a simple cluster statistic for 2007–2010 that resulted in an expected 4–7 grid cells annually that contained two wolf pack centroids. In contrast, 2023, a typical year of the later period, 22 grid cells contained two centroids, and 2 cells contained three centroids. This 5-fold increase in clustered cells is likely due to double-counting errors (culprit #2 above). We strongly recommend that a transparent analysis and validation of this type be conducted and presented.

In addition to the absence of a confirmatory centroid model, iPOM's deficient and static covariate model (culprit #3 above) further limits the ability of iPOM to correct for the significant overestimation biases. Thus, we strongly recommend including the known causal covariates needed to correct for them. Also, updating the dynamic covariate values every year would better account for changing wolf distributions in the fall and allow wolf abundance to respond to annual changes in biological, climate, and landscape factors that affect wolf occupancy. We also strongly recommend tracking a sufficient sample of collared pack wolves (marked territories) to provide the second-method confirmation step required for their centroid model. This could correct for other problems we have identified, including double-counting errors, sampling biases, and informed colonization and extinction parameters. At a minimum, iPOM methods need to include tracking data from a collared subsample of pack wolves in three or four adjacent territories in three to four clusters across Montana to test and verify numerous critical assumptions and provide information on mortality during the fall survey. Another obvious solution to reduce bias is to move the wolf pack survey to the end of winter when double-counting risks are lower and annual mortality has mostly occurred. Then, for several reasons, trained volunteers and wolf specialists should be included who properly conduct designed surveys of wolf observations and wolf pack size during the standard winter period. This could also include the collection of scat for DNA analysis, snow-tracking, and individually distinguishable packs. Snow-track sampling reveals pack behavior, recovers scat, is ergodic, and matches home range determinations from tracking radio-collared red fox [80].

6. Territory size model

Here, we provide a more brief assessment of model bias in the TS model because Creel [55] found an overestimation bias in abundance with a reanalysis of the TS predictions versus empirical estimates [39]. Creel's analysis demonstrated that iPOM's TS simulator output deflates TS by 35%, thereby inflating wolf abundance by the same amount. Using the TS simulator in iPOM is a fundamental violation of scientific inference because it does not ingest annual sample data from the target population of inference. It is neither an estimator nor a prediction and should be considered an uncalibrated index producing a nearly yearly constant value since 2012 ([54], Figure 4). It is non-mechanistic demographically and does not include annual mortality rates, which significantly affect TS, PS, NP, and the LW rate. Although we do not recommend using the outputs of their simulation model, we do not understand why Sells et al. [16] use the TS output when it also produces NP and AO, both used to determine abundance. Because the TS simulator depends on the total number of wolf packs and PS, it is circular logic to use the average territory size to determine NP in iPOM. iPOM's TS model diverged further from empirical reality by developing a density identifier formula to "identify the approximate degree of competition each year" and "to avoid rerunning simulations every year" [16]. This is, by definition, another ad hoc correction factor occurring outside the original modeling procedures. Because the formula is directly tied to territorial centroids, any error or bias in hunter surveys or the centroid model is propagated through TS and abundance. Finally, it would behoove Sells et al. [16] to reanalyze and correct Creel's [55] verification of the bias in predicting TS.

7. Pack size model

Similar to TS, we briefly assess the model biases of iPOM's PS model because of Creel's evaluation [50]. An implicit assumption in iPOM's PS model is that it responds to impacts that increase or decrease pack size. This assumption is violated because their PS model does not ingest annual data on PS or the covariates that are known to affect it, particularly mortality. An increase in adult-caused mortality directly causes a decrease in pack size [54], but their PS model does not include this known effect [59,60]. In southern Montana and northern Wyoming, the recent increase in wolf-killing reduced pack size in five of six packs and caused the dissolution of two packs [19]. In addition, colonizing packs after dissolution from either natural or human-caused mortality often results in an initial pack size of only two that year. Collectively, these errors and omissions result in two effects: (1) an overestimation of PS, which directly overestimates abundance, and (2) a PS model that does not respond to the certainty of pack size reduction in the late fall and winter. Figure 5 in Parks et al. [54] shows no significant change after 2016, where it only varied between 5.3 and 5.4 wolves per pack. The data they used in their regression analysis were from an earlier period when mortality, especially from human killing [19], was less than that in later years. This renders their model and validation irrelevant to 2018 to 2024, when they had insufficient empirical data to estimate PS. Furthermore, their PS covariate model is deficient (see Section 10) and resulted in an expected constant output from 2012 to 2017 [54], a period in which the observed mean pack size changed by ~25%. Creel [55] also demonstrated analysis flaws similar to the regression error in TS. Due to high uncertainty in the regression slope, Sells et al. [16] failed to show that their PS model could make meaningful predictions. We also note that their Poisson regression response variable assumes values of 0, 1, 2, and more; however, this distribution should be truncated at 2 or more to match the definition of a pack. As a result of these solvable problems, it appears their PS model is invalid, does not respond to known annual changes, and overpredicts PS and abundance.

8. iPOM's use of ad hoc parameters

iPOM uses numerous ad hoc parameters and decisions (adjustments, uncalibrated indices, assumed constant values) spread throughout its models. An ad hoc parameter, often referred to as a correction factor, in statistical models is defined as a parameter specifically introduced to adjust or address a particular issue or question arising from the data analysis, rather than being a standard or pre-defined part of the model framework. The same goes for a subjective decision. These forms of model incoherency lead to various problems, including bias in model selection and comparison, model overfitting, model instability, and interpretation difficulty, and cast serious doubt on their ability to make valid inferences [81–83]. We highly recommend conforming to Anderson [44], who reviews the serious problems when using convenience sampling and indices in wildlife studies. Any proxy variable or index used in iPOM covariate models should be calibrated with the true covariate of interest using, for example, double-sampling theory [10]. If not, it produces unaccounted variance and likely bias in their models. At the end of the Supplementary Materials, we provide a partial list of the ad hoc parameters and methods we found.

9. Combining components to predict abundance

We cannot calculate the magnitude of the overestimation bias in iPOM's prediction of wolf abundance because we do not have, or were not allowed, the needed empirical data. Instead, we provide a qualitative appraisal by assuming a realistic inflation of each main submodel and its multiplicative contribution to the overall overestimation of abundance to offer a range of probable bias. For TS, we use a −35% bias based on Creel's [55] analysis. For PS, we chose a minimal +13% bias (4.7 instead of 5.3) because their annual PS values do not reflect biological reality. For example, increased human mortality before and during hunter surveys (including pack dissolution) would lower the PS as in Sells et al. [39]. Then, if we choose, say, a +33% bias for AO due to the numerous errors and biases we documented (closure violations, double-counting, and a deficient covariate model) that cannot be fully accounted for, we arrive at a 2× overestimation bias of wolf abundance. This means, for example, if the true wolf population were 550, iPOM would predict 1100. The overestimation bias could be higher but would not likely exceed 3.5× because double-counting wolf packs is unlikely to be more than double (50% of, say, 120 packs results in a minimum wolf abundance of 300 wolves (60 packs × a PS of 5)). We provide these calculations to encourage and guide Sells et al. [16] towards future corrections, improvements, and simulations to determine the magnitude of the bias.

Furthermore, iPOM has an additional reliability problem because no empirical or independent validation of iPOM's models and abundance predictions was provided. Although Sells et al. [16] compared their predictions to the POM and the minimum verified packs, that is not validation nor verification because one cannot use biased apples to validate biased oranges. This circular problem parallels the iPOM methodology developed by Rich et al. [21], who claimed that hunter survey data offers an opportunity to monitor wolves in Montana but provided no validation yet suggested that occupancy models using hunter sightings needed monitoring to verify presence. We similarly reject the assertion that iPOM's abundance prediction has to be higher than the verified minimum number of known packs. Like the centroid determinations, minimum counts without tracking collared wolves in packs are plagued with double-counting and other errors. Finally, an important assumption—the independence of the models—is violated because the TS and PS models are linked to centroid determinations in the AO model, and the TS model includes data used in the AO model.

We also use a simple assessment of iPOM's output to test whether its use of (1) static and deficient covariate submodels, (2) constant ad hoc parameters that actually vary annually, and (3) substantial changes in methodologies and input data (e.g., centroids) causes iPOM to produce relatively constant abundance estimates after 2016 when wolf abundance peaked. Visual inspection of the reported results on the AO, TS, and PS for 2007 to 2023 ([54], Figures 3–5) raises red flags and supports our contention and that of Creel [55]. Two distinct patterns are apparent: (1) dynamical changes during 2007–2016 when it appears that a proper centroid model was used and (2) relatively little change from 2012 to 2023, when iPOM's methods changed substantially, a radio-marking and tracking program for determining abundance was eliminated, and wolf mortality increased substantially due to public hunting and trapping. As expected, in the last 12-year period (2012–2023), both TS and PS are statistically constant, varying by around 2–3% [54]. AO did show some variability that was governed by two factors: (1) annual hunter surveys and MFWP wolf specialists determining the centroids and (2) an ad hoc density indicator formula tied to the centroids. Nonetheless, the overall wolf abundance did not change in six of seven regions from 2012 to 2023 ([54], Figure 6).

10. iPOM's variance

We found numerous problems with iPOM's estimates of variance used to report the confidence interval for statewide wolf population abundance. Major problems were two-fold: (1) iPOM's method for estimating variance is incorrect because it does not estimate abundance but predicts it and (2) the exclusion of components of variance (input variables), which resulted in a severe underreporting bias and the inability of iPOM to detect a change in abundance. They also made a mathematical error in the variance in their lone wolf rate, LW (Supplementary Materials). Each of these problems causes underreporting (bias), which, when combined, results in a severely underreported confidence interval.

First, proper iPOM predictions have wider prediction intervals than the incorrect confidence intervals Sells et al. [16] used for parameter estimation. Although we were unable to obtain the data sources for each of the iPOM's stated main variables (AO, TS, PS, and LW) to calculate a proper prediction interval, we started with the derivation of variance (and CVs) for each primary input variable as a reality check on possible underreporting biases. We then calculated a Delta method variance to approximate the relative magnitude of the exclusion bias in iPOM's variance used to report the confidence interval (CI) for wolf abundance (Supplementary Materials).

iPOM's prediction of wolf abundance is a random variable and not an estimated parameter. Estimation and prediction are distinct but related concepts in statistical modeling, each serving different purposes and using different methods. As such, there is a substantial difference between estimating a parameter (e.g., abundance from a design-based survey sampling) and predicting a random variable (e.g., abundance for a model-based design that often includes covariates). Estimation focuses on estimating unknown parameters of a model using existing data. Prediction involves a prognostic outcome where the model is constructed to predict for different conditions across space and/or time and uses new data inputs (e.g., relevant covariates) not included in the original sampled data. Also, when estimating a parameter, such as the population abundance for a particular area (e.g., Montana), as the sample size increases (e.g., more individuals are counted), the information about a parameter increases, and the uncertainty about its value decreases; if all animals are counted, then the uncertainty decreases to zero. However, for predictions of population size, as the sample size for the sampled area increases, uncertainty about the prediction for unsampled areas does not decrease to zero. The same holds for iPOM's prediction of total abundance. See [84,85] for a detailed discussion on predicting versus estimating the total of a finite population.

Second is exclusion bias. Logically, excluding input variables when determining variance leads to underestimation of uncertainty, whether calculated directly, approximated linearly (e.g., the Delta method), or calculated using Monte Carlo simulation [86]. Including all key input variables in iPOM—AO, TS, PS, and LW and their covariate submodels—is necessary to capture uncertainty propagation across space and time. Such exclusion errors will result in severe underreporting (bias) of the prediction intervals. Sells et al. [16] omitted the variance in (1) the entire TS model, (2) the entire covariate models, (3) the covariates in the AO and PS submodels, (4) numerous ad hoc correction factors and methods, and (5) the covariance between their main models. Even if the application of a credible interval were appropriate, the uncertainty (variance) of iPOM's abundance estimate would still result in a misleading underestimation of the true variance in wolf abundance. According to scientific principles, a credible interval would need to include at least the major components of variance contained in the random variables of their multiple models and multiple submodels by constructing a joint posterior distribution over all relevant parameters [87].

An example of omitting variation in their covariate models is the exclusion of the variance in their ungulate density index and human density, which are based on sampling and estimates for small areas that are then used to project the predicted density across larger areas. At a minimum, estimates of the variance in or distributions of the ungulate density index for different regions and a variance for human densities should be included in the territory size model. More problematic for their TS predictions is that their highly theoretical territory formation model is based on model agents assessing whether territories were economical and adjusting TS according to model rules about theoretical competition costs due to theoretical neighboring wolf packs. iPOM's abundance estimate did not include these substantial variance components, which consists of rules with no reported empirical data; the rules were calibrated by matching the model results with desired wolf densities. Similar exclusion errors occurred in their PS and later AO models.

To determine the minimum magnitude of the underreporting bias in iPOM's variance used to determine the CI [16], we include the variance in empirically derived input values using the Delta method [88]. As a reality check, to start, we examined the sensitivity of iPOM's variance to exclusion bias by adding the proper variance in a model component, LW, using the empirical samples in Sells et al. [16]. We concluded that iPOM's CI was unrealistic because the CV for LW at 10.2% was larger than iPOM's CV of 5.5% for wolf abundance (Supplementary Materials). Based on this, we proceeded to examine the main model components, AO, TS, and PS.

Although we could not obtain the actual sample data, we evaluated the exclusion biases in iPOM's variance (and CI) estimate further by first determining the components of variance in its main input variables by (1) using calculations for AO as we did for NP, (2) extracted empirical data values by reading them from Figures 1.8 and 1.16 in Sells et al. [39] for TS and PS, and (3) using the LW variance above (Supplementary Materials). The resultant CVs for their main model components AO, TS, PS, and LW were 4%, 72%, 45.3%, and 10.2%, respectively, compared to iPOM's CV for wolf abundance at 5.5%.

We then proceeded to estimate the Delta method of variance (Supplementary Materials) for wolf abundance using iPOM's variance for NP (var = 79.7 for 161 packs) and the LW proportions (var = 0.039, mean = 1.120), combined with our variance estimate for 47 pack size samples (var = 6.5, mean = 5.63). The result was an abundance of 1016 wolves with a CI of 0 to 2243 and a CV of 45.3% (an 8.4× underreporting bias), demonstrating iPOM's inability to detect any change in wolf abundance. The true variance (and underreporting bias) in wolf abundance would be even higher because our minimum approximation does not include (1) a proper prediction interval, (2) the variance in the excluded covariate submodels, (3) covariance due to violation of component model independence, (4) spatial and temporal mismatch of the variables, and (5) uncalibrated indices and ad hoc correction factors. Also, iPOM's variance for NP does not appear to include the variance from TS because its CV (72%) is much larger than 5.5%. However, we recommend, of course, that Sells et al. [16] transparently calculate and share a proper prediction interval, along with all of the data.

11. Coherence and reproducibility

The question "how good is your science?"—whether experiments or statistical models—can be answered by "how well it predicts". This also measures the strength of inference [89]. For statistical models, predictions are evaluated by determining bias (assumption testing) and variance (**Figures 1 and 2**), as well as model coherence and reproducibility. Without an assessment of these attributes, iPOM can neither be considered scientific nor its predictions reliable.

11.1. Coherence

Although model coherency is an essential attribute for a complex model like iPOM, it has several different but related meanings. In predictive models, coherency ensures the reliability of statistical inference through alignment with the study objectives and design, consistency in data aggregation, and data sharing between submodels and across levels [90]. The concept also includes covariate models in probabilistic judgments in Bayesian ecological modeling approaches [91], inference in evolution [92], and AI language models [93]. In population modeling, hierarchical and integrated population models are explicitly coherent because they are structurally designed with integration mechanisms across domains and levels, allowing for consistent integration and harmonization of different datasets within a single inferential framework [28,30]. Taken together, a coherent statistical model reflects how samples can be consistently brought together with other information within a model or between models under one broader analytic framework (e.g., iPOM). Thus, coherency is an intentional, consistent, and logically structured approach where all model components and data align and flow harmoniously from start to finish (design, sampling, data model, process model, explicit and implicit assumptions, covariates, and output). Based on these characteristics, iPOM lacks nearly every feature and step in the structure and process needed to develop a coherent model. iPOM's lack of an explicit mathematical equation—a stochastic model of the sample data and sampling process—that defines each unknown parameter and its corresponding spatial and temporal units is indicative of an incoherent model. Rather, iPOM attempts to combine, not integrate, model components.

The most flagrant violation of coherency is iPOM's ingestion of information from outside the population for which it is making inference. For example, iPOM ingests values from the TS simulator, not empirical sample values from Montana's wolf population. Such a mismatch causes bias and results in misleading inference [11]. This also holds for iPOM's Monte Carlo simulations because iPOM's data resampling efforts include data from populations outside of Montana or from a different time period [16]. Second, iPOM's submodels are incoherent within themselves. The TS and PS models (and LW rate) do not ingest concurrent annual inputs of empirical samples from Montana's wolf population. If iPOM's AO model dynamic [12] had included (1) a coherent link to a properly designed and sampled data model, (2) a centroid model from a sufficiently large sample of representatively sampled radio-marked wolf pack members, and (3) a non-deficient covariate model of known, causal covariates, it would be considered coherent but only as a stand-alone submodel disconnected to the TS and PS models. Finally, iPOM's covariate models do not ingest annual inputs from surveyed areas, and inferential yearly predictions are not valid. Because of this, it is no surprise that iPOM produces relatively constant values yearly (since 2016) but with some variation due to flawed hunter surveys and subjective centroid locations. We found numerous cases of spatial and temporal mismatch within iPOM's covariate models and between them and the response variable. Such mismatch or oblique problems are also pointed out by Creel [55] and lead to unreliable and potentially misleading conclusions because the covariates may not accurately capture the conditions influencing the response variable at the time and location of the observations [94,95].

11.2. Reproducibility

Reproducibility is a fundamental hallmark of science related to falsifiability. Just as scientific hypotheses and theories must be subject to falsification or considered unscientific, a research finding must be repeatable, or it cannot be regarded as reliable. Much attention has been paid to failed efforts at replication [96,97]. Here, we address the prerequisites for the replication attempts applied to iPOM's findings [16].

Before a specific finding can be replicated and confirmed by qualified investigators following every step of the original methods, reproducibility demands two tests of the original findings [98–100]. The first test is whether the procedures were described so that a replication effort stands a chance of succeeding. If the methods are incomplete, unclear, or impossible to repeat, one cannot succeed in replication. IPOM fails this, as Sells et al. [16] include subjective methods (hand-drawn centroids with subjective placements) or unexplained methods (e.g., hunter surveys). The second test is whether the data underlying the original finding exist. If the data does not exist, they cannot support the original claim. Both tests fail for Sells et al. [16], yet the missing data concern us the most. The lead author of this paper had to pursue multiple private communications with the senior author Sells and other coauthors to elucidate the missing steps and data, yet guesswork remains over 2 years since those communications began. Whether the methods can be explained remains uncertain and is undoubtedly unanswered by the published article. Therefore, we call for a correction or editorial flag of caution about Sells et al. [16] at a minimum. Correction may give way to a need for retraction if the second test also fails, as described next.

Essential input data in Sells et al. [16] are hunters' observations of wolves. They represent the primary source of empirical observations of the current presence of wolves on the Montana landscape. Even though such observations cannot be verified to be true wolves, the survey method used to collect those data has not been sufficiently described and is therefore irreproducible. As such, one cannot rule out, or correct for, double- or even triple-counting. We do not know the standardized methods of collection or quality control steps. Worse, the survey data do not exist, according to communication from MFWP. In sum, the primary empirical data based on which wolf abundance was estimated in Sells et al. [16] appear not to exist anymore. If Sells et al. [16] do not confirm to the editors that the data exist, this would be grounds for retraction according to journal policies and the Ecological Society of America's code of ethics for authors. It also frustrates our efforts at replication and improvement. That claim of open data also demands a correction or published notice of concern by the editors. Therefore, we conclude that iPOM as it currently reads is irreproducible and unscientific and merits retraction.

12. Conclusions

For a variety of reasons, wolves are a poor candidate species for the application of occupancy modeling. Their high mobility and wide-ranging behavior ensure violation of sensitive assumptions that lead to overestimation bias, especially geographic closure. Furthermore, their complex social behavior, adding to already high spatial and temporal heterogeneity as they traverse complex landscapes, makes it difficult at best to correct for detection errors with covariate models. Yet tantamount to these issues is the fundamental mismatch between a complex statistical model proposed for population demography—iPOM's abundance over time—and the incorporation of two spatial models developed for non-demographic applications. Occupancy modeling was designed to determine species distributions and habitat relations, whereas iPOM's TS simulator is a theoretical agent-based model for simulating territorial behavior. It is important to note that Smith and Boyd (the latter co-author with Sells et al. [16]) concluded that iPOM should not be used to determine wolf abundance and that alternative methods like genetic capture–recapture should be applied [101].

The shoe-horning of iPOM's main models resulted in an incoherent mash-up of model components that lack several foundational principles governing inference, any one of which invalidates its use in wildlife decision-making. Based on our evaluation criteria, extensive assumption testing, and the concepts governing valid inference, there is overwhelming evidence that iPOM is unreliable. A federal court agrees with this conclusion on largely independent and qualitative grounds (Center for Biological Diversity et al. and Western Watersheds Project et al. v US Fish and Wildlife Service et al. 2025. U.S. District Court for the District Of Montana, 9:24-cv-00086-DWM Doc 98, hereafter CBD & WWP v FWS 2025). Although we could not estimate precisely the magnitude of the bias in iPOM's abundance predictions, the evidence indicates a severe overestimation bias. At this point, we do not know the wolf population size (abundance) in Montana, and iPOM's variance and CI are so severely underreported (at least an 8.4× bias) that iPOM cannot detect a change in abundance. This results in the worst-case scenario for managers and wolves: decision-makers are misled because Sells et al. [16] claim accurate (low bias) and precise estimates (**Figure 1A**), when iPOM yields neither (**Figure 1D**). Sells et al. [16] also arrived at an uncommon situation in statistical modeling where the addition of covariates failed to minimize bias (**Figure 2**), thereby increasing both the variance and model error. iPOM further detracted from biological reality by (1) using an incorrect statistical analysis for their PS and TS models which, again, resulted in overestimation bias; (2) using numerous ad hoc correction factors [44]; and (3) excluding substantial sources of variance in their input variables that ensured that iPOM cannot detect a change if abundance declines to less than 150 individuals. Again, the court (in CBD & WWP v FWS 2025) ordered the FWS back to the drawing board to use the best available science to address the adequacy of state regulatory mechanisms. The court expressed concern over the inadequacy of Montana's regulatory mechanisms to keep wolves from being put back on the federal list of endangered species.

IPOM's lack of the sampling design needed for reliable predictions is at odds with 21st century standards and strong inference. At its core, iPOM's goal was to use hunter information as its primary data source to estimate the fraction of a grid cell that belongs to a singularly identifiable wolf territory, yet the "sample" data are from surveys of deer and elk hunters. If those data can be recovered, they nevertheless come from a set of untrained individuals of unknown independence from each other, who had to recollect the spatial and temporal accuracy of their observations of assumed wolf pack members. They (1) inadvertently and incompletely sample grid cells that are too large, (2) miss entire grid cells, and (3) fail to provide the representative sampling needed to estimate the space use of wolves belonging to stable territorial packs during the late fall 5-week sampling period. IPOM's approach cannot correct these sampling biases, even with the crucial confirmation step (proper centroid model) from a sufficient sample of collared wolves from confirmed packs at that time. Moreover, the difficulty classifying wolves as pack members and their natural spatial instability in late fall are amplified by the killing and disintegration of packs before and during the survey period.

Although the original application of iPOM's AO model during the early 2007–2015 period appeared to correct partially for the three major overestimation biases identified (false-positive errors, closure violation, and resolution), an insufficient number of radio-collared wolves and increased mortality after 2015 ensured violation of the highly sensitive confirmation step and the sensitive closure assumption, respectively. Together, this resulted in severe overestimation of abundance. In addition, the potentially corrective covariate model was deficient and contained static covariates, which limited its ability to correct for overestimation biases and resulted in a constant annual output after 2015, when the environmental conditions changed. Here, too, the CBD & WWP v FWS 2025 court chided the FWS for relying on population size forecasts that did not make use of all of the available information on Montana's wolves through 2023.

Sells et al. [16] did achieve a part of their primary goal to substantially reduce their reliance upon expensive empirical field sampling of wolves. However, in doing so, they ensured the failure of their overriding goal to produce reliable predictions of wolf abundance. Even more problematic is their claim [16] that valid inference can be made without including sample data from the population of interest [55]. There will always be significant model errors and weak inferences when information comes from sources not belonging to the sampled population. Perhaps the most considerable lacuna is iPOM's exclusion of annual mortality, which is known to affect the primary input variables, AO, TS, PS, and LW, and demographic rates of wolves, such as dispersal, emigration, reproduction, and recruitment. These demographic forces might be especially strong in late fall when the surveys are conducted.

Regardless of whether or not iPOM can achieve reliability, we strongly recommend comparing its costs and reliability to alternative methods. Any comparison should be weighed against the economic benefits minus the costs of large carnivore populations, especially the ecological function of wolves in naturally forming larger packs [19]. In a recent review of population abundance models, Iijima [102] agreed with Royle and Dorazio [103] and Kery and Royle [104] that recent advances in hierarchical modeling should become a fundamental standard framework for the development, testing, and application of abundance methods because they provide a coherent structure and process. As a result, and because of the exceptional difficulties in monitoring carnivores and the need for reliability, recent investigators are simultaneously combining two independent methods for abundance estimation, one of which requires individual identification (e.g., [105]). Even SECR models, considered the recent gold standard for estimating carnivore abundance, can not be reliably used without at least a substantial subsample of marked individuals [106]. There are also reliable and robust capture–recapture models that can be applied using various methods that do not require physical restraint of wild individuals [107,108]. Individual marks can be derived from DNA analysis and unique natural markings from confirmed photos and resighting observations. With low-cost delivery systems, there are also biochemical scat markers such as iophenoxic acid [109], chlorinated benzenes [110], and isotopes, both stable and unstable [111]. Additional research could result in the delivery of visible dye marks and patterns using drones or autonomous ground devices. In particular, non-invasive scat detection may be well worth revisiting in light of genetic, laboratory, and field innovations that have reduced the costs of individual detection [112]. A recent method for successful genetic mark–recapture for wolves [113,114] is currently being tested in Montana, with encouraging preliminary results. Additional biological information valuable to wolf management and conservation comes from non-invasive techniques identifying individuals [115–119].

Funding: This research was supported by the Jodar Family Foundation and the Rangeland Foundation.

Author contributions: R.C. took the lead in writing. R.C., M.C., and A.T. for conceptualization. R.C. and M.C. for analysis. M.C. and A.T. led the sections on variance and reproducibility, respectively. R.C. and A.T. submitted and revised reviewers' comments.

Conflict of interest: The authors declare that they have no competing interests.

Data availability statement: We extracted data from existing published papers [16, 39, 54, 69] for our evaluation review. More detailed analysis results are included in Supplement S1.

Institutional review board statement: Not applicable.

Informed consent statement: Not applicable.

Supplementary materials

The supplementary materials are available at https://doi.org/10.20935/xxx. References [120–123] are cited in the supplementary materials.

Additional information

Received: 2025-04-25

Accepted: 2025-09-23

Published: yyyy-mm-dd

Academia Biology papers should be cited as *Academia Biology 2025*, ISSN 2837-4010, https://doi.org/10.20935/xxx. The journal's official abbreviation is *Acad. Biol.*

Publisher's note

Academia.edu Journals stays neutral with regard to jurisdictional claims in published maps and institutional affiliations. All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Crabtree RL, Conner MC, Treves A. Misleading overestimation bias in methods to estimate wolf abundance that use spatial models. Academia Biology. 2025;in press. https://doi.org/10.31220/agriRxiv.2023.00215.

Copyright

References

1. Nichols JD, Runge MC, Johnson FA, Williams BK. Adaptive harvest management of North American waterfowl populations: a brief history and future prospects. J Ornithol. 2007;148:343–9. doi: 10.1007/s10336-007-0256-8
2. Williams BK, Brown ED. Double-loop learning in adaptive management: the need, the challenge, and the opportunity. Environ Manag. 2018;62:995–1006. doi: 10.1007/s00267-018-1107-5
3. Estes JA, Terborgh J, Brashares JS, Power ME, Berger J, Bond WJ, et al. Trophic downgrading of planet Earth. Science. 2011333(6040):301–6. doi: 10.1126/science.1205106
4. Ripple WJ, Estes JA, Beschta RL, Wilmers CC, Ritchie EG, Hebblewhite M, et al. Status and ecological effects of the world's largest carnivores. Science. 2014;343(6167):1241484. doi: 10.1126/science.1241484
5. Speldewinde PC, Slaney D, Weinstein P. Is restoring an ecosystem good for your health? Sci Total Environ. 2015502:276–9. doi: 10.1016/j.scitotenv.2014.09.028
6. Gutierrez-Arellano C, Mulligan M. A review of regulation ecosystem services and disservices from faunal populations and potential impacts of agriculturalisation on their provision, globally. Nat Conserv. 2018;30;1-39. doi: 10.3897/natureconservation.30.26989
7. Raynor JL, Grainger CA, Parker DP. Wolves make roadways safer, generating large economic returns to predator conservation. Proc Natl Acad Sci USA. 2021;11822):e2023251618. doi: 10.1073/pnas.2023251118
8. Thornton H. Ecosystem restoration and species recovery benefit people and planet. UN chronicle. 2022 [accessed on 1 May 2024]. Available from: https://www.un.org/en/un-chronicle/ecosystem-restoration-and-species-recovery-benefit-people-and-planet.
9. Hoeks S, Huijbregts MA, Busana M, Harfoot MB, Svenning JC, Santini L. Mechanistic insights into the role of large carnivores for ecosystem structure and functioning. Ecography. 2020;43(12):1752–63. doi: 10.1111/ecog.05191
10. Cochran WG. Sampling techniques. Hoboken (NJ): John Wiley & Sons; 1977.
11. Williams BK, Brown ED. Sampling and analysis frameworks for inference in ecology. Methods Ecol Evol. 2019;10(11:1832–42. doi: 10.1111/2041-210X.13279
12. Miller DA, Nichols JD, Gude JA, Rich LN, Podruzny KM, Hines JE, et al. Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data. PLoS ONE. 2013;8(6):e65808. doi: 10.1371/journal.pone.0065808
13. Alonso RS, McClintock BT, Lyren LM, Boydston EE, Crooks KR. Mark-recapture and mark-resight methods for estimating abundance with remote cameras: a carnivore case study. PloS ONE. 201510(3):e0123032. doi: 10.1371/journal.pone.0123032
14. Karanth KU, Nichols JD. Estimation of tiger densities in India using photographic captures and recaptures. Ecology. 1998;79(8):2852–62. doi: 10.1890/0012-9658(1998)079[2852:EOTDII]2.0.CO;2
15. Mills LS, Whiteley AR, Tourani M. Conservation of wildlife populations: applications of ecological, evolutionary, and genetic concepts. 3rd ed. Oxford: Oxford University Press; 2025.
16. Sells SN, Podruzny KM, Nowak JJ, Smucker TD, Parks TW, Boyd DK, et al. Integrating basic and applied research to estimate carnivore abundance. Ecol Appl. 2022;32(8):e2714. doi: 10.1002/eap.2714
17. Boitani L, Mech LD, editors. Wolves: behavior, ecology, and conservation. Chicago (IL): University of Chicago Press; 2010.
18. Mitchell MS, Ausband DE, Sime CA, Bangs EE, Gude JA, Jimenez MD, et al. Estimation of successful breeding pairs for wolves in the Northern Rocky Mountains, USA. J Wildl Manag. 2008;72(4):881–91. doi: 10.2193/2007-157
19. Cassidy KA, Borg BL, Klauder KJ, Sorum MS, Thomas-Kuzilik R, Dewey SR, et al. Human-caused mortality triggers pack instability in gray wolves. Front Ecol Environ. 2023;21(8):356–62. doi: 10.1002/fee.2597
20. Crabtree RL, Sheldon JW, Wang Y. Monitoring and modeling environmental change in protected areas: integration of focal species populations and remote sensing. In: Remote sensing of protected lands. Boca Raton (FL): CRC Press; 2012. p. 495–524.
21. Rich LN, Russell RE, Glenn EM, Mitchell MS, Gude JA, Podruzny KM, et al. Estimating occupancy and predicting numbers of gray wolf packs in Montana using hunter surveys. J Wildl Manag. 2013;77(6):1280–9. doi: 10.1002/jwmg.562
22. Stauffer GE, Roberts NM, Macfarland DM, Van Deelen TR. Scaling occupancy estimates up to abundance for wolves. J Wildl Manag. 2021;85(7):1410–22. doi: 10.1002/jwmg.22105
23. Ausband DE, Lukacs PM, Hurley M, Roberts S, Strickfaden K, Moeller AK. Estimating wolf abundance from cameras. Ecosphere. 2022;13(2):e3933. doi: 10.1002/ecs2.3933
24. Fuller MR, Fuller TK. Radio-telemetry equipment and applications for carnivores. In: Carnivore ecology and conservation: a handbook of techniques. Vol. 162. Oxford: Oxford University Press; 2012. p. 168.
25. Elliot NB, Gopalaswamy AM. Toward accurate and precise estimates of lion density. Conserv Biol. 2017;31(4):934–43. doi: 10.1111/cobi.12878
26. Treves A, Santiago-Ávila FJ. Estimating wolf abundance with unverified methods. Acad Biol. 2023;1(2):1–10. doi: 10.20935/AcadBiol6099

27. Sells SN, Mitchell MS. The economics of territory selection. Ecol Model. 2020;438:109329. doi: 10.1016/j.ecolmodel.2020.109329
28. Schaub M, Kéry M. Components of integrated population models. In: Integrated population models, theory and ecological applications with R and JAGS. Cambridge (MA): Academic Press; 2022.
29. Horne JS, Ausband DE, Hurley MA, Struthers J, Berg JE, Groth K. Integrated population model to improve knowledge and management of Idaho wolves. J Wildl Manag. 2019;83(1):32–42. doi: 10.1002/jwmg.21554
30. Plard F, Fay R, Kéry M, Cohas A, Schaub M. Integrated population models: powerful methods to embed individual processes in population dynamics models. Ecology. 2019;100(6):e02716. doi: 10.1002/ecy.2715
31. Bailey LL, MacKenzie DI, Nichols JD. Advances and applications of occupancy models. Methods Ecol Evol. 2014;5(12):1269–79. doi: 10.1111/2041-210X.12100
32. Rota CT, Fletcher Jr RJ, Dorazio RM, Betts MG. Occupancy estimation and the closure assumption. J Appl.Ecol. 2009;46(6):1673–81. doi: 10.1111/j.1365-2664.2009.01734.x
33. Otto CR, Bailey LL, Roloff GJ. Improving species occupancy estimation when sampling violates the closure assumption. Ecography. 2013;36(12):1299–309. doi: 10.1111/j.1600-0587.2013.00137.x
34. Miller DA, Nichols JD, McClintock BT, Grant EH, Bailey LL, Weir LA. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology. 2016;92(7):1422–8. doi: 10.1890/10-1396.1
35. MacKenzie DI, Nichols JD, Hines JE, Knutson MG, Franklin AB. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology. 2003;84(8):2200–7. doi: 10.1890/02-3090
36. Royle JA, Link WA. Generalized site occupancy models allowing for false positive and false negative errors. Ecology. 2006;87(4):835–41. doi: 10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2
37. Nimon KF. Statistical assumptions of substantive analyses across the general linear model: a mini-review. Front Psychol. 2012;3:322. doi: 10.3389/fpsyg.2012.00322
38. Wilensky U. NetLogo (5.0) [computer software]. Evanston (IL): Center for Connected Learning and Computer-Based Modeling, Northwestern University; 1999.
39. Sells SN, Keever AC, Mitchell MS, Gude J, Podruzny K, Inman R. Improving estimation of wolf recruitment and abundance, and development of an adaptive harvest management program for wolves in Montana. Final report for federal aid in wildlife restoration grant W-161-R-1. Helena (MT): Montana Fish, Wildlife and Parks; 2020.
40. Chambert T, Grant EH, Miller DA, Nichols JD, Mulder KP, Brand AB. Two-species occupancy modelling accounting for species misidentification and non-detection. Methods in Ecology and Evolution. 2018 Jun;9(6):1468-77.
41. Greenwood JJ. Citizens, science and bird conservation. J Ornithol. 2007;148(Suppl. 1):77–124. doi: 10.1007/s10336-007-0239-9
42. Arazy O, Malkinson D. A framework of observer-based biases in citizen science biodiversity monitoring: semi-structuring unstructured biodiversity monitoring protocols. Front Ecol Evol. 2021;9:693602. doi: 10.3389/fevo.2021.693602
43. Van Strien AJ, Van Swaay CA, Termaat T. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. J Appl Ecol. 2013;50(6):1450–8. doi: 10.1111/1365-2664.12158
44. Anderson DR. The need to get the basics right in wildlife field studies. Wildl Soc Bull. 2001;29:1294–7.
45. Daru BH, Rodriguez J. Mass production of unvouchered records fails to represent global biodiversity patterns. Nat Ecol Evol. 2023;7(6):816–31. doi: 10.1038/s41559-023-02047-3
46. Altwegg R, Nichols JD. Occupancy models for citizen-science data. Methods Ecol Evol. 2019;10(1):8–21. doi: 10.1111/2041-210X.13090
47. Treves A, Artelle KA, Darimont CT, Parsons DR. Mismeasured mortality: correcting estimates of wolf poaching in the United States. J Mammal. 2017;98(5):1256–64. doi: 10.1093/jmammal/gyx052
48. Fuller TK. Population dynamics of wolves in north-central Minnesota. Wildl Monogr. 1989;(1):3–41.
49. Jimenez MD, Bangs EE, Boyd DK, Smith DW, Becker SA, Ausband DE, et al. Wolf dispersal in the rocky mountains, Western United States: 1993–2008. J Wildl Manag. 2017;81(4):581–92. doi: 10.1002/jwmg.21238
50. Roffler GH, Pilgrim KL, Williams BC. Patterns of wolf dispersal respond to harvest density across an island complex. Animals. 2024;14(4):622. doi: 10.3390/ani14040622
51. Mech LD. Unexplained patterns of grey wolf Canis lupus natal dispersal. Mammal Rev. 2020;50(3):314–23. doi: 10.1111/mam.12198
52. Benson JF, Keiter DA, Mahoney PJ, Allen BL, Allen L, Alvares F, et al. Intrinsic and environmental drivers of pairwise cohesion in wild Canis social groups. Ecology. 2025;106(1):e4492. doi: 10.1002/ecy.4492
53. Mech L. The wolf: the ecology and behavior of an endangered species. Garden City (NY): The Natural History Press; 1970.
54. MacKenzie D. Occupancy modelling—home range vs grid cell size. 2018 [accessed on 16 March 2024]. Available from: https://www.youtube.com/watch?v=6TtVXy91AY4.
55. Parks MK, Podruzny K, Cole W, Parks T, Lance N, Zielke S, et al. Montana gray wolf conservation and management 2023 annual report. Helena, Montana: Montana Fish, Wildlife & Parks; 2024.
56. Creel S. Methods to estimate population sizes of wolves in Idaho and Montana. Comments in federal register 86 FR 51857, “endangered and threatened wildlife and plants; 90-Day finding for two petitions to list the gray wolf in the western united states”. 2023 [accessed on 20 October 2023]. Available from: https://www.regulations.gov/comment/FWS-HQ-ES-2021-0106-

49075.
57. Brack IV, Kindel A, Oliveira LF. Detection errors in wildlife abundance estimates from Unmanned Aerial Systems (UAS) surveys: synthesis, solutions, and challenges. Methods Ecol Evol. 2018;9(8):1864–73. doi: 10.1111/2041-210X.13026
58. Terletzky PA, Koons DN. Estimating ungulate abundance while accounting for multiple sources of observation error. Wildl Soc Bull. 2016;40(3):525–36. doi: 10.1002/wsb.672
59. Strickfaden KM, Fagre DA, Golding JD, Harrington AH, Reintsma KM, Tack JD, et al. Dependent double-observer method reduces false-positive errors in auditory avian survey data. Ecol Appl. 2020;30(2):e02026. doi: 10.1002/eap.2026
60. Brainerd SM, Andrén H, Bangs EE, Bradley EH, Fontaine JA, Hall W, et al. The effects of breeder loss on wolves. J Wildl Manag. 2008;72(1):89–98. doi: 10.2193/2006-305
61. Borg BL, Brainerd SM, Meier TJ, Prugh LR. Impacts of breeder loss on social structure, reproduction and population growth in a social canid. J Anim Ecol. 2016;84(1):177–87. doi: 10.1111/1365-2656.12256
62. Moore CC. Ergodic theorem, ergodic theory, and statistical mechanics. Proc Natl Acad Sci USA. 2015;1127):1907–16. doi: 10.1073/pnas.1421798112
63. Ruth TK, Smith DW, Haroldson MA, Buotte PC, Schwartz CC, Quigley HB, et al. Large-carnivore response to recreational big-game hunting along the Yellowstone National Park and Absaroka-Beartooth Wilderness boundary. Wildl Soc Bull. 2003;31:1650–61.
64. Noonan MJ, Tucker MA, Fleming CH, Akre TS, Alberts SC, Ali AH, et al. A comprehensive analysis of autocorrelation and bias in home range estimation. Ecol Monogr. 2019;89(2):e01344. doi: 10.1002/ecm.1344
65. Peris A, Closa F, Marco I, Acevedo P, Barasona JA, Casas-Díaz E. Towards the comparison of home range estimators obtained from contrasting tracking regimes: the wild boar as a case study. Eur J Wildl Res. 2020;66:32. doi: 10.1007/s10344-020-1370-7
66. Rich LN, Mitchell MS, Gude JA, Sime CA. Anthropogenic mortality, intraspecific competition, and prey availability influence territory sizes of wolves in Montana. J Mammal. 2012;93(3):722–31. doi: 10.1644/11-MAMM-A-079.2
67. Joseph MB, Preston DL, Johnson PT. Integrating occupancy models and structural equation models to understand species occurrence. Ecology. 2016;97(3):765–75. doi: 10.1890/15-0833.1
68. U.S. Fish and Wildlife Service. Species status assessment for the Gray Wolf (*Canis lupus*) in the Western United States. Version 1.2. Lakewood (CO): U.S. Fish and Wildlife Service; 2023; 362p.
69. Kittle AM, Anderson M, Avgar T, Baker JA, Brown GS, Hagens J, et al. Landscape-level wolf space use is correlated with prey abundance, ease of mobility, and the distribution of prey habitat. Ecosphere. 2017;8(4):e0178363. doi: 10.1002/ecs2.1783
70. Fuller, TK. Mech, LD, and Cochrane, JF. Wolf population dynamics. In: Mech LD, Boitani L, editors. Wolves: behavior, ecology, and conservation. Chicago (IL) & London: University of Chicago Press; 2003. p. 322.
71. Moorcroft PR, Lewis MA. Mechanistic home range analysis. (MPB-43). Princeton (NJ): Princeton University Press; 2013.
72. Moorcroft PR, Lewis MA, Crabtree RL. Mechanistic home range models capture spatial patterns and dynamics of coyote territories in Yellowstone. Proc R Soc B Biol Sci. 2006;273(1694):1651–9. doi: 10.1098/rspb.2005.3439
73. Muhly TB, Johnson CA, Hebblewhite M, Neilson EW, Fortin D, Fryxell JM, et al. Functional response of wolves to human development across boreal North America. Ecol Evol. 2019;9(18):10801–16. doi: 10.1002/ece3.5600
74. Milakovic B, Parker KL, Gustine DD, Lay RJ, Walker AB, Gillingham MP. Habitat selection by a focal predator (*Canis lupus*) in a multiprey ecosystem of the northern Rockies. J Mammal. 2016;92(3):568–82. doi: 10.1644/10-MAMM-A-040.1
75. Mladenoff DJ, Sickley TA, Wydeven AP. Predicting gray wolf landscape recolonization: logistic regression models vs. new field data. Ecol Appl. 1999;9(1):37–44. doi: 10.1890/1051-0761(1999)009[0037:PGWLRL]2.0.CO;2
76. Ferguson PF, Conroy MJ, Hepinstall-Cymerman J. Occupancy models for data with false positive and false negative errors and heterogeneity across sites and surveys. Methods Ecol Evol. 2016;6(12):1395–406. doi: 10.1111/2041-210X.12442
77. Chambert T, Miller DA, Nichols JD. Modeling false positive detections in species occurrence data under different study designs. Ecology. 2016;96(2):332–9. doi: 10.1890/14-1507.1
78. McClintock BT, Bailey LL, Pollock KH, Simons TR. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. Ecology. 2010;91(8):2446–54. doi: 10.1890/09-1287.1
79. McClintock BT, Bailey LL, Pollock KH, Simons TR. Experimental investigation of observation error in anuran call surveys. J Wildl Manag. 2010;74(8):1882–93. doi: 10.2193/2009-321
80. Clare JD, Zuckerberg B, Townsend PA. A generalized observation confirmation model to account for false positive error in species detection-nondetection data. bioRxiv; 2018 [accessed on 5 March 2025]. Available from: https://www.biorxiv.org/content/10.1101/422527v2.abstract.
81. Van Etten KW, Wilson KR, Crabtree RL. Habitat use of red foxes in Yellowstone National Park based on snow tracking and telemetry. J Mammal. 2007;88(6):1498–507. doi: 10.1644/07-MAMM-A-076.1
82. Koushanfar F, Shamsi D. The challenges of model objective selection and estimation for ad-hoc network data sets. Lect Notes-Monogr Ser. 2009;57:333–47. doi: 10.1214/09-LNMS5720.
83. Majda AJ, Yuan Y. Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems. Discret Contin Dyn Syst-Ser B. 2012;17(4):1333. doi: 10.3934/dcdsb.2012.17.1333
84. Yung G, Lin X. Validity of using ad hoc methods to analyze secondary traits in case-control association studies. Genet Epidemiol. 2016;40(8):732–43. doi: 10.1002/gepi.21994
85. Royall RM, Cumberland WG. Variance estimation in finite population sampling. J Am Stat Assoc. 1978;73(362):351–8. doi:

10.1080/01621459.1978.10481581
86. Valliant R, Dorfman AH, Royall RM. Finite population sampling and inference: a prediction approach. New York (NY): Wiley; 2000.
87. Raychaudhuri S. Introduction to Monte Carlo simulation. Proceedings of the 2008 Winter Simulation Conference; 2008 Dec 7–10; Miami, FL, USA. Piscataway (NJ): IEEE; 2018. p. 91–100.
88. Pick JL, Kasper C, Allegue H, Dingemanse NJ, Dochtermann NA, Laskowski KL, et al. Describing posterior distributions of variance components: problems and the use of null distributions to aid interpretation. Methods Ecol Evol. 2023;14(10):2557–74. doi: 10.1111/2041-210X.14200
89. Cox C. Delta method. In: Encyclopedia of biostatistics. Hoboken (NJ): John Wiley & Sons, Ltd.; 2005.
90. Platt JR. Strong Inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. Science. 1964;146(3642):347–53. doi: 10.1126/science.146.3642.347
91. Li N, Lee RD. Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method. Demography. 2005;42(3):575–94. doi: 10.1353/dem.2005.0021
92. Zhu JQ, Newall PW, Sundh J, Chater N, Sanborn AN. Clarifying the relationship between coherence and accuracy in probability judgments. Cognition. 2022;223:105022. doi: 10.1016/j.cognition.2022.105022
93. Templeton AR. Coherent and incoherent inference in phylogeography and human evolution. Proc Natl Acad Sci USA. 2010;107(14):6376–81. doi: 10.1073/pnas.0910647107
94. Acciai A, Guerrisi L, Perconti P, Plebe A, Suriano R, Velardi A. Narrative coherence in neural language models. Frontiers in Psychology. 2025 Apr 1;16:1572076 1572076. doi: 10.3389/fpsyg.2025.1572076
95. Huque MH, Bondell HD, Ryan L. On the impact of covariate measurement error on spatial regression modelling. Environmetrics. 2014;25(8):560–70. doi: 10.1002/env.2305
96. Lin GE, Justin XT, Zhang H, Hongyue WA, Hua HE, Gunzler D. Modern methods for longitudinal data analysis, capabilities, caveats and cautions. Shanghai Arch Psychiatry. 2016;28(5):293. doi: 10.11919/j.issn.1002-0829.216081
97. Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: A tragedy of errors. Nature. 2016;530(7588):27–9. doi: 10.1038/530027a
98. Baker M, Brandon K. Is there a reproducibility crisis? A nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help. Nature. 2016;533:452–4. doi: 10.1038/533452a
99. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2016;349(6251):aac4716. doi: 10.1126/science.aac4716
100. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med. 2016;8(341):341ps12. doi: 10.1126/scitranslmed.aaf5027
101. de Haas G. A comparative analysis of clustering quality based on internal validation indices using datasets with different characteristics. In: Advances in artificial intelligence and data engineering. Berlin/Heidelberg: Springer; 2021.
102. Smith DW, Boyd D. Guest column: wolf management plan should be informed by science. Bozeman Daily Chronicle. 2023 Jun 17.
103. Iijima H. A review of wildlife abundance estimation models: comparison of models for correct application. Mammal Study. 2020;45(3):177–88. doi: 10.3106/ms2019-0082
104. Royle JA, Dorazio RM. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Amsterdam: Elsevier; 2008.
105. Kéry M, Royle JA. Applied hierarchical modeling in ecology: analysis of distribution, abundance and species richness in R and BUGS: volume 2: dynamic and advanced models. Cambridge (MA): Academic Press; 2020.
106. Gervasi V, Aragno P, Salvatori V, Caniglia R, De Angelis D, Fabbri E, et al. Estimating distribution and abundance of wide-ranging species with integrated spatial models: opportunities revealed by the first wolf assessment in south-central Italy. Ecol Evol. 2024;14(5):e11285. doi: 10.1002/ece3.11285
107. Green AM, Chynoweth MW, Şekercioğlu ÇH. Spatially explicit capture-recapture through camera trapping: a review of benchmark analyses for wildlife density estimation. Front Ecol Evol. 2020;8:563477. doi: 10.3389/fevo.2020.563477
108. Stenglein JL, Waits LP, Ausband DE, Zager P, Mack CM. Efficient, noninvasive genetic sampling for monitoring reintroduced wolves. J Wildl Manag. 2010;74(5):1050–8. doi: 10.2193/2009-305
109. López-Bao JV, Godinho R, Pacheco C, Lema FJ, García E, Llaneza L, et al. Toward reliable population estimates of wolves by combining spatial capture-recapture models and non-invasive DNA monitoring. Sci Rep. 2018;8(1):2177. doi: 10.1038/s41598-018-20675-9
110. Saunders G, Harris S, Eason CT. Iophenoxic acid as a quantitative bait marker for foxes. Wildl Res. 1993;20(3):297–302. doi: 10.1071/WR9930297
111. Johnston JJ, Windberg LA, Furcolow CA, Engeman RM, Roetto M. Chlorinated benzenes as physiological markers for coyotes. J Wildl Manag. 1998;62:410–21. doi: 10.2307/3802307
112. Crabtree RL, Burton FG, Garland TR, Cataldo DA, Rickard WH. Slow-release radioisotope implants as individual markers for carnivores. J Wildl Manag. 1989;53:949–54. doi: 10.2307/3809594
113. Grimm-Seyfarth A, Harms W, Berger A. Detection dogs in nature conservation: a database on their world-wide deployment with a review on breeds used and their performance compared to other methods. Methods Ecol Evol. 2021;12(4):568–79. doi: 10.1111/2041-210X.13560

114. Roda F, Sentilles J, Molins C, Duchamp C, Hansen E, Jean N. Wolf scat detection dog improves wolf genetic monitoring in new French colonized areas. J Vertebr Biol. 2021;69(3):20102–1. doi: 10.25225/jvb.20102
115. Vynne C, Skalski JR, Machado RB, Groom MJ, Jácomo AT, Marinho-Filho JA, et al. Effectiveness of scat-detection dogs in determining species presence in a tropical savanna landscape. Conserv Biol. 2011;25(1):164–62. doi: 10.1111/j.1523-1739.2010.01581.x
116. Sands J, Creel S. Social dominance, aggression and faecal glucocorticoid levels in a wild population of wolves, Canis lupus. Anim Behav. 2004;67(3):387–96. doi: 10.1016/j.anbehav.2003.03.019
117. Vonholdt BM, Stahler DR, Smith DW, Earl DA, Pollinger JP, Wayne RK. The genealogy and genetic viability of reintroduced Yellowstone grey wolves. Mol Ecol. 2008;17(1):252–74. doi: 10.1111/j.1365-294X.2007.03468.x
118. Vonholdt BM, Stahler DR, Brzeski KE, Musiani M, Peterson R, Phillips M, et al. Demographic history shapes North American gray wolf genomic diversity and informs species' conservation. Mol Ecol. 2024;33(3):e17231. doi: 10.1111/mec.17231
119. Stansbury CR, Ausband DE, Zager P, Mack CM, Waits LP. Identifying gray wolf packs and dispersers using noninvasive genetic samples. J Wildl Manag. 2016;80(8):1408–19. doi: 10.1002/jwmg.21136
120. Bassing SB, Ausband DE, Mitchell MS, Lukacs P, Keever A, Hale G, et al. Stable pack abundance and distribution in a harvested wolf population. J Wildl Manag. 2019;83:577–590. doi: 10.1002/jwmg.21616
121. Cooch E, White G. Program MARK: a gentle introduction. 19th ed. Fort Collins (CO): Colorado State University; 2019 [accessed on 4 August 2024]. Available from: http://www.phidot.org/software/mark/docs/book/.
122. Powell LA. Approximating variance of demographic parameters using the Delta method: a reference for avian biologists. Condor. 2007;109:949–54. doi: 10.1093/condor/109.4.949
123. Williams BK, Nichols JD, Conroy MJ. Analysis and management of animal populations. San Diego (CA): Academic Press; 2002.

Supplementary materials

Evaluating the variance in estimates of wolf population size (abundance) in the state of Montana for the integrated patch occupancy model (iPOM)

1. Summary

The integrated patch occupancy model (iPOM) was created to estimate the abundance (population size) of wolves (N_{wolves}), where a patch is a wolf management area that is 'sampled' by hunters [1]. For each patch, abundance estimates were derived by combining the probability of patch occupancy, predicted territory size, predicted number of packs, predicted pack sizes, and a lone wolf rate. The total abundance (N_{wolves}) in a given area is the sum of all fractional abundance estimates from each patch (600 km^2 grid cells). Here, the variance, reported as a confidence interval (CI) in iPOM [1], is evaluated for estimates of the total N_{wolves} in Montana, with information presented in two reports [2,3]. Since iPOM was used to project population size from sampled areas to the entire state, prediction intervals should be used rather than confidence intervals. Because we could not obtain their data to reproduce their CI or produce a prediction interval, we evaluate the accuracy of their confidence intervals here. This evaluation is not an attempt to estimate the exact variance in wolf abundance but rather a common-sense check to determine whether the CI reported in the report was realistic. The main findings of this work were

- iPOM uses mechanistic and retrospective predictive models to estimate N_{wolves} in Montana. These models are inappropriate for projecting predicted point estimates of wolf population size across the state from sampled areas much smaller than the entire state.
- If iPOM is to be used to project estimates of N_{wolves} across the entire state and/or forward a year(s) to set harvest regulations, a more general approach that uses distributions of the territory size and pack size should be used to include the uncertainty in the predictions. This is akin to how typical population projection modeling is undertaken, such as for population viability analyses. The intervals from this approach are prediction intervals.
- iPOM omits numerous components of variation from the variance estimate used to report a confidence interval. Such exclusion biases result in a severe underreporting of uncertainty (variance).
- Based on a Delta method evaluation, the confidence intervals for the estimated N_{wolves} in Montana are unrealistically small and severely underreported.
- Based on the confidence intervals presented, the approximate coefficient of variation for N_{wolves} in Montana averaged at about 5.5%. The Delta method indicates that the coefficient of variation should be at least 47% (≥ 8.5x)
- For example, the estimate of N_{wolves} and its confidence interval for 2010 were 1145 (1024–1280). If the coefficient of variation were within the 47% range indicated by the Delta method, the confidence interval would be approximately 85–1946 wolves.

Given this high uncertainty in variance, iPOM cannot achieve detection if the population drops below 150. It is doubtful that iPOM can detect any change in abundance.

- In particular, the mechanistic model for territory size contains a multitude of untested assumptions and nonexistent data that make it unreliable and prone to producing unrealistically small estimates of variance.
- Due to mathematical errors, the application of the wrong variance measure, and the exclusion of the variance in numerous components (input variables), iPOM produces unreliable results incapable of management decision-making.

Here, we test and approximate the confidence intervals in iPOM's estimate of N_{wolves} by calculating the Delta method variance used to report confidence intervals (CIs). This evaluation is not an attempt to estimate the exact variance in N_{wolves} but rather a common-sense reality check and a determination of the approximate variance for N_{wolves}. However, even if the CIs are realistic, this is not the correct interval for N_{wolves}. A prediction interval should be used because N_{wolves} is projected across the entire state based on smaller sampled areas.

We started by calculating the variance (and CVs) in each of their four main input variables [1], AO, TS, PS, and LW, for the initial assessment and a comparison to iPOM's CV in the wolf abundance from the CI that they reported. We then use an average year's variance for iPOM's NP and LW (corrected) and the empirical values for PS from Sells et al. [2] to approximate the minimum variance in N_{wolves} (and resultant CI) using the Delta method [4]. The Delta method is used when an estimate (wolf population abundance in this case) is a function of multiple estimates; the sampling variance of the new parameter is also a function of the sampling variances in the component estimates. The statistical terms used in this report include the coefficient of variation (CV), standard error (se), and variance in the estimates of wolf population abundance, which are related as

$$se = \sqrt{variance},$$

$$CV = \frac{se\ (estimate)}{estimate},$$

where est = the estimate. So, for N_{wolves}, the CV is $CV(\hat{N}_{wolves}) = \frac{se(Nwolves)}{\hat{N}_{wolves}}$.

Because no estimates of variance were provided, we used CIs to approximate *se* as

$$se = \frac{(UCI - LCI)}{(2\ \times\ 1.96)}$$

where UCI = the upper CI and LCI = the lower CI. This equation is based on an assumption that the CI is symmetrical; i.e., $\text{UCI}(Nwolves) = \hat{N}_{wolves} + 1.96 \times se(\hat{N}_{wolves}), LCI(\hat{N}_{wolves}) = \hat{N}_{wolves} - 1.96 \times se(\hat{N}_{wolves})$. This is not exactly true for iPOM because estimates of wolf abundance are based on distributions of sampled packs (through occupancy and territory size) and pack size, which are not symmetrical. However, approximating *se* from the 95% CI will be close and sufficient for a reality check of the variance estimates presented in the final report. That is, the CV based on the Delta method will not be exactly the same as that from the simulation approach iPOM uses, but they should be relatively close. If anything, variance and CIs based on the Delta method should be biased low because they are based on symmetrical distributions and may miss some of the larger unsymmetrical numbers in distributions.

The variance in $N_{wolves,}$ which is directly related to the CIs for N_{wolves}, was estimated from estimates of patch occupancy, territory sizes, and pack sizes, each with its variance, and this variance was propagated in the multiplication and division required to estimate N_{wolves}. iPOM's N_{wolves} is predicted using a combination of retrospective and mechanistic models. The final equation used to predict N_{wolves} in the iPOM report [2] was

$$N_{wolves} = N_{packs}\ x\ pack_{size}\ x\ lone_{rate}$$

where N_{packs} is the estimated *NP*, $pack_{size}$ is the estimated *PS*, and $lone_{rate}$ accounts for non-pack lone and dispersed wolves or *LW*. Note that the estimates for *NP* and *PS* are based on additional sub-equations, yet we used this final equation as a common-sense reality check on the variance estimates of N_{wolves} in iPOM. The CI derived from the Delta method variance approximates but would be less than the prediction interval (PI) for iPOM. We began by estimating the variance in N, the estimated number of wolves, using the Delta method based on the final equation above. The Delta method equation for the variance in $\hat{N}_{wolves}$ with the iPOM formula above, assuming independence (no covariance between estimates of NP, PS, and LW), is

$$Var\ \hat{N} = var\ (\widehat{NP})x(\widehat{PS}\ x\ \widehat{LW})^2\ + var\ (\widehat{PS})x(\widehat{NP}\ x\ \widehat{LW})^2 + \ var(\widehat{LW})x(\widehat{NP}\ x\ \widehat{PS})^2$$

In Table 1.7, Sells et al. [2] provided annual estimates of NP and its 95% CI. Because estimates of variance were not provided, we calculated the approximate *se* for NP as described above. The resulting approximations of *se* were used to estimate the variance in the estimates (i.e., variance = se^2) used in the Delta equation. The estimate for LW was modeled for all years as having a mean of 1.125 and a standard deviation of 2.25]. This was incorrect; the lone wolf rate was estimated from Table 6.1 of Fuller et al. [5], and Sells et al. [2] added 1 to the proportions of non-resident wolves and incorrectly applied the variance in the proportion to the larger value. They should have used the proportions directly. The proportions from Table 6.1 below yield a mean lone wolf rate = 0.120, a variance = 0.039, and a CV = 10.2%, which were used in the Delta method calculations.

Table S1. The proportion of non-resident wolves (i.e., proportion of lone wolves in the population) from Table 6.1 in Fuller et al. [5].

% Non-resident	Prop non-res
14	0.14
10	0.1
20	0.2
7	0.07
9	0.09
9	0.09
10	0.1
12	0.12
13	0.13
16	0.16

Note, the standard deviation of the LW estimate in this case is *se*; we use the term se; both terms are used in the literature, but *se* is appropriate for this proportion (i.e., the standard deviation of an estimate is *se*).

iPOM's annual estimate of *PS* drew a random value from the group size distribution, based on the predictive PS model for each iPOM grid cell (patch) each year. The resulting mean and *se* for each grid cell for each year were apparently used to create gamma distributions that were spatially and temporally explicit estimates of pack size and its uncertainty. For a statewide estimate, and because iPOM used empirical samples for LW [5], we extracted one representative year by rounding 1/14 of the column values of the empirical pack sizes from their data in Figure 1.15 [2], as illustrated here as Figure S1:

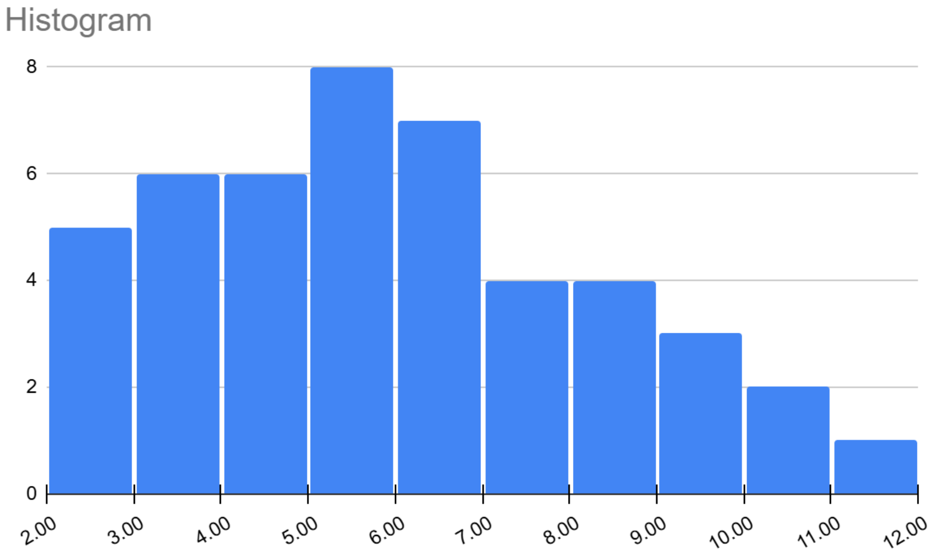

Figure S1. Empirical pack sizes modified from Figure 1.15 in Sells et al. [2]. The mean was 5.63 with a variance of 6.5.

We calculated the variance in the area occupied (AO) using the same method as that for NP, using Table 1 in Parks et al. [3] for the years 2007–2023. The mean AO was 67,812 km² with a variance of 7,206,212. We extracted the 33 empirical data values from Figure 1.8 [2] for TS. They were (in km²) 180, 210, 250, 250, 250, 270, 270, 270, 280, 350, 360, 390, 420, 460, 460, 480, 480, 510, 530, 540, 620, 620, 630, 650, 810, 830, 840, 860, 910, 1030, 1040, 1840, and 2230. The calculated mean of the TS was 609.7 km², and its variance was 190,870. We took the NP for 2010 at 161 with a variance of 79.7. We used an LW of 1.12 and a variance = 0.0132. The resultant CVs for the AO, TS, PS, and LW were 4%, 72%, 45.3%, and 10.2%, respectively, compared to iPOM's CV for wolf abundance at 5.5%.

Thus, the predicted $\hat{N}$ = $NP \times PS \times LW$ = 161 × 5.63 × 1.12 = 1015 wolves. We then took the means and variances for each of the three variables and calculated $Var\ \hat{N}$ using the Delta method formula above.

$$Var\ \hat{N} = 79.7 \times (5.63\ \times\ 1.12)^2 + 6.5 \times (161\ \times\ 1.12)^2 + 0.132\ \times\ (161\ \times\ 5.63)^2$$

$$Var\ \hat{N} = 225{,}364$$

$$\sim se\ \left(\hat{N}_{wolves}\right) = 475$$

$$\sim CV\left(\hat{N}_{wolves}\right) = 46.8\%$$

$$\sim 95\%\ \text{CI} = 85\text{–}1946\ \text{wolves}$$

2. Results

Overall, CV($\hat{N}_{wolves}$) from Table 1.7 in [6] averaged 5.5%, while the CV based on the Delta method that we calculated was 46.8%. Remember the CV from Table 1.7 was estimated as $CV(\hat{N}_{wolves}) = \frac{SE(\hat{N}_{wolves})}{\hat{N}_{wolves}}$, with $SE(\hat{N}_{wolves}) = \frac{(UCI(\hat{N}_{wolves}) - LCI(\hat{N}_{wolves}))}{(2\times 1.96)}$, where $\hat{N}_{wolves}$, UCI, and LCI were labeled as Wolves in Table 1.7. $CV(\hat{N}_{wolves})$ would be, on average, at least 8.4× higher based on the Delta method compared to that for iPOM (Table 1). Based on our initial Delta method assessment, we concluded that their CIs were unrealistic because numerous components of variation were omitted, in addition to the problems above. Proper analysis and inclusion of the empirical data sources for LW and PS increased the CV. Yet we left out components of variance that, if properly included, would result in an even higher variance and wider CI.

3. Discussion

Although iPOM's CI is not the correct interval for the estimate of $\hat{N}wolves$, the Delta method indicates that their CI is wrong—the CV for the prediction of $\hat{N}$ is a fraction of what it should be. Even compared to other estimates for similar studies, the CIs on wolf abundance for the large scale of Montana (area = 380,832 km²) are unrealistically low. For example, a similar approach using home range size, mean pack size, and occupancy estimates was used to scale up and estimate wolf abundance for the known wolf range in Wisconsin (area = 91,000 km²; Stauffer et al. [7]). The CV of their estimate was approximately 12.1% (based on 95% credible intervals), and this was for a much smaller area than the state of Montana (the CV typically increases with population size; [8]). The same approach was used at the much smaller scale of 30,000 km² in southwestern Alberta, Canada [9]. Even for this smaller and intensively sampled population of wolves (with a population size of approximately 160), the CVs ranged from 9% to 10% for the 3 years of that study. These two studies, which employed the same methods as that in the final report [2], also indicate that iPOM's CIs of wolf population abundance, which averaged 5.5% between 2007 and 2019 for the entire state of Montana, are too low.

In addition, the CIs used for pack size (PS) are unrealistically small for projection models. For example, from the graphs in Figure 1.19 in 2010, the 95% CI for the Northern Rockies is approximately 5.4–5.8. This means all wolf packs in the Northern Rockies have 5–6 members. Based on the observed annual pack sizes in Figure 1.15, this is not biologically realistic. For example, for 2010, observed pack sizes varied from 2 to 11, with most being between 4 and 7. Although the Northern Rockies is a subset of all packs observed, there is no reason that the variability in this region should be so restricted in size. iPOM uses values much lower than what is observed. iPOM's mixed model, developed from data for a limited area, is being used to predict pack size over a large area, for which there has been no validation. Another approach would be to generate a distribution of pack sizes based on the covariate averages over the entire patch (or cell) and across years. This would represent the uncertainty of predicting pack sizes better.

More importantly, the values and variance used for territory size (TS), from which NP is derived, are based on virtually no data and many untested assumptions. This type of model is not appropriate for projecting abundance estimates with reality. Mechanistic models may be suitable for relative comparisons but not for estimating or predicting population sizes (abundance). For territory sizes, using mixed models and a similar approach to that recommended for pack sizes would produce more realistic variance for projecting population sizes into the future to set wolf harvest numbers.

The Delta method is based on a first-order Taylor series expansion of the transformed function (e.g., multiplied and divided estimates), which may perform well if the function is highly nonlinear (e.g., an exponential function used) over the range of values being examined [10]. However, in this situation, where the function multiples three variables, this is unlikely to be an issue. Other issues with the Delta method used here were that the covariance was not included. However, the addition (or subtraction if the covariance between estimates is negative) is usually a small change. Overall, the Delta method is likely to underestimate the variance slightly in this situation because large values are not accounted for in symmetrical CIs.

In this discussion, we have left out other factors that would inflate the variance (and CI). They are (1) omission of the variance in the covariate models and (2) the inclusion of ad hoc variables, ad hoc methods, and correction factors. For the latter, we found numerous examples of these in iPOM [1], including

1. Harvest intensity covariates are used as proxies for annual mortality in their TS and PS models. We cannot know whether these corrections are calibrated or correlated with their pack size data. Similarly, their PS model uses uncalibrated spatial density indices as a proxy for prey abundance.
2. Creel [11] points out another case where an ad hoc spatial adjustment factor is used to adjust the prey densities in the TS model.
3. iPOM's density identifier formula (see Section 6).
4. iPOM's use of a territory overlap index. This factor, like their density identifier formula, biologically varies annually, yet they include no annual empirical inputs from wolves.
5. iPOM incorrectly assumes that LW is constant when it is a biologically dynamic variable, as seen in numerous studies [5], including individual extra-territorial movements of long durations and long distances [12]. It varies with mortality, especially when a pack is dissolved due to human killing.

References

1. Sells SN, Podruzny KM, Nowak JJ, Smucker TD, Parks TW, Boyd DK, et al. Integrating basic and applied research to estimate carnivore abundance. Ecol Appl. 2022;32(8):e2714. doi: 10.1002/eap.2714
2. Sells SN, Keever AC, Mitchell MS, Gude J, Podruzny K, Inman R. Improving estimation of wolf recruitment and abundance, and development of an adaptive harvest management program for wolves in Montana. Final report for federal aid in wildlife restoration grant W-161-R-1. Helena (MT): Montana Fish, Wildlife and Parks; 2020.
3. Parks MK, Podruzny K, Cole W, Parks T, Lance N, Zielke S, et al. Montana gray wolf conservation and management 2023 annual report. Helena, Montana: Montana Fish, Wildlife & Parks; 2024.
4. Cox C. Delta method. In: Encyclopedia of biostatistics. Hoboken (NJ): John Wiley & Sons, Ltd.; 2005.
5. Fuller, TK. Mech, LD, and Cochrane, JF. Wolf population dynamics. In: Mech LD, Boitani L, editors. Wolves: behavior, ecology, and conservation. Chicago (IL) & London: University of Chicago Press; 2003. p. 322.
6. Miller DA, Nichols JD, McClintock BT, Grant EH, Bailey LL, Weir LA. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology. 201192(7):1422–8. doi: 10.1890/10-1396.1
7. Stauffer GE, Roberts NM, Macfarland DM, Van Deelen TR. Scaling occupancy estimates up to abundance for wolves. J Wildl Manag. 2021;85(7):1410–22. doi: 10.1002/jwmg.22105
8. Williams BK, Nichols JD, Conroy MJ. Analysis and management of animal populations. San Diego (CA): Academic Press; 2002.
9. Bassing SB, Ausband DE, Mitchell MS, Lukacs P, Keever A, Hale G, et al. Stable pack abundance and distribution in a harvested wolf population. J Wildl Manag. 2019;83:577–590. doi: 10.1002/jwmg.21616
10. Cooch E, White G. Program MARK: a gentle introduction. 19th ed. Fort Collins (CO): Colorado State University; 2019 [accessed on 10 October 2024]. Available from: http://www.phidot.org/software/mark/docs/book/.
11. Creel S. Methods to estimate population sizes of wolves in Idaho and Montana. Comments in federal register 86 FR 51857, "endangered and threatened wildlife and plants; 90-Day finding for two petitions to list the gray wolf in the western united states". 2023, available from: https://www.regulations.gov/comment/FWS-HQ-ES-2021-0106-49075.
12. Royall RM, Cumberland WG. Variance estimation in finite population sampling. J Am Stat Assoc. 1978;73(362):351–8. doi: 10.1080/01621459.1978.10481581