Inter-observer variance and agreement of wildlife information extracted from camera trap images

Theresa Zett \(^1\) · Ken J Stratford \(^2\) · Florian J Weise \(^2\)

Received: 11 June 2021 / Revised: 8 August 2022 / Accepted: 29 August 2022 / Published online: 10 September 2022 © The Author(s) 2022

Abstract

Camera traps are a popular tool in terrestrial wildlife research due to their low costs, easy operability, and usefulness for studying a wide array of species and research questions. The vast numbers of images they generate often require multiple human data extractors, yet accuracy and inter-observer variance are rarely considered. We compared results from 10 observers who processed the same set of multi-species camera trap images (\(n=11,560\)) from seven sites. We quantified inter-observer agreement and variance for (1) the number of mammals identified, (2) the number of images saved, (3) species identification accuracy and the types of mistakes made, and (4) counts of herbivore groups and individuals. We analysed the influence of observer experience, species distinctiveness and camera location. Observers varied significantly regarding image processing rates, the number of mammals found and images saved, and species misidentifications. Only one observer detected all 22 mammals (range: 18–22, \(n=10\)). Experienced observers processed images up to 4.5 times faster and made less mistakes regarding species detection and identification. Missed species were mostly small mammals (56.5%) while misidentifications were most common among species with low phenotypic distinctiveness. Herbivore counts had high to very high variances with mainly moderate agreement across observers. Observers differed in how they processed images and what they recorded. Our results raise important questions about the reliability of data extracted by multiple observers. Inter-observer bias, observer-related variables, species distinctiveness and camera location are important considerations if camera trapping results are to be used for population estimates or biodiversity assessments.

Keywords Camera trap · Intra-class correlation coefficient · Image classification · Inter-observer agreement · Accuracy · Mammal

Communicated by Lilly O Rodriguez.

\(^{1}\) School of Animal, Rural and Environmental Sciences, Nottingham Trent University, NG25 0QF Brackenhurst, Southwell, Nottinghamshire, UK

\(^{2}\) Ongava Research Centre, Private Bag 12041, Windhoek, Namibia
Introduction

When remote wildlife cameras became commercially available in the early 1990s (Kucera and Barrett 1993) their favourable characteristics as a stand-alone research tool led to a revolution in the monitoring of terrestrial wildlife. Animals could now easily be recorded non-invasively anywhere in the world (Kucera and Barrett 2011; Meek et al. 2015). Cameras have since been widely deployed to answer management and scientific questions alike, applicable to an immense spectrum of inquiry that ranges from simple species detection to animal-environment interactions and ultimately global biodiversity changes (Swann and Perkins 2014; Trolliet et al. 2014; Steenweg et al. 2017).

Camera traps combine several useful characteristics and, hence, have important advantages over most other terrestrial monitoring tools. They are affordable, enabling more extensive and intensive sampling strategies. Their operation does not require specialised skills or training and they cause little disturbance to wildlife. Cameras can be left in remote areas for prolonged time periods operating almost independently. Camera trapping, thus, is a cost-effective way of improving both detection probability of cryptic species and the number of observations obtained, creating permanent records that yield irrefutable evidence of a wide array of species and their behaviours (De Bondi et al. 2010; Welbourne et al. 2015; Caravaggi et al. 2017; Wearn and Glover-Kapfer 2019; Thomas et al. 2020).

It is no surprise then that camera trap studies have increased substantially in many parts of the world (Meek et al. 2015; Agha et al. 2018), often focussing on more than a single species (Burton et al. 2015). Camera traps are now probably used more than any other research tool in terrestrial wildlife studies. Their popularity and functional versatility, however, create a luxury problem. Large quantities of images are generated, and the relevant data contained in these images still needs to be extracted in a reliable fashion. Since cameras can record quick image successions from a single trigger event, devices deployed in areas of high wildlife activity may easily generate up to 40 000 images per day. It is no longer unusual that research programmes operate camera grids with, sometimes, hundreds of devices across entire landscapes (Swanson et al. 2015; McShea et al. 2016). With vast amounts of images being recorded, the time required to review the available footage also increases, thus demanding more resources to extract data. It follows that most research programmes need to rely on multiple observers to process camera trap information efficiently.

Ongava Research Centre (ORC) in northern Namibia faces a similar challenge. Here, dozens of cameras are deployed regularly on a private reserve to study wildlife ecology and distributions (Edwards et al. 2018; Stratford et al. 2019), to assess camera utility and performance (Stratford and Naholo 2017), and to assist with wildlife management questions on the reserve. For example, over the last decade approximately 50 cameras have been set at permanent waterholes every year as part of the reserve’s wildlife population monitoring programme, routinely generating upwards of 800 000 images within only a few weeks. In order to facilitate the process of tagging, storing, and analysing these magnitudes of image data, ORC developed a software system that provides an optimised user interface. Termed GDMS (Geo-Data Management System), the primary interface for processing images was designed to maximise user efficiency by providing methods for rapid image handling, in particular moving between groups of images. It has been particularly helpful at processing so-called “empty triggers”, image sequences triggered by moving vegetation that contain no relevant wildlife, but can contribute considerably to the number of images needing to be
managed. Users record keywords for images, typically species, with simple pre-configured on-screen buttons (Fig. 1). In case of uncertainty, the GDMS enables users to ‘flag’ images for independent verification by other observers. All images and their meta-data (camera types, locations, configurations, etc.) are then stored in a database and can be accessed using a range of filters and sort criteria. This way, ORC has processed >21.7 million images since 2009, resulting in a database now housing more than 7.6 million keyworded wildlife images that support a variety of analyses. As is customary elsewhere, ORC has depended on multiple data extractors to process and classify image contents.

In principle, the mass-extraction of wildlife information from camera trap images is no different than recording wildlife with any other method. The measurement taken needs to be accurate and sources of recording bias must be minimised. The obvious advantage of camera trapping is that a permanent record exists, one that can be assessed without time constraints in the comfort of an office and be reviewed independently by multiple observers. Despite an objective, static wildlife record, the question remains whether different observers indeed extract the same information from images. Previous research suggests not, showing that inter-observer variance can greatly affect the reliability of results obtained when multiple human observers are expected to perform the same task. Studies found considerable inconsistencies in the information obtained, be it during the identification of mammalian species or unique individuals from camera trap footage (Gooliaff and Hodges 2018; Johansson et al. 2020), or the classification of population characteristics (Newbolt and Ditchkoff 2019). It is also clear that different factors such as camera type (Randler and Kalb 2018) and observer experience can influence the results (Burns et al. 2018; Katrak-Adefowora
et al. 2020) and that accuracy may differ with species size and phenotypic distinctiveness (Potter et al. 2019). A variety of mistakes can occur, for instance observers missing species entirely or misidentifying them. Even among trained observers, variance can be high (e.g. Meek et al. 2013), especially if similar-looking wildlife occur in the same dataset (Gooliaff and Hodges 2018).

Different statistical tools are available to measure agreement between observers and account for variance in observer accuracy (Watson and Petrie 2010; Koo and Li 2016; Ranganathan et al. 2017). A random sample of 70 mammal camera trap studies published over the last decade revealed that 43% of articles reported multiple observers extracting information from images for data analysis, yet without measuring reliability between them, while most studies did not explicitly acknowledge whether multiple observers were involved in data extraction, or how this might have influenced the results. Only four studies explained how the risk of misidentifications was mitigated (Supplementary Information 1). The potential value of camera trapping information for biodiversity conservation is beyond doubt (Steenweg et al. 2017; Wearn and Glover-Kapfer 2019). Quantifying inter-observer differences is paramount, however, if the resulting data are utilised for population estimates (Després-Einspenner et al. 2017), species mapping or biodiversity assessments that may lead to conservation policy decisions and prioritisation of limited conservation funds, for example.

In order to understand inter-observer variance in greater detail, we conducted an experiment, asking observers with varying degrees of camera trapping and wildlife identification experience to process the same set of multi-species camera trap images. We aimed to determine inter-observer variance and agreement in terms of: (1) the number of relevant images stored, (2) the number of mammal species identified, and (3) the types and frequencies of mistakes made. We further compared differences in how observers counted herbivore groups and individuals while exploring the influence of observer experience, species distinctiveness and camera location on the results. Finally, we assessed causes of inter-observer variance and how removal of outliers may influence agreement and the accuracy of the results obtained.

Materials and methods

Experimental set-up and image processing

To assess inter-observer agreement and variance of mammal species detections extracted from camera trap images, we conducted a multi-rater experiment, asking 10 in-house observers (based at ORC) to process the same set of images following a set of guidelines and tasks (Supplementary Information 2). Observers included three senior researchers and two research technicians with camera trapping experience, as well as five student interns, of which only one had previously worked with camera traps on a regular basis. Although not permanently employed, we included student interns in this experiment based on the premise that data extraction tasks are often relegated to assistant staff when large amounts of information need to be processed, and these tasks are increasingly outsourced to layman audiences (e.g. McShea et al. 2016; Swanson et al. 2016; Willi et al. 2019; Katrak-Adefowora et al. 2020). Based on their experience (no. of years) with both camera trap studies and
identification of Namibian wild mammals, we classified observers into three experience categories. Because years of experience with camera trapping and species identification were strongly correlated (Supplementary Fig. 2; \( r_s = 0.969, n=10, p<0.001 \)), we calculated the average between the two for each participant for final experience classification, being: novices (0 ≤ 1 years), semi-experienced (1–5 years), and experienced observers (>5 years).

We utilised a sub-set of camera trap data recorded as part of Ongava Game Reserve’s 2019 annual wildlife monitoring programme. From dry season waterhole monitoring, which comprised a total of 863 591 camera trap images recorded with 51 cameras at 12 permanent waterholes over 2.5 weeks in September 2019, we randomly selected a sample of seven camera locations containing a total of 11 560 images for our experiment. Our sample was recorded with seven cameras (models: Bushnell HD x 2, Bushnell Core x 2, Reconyx HC500 \( \times 2 \), and one Reconyx HF2X Hyperfire2) positioned at six of the waterholes. Regardless of the specific model, Bushnell cameras were set to record three images per burst with a minimum delay interval of 15 s between bursts, whereas Reconyx cameras recorded 10 images per burst with a minimum delay interval of 30 s between bursts. All cameras were set to ‘high’ detection sensitivity and highest possible image quality, which differed between models. There were no differences between day- and night-time recording settings. All cameras had a different field of view and, hence, we treated the two units positioned at the same waterhole independently, hereafter reported as stations 6a and 6b. Sample data spanned a total of 271.5 deployment hours, including both day and night images. The image data contained 22 different mammal species of varying body sizes and phenotypic characteristics (Supplementary Information 3). To ensure correct comparisons with observer responses, we carefully scanned the entire image set four times to determine all species present and their respective identifications.

Following the set of instructions, each observer processed camera trap images independently, extracting mammal information using the GDMS software. The screen interface (Fig. 1) automatically loads image frames from a chosen image folder and observers were asked to assess each frame independently, not inferring species identification from images they had seen before or afterwards. Prior to data processing, all participants were trained to use the GDMS screen interface and its functionalities. Observers saved image frames that contained mammal species, assigning the relevant species keywords to each frame, including multiple species occurring in the same frame. Observers deleted those frames that did not contain mammal species. We instructed observers to assign species names only if identification was certain. Focal species included all wild mammals occurring on Ongava Game Reserve, regardless of their body size and phenotypic characteristics, or their size and location in the image frame. If detected species were not listed for keywording in the GDMS’s screen interface, observers recorded their species names separately and keyworded and saved frames as “mammal”. If animals were only partially visible in a given frame, observers assigned the keyword “mammal” or the respective species name if unambiguous identification was possible from the visible body parts. Observers “queried” image frames in which they could detect mammalian wildlife but were uncertain about accurate species identification. The ‘query’ function enables images to be saved and re-evaluated by other observers at a later stage. Observers also recorded the time spent on processing each of the seven image folders. There was no time restriction on data extraction and observers independently decided when and for how long they processed imagery. Observers were not allowed to consult other researchers but could utilise any other resources to aid their iden-
Biodiversity and Conservation (2022) 31:3019–3037

The choice of method to assess inter-observer reliability depends on the number of observers (raters), the type of response variable assessed and its statistical distribution (Mitchell 1979; Watson and Petrie 2010; Koo and Li 2016; Ranganathan et al. 2017). We used Anderson-Darling tests (Stephens 1974) to assess whether response data were normally distributed and selected parametric and non-parametric statistics accordingly.

We compared the number of mammal images saved by observers using a Friedman test (Friedman 1937, 1939), the non-parametric alternative to the one-way ANOVA with repeated measures, with observers as treatments and camera stations as blocks. The results are reported adjusted for ties. Because some mammals occurred at several of the seven locations, we determined the ratios of correct, incorrect, and missed species for each camera station by calculating the proportion of identified species per observer and camera station to the number of possible correct identifications for that station. We then compared these ratios using one-way ANOVAs. We compared the total number of species correctly detected between observers with a Chi-Square Goodness-of-Fit test.

Despite limited sample sizes, we considered the influence of observer experience and camera station on the results. Data between camera stations were compared with one-way ANOVAs, each for the ratio of correct vs. incorrect identifications and also missed species, i.e. the mean of all observer ratios for each station, as well as for the mean image processing rates (images/minute) between stations.

We determined observer agreement for herbivore herd and individual counts, as well as mean herd size estimates, using Intra-Class Correlation Coefficients (hereafter ICC, Fisher 1958; Bartko 1966). ICCs were computed between observers as single scoring, multi-rater absolute agreement measures using a two-way random effects model, with camera stations as repeats and observers as factors (Koo and Li 2016). Camera stations where a particular focal herbivore was not present were removed from the analysis. For each herbivore estimate, we screened ICC results for outliers using Bland-Altman plots (Bland and Altman 1986). Our interpretations of ICC scores followed the scale proposed by (Koo and Li 2016), with: values < 0.5 = poor agreement, 0.5–0.75 = moderate agreement, 0.75–0.9 = good agreement, and values > 0.9 = excellent agreement. Due to limited sample sizes, we deemed a conservative scale more appropriate than the more lenient ones proposed by (Cicchetti and Sparrow 1981) or (Shrout and Fleiss 1979).
We used “queried” images, those for which observers could not ascertain species identification with absolute certainty, as a proxy for observer confidence. The total number of images queried was compared between all observers with a Chi-Square Goodness-of-Fit test with a single variable (observer).

We quantified each observer’s mistakes made during species detection and identification, categorising mistakes as either ‘species missed’: a species occurred but was not recorded, or ‘misidentification’: a mammal was recorded but identified wrongly, also including confusions between similar-looking species. Since their accurate identification is nearly impossible from camera trap images, we made no distinction between the Cape rock hyrax (*Procavia capensis*) and the Kaokoveld rock hyrax (*Procavia capensis welwitschii*), nor between the South African ground squirrel (*Xerus inauris*) and the Damara ground squirrel (*Xerus princeps*).

To assess the influence of phenotypic similarity on misidentifications, we computed a species distinctiveness index for all terrestrial mammals known to occur on Ongava Game Reserve. We based species distinctiveness on the combination of three criteria, being: (1) a distinctiveness score that compared 10 morphological features between species; (2) a uniqueness score that compared unique morphological traits; and (3) a taxonomic score that accounted for the number of species in a given taxonomic group (adjusted from Potter et al. 2019). Their detailed calculation is given in Supplementary Information 4. For each species, we aggregated the results obtained for each criterion into a cumulative index ranging from 0 to 29 and calculated distinctiveness as a percentage ratio to indicate high (1.0) or low (0.0) distinctiveness from other species (Supplementary Information 5).

Prior to analyses, all data were assembled and summarised with Microsoft Excel. Analyses were computed with RStudio v1.4.1106 (RStudio Team 2020), Minitab v19.2020.1 (2010), and JMP Pro v15.1.0 (1989). We calculated ICC scores with package *irr* v. 0.84.1 (Gamer et al. 2019) and plotted graphs with *ggplot2* (Wickham 2016) and *ggpubr* (Kassambara 2020).

### Results

#### Data Processing

Ten observers with varying degrees of experience independently processed the total of 11 560 images recorded at seven camera stations. Participants spent a mean of 10.2 h±1.5 h S.E. processing imagery, with a high degree of variance between individual observers (range: 4.23 – 18.82 h, $S^2=19.02$). The most experienced observer processed imagery nearly 4.5 times faster than the least experienced one. Faster processing resulted in fewer images saved (Fig. 2a) and processing rates significantly increased with observer experience (Fig. 2b). Image processing rates differed significantly among observers ($f_{9,60} = 2.75$, $p=0.009$) and mean image processing rates varied significantly between different camera stations ($f_{6,63} = 3.11$, $n=7$, $p=0.010$; Table 1).

The total number of images that observers saved as containing mammalian wildlife differed significantly (Table 2). Additionally, observers appeared to assign species keywords with different degrees of confidence. The number of images queried (i.e. species identification could not be assigned with absolute certainty) ranged from 0 to 1 863 (median=355,
revealing a high degree of variance ($S^2 = 445,663.57$). Querying differed significantly between observers ($\chi^2 = 7,243.94; d.f. = 9; p < 0.001$) and significantly decreased with increasing observer experience ($r_s = -0.791, \ n = 10, \ p = 0.006$). While the four most experienced observers did not query species identification in any of the images, novices and semi-experienced observers queried species identification regularly (median = 797 images, range: 61–1,863, $n=6$).

Table 1 Correct mammal detections, frequency of identification mistakes, and image processing rates across seven camera stations. Results are pooled for all 10 observers. Stations 6a and 6b reflect two cameras deployed at different positions at the same waterhole

| Camera station | Correct detections* | Misidentifications | Missed species | Mean rate (±S.E.) | Median (range) |
|----------------|---------------------|--------------------|----------------|------------------|---------------|
| Station 1      | 70 (100%)           | 0                  | 0              | 17.1 (±3.43)     | 15.3 (4.9–34.4) |
| Station 2      | 108 (98.2%)         | 5                  | 2              | 28.0 (±4.87)     | 25.5 (10.2–55.8) |
| Station 3      | 100 (100%)          | 2                  | 0              | 26.1 (±4.39)     | 22.6 (13.3–55.0) |
| Station 4      | 94 (85.5%)          | 4                  | 16             | 17.4 (±2.74)     | 35.7 (7.9–35.7)  |
| Station 5      | 76 (95.0%)          | 0                  | 4              | 37.0 (±5.69)     | 30.6 (14.3–63.1) |
| Station 6a     | 49 (98.0%)          | 2                  | 1              | 20.0 (±2.79)     | 17.6 (9.4–35.7)  |
| Station 6b     | 50 (100%)           | 1                  | 0              | 26.3 (±3.19)     | 27.9 (10.1–40.0) |
| Total          | 547 (96.0%)         | 14                 | 23             | 24.6 (±1.65)     | 21.1 (4.9–63.1)  |

* Frequency represents the cumulative number of correct detections by all 10 observers and percentage value shows the proportion of these in relation to all 750 possible correct detections (no. of mammals occurring x 10 observers)
Only one observer (a novice) detected all 22 mammalian wildlife species present in the entire dataset. The number of species found by observers varied between 18 and 22 (Table 2), with a mean overall detection rate of 92.7% ± 1.5% S.E. (range: 81.8 – 100%, S^2=25.1). Of the 57 possible correct species identifications (incl. repeat occurrences of the same species at different camera stations), observers detected a median of 55 species with a range of 51–56 (S^2=2.16), or 96.0% of all possible correct detections (n=570; Table 1). Camera station had a significant effect on the accuracy of correct species detections as well as missed species (both f_{6,63}=1.01, p=0.443). Camera station did not significantly influence species misidentifications (f_{6,63}=1.01, p=0.443).

None of the observers correctly identified all mammalian wildlife across the seven sub-datasets. Observers missed between 0 and 4 species (median=2.0, S^2=1.04, n=10) on 23 instances across the seven stations (Table 1). Missed species mostly were small mammals (56.5%) with body masses <5 kg. Five observers also misidentified mammals on a total of 14 instances (median=0.5, range: 0–4, S^2=2.21, n=10), an error occurring mainly among least and semi-experienced observers (n=13, 92.9%). In some instances (see Table 1 camera station 3), observers correctly detected and identified mammals in an image sequence while also misidentifying the same species during a separate detection.

Using species distinctiveness scores as a guideline, confusions between least phenotypically distinct species, notably between the two rhino species, the two zebra species as well as two species of dwarf antelopes with similar phenotypic appearance, represented the majority of misidentification mistakes (n=12, 52.2%) (Supplementary Information 6). The total number of mistakes occurring during species detection and identification declined with increasing observer experience (Fig. 3; r_s=-0.874, n=10, p=0.001). Experienced observers that processed data faster mainly overlooked small species, whereas less experienced observers frequently missed and misidentified mammalian wildlife (Fig. 3). The proportion of misidentifications in relation to missed species mistakes increased with declining experience (Fig. 3).

### Counting herbivore individuals and herds

The individual, herd and resulting mean herd size estimates of 10 observers were characterised by high to very high sample variances (Table 3). Depending on the species, individual
counts differed by a magnitude of 140 – 210%, herd counts by 180 – 320%, and mean herd size estimates by 140 – 170% (Table 3). Novices recorded the most variable estimates for 14 of the 15 assessments (93.3%), followed by semi-experienced observers, whereas experienced observers generally produced the most accurate estimates (Fig. 4).

Table 3 Variability of common herbivore individual counts, herd counts and mean herd size estimates by 10 observers using camera trap images from seven sampling stations. Reported variances are sample variances. Difference factor reports the order of magnitude between highest and lowest estimate.

|                     | Plains zebra | Mountain zebra | Impala | Kudu     | Oryx     |
|---------------------|--------------|----------------|--------|----------|----------|
| Number of individuals |              |                |        |          |          |
| Median (Range)       | 150.5 (114–234) | 126.0 (108–169) | 533.0 (453–788) | 187.5 (162–252) | 64.5 (58–82) |
| Difference factor    | 2.1          | 1.6            | 1.7    | 1.6      | 1.4      |
| Variance ($S^2$)     | 1 015.3      | 344.2          | 9 988.0| 581.6    | 51.4     |
| Number of herds      |              |                |        |          |          |
| Median (Range)       | 33.0 (24–66) | 32.0 (25–49)   | 77.5   | 44.5     | 22.0     |
| Difference factor    | 2.8          | 2.0            | 3.2    | 2.1      | 1.8      |
| Variance ($S^2$)     | 132.4        | 48.5           | 1 204.5| 117.9    | 28.4     |
| Mean herd size       |              |                |        |          |          |
| Median (Range)       | 4.4 (3.9–5.6)| 3.5 (2.6–4.5)  | 7.0    | 3.7      | 3.1      |
| Difference factor    | 1.4          | 1.7            | 1.5    | 1.4      | 1.5      |
| Variance ($S^2$)     | 0.24         | 0.25           | 0.78   | 0.1      | 0.1      |

Fig. 3 Proportional contribution of missed species and misidentification mistakes to overall error structure classified by observer experience. Bold values atop bars show the total number of mistakes for each experience category.
Fig. 4 Boxplots showing the variance in herbivore estimates recorded from camera trap images classed according to observer experience for (a) individuals counted, (b) herds counted, and (c) estimated mean herd sizes. Black dots represent outliers.
Considering the full dataset for analysis, inter-observer agreement for herbivore estimates was mostly moderate \( (n=6) \), followed by good \( (n=5) \) and excellent estimate agreement \( (n=4) \) (Fig. 5). In terms of estimate precision, the highest exact agreement for any estimate was Oryx herd count, recorded as the same value by four observers. Herbivore estimates and their corresponding ICC agreement scores significantly differed for all individual counts, herd counts and mean herd size comparisons (all F-Tests \( p<0.001 \); see Supplementary Information 7).

Screening of count data revealed that high estimate variances, and hence observer disagreement, were mainly attributable to statistical outliers representing the estimates recorded by a novice observer, contributing 13 outliers to the 15 assessments. We, therefore, also compared inter-observer agreement using both the entire dataset (10 observers) and a reduced dataset (9 observers). Removal of this observer from the analysis increased ICC agreement scores for 14 of the 15 assessments and resulted in higher ICC accuracy. ICC agreement quality changed for eight estimates (53.3%) following outlier removal. Inter-observer agreement improved by one category in six instances (moderate-good: 3; good-
excellent: 3) and by two categories for another estimate (moderate-excellent), while the quality rating of one estimate decreased from good to moderate (Fig. 5).

Discussion

Our small-scale experiment raises important questions about the reliability of data extracted by multiple human observers from the same set of camera trap imagery. Observers considerably differed in how they processed imagery, how many data they saved, how confidently they assigned species identifications, how many species they detected overall, and how they interpreted processing rules. Observers also differed in their ability to identify mammalian wildlife accurately, in the number and types of mistakes that occurred and, ultimately, in their estimates of common herbivore individuals and herds seen. Most inter-observer comparisons yielded significant differences and only moderate to good agreement. With few exceptions, the extracted data were characterised by high degrees of variance, with the magnitude of estimate differences ranging in the order of several 100%. These differences are particularly worrying when considering that observers extracted data from near optimal recording conditions; 20 of the 22 focal species had medium to large body sizes (Supplementary Information 4), and the animals were positioned in open areas close to cameras, remaining stationary at waterholes for prolonged periods. Also, we imposed no time limits on data processing and observers had access to a range of resources aiding species identification. Our findings, thus, stress the importance of comparing agreement in multi-observer camera trap studies and, where necessary, the need to truncate unreliable responses.

Observer experience appeared to be one key variable influencing the results. Experienced observers assigned species labels with a very high degree of confidence and accuracy and their herbivore estimates had the highest accuracy overall. They processed images faster and mainly overlooked small-bodied mammals, such as the Rock hyrax, if these were located away from the images’ centre and focal point of animal activity. Less experienced observers, on the other hand, processed data more slowly, yet they assigned species identifications with less confidence, regularly missing and confusing similar-looking mammals, resulting in lower accuracy in species detection and correct identification. Their estimates of herbivore numbers and herds also revealed greater variance. Only using experienced observers for image classification may, therefore, improve accuracy and confidence in the results. However, many research projects rely on (untrained) volunteers for these time-intensive tasks in particular and the number of images obtained during large-scale multi-camera surveys makes it difficult to refer these tasks solely to experienced data processors. It is noteworthy, however, that only one observer, a novice data extractor, detected all 22 mammals present in the entire dataset. During post-experiment feedback, novice and semi-experienced observers clearly pointed out that fatigue was a challenge, especially after multiple hours of continuous data extraction (cf. Oliveira-Santos et al. 2010). This may contribute to an explanation of higher results variance and the number of mistakes that occurred, especially misidentifications. While detecting and counting animals from static images may seem easy, experiments show that human concentration decreases significantly within 30 min of performing intensive cognitive tasks, leading to mental fatigue and increased error rates (Slimani et al. 2018). In our study, observers dealt differently with fatigue. While some stopped image processing and resumed tasks at a later stage (sometimes days later), others continued
to finish a particular image sequence. Given the deleterious effects of mental fatigue on human attention (Boksem et al. 2005; Faber et al. 2012), maximum time thresholds for data extraction may be needed to ensure greater reliability. Additional studies are needed to examine the impact of fatigue on camera trapping results.

Most identification mistakes occurred between species with least phenotypic distinctiveness, for instance when observers confused Plains and Hartmann’s mountain zebras. This finding is consistent with other camera trap studies showing that accurate mammal classifications are difficult to obtain where a range of similar-looking species occur sympatrically (Gooliaff and Hodges 2018), or individuals of a particular species exhibit little distinctiveness (Güthlin et al. 2014; Newbolt and Ditchkoff 2019). Following our study, observers commented that animal orientation in the image frame (broad side vs. frontal vs. rear views) as well as ‘crowdedness’, i.e. the occurrence of large numbers of animals in the same image frame, also made accurate identification very difficult. Estimates of Black-faced impala, the species that occurred in the largest social groups at various recording stations, consistently had the highest variance. Group size estimates appeared challenging for any herbivore occurring in gregarious herds, confirming that camera trap-based abundance estimates of non-distinct social species need to be treated with great caution as the cameras’ limited field of view and the complexity of animal movements obstruct an accurate assessment (Stratford and Naholo 2017).

Although based on a small sample of only seven recording stations, our results give evidence that camera location influenced data extraction reliability. This is probably associated with the specific contents and image characteristics recorded at each locality, as background habitat, species compositions and levels of animal activity differed between waterholes. Habitat structure, especially in relation to “openness” has been shown to influence detection and identification accuracy and might have contributed to differing results between camera stations in our study (Wearn and Glover-Kapfer 2019; Egna et al. 2020). Observer image processing rates, mammal detection and identification accuracies, as well as the types and frequency of mistakes indeed varied significantly across the different camera stations. In addition to a diverse range of observer-related influences on classification accuracy, further complications may arise from technical factors influencing image quality, such as image complexity, motion blur and light conditions due to flash settings or varying natural light and contrast between day- and night-time (Rovero et al. 2013; Zheng et al. 2016) (Table 4). Finally, observers also interpreted data extraction rules differently, resulting in greater variance. The novice observer whose estimates frequently contributed extreme outliers adjusted definitions of social group and independent detection events throughout the experiment, indicating that any ambiguity in the data extraction tasks could contribute to higher variance. Inter-observer agreement improved considerably when this participant was omitted from the analyses.

While surveyors can choose from a variety of image management software options to optimise data handling procedures (Young et al. 2018; Greenberg et al. 2019), the final classification of image content still mostly depends on human assessment. Clearly, individual observers can have profound influences on the results. Low agreement between data extractors, however, can lead to skewed population demographics (Newbolt and Ditchkoff 2019) and biased survivorship and population estimates (Morrison et al. 2011; Van Horn et al. 2014; Johansson et al. 2020), thus potentially misguiding conservation measures and policy.
A rapidly evolving suite of automated image processing software aims to mitigate the concerns associated with manual data extraction. These algorithms can overcome certain human biases and issues associated with fatigue, thus improving the precision of results by offering predictable, standardized error rates. However, most, if not all, models still require manual preparation of a training dataset, which is undertaken by both experts (Tabak et al. 2019; Whytock et al. 2021) and/or untrained volunteers (Norouzzadeh et al. 2018, 2019; Willi et al. 2019). Human error is, therefore, not wholly eliminated from the process and at least some of the risks stemming from low classification agreement and false classifications remain. Lower accuracies for out-of-sample data and species with less images in the training dataset (Tabak et al. 2019; Willi et al. 2019; Schneider et al. 2020) suggest that unexpected and rare species may either not be detected or classified incorrectly (Whytock et al. 2021).

Despite fast processing speeds being achieved by some algorithms (e.g. 2 000 images / minute, Tabak et al. 2019), efficiency considerations also need to take into account time and resources required for software development and human classification of training datasets, especially if individual models are not transferable between species and sites. Algorithm classification accuracies and image processing rates in multi-species studies still vary greatly (Supplementary Information 8) and many were similar to those obtained by expert observers using the GDMS. Using models to detect empty image frames and classify common species, however, could significantly reduce the workload of human observers and standardize error rates to predictable levels. Optimizing the processing workflow of large-scale studies with hundreds of thousands or millions of images could involve a hybrid solution, in which trained software models detect animals in image frames but only classify the most common ones, while less common and uncertain records are queried for independent assessment by experienced human observers.

Until machine learning methods can provide reliable automated detection and identification tools for a large variety of common and uncommon species in different habitats (see Willi et al. 2019; Miao et al. 2019; Schneider et al. 2020), much of the time-consuming data extraction work will continue to be the responsibility of human observers. In these studies, the issue of inter-observer reliability can be addressed in different ways. For instance, online on-demand training provides resources to inexperienced data extractors, measurably improving the accuracy of image classifications (e.g. Katrak-Adefowora et al. 2020). Other approaches to reduce the impact of misclassifications may include the omission of records.
causing disagreement (Rode et al. 2021), multi-observer validation of image content (Swanson et al. 2016), or the selective flagging of doubtful records for independent review by more experienced assessors. Here, we demonstrate that simple outlier screening and truncation can improve reliability. Based on our results, we emphasise the importance of assessing classification agreement in multi-observer camera trap studies, especially if a range of similar wildlife species is monitored simultaneously.

The issue of inter-observer reliability is not unique to camera trap studies. It pertains to any field of research inquiry where multiple human observers are expected to assess the attributes of the same ‘object’ and classify their observations. The problems that arise using multiple observers for data extraction can be highly context specific and, in case of camera traps, a variety of factors influence what humans detect and manage to identify. Our results revealed great inconsistencies in some of the most common measurements obtained with camera trapping: the detection and accurate identification of animals if they occurred (occurrence and biodiversity studies), resulting in both false negatives and false positives, as well as high variances in population size parameters of common social species. Simply put, we do not all see or count the same in a given set of camera trap images, even under favourable conditions. While samples sizes prevented us from testing how the different factors that influence content classification accuracy might interact, our experiment reflected realistic research circumstances in terms of observer selection, the images being processed, the tasks assigned, and the software utilised. Observer performance and associated biases deserve greater attention in multi-species studies, and particularly so if these are founded on mass data extraction by different observers (Cruickshank et al. 2019).

Acknowledgements We sincerely thank all observers who participated in our experiment and Simeon Naholo for assistance during field data collection. We specially thank Stéphanie Périquet for valuable contributions to the methods and study concept, as well as pre-publication comments. We thank the Directors of Ongava Game Reserve.

Author contributions Theresa Zett: Formal Analysis (lead), Writing – original draft (equal), Writing – review & editing (equal), Visualization (equal). Ken Stratford: Data curation (equal), Software (lead), Writing – original draft (equal), Writing – review & editing (equal). Florian Weise: Conceptualization (lead), Data curation (equal), Formal Analysis (equal), Visualization (equal), Writing – original draft (equal), Writing – review & editing (equal).

Funding This study was funded by the Ongava Research Centre.

Data Availability The datasets generated and analysed during the study contain species at high risk of poaching and are therefore not publicly available. Data can be shared upon special, reasonable request to the corresponding author.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics Data collection was non-invasive and required no specific authorisation.

Consent to participate Informed consent was obtained from all observers participating in the study.
References

Agha M, Batter T, Bolas E et al (2018) A review of wildlife camera trapping trends across Africa. Afr J Ecol 56:694–701. https://doi.org/10.1111/aje.12565

Bartko J (1966) The Intraclass Correlation Coefficient as a Measure of Reliability. Psychol Rep 19:3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Beery S, Van Horn G, Perona P (2018) Recognition in Terra Incognita. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp 472–489

Bland J, Altman D (1986) Statistical Methods For Assessing Agreement Between 2 Methods Of Clinical Measurement. Int J Nurs Stud 47. https://doi.org/10.1016/j.ijnurstu.2009.10.001

Boksem M, Meijman T, Lorist M (2005) Effects of mental fatigue on attention: An ERP study. Brain Res Cogn Brain Res 25:107–116. https://doi.org/10.1016/j.cogbrainres.2005.04.011

Burns P, Parrott M, Rowe K, Phillips B (2018) Identification of threatened rodent species using infrared and white-flash camera traps. Aust Mammal 40. https://doi.org/10.1071/AM17016

Burton C, Neilson E, Moreira-Arce D et al (2015) REVIEW: Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. J Appl Ecol 52. https://doi.org/10.1111/1365-2664.12432

Caravaggi A, Banks PB, Burton AC et al (2017) A review of camera trapping for conservation behaviour research. Remote Sens Ecol Conserv 3:109–122. https://doi.org/10.1002/rse2.48

Cicchetti D, Sparrow S (1981) Developing criteria for establishing the interrater reliability of specific items in a given inventory. Am J Ment Defic 86:127–137

Cruickshank S, Bühler C, Schmidt B (2019) Quantifying data quality in a citizen science monitoring program: False negatives, false positives and occupancy trends. https://doi.org/10.1111/esp2.54

De Bondi N, White J, Stevens M, Cooke R (2010) A comparison of the effectiveness of camera trapping and live trapping for sampling terrestrial small-mammal communities. Wildl Res 37:456–465. https://doi.org/10.1071/WR10046

Després-Einspenner M-L, Howe E, Drapeau P, Kühl H (2017) An empirical evaluation of camera trapping and spatially explicit capture-recapture models for estimating chimpanzee density. Am J Primatol 79:22647. https://doi.org/10.1002/ajp.22647

Edwards S, Portas R, Hanssen L et al (2018) The spotted ghost: Density and distribution of serval Leptailurus serval in Namibia. https://doi.org/10.1111/aje.12540

Faber L, Maurits N, Lorist M (2012) Mental Fatigue Affects Visual Selective Attention. PLoS ONE 7:e48073. https://doi.org/10.1371/journal.pone.0048073

Fisher RA (1958) Statistical Methods for Research Workers. Temporary Publisher

Friedman M (1937) The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. J Am Stat Assoc - J AMER Stat ASSN 32:675–701. https://doi.org/10.1080/01621459.1937.10503522

Friedman M (1939) A Correction: The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. https://miltonfriedman.hoover.org/objects/57257/a-correction-the-use-of-ranks-to-avoid-the-assumption-of-no. Accessed 6 May 2021

Gamer M, Lemon J, Fellows I, Sing P (2019) irr: Various Coefficients of Interrater Reliability and Agreement. Version R package Version 0.84.1 URL https://CRAN.R-project.org/package=irr

Gooliaff T, Hodges K (2018) Measuring agreement among experts in classifying camera images of similar species. Ecol Evol 8. https://doi.org/10.1002/ece3.4567

Greenberg S, Godin T, Whittington J (2019) Design patterns for wildlife-related camera trap image analysis. Ecol Evol 9. https://doi.org/10.1002/ece3.5767
Steenweg R, Hebblewhite M, Kays R et al (2017) Scaling-up camera traps: monitoring the planet’s biodiversity with networks of remote sensors. Front Ecol Environ 15. https://doi.org/10.1002/fee.1448

Stephens MA (1974) EDF Statistics for Goodness of Fit and Some Comparisons. J Am Stat Assoc - J AMER Stat ASSN 69:730–737. https://doi.org/10.1080/01621459.1974.10480196

Stratford K, Naholo S (2017) Can camera traps count game? Namib J Environ 1:B–31

Stratford K, Stratford S, Périquet S (2019) Dyadic associations reveal clan size and social network structure in the fission–fusion society of spotted hyaenas. Afr J Ecol 58. https://doi.org/10.1111/aje.12641

Swann DE, Perkins N (2014) Camera trapping for animal monitoring and management: A review of applications. In: Camera Trapping Wildlife Management and Research. CSIRO Publishing, pp 3–11

Swanson A, Kosmala M, Lintott C et al (2015) Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. Sci Data 2:150026. https://doi.org/10.1038/sdata.2015.26

Swanson A, Kosmala M, Lintott C, Packer C (2016) A generalized approach for producing, quantifying, and validating citizen science data from wildlife images: Citizen Science Data Quality. https://doi.org/10.1111/cobi.12695. Conserv Biol 30:

Tabak MA, Norouzzadeh MS, Wolfson DW et al (2019) Machine learning to classify animal species in camera trap images: Applications in ecology. Methods Ecol Evol 10:585–590. https://doi.org/10.1111/2041-210X.13120

Thomas M, Baker L, Beattie J, Baker A (2020) Determining the efficacy of camera traps, live capture traps, and detection dogs for locating cryptic small mammal species. Ecol Evol 10. https://doi.org/10.1002/ece3.5972

Trolliet F, Huynen M-C, Vermeulen C, Alain H(2014) Use of camera traps for wildlife studies A review. Biotechnol. Agron. Soc. Environ.18

Van Horn R, Zug B, Lacombe C et al (2014) Human visual identification of individual Andean bears Tremarctos ornatus Human visual identification of individual Andean bears Tremarctos ornatus. Wildl Biol 20:291–299. https://doi.org/10.2981/wlb.00023

Watson P, Petrie A (2010) Method agreement analysis: A review of correct methodology. Theriogenology 73:1167–1179. https://doi.org/10.1016/j.theriogenology.2010.01.003

Wearn OR, Glover-Kapfer P (2019) Snap happy: camera traps are an effective sampling tool when compared with alternative methods. R Soc Open Sci 6. https://doi.org/10.1098/rsos.181748

Welbourne D, Macgregor C, Paull D, Lindenmayer D (2015) The effectiveness and cost of camera traps for surveying small reptiles and critical weight range mammals: A comparison with labour-intensive complementary methods. Wildl Res 42:414–425. https://doi.org/10.1071/WR15054

Whytock RC, Świeżewski J, Zwerts JA et al (2021) Robust ecological analysis of camera trap data labelled by a machine learning model. https://doi.org/10.1111/2041-210X.13576. Methods Ecol Evol n/a

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Version R package Version 3.3.3. Springer-Verlag, New York. URL https://cran.r-project.org/web/packages/ggplot2/index.html

Willi M, Pitman RT, Cardoso AW et al (2019) Identifying animal species in camera trap images using deep learning and citizen science - Wiley Online Library. Methods Ecol Evol

Young S, Rode-Margono J, Amin R (2018) Software to facilitate and streamline camera trap data management: A review. Ecol Evol 8. https://doi.org/10.1002/ece3.4464

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.