Condition metrics for Elm Forest and Saxaul ecosystems in the Gobi Desert

S.J. Sinclair, O. Avirmed, K. Batpurev, P.A. Griffioen, M.D. White and Kirk Olson

February 2020

Arthur Rylah Institute for Environmental Research
Technical Report Series No. 310

Arthur Rylah Institute for environmental research

WCS

Оюу Толгой

VICTORIA State Government | Environment, Land, Water and Planning

Acknowledgment of traditional owners in Victoria

In Victoria, we acknowledge and respect Victorian Traditional Owners as the original custodians of Victoria's land and waters, their unique ability to care for Country and deep spiritual connection to it. We honour Elders past and present whose knowledge and wisdom has ensured the continuation of culture and traditional practices.

We are committed to genuinely partner, and meaningfully engage, with Victoria's Traditional Owners and Aboriginal communities to support the protection of Country, the maintenance of spiritual and cultural practices and their broader aspirations in the 21st century and beyond.

In the Gobi Desert, we acknowledge and respect nomadic pastoralists as the custodians and managers of the land. We honour Elders past and present whose knowledge and wisdom has ensured the continuation of culture and traditional practices.

We are committed to genuinely partner, and meaningfully engage, with local herding communities in the Gobi to support the protection of the land, the maintenance of spiritual and cultural practices and their broader aspirations in the 21st century and beyond.

Arthur Rylah Institute for Environmental Research
Department of Environment, Land, Water and Planning
PO Box 137
Heidelberg, Victoria 3084
Phone (03) 9450 8600
Website: www.ari.vic.gov.au

Citation: Sinclair, S.J., Avirmed, O., Batpurev, K., Griffioen, P.A., White, M.D. and Olson, K. (2020). Condition metrics for Elm Forest and Saxaul ecosystems in the Gobi Desert. Report for Oyu Tolgoi and Arthur Rylah Institute for Environmental Research Technical Report Series No. 310. Department of Environment, Land, Water and Planning, Heidelberg, Victoria.

Front cover photo: Siberian Elm trees growing along an ephemeral watercourse, near Khanbogd, Mongolia (Steve Sinclair).

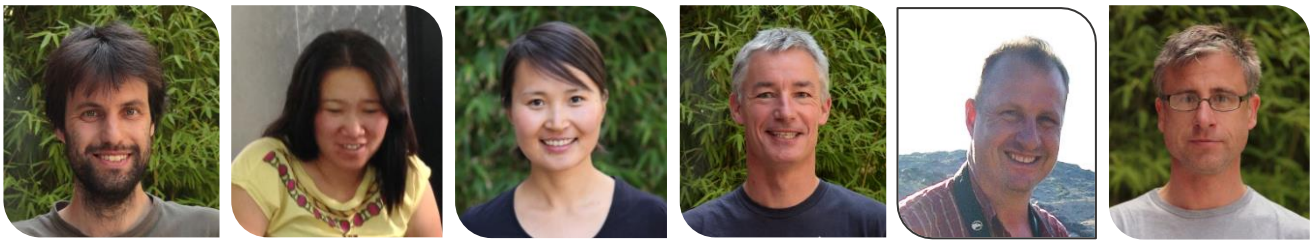
ISSN 1835-3827 (print)
ISSN 1835-3835 (pdf))

Accessibility

If you would like to receive this publication in an alternative format, please telephone the DELWP Customer Service Centre on 136 186, email customer.service@delwp.vic.gov.au or contact us via the National Relay Service on 133 677 or www.relayservice.com.au. This document is also available on the internet at www.delwp.vic.gov.au

Condition metrics for Elm Forest and Saxaul ecosystems in the Gobi Desert

Steve J. Sinclair[1], Otgonsuren Avirmed[2], Khorloo Batpurev[1], Peter A. Griffioen[1], Kirk Olson and Matt D. White[1]



[1]Arthur Rylah Institute for Environmental Research
123 Brown Street, Heidelberg, Victoria 3084

[2]Wildlife Conservation Society, Mongolia Country Program
San Business Center-201, Amar Street-29, Small ring road, Sukhbaatar district, Post-20A, Box-21, Ulaanbaatar-14200, Mongolia

Arthur Rylah Institute for Environmental Research
Technical Report Series No. 310

Arthur Rylah Institute for Environmental Research
Department of Environment, Land, Water and Planning
Heidelberg, Victoria

Acknowledgements

This project was funded by Oyu Tolgoi (OT), Mongolia. We thank Samdanjigmed Tulganyam and James Hamilton for their support. This was a collaborative project between ARI and Wildlife Conservation Society, Mongolia Country Program (WCS). We thank Enkhtuvshin Shiilegdamba (Director WCS Mongolia) for her support. The work depended on the generosity of over many stakeholders. We are greatly appreciative of their efforts. The project also benefitted from the support of Kim Lowe, Tim O'Brien, Graeme Newell and Matthew Bruce at ARI.

Contents

Tables

Figures

Summary

Context:

Ecological change in the Gobi Desert of Mongolia must be monitored to understand the impacts of changing land use, including the Oyu Tolgoi (OT) mining project. To address this need, metrics were created in 2018 which quantify the condition of five rangeland ecosystems (Avirmed et al. 2018). That work was undertaken as a collaboration between Wildlife Conservation Society (WCS Mongolia and the Arthur Rylah Institute for Environmental Research (ARI). That work, like the current project, was funded by OT.

The 2018 metrics were created by collecting stakeholder opinions about the condition of a wide range of computer-generated sites, and then using these opinions to create models of stakeholder condition score. These models can be applied as metrics. They take field-measured data from a site of interest and return a condition score. Given the stakeholder-driven approach, the metric score explicitly reflects stakeholders' views. While Avirmed et al. (2018) created metrics for five ecosystems, the original project scope only allowed three priority ecosystems to be tested using field data. The metrics for the other two ecosystems (Elm Forest and Saxaul) remained draft metrics.

Aims:

We aimed to test and refine the condition metrics for Elm Forest and Saxaul, with the aim of producing final metrics fit for field monitoring.

Methods:

We tested the existing Elm Forest and Saxaul metrics using the same tests used by Avirmed et al. (2018), and diagnosed the issues we found. We implemented three strategies to improve the metrics:

- Newly available field-derived evaluation data were added to the training set,
- The computer-generated training sites were culled to remove sites with 'unrealistic' attributes. This reduction was guided by stakeholder opinions about the 'plausibility' (likelihood of encountering) each site.
- The training data were stratified and weighted, to ensure that all models encountered data from a range of sites, with a focus on sites with realistic attributes.

We tested the new metrics using a cross validation approach. Some field data were used in the above re-modelling strategy, while the remaining sites were withheld for testing. We repeated this ten times, with different sub-samples of test data. We appraised model improvement by the difference in r^2 between the old and new metrics.

Results:

When tested against stakeholder field evaluations, the Elm Forest and Saxaul metrics created in 2018 were positively related to stakeholder evaluations, and were thus capable of providing condition scores consistent with stakeholder expectation. The Elm Forest and Saxaul metrics did not perform as well as the metrics previously created for the three desert ecosystems. We suggest this was caused by-

1) a lack of consensus among stakeholders for these ecosystems, making it more difficult to model the consensus. We suggest that it is inherently more difficult for stakeholders to conceptualise condition change in these systems, and
2) bias in the original training data. The computer-generated sites were skewed towards sites with unrealistically high cover and richness values. Real field sites resemble only a small range of the data used in training, at the lower end of the score range. This presumably resulted in the models returning low scores and having low resolution when dealing with the attributes of real sites.

The strategy we applied to improve the metrics clearly improved the Elm Forest metric but resulted in a Saxaul metric which was not demonstrably better, and was likely overfit to the field training data.

Conclusions and implications:

We conclude that the final metrics presented here are fit for monitoring, for both Elm Forest and Saxaul. The Saxaul model is likely to be over-fit to the training data (meaning that the model's ability to extrapolate beyond the training data is lower than expected) but we argue that this is not a major cause for concern.

1 Introduction

1.1 The need to evaluate ecosystem condition in the Gobi Desert

The Gobi Desert in Inner Asia is an arid region in the rain-shadow of the Himalaya, in southern Mongolia and northern China. It experiences some of the most extreme weather conditions on Earth. Annual rainfall often falls short of 50 mm. Winter temperatures routinely drop below -30ºC and daily maxima may not exceed minus 10ºC for weeks.

Despite the extreme conditions, the Gobi has been populated for tens of thousands of years, and its people have adapted to changes in climate over that period (Owen et al., 1998). Nomadic pastoralism has been the dominant land use for millennia, but over the last thousand years there have been profound changes in social organisation. Pastoral practices became more concentrated and regulated under Mongol and Manchu rule, with complex systems controlling the wealth and movement of nomadic families (Fernández-Giménez 1999). The spread of communism in the 1920s saw a profound upheaval, with traditional administrative structures abolished, and livestock confiscated. By the 1950s, most livestock were tended by collectives, some traditional pastoral knowledge was lost, nomadic migrations were curtailed, many wells were established, and supplementary feeding became commonplace (Fernández-Giménez 1999, Addison et al. 2012). In the 1990s, communism collapsed, and livestock were again privatised. These events ushered in the current era, which has seen an increase in economic inequality, and non-traditional patterns of grazing (Fernández-Giménez 1999). Livestock numbers, particularly goats, have increased substantially since the 1960s (Bedunah and Schmidt 2000, Addison et al. 2012; Tuvshintogtokh and Ariungerel 2013; Rao et al. 2015).

As land use changes, land managers must understand and evaluate the changes, to guide land use planning, assessments of environmental harm and mitigation and the evaluation of management outcomes (Aronson et al. 1995, Brownlie and Treweek 2016). This is difficult because ecosystem change usually involves changes to numerous components (Folke et al., 2002, Walsh and Carpenter, 2016) including processes (e.g. soil erosion, nutrient cycling, etc), habitat structures (e.g. vegetation form, height and density, soil structure) and the abundance and richness of many species. This complexity makes it difficult to evaluate whether a given change represents a net positive or negative change, and to what degree.

In the Gobi, for example, increased grazing pressure is generally thought to induce the loss of palatable species, the increase of non-palatable species, the overall loss of vegetation cover, and soil erosion. These changes are routinely perceived as a loss of condition (Tserendash and Erdenebaatar 1993, Shiping & Yonghong 1999, Fernández-Giménez and Allen-Diaz 2001, Stumpp et al. 2005, Zhou et al. 2005, Yong-Zhong et al. 2005, Lkhagva et al. 2013, Narantsetseg et al. 2015, Jamiyansharav et al. 2018), however the relative importance of the different elements, their causes and interactions are debated (Lambin et al. 2001, Jamsranjav et al. 2018).

Recently, mining has surpassed pastoralism as the major economic activity in the Gobi. In 2010, construction began on the Oyu Tolgoi (OT) mine project. A Comprehensive Environmental and Social Impact Assessment (OT 2012) identified direct and indirect environmental impacts of the operation, and proposed ways to minimize and manage those impacts as well as to maximise its benefits, including offset activities delivered through the ‘Sustainable Cashmere Project’, which aims to reduce inappropriate grazing pressure. Rangelands are being used as a surrogate for some biodiversity features, and it is agreed that their ‘condition’ will be monitored over time (OT 2012).

Together, concerns about over-grazing and the mandated requirement to monitor rangelands under the OT offset program have made it imperative that ecological condition is understood and quantified.

1.2 What is ecosystem condition?

1.2.1 The concept of condition

The evaluation of complex ecological change can be assisted by the simplifying concept of 'condition', measured by a condition metric. A condition metric is an algorithm which takes multiple attributes and combines them into a single number (condition score). Despite the importance of 'condition' in conservation biology, the construction of metrics capable of actually quantifying condition remains controversial. There are several overlapping areas of debate:

- What ecological attributes should be used to characterise an ecosystem? (grass cover? ant abundance?)
- How do these attributes relate to condition? (how much grass cover is best? can there be too much?)
- How do these attributes relate to each other (is grass species richness as important as grass cover? Are these attributes interactive?)
- Is there only a single desired state? (is a shrubland as valuable as a grassland in a particular context?)
- How do the attributes relate to the "desired state"? (is there one optimal grass cover?)
- Do naturally reversible fluctuations in the attributes represent condition fluctuation? (do seasonal change, or responses to disturbance represent degradation?)
- Should condition measures allow direct comparison between ecosystems? (is there a 'common currency' that expresses condition in a steppe as well as a jungle?)

These questions are controversial because they are ultimately subjective. Science and measurement alone cannot resolve them without interpretation and judgement by people. Ecological condition is unavoidably subjective (Daniel & Vining 1983, Keith & Gorrod 2006, Sinclair et al. 2015, 2018).

Subjectivity can be addressed in two main ways:

- Consultation to develop collective opinions, which gain credibility from their 'democratic' origins (Oliver et al. 2007; Wood and Lavery 2000; Venables & Boon 2016).
- Construction of repeatable methods that allow evaluations to be made repeatedly using the same criteria, which confers credibility from transparency and consistency (Gibbons & Freudenberger 2006).

These approaches may be combined, such that stakeholder consultation leads to the creation of a repeatable metric.

1.2.1 Our definition of condition

We use the following definition of condition that is consistent with most published studies (Oliver et al. 2002, Parkes et al. 2003; Buckland et al. 2005, Gibbons & Freudenberger 2006, Stoddard et al. 2006; Sinclair et al. 2018).

Ecological condition measures the retention (or loss) of the ecological attributes that characterise an ecosystem in its desired state.

Our method for metric construction acknowledges that the attributes and the desired state are subjective, and asks each stakeholder to contribute their own personal ideas, which are incorporated into a collectively-defined consensus. The following constraints were placed on the conception of condition, and these were explained to all stakeholders.

Condition may include elements of "quality", "intactness", "health" or "conservation value". It may include consideration of the following factors (to any degree):

- The value of the site in providing key ecological functions,
- The provision of habitat for the wildlife of the ecosystem,
- The provision of habitat for the plants of the ecosystem,
- The stabilisation of the soil,
- The value of the site as an example of its type,
- The abundance of particularly important species or life-forms,
- How important the site should be for conservation / protection,

- The degree to which the site resembles a site that has suffered no loss of condition,
- How much a well-informed (expert) stakeholders “likes” the site.

The following considerations are not included in our conception of condition (although their importance in other contexts is acknowledged):

- The personal wealth that could be derived from the site (livestock or money),
- The value of the site for any other purpose other than as an example of its ecosystem type,
- The likely future for the site (whether good or bad),
- The cost of rehabilitating the site.

1.3 Context: Preceding work

This report is best read as a continuation of prior work described in Avirmed et al. (2018; a former Technical Report from the Arthur Rylah Institute (No. 289)). This section briefly summarises that work.

1.3.1 Variable selection

Variables that express the condition of sites were selected based on stakeholder consultation. These mostly relate to plant cover and plant species richness. The appropriateness of the variables was later tested quantitatively, by comparing stakeholder evaluations of real sites (without reference to the variable set) with stakeholder evaluations of the same sites in a workshop context, where the sites were abstracted and described only by the site variables. The variable sets were robust for the three ecosystems tested.

1.3.2 Generation of a set of hypothetical sites to gather evaluation data

We used the selected variables to describe a large set of hypothetical (computer-generated) sites (n=125 for Elm Forest, n=135 for Saxaul). These sites were created in order to describe the full range of condition states for each ecosystem, including many different combinations of the site variables (combinations of high and low cover and richness for various plant groups). These sites were used to prompt experts to provide their evaluations of condition.

1.3.3 Stakeholder participation

Our metrics were created from stakeholder opinion, with the aim that the metrics speak for the stakeholder group. This is crucial, given the role of both scientific and traditional knowledge in understanding the dynamics of Mongolian rangelands (Fernández-Giménez and Allen-Diaz, 2000). Stakeholders were selected by WCS, in consultation with OT. They were required to be very familiar with the composition and dynamics of Gobi Ecosystems, and the management challenges they face. They were deliberately chosen to represent a wide range of experience and expertise, and included botanists, animal ecologists, nomadic pastoralists and conservation managers and policy-makers.

A self-assessed stakeholder questionnaire covering many different topics was used to show the expertise possessed by the stakeholders. It is essential that the stakeholder population is described, so that it is transparent which collective opinion is represented.

In a workshop context, stakeholders were introduced to the concept of condition (as described above). Then, a small selection of the computer-generated sites (n=15) was presented to them as descriptions on paper cards. Each stakeholder was asked to provide a condition score for each site, on a scale of 0-100.

We used outlier detection to identify stakeholders whose evaluations contrary to the consensus, and we removed the data from these people from the dataset. The final dataset included the evaluations of 74 stakeholders for Elm Forest and 78 for Saxaul.

1.3.4 Metric creation

We sought a metric for each ecosystem that spoke for the collective opinion of all stakeholders. The opinions of stakeholders were explicitly used to create each metric. Their scores for each site (dependent variables) and the variables describing that site (independent variables) were then used to train models (an ensemble of bagged regression trees (Breiman et al. 1984, Blockeel et al. 1999)) that aimed to predict the quality score from the measured variables. The models were converted directly into metrics for each system.

The method was chosen because it has several advantages over other methods, such as weighted combinations. These include:

- There is an explicit recognition in the method that the concept of ecological quality is subjective, and is derived from human preferences.
- The means of blending the multiple variables is driven by data, and is transparent and repeatable.

- Allowing each stakeholder to envisage their own “desired state” (rather than having one defined by the project), within the limits of the variables provided, effectively introduces multiple desired states into the metric, partially overcoming the problems of natural fluctuations and between-site variation.

It is important to note that the use of opinion in this context is not in lieu of other empirical data; as no such data could conceivably be obtained. The stakeholder evaluations are the primary data and must not be considered ‘placeholders’ until better data fills the void.

1.4 Aims of the current work

This project aimed to:

- Evaluate the Elm Forest and Saxaul metrics produced by Avirmed et al. (2018),
- If necessary, refine those metrics to make them fit for use,
- Present the final metrics in a format that allows them to be implemented for monitoring.

Consistent with Avirmed et al. (2018), we consider that a metric that is ‘fit for use’ is one that-

- can distinguish sites of different condition, as perceived by stakeholders in the field, including sites at the extreme ends of the condition spectrum,
- calculate scores using data that is easily derived from field plots, which can be completed by any moderately skilled botanist within 1 hour,
- can detect changes related to land-use change over multi-year periods,
- is not unduly influenced by natural and short-term fluctuations,
- is supported and justified by good data,
- is explicitly linked to the views of stakeholders,
- is tested on field data, and
- can facilitate comparisons of condition both within and between ecosystems.

The metrics are NOT designed to-

- explicitly evaluate habitat for any species of plant or animal (although habitat quality for wildlife does contribute to the conception of condition),
- explicitly consider values associated with rare or threatened species (although the distribution of some rare species may be related to condition),
- consider the area or spatial extent of sites,
- consider the spatial arrangement or context of sites,
- be calculable from remotely sensed data (although it is important to work towards this, as discussed in Avirmed et al. (2018)).

1.5 The ecosystems covered by the metric

1.5.1 Elm Forest

The Elm Forest ecosystem is restricted to ephemeral sandy or pebbly watercourses (sayrs) which occasionally flood, and where groundwater is always available (Wesche et al. 2011). The ecosystem is characterised by the presence of Siberian Elm (*Ulmus pumila*) (Figure 1) which form a patchy canopy (known locally as 'forest', and referred to as such in this report, although not meeting some global definitions of forest based on canopy cover). The ground-level vegetation is very sparse or almost absent, with occasional shrubs (e.g. *Nitraria sibirica* (Nitrariaceae)), forbs and grasses.

It is suspected that Siberian Elm was once more widespread and numerous within this niche, and that it has been depleted by human land use. Trees are sometimes harvested, and livestock prevent the recruitment of new stems. The species probably has the potential to expand along sayrs and increase its local density, if human impacts were relaxed (Wesche et al. 2011). Consequently, it may be unclear whether a treeless portion of a sayr is former or potential Siberian Elm habitat, making the fine-scale delineation of this ecosystem difficult.

Siberian Elm sometimes occurs outside the river bed habitat described here, such as in rocky gorges (in the Gobi) or in areas with higher rainfall (outside the Gobi) (Wesche et al. 2011). These other occurrences are beyond the scope of this work, and those ecosystems are not served by the metric developed here.

Figure 1. An example of Elm Forest, with a canopy of Siberian Elm (*Ulmus pumila*).

1.5.2 Saxaul

The Saxaul ecosystem is defined by the dominance of a single species of shrub: Saxaul (*Haloxylon ammodendron*), which may grow to over 4 m in height (Figure 2). This species is extremely tolerant of environmental extremes, including salinity, sand burial and both extended droughts and waterlogging or flooding (Xu et al. 2014). Few other species in inner Asia tolerate these extreme conditions and therefore Saxaul often occurs with little other vegetation. When other species are present, they include a range of Chenopod shrubs, along with other drought tolerant species such as *Calligonum mongolicum* (Polygonaceae) and *Zygophyllum xanthoxylon* (Zygophyllaceae).

Despite this tolerance, seedlings require moisture, and recruitment occurs only occasionally, in wet years and in habitats where water collects (Fa-min et al. 2003). Several distinct geomorphic contexts provide the combination of conditions that allow Saxaul to dominate, including alluvial sand plains with groundwater access, stony floodways or flood-outs, saltpans and clay-beds of ephemeral lakes.

Saxaul is considered an important species because it is harvested for use by people (fuel, dyes and medicines), because it binds sand in places where few other species occur (Zou et al. 2010), and because it provides important habitat for several wildlife species (Maclean 1996).

Figure 2. An example of Saxaul, showing the dominance of Saxaul (*Haloxylon ammodendron*).

1.5.3 Why are these ecosystems the focus of additional work?

There are two main reasons why the Elm Forest and Saxaul metrics were treated as draft metrics in Avirmed et al. (2018), necessitating further work to test and complete them.

First, the needs of the OT offset scheme are focussed on the other desert ecosystems, which are widespread and common. The limited field time in the prior project was best spent dealing with these priority systems in the field season of 2018, leaving the Elm Forest and Saxaul ecosystems untested.

Second, both the Elm Forest and Saxaul ecosystems were considered likely to be problematic for the creation of condition metrics, meaning that both systems were suspected to require focussed attention. There are several reasons for this, including the following:

- Both systems generally support few species and have very low vegetation cover (often <10% total cover). Systems with very low cover present problems for the conceptualisation of condition. This is best illustrated by imagining an extreme case, such as a desert (or an ocean bed) with shifting sand and no natural vegetation. In such a scenario it is difficult to imagine the difference between a high and a low condition site, from the point of view of vegetation or habitat structure. There are simply too few measurable ecological attributes that can incur a condition loss, and few means of distinguishing change. Both the Elm Forest and Saxaul systems are approaching this situation.
- Elm Forest and Saxaul are relatively rare in the landscape, and most stakeholders are expected to have more limited experience of their ecology than the more-widespread desert ecosystems.

2 Methods

2.1 Variables to express and measure condition

Avirmed et al. (2018) selected sets of site variables, with the intention that they enable satisfactory evaluation of site condition. Variables were selected which-

- describe the main features of the vegetation of the ecosystem (i.e. dominant species and lifeforms),
- are likely to respond to the main pathways of degradation and recovery (e.g. grazing regimes),
- do not experience substantial short-term (weeks, months) fluctuations which may obscure longer-term (years) processes of degradation and recovery, and
- could be quantified easily during a single site visit of <1 hour.

A description of the variable selection process and an ecological rationale for the inclusion of each variable was presented in Avirmed et al. (2018). The final sets of variables selected for each ecosystem are shown in Table 1. The variable sets differ between systems, reflecting their different composition and ecology. Some of the variables are nested (e.g. 'Cover of shrubs' is a subset of 'Cover of all vegetation'), and some variables are closely correlated (e.g. 'Cover *Haloxylon ammodendron*' and 'Density *Haloxylon ammodendron'*.). Correlation and nestedness are not problems for the modelling approaches used here.

Table 1. The variables used to assess condition for each ecosystem.

Variable	Elm Forest	Saxaul
Total vegetation cover	✓	✓
Cover all shrubs	✓	✓
Richness all shrubs	✓	✓
Cover all perennial grasses and sedges	✓	✓
Richness all grasses and sedges	✓	✓
Cover perennial forbs	✓	✓
Richness all forbs	✓	✓
Cover of all Succulent shrubs		✓
Cover of litter		✓
Max height exposed roots pedestals		✓
Cover *Ulmus pumila*	✓	
Density adult *Ulmus pumila*	✓	
Density juvenile *Ulmus pumila* (suppressed)	✓	
Density juvenile *Ulmus pumila* (escaped)	✓	
Density sapling *Ulmus pumila*	✓	
Cover *Haloxylon*		✓
Density large *Haloxylon*		✓
Total number of variables	**12**	**12**

The terms used to define the site variables for the Elm and Saxaul ecosystems are explained below. The species groups are not mutually exclusive (some species belong to multiple groups).

- Shrubs: Dicotyledonous plants (of any family) which form perennial, above-ground woody stems. Such stems have secondary thickening and can be “snapped”.
- Forbs: Any species of angiosperm (monocot or dicot) that is not a shrub, and not a member of the Poaceae or Cyperaceae. This group also includes sub-shrubs (also called semi-shrubs) such as *Anabasis brevifolia*. It also includes the onion family (*Allium* sp.).
- Grasses and sedges: Any species in the families Poaceae (grasses) or Cyperaceae (Sedges).
- Perennial (forbs / grasses and sedges): Any species which is not annual. This group includes biennials and species which may be facultatively annual under harsh conditions.
- Succulent species: Any species of dicot (shrub of forb) which has thickened, fleshy foliage that is “juicy”, including *Haloxylon ammodendron.*
- Large *Haloxylon*. Any individual specimen of *Haloxylon ammodendron* that exceeds 1.5 m in total height. Only living specimens are included. In cases where plants have defoliated, a judgement must be made as to whether the plants remain alive (and they are included), or they have died (and they count only as litter).
- Adult *Ulmus pumila*. Any individual specimen of *Ulmus pumila* that exceeds 2.5 m in total height.
- Juvenile *Ulmus pumila* (suppressed). Any individual specimen of *Ulmus pumila* that is between 0.5 m and 2.5 m in total height, and is experiencing browsing by animals, such that it has many growth points, none of which are forming a new leader / future trunk.
- Juvenile *Ulmus pumila* (escaped). Any individual specimen of *Ulmus pumila* that is between 0.5 m and 2.5 m in total height that has one or a few extended recent branches that are likely to exceed 2.5 m and form a future trunk.
- Sapling *Ulmus pumila*. Any individual specimen of *Ulmus pumila* that is less than 0.5 m in total height.
- Litter. Any plant material that is dead, including material detached from the plant on which it formed (e.g. discarded leaves, twigs, etc.) and whole dead plants.
- Cover. Projective foliage cover. i.e. the shadow cast by the species (including all leaves, branches, trunk, etc., but not double-counting overlapping cover).
- Density: Density refers to the number of the item per 900 m^2 plot.
- Richness: The count of species within the 900 m^2 plot.
- Exposed roots/pedestals. Roots which formed below ground, but have been exposed by the erosion of soil. The height is measured vertically, from the root / trunk boundary, to the point at which the lowest root is concealed by soil. The variable measures the highest example that can be found in the plot (not the mean).

2.2 Field sampling protocol for Saxaul

2.2.1 Plot design

The procedure described here is recommended for use in future sampling.

At each site, a 30 x 30 m (900 m^2) square plot was laid out. Each corner of the plot was marked with a flag. The plot was sampled using the methods described below. Every plot was sampled in less than 1 hour.

2.2.2 Sampling plant species cover and litter cover

Within this plot, 4 parallel tape measures were laid out, crossing the plot at 6 m, 12 m, 18 m and 24 m. Each of these tape measures defined a point intercept sampling line. 120 sampling points were distributed evenly along each line, spaced every 0.25 m (commencing at 0.25, ending at 30.0), totalling 480 points per plot. The plot design in shown in Figure 3.

At each point, a narrow steel pin was held vertically, and any plant species or organic litter in contact with the pin was recorded. Multiple species (and litter) were recorded at a single point, but each species was only recorded once per point (i.e. we did not quantify overlapping cover). We calculated the cover of each species (and litter) individually using the following formula:

Percentage cover of species = (# points species recorded / 480) x 100

This species-specific cover data was used to calculate all of the cover-based variables (e.g. Cover of all shrubs), by summing the covers of all species in each lifeform category (It is assumed that the generally low overall vegetation cover in the Gobi Desert permits this approach, without a correction for overlapping cover between species, as would be required for some vegetation types, such as a multi-layered rainforest).

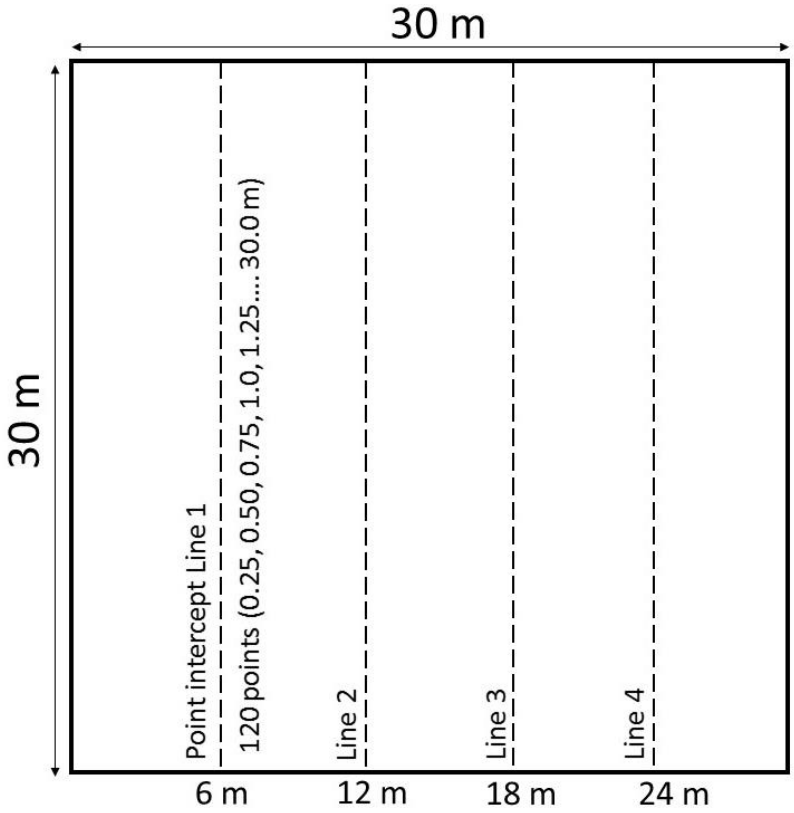


Figure 3. The plot method used to sample vegetation in the field.

2.2.3 Sampling species richness

Species richness refers to the count of species present in a defined area (here, 900 m^2). Point intercept methods are unreliable for quantifying species richness, because they only sample a relatively small area of the plot (the points), and rare species are routinely missed (Godínez-Alvarez et al. 2009). In order to sample species richness, we employed a 10-minute timed search of the plot. The timed search was undertaken by a single experienced botanist (in this case S. Jambal, WCS), recording all vascular plant species, regardless of their cover. Richness values for each of the lifeforms was calculated by simply counting the number of species in each lifeform.

2.2.4 Measuring the density of 'Large *Haloxylon*'

The 10-minute search of the plot also includes a count of all *Haloxylon* plants >1.5 m tall. This count provides the measurement for 'Density of large *Haloxylon* plants'.

2.2.5 Sampling the maximum height of roots exposed by soil loss

To quantify the maximum height of roots exposed by soil loss, a single observer checked the root systems of all shrubs in the plot. For shrubs where some of the root system was exposed by soil loss, the vertical distance between the root-shoot junction and the point of contact with the current soil level was measured (Figure 4). The maximum distance found on any shrub in the plot was recorded. This process was easily completed within the 10-minute search time allotted to the botanical observer.

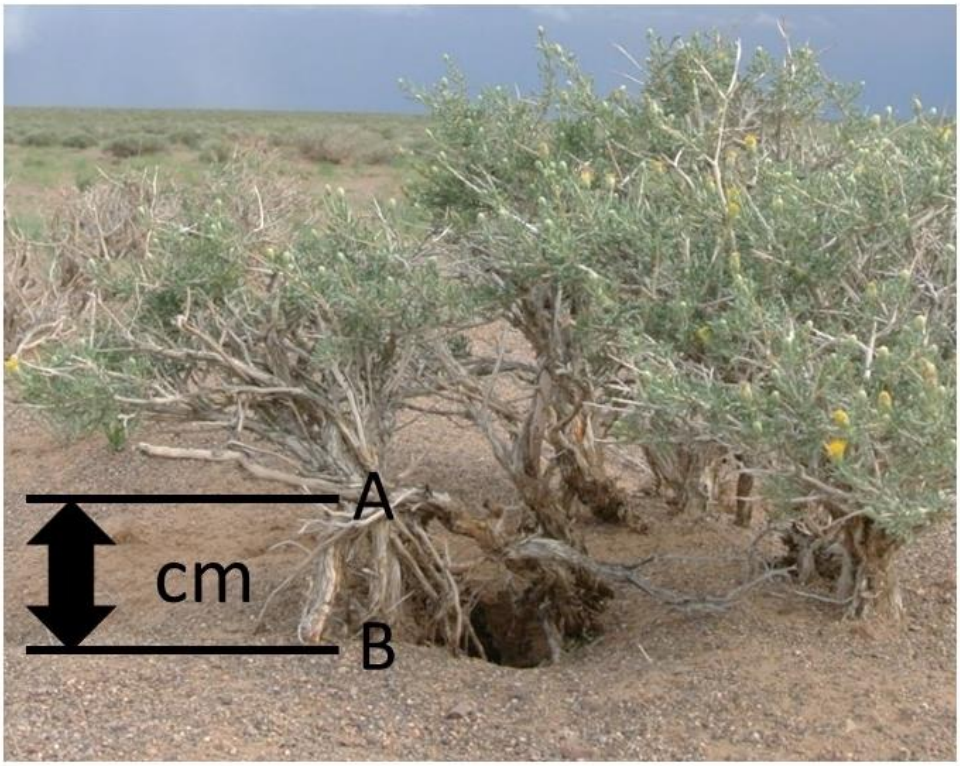


Figure 4. Measurement of roots exposed by soil loss.

The measurement is the vertical distance between the root-shoot junction visible on a plant (A) and the junction between the plant's root system and the soil level (B), in centimetres. The example shown uses *Brachanthemum gobicum* (Asteraceae).

2.3 Field sampling protocol for Elm Forest

2.3.1 Plot design

Plots should sample 900 m^2, consistent with the scale of sampling for the other ecosystems, however the square plot design (30 x 30 m), advocated by Avirmed (2018) and suitable for Saxaul, is not suited to Elm Forest monitoring because the Elm Forest ecosystem often occurs in bands narrower than 30 m. Instead, a method allowing variable plot shapes is recommended. All plots should be placed entirely within the geomorphic context occupied by the Elm Forest ecosystem (i.e. river beds capable of supporting Siberian Elm Trees), and must not include any different surrounding habitat.

- If a 30 x 30 m (900 m^2) square plot fits within the Elm Forest context at the plot location, the plot design recommended in Methods 2.4 should be used.
- If a 30 x 30 m square plot does not fit within the Elm Forest context at the plot location, a 15 x 60 m (900 m^2) rectangular plot should be used, aligned in any direction to fit within the band of Elm Forest.
- If a 15 x 60 m rectangular plot does not fit within the Elm Forest context, the plot location should be rejected, and another sampling location should be found.

2.3.2 Sampling plant species cover and litter cover

If a 30 x 30 m plot was used, plant and litter cover should be measured as described above for Saxaul. If a 15 x 60 m plot was used, 2 point-intercept lines 60 m in length should be established, each with 240 points. These lines should meet the short (15 m) ends of the plot at 5 and 10 m. In each case, overhead canopy of Siberian Elm that lines up with the point location must be counted as a 'touch'.

2.3.3 Sampling species richness

A single experienced botanist should spend 10 minutes within the 900 m^2 plot (regardless of its shape), recording all vascular plant species, regardless of their cover. Richness values for each life forms are calculated by simply counting the number of species in each lifeform.

2.3.4 2.3.4 Measuring the density of *Ulmus*

The following method should be competed separately for Adult, Juvenile (suppressed) and Juvenile (escaped) categories, as explained in Figure 5.

1. The number of *Ulmus* should be counted within the 900 m^2 plot (regardless of its shape).
2. If there are 5 or more *Ulmus* counted, no further counting is required, and the 'density' is simply the number counted.
3. If fewer than 5 *Ulmus* are found in the plot (which is frequently the case), then an ever-expanding radius around the plot centre must be searched until 5 adult *Ulmus* are found (This can be done on the GIS or in the field, see Figure 1). The density is then calculated as follows:

   $$\textbf{Density (per 900m}^2\textbf{)} = 5 / ((\pi R^2 / 900) / E)$$

   Where R represents the distance between the plot centre and the outermost of the 5 adult *Ulmus*, and E represents the area of Elm Forest ecosystem within the circle defined by radius R
4. If fewer than 5 adult *Ulmus* are found before R extends beyond 100 m, then the search for adult *Ulmus* should stop, and the density calculated as follows:

   $$\textbf{Density (per 900m}^2\textbf{)} = n / (34.9 / E)$$

   Where n represents the number of Adult Ulmus found within the 100 m radius.

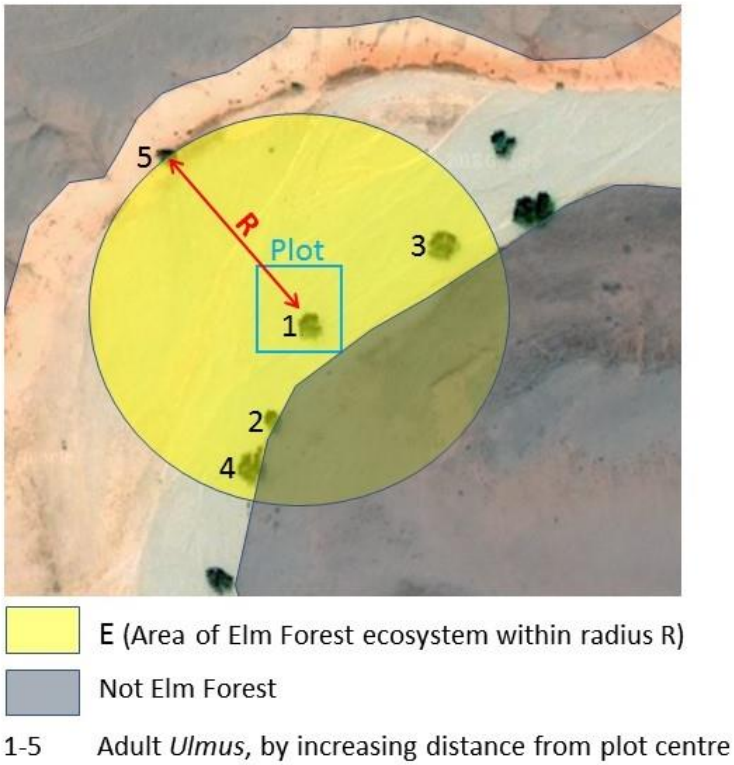


Figure 5. Diagram showing the method for calculating *Ulmus* density around the plot.

2.4 Gathering field validation data

To create a robust dataset for field testing of the metrics, we selected (for both Elm Forest and Saxaul) a set of nine field locations. These were subjectively selected to sample a broad range of the condition spectrum: we deliberately targeted the most intact sites we could find, along with sites that had been more or less degraded in various ways. The same subjective approach was used in Avirmed et al. (2018). The Saxaul sites were visited on the 13th - 18th August, 2019. The Elm sites were visited on the 15th and 16th September. 2019. Appendix 1 provides details of these sites.

At each site, we marked out a plot using red poles to mark the corners. We took the 6 stakeholders to each plot and asked them to inspect the site. We asked them to provide an evaluation of the condition of each site on a scale of 0 (all value lost) to 100 (the best imaginable site). We provided guidance on how to conceive condition, by explaining the condition concept described above (Introduction 1.2.2). We did not tell the stakeholders which variables to incorporate, nor how to integrate them into a score. All stakeholders provided their own scores without consultation with others.

Once the experts had provided their evaluation score, we collected field data from the plots, using the methods described above.

2.5 Gathering additional training data for Elm Forest

We sought additional data from the Elm Forest ecosystem. We did not have budget allowance to permit further workshops or field trips. We were, however, able to gather some assessment data from 360^{o} photographs of the Elm field sites. During each field assessment, we took a 360^{o} photograph centred at the middle of each plot, which showed the vegetation in sufficient detail to identify most species and see their cover on the ground (using a Samsung Gear 360 CMOS 8.4MP x2 / F2.2 lens (Default output pixel count equivalent to 15 MP)). An example of a portion of such a photo is shown in Figure 6. By email, we asked a large pool of experts to examine these photos and provide an evaluation of the site's condition. We provided the same guidance to the stakeholders on how to conceive and evaluate condition as we did for the actual field assessment.

Figure 7 shows that the evaluations of the nine field sites provided by one set of observers in the field correlates moderately strongly ($r^2 = 0.57$) with the evaluations of the same sites using the 360^{o} photographs, by a different set of stakeholders. This confirms that the photographs convey some useful information related to ecological condition. We have used these photograph assessments as additional training data for the models described below. We did not, however, use these data as test data, since the correlation was of moderate strength only. We assume that field inspection allows stakeholders to form a better view of the site, and we wanted to maintain the integrity of the test data.

Figure 6. An excerpt from one of the photographs used to gather evaluations.

The red post marks one corner of the field plot. This image has been cropped from the original image for display in this document.

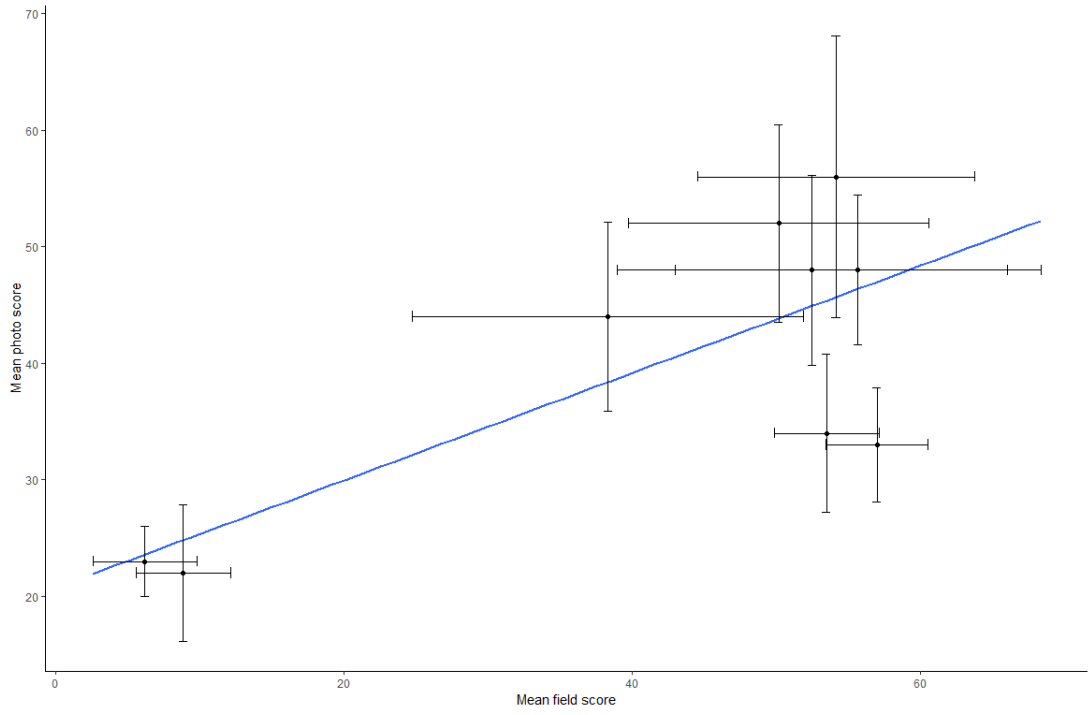


Figure 7. Relationship between Elm Forest evaluations made in the field and via photographs

2.6 The stakeholder group who provided validation data

The prior report of Avirmed et al. (2018) provided a detailed summary of the stakeholders who contributed evaluations to the original metrics (n=74 for Elm Forest, n=78 for Saxaul). The stakeholder group who contributed to field evaluations for these ecosystems is summarised in Table 2, using the same four categories of expertise established in Avirmed et al. (2018). Two of the seven stakeholders in Table 2 contributed to the previous work.

Table 2. The stakeholders who contributed to the Elm Forest and Saxaul field evaluations.

Primary area of expertise	Number of stakeholders		
	Male	Female	Total
Pastoralist	1	1	**2**
Specialist – Botany	2	-	**2**
Specialist - Wildlife	1	-	**1**
Conservation (Policy and practice)	1	1	**2**
Total	**5**	**2**	**7**

2.7 Testing the draft metrics

We subjected the existing Elm Forest and Saxaul metrics to similar performance tests to what we used for the Desert ecosystems (Avirmed et al. 2018). We used two related tests, both relying on graphical displays for their full interpretation.

1) We assessed the ability of the metric to predict the consensus score among a group of expert stakeholders for a set of field sites. The field sites were not used to train the metrics, and are a true 'hold out' set.
   We treated the metric as another stakeholder among the human evaluators. We plotted the position of each evaluator (human and metric) on an ordination in 'evaluation space', defined by the scores given to all sites. A successful metric would be expected to cluster with the group of human stakeholders, and close to the centre of this cluster.
   To provide context we also created and plotted 100 random (uninformed) evaluators. The meaningful evaluations would be expected to occur in a small subset of the possible evaluation space, in an area distinct and more compact than the random evaluation space.
2) We used the same data to demonstrate metric performance in a second way. We took a single stakeholder and regressed their site evaluations with the median site evaluation from all other evaluators (human and the metric). We did this for each stakeholder in turn, including the metric (which was compared to the humans). An evaluator that was perfectly aligned to the consensus view would be expected to plot a line intersecting (0,0) with a slope of 1.0 (i.e. 45^{o}).

2.8 Refining the metrics

2.8.1 Improving the training dataset

As described in more detail below (Results 3.1.1), we determined that the metrics would be improved by using a modified training set that was more 'focussed' on realistic sites. We did not have a budget allowance for further field work or workshop consultation, so we created such a training set by modifying the original dataset as follows:

1. We added field-based evaluations into the training set. These were not available in 2018 when the first metrics were made. By definition, field sites are real. As described below, we withheld some of the field sites from the training data for use in testing; but included all of the field sites in the training set for the final model.
2. For Elm forest, we added the new 360^{o} photo evaluations to the training set.
3. We culled the computer-generated sites to remove 'implausible' sites. We defined such sites for Elm Forest as those cards which had:
    - more than 20% of stakeholder flag the card as being 'implausible' in the workshop, or
    - total vegetation cover > 60%, or
    - *Ulmus* cover >25%, or
    - *Ulmus* adult density >20, or
    - grass richness >7 species, or
    - forb richness >15.

    We defined implausible sites for the Saxaul ecosystem in a comparable manner, as those sites which had:
    - more than 20% of stakeholder flag the card as being 'implausible' in the workshop, or
    - total vegetation cover > 60%, or
    - Saxaul cover >30%
    - Large Saxaul density >150, or
    - Grass richness >12 species.

    For both ecosystems, we retained the low calibration sites (which had no cover of any species) and the lowest-scoring of the three 'high quality' calibration sites, despite the fact that many stakeholders felt such a site to be implausible (See Avirmed et al. 2018 for an explanation of calibration).
4. We re-scaled the evaluation scores for the remaining computer-generated sites. This was done to remove the influence of the implausible sites in dampening the scores for realistic sites. We multiplied by a re-scaling factor that would make the mean of the retained workshop evaluations (minus the calibration cards) the same as the mean of the field evaluations.
    - For Elm Forest, we re-scaled by multiplying the original card set by 1.35.
    - For Saxaul, we re-scaled by multiplying by 1.39.

2.8.2 Improving the modelling approach

The basic approach used by Avirmed et al. (2018)- modelling the score using ensembles of regression trees- was considered appropriate to produce metrics in the given context (as discussed in Sinclair et al. 2015, 2018). We used regression trees again here but made some refinements to their use.

We wanted to ensure that each model in the ensemble was trained on sites that spanned a wide range of variation in the site attributes (and, consequently, also in the condition spectrum), and to simultaneously ensure that that models were 'focussed' on sites considered 'realistic'. To this end, we introduced strata into the training data according to vegetation characteristics, and then weighted these strata by requiring the models to use set numbers of evaluations from each of these strata. Table 3 defines these strata and shows how they were weighted.

Table 3. Summary of the strata and weights employed in the training dataset.

Ecosystem	Description of stratum	Number of sites	Number of evaluations	Number selected for each model in emsemble	Relative Weight
Elm	Field and 360º photo assessments	9	90	70	0.8
Elm	Low calibration card	1	79	5	0.06
Elm	No Elm Cover	3	19	15	0.8
Elm	Elm cover up to 5% AND total cover up to 25%	22	146	100	0.7
Elm	Elm cover >5% OR total cover >25%	38	251	75	0.3
Elm	High calibration card	1	73	5	0.06
Elm	**TOTAL**	**74**	**658**	**270**	**-**
Saxaul	Field assessments	9	62	50	0.8
Saxaul	Low calibration card	1	77	5	0.06
Saxaul	No Saxaul cover	15	100	30	0.3
Saxaul	Total cover up to 15%, and some Saxaul	42	254	180	0.7
Saxaul	Total cover >15%, and some Saxaul	46	313	80	0.25
Saxaul	High calibration card	1	77	5	0.06
Saxaul	**TOTAL**	**114**	**883**	**350**	**-**

2.8.3 Cross validation

By adding the field-based evaluation data to the training set we used the only data available that could serve as a hold-out test set. While this may improve the metrics, it prevents us from testing the final metrics against independent standards.

To test whether the refinement strategy above had positive effects on the metrics, we used a cross validation process. We used the strategy described above (Methods 2.81. 2.82), but held out 5 of the 9 field sites, chosen at random. We made an ensemble of 30 trees to predict the score. We made 10 such ensembles, each time holding out a different random selection of 5 field sites.

We then performed the following tests:

- The original metric (Avirmed et al. 2018) against each of the ten sets of test data (each with 5 hold-out sites)
- The new metric using the strategy described above (the ‘All-in” model, without any field sites held out) against each of the ten sets of test data, and
- The ten cross validation models, each against its own test set, so that no model was trained with the test data.

We examined the coefficients of determination (r^2) for all models. As explained in Avirmed et al. (2018) r^2 is a poor absolute representation of model performance in this context (because each X value has multiple true Y values in the evaluation set), but does enable a clear comparison between one model and another.

If the cross-validation models had higher r^2 values than the previous metric, this provides positive evidence that the strategy improved metric performance.

Cross validation was only used in testing. The final models presented for use are trained on all the data, without a hold-out portion.

2.8.4 Summary of strategy for refining the models

The strategy to refining the models, involving changes to the training data (Methods 2.8.1), stratification of the training data (2.8.2), and cross validation (2.8.3) is explained with the aid of the diagram in Figure 8.

The diagram shows three columns representing the different data strategies:

- The 2018 strategy, employing only the unweighted training set, entirely composed of computer-generated sites, that spanned from plausible to implausible,
- The 10-fold cross-validation step, showing that the computer-generated data has been culled, and the remaining data weighted by strata; and showing that the field data have been used in multiple permutations for training and cross validation testing,
- The final metrics, showing that the computer-generated data has been culled, and the remaining data weighted by strata; and all the field data used for training.

We acknowledge that the culling, re-scaling, stratification and weightings described above are all arbitrary. These strategies were developed on a trial and error basis. Despite this subjective approach, the impact of these decisions was examined quantitatively (see Results 3.2.3).

We note that no further re-scaling of the metrics is applied to the Elm Forest and Saxaul metrics (cf. Avirmed et al. 2018), given we have already rescaled the metrics to match field evaluations.

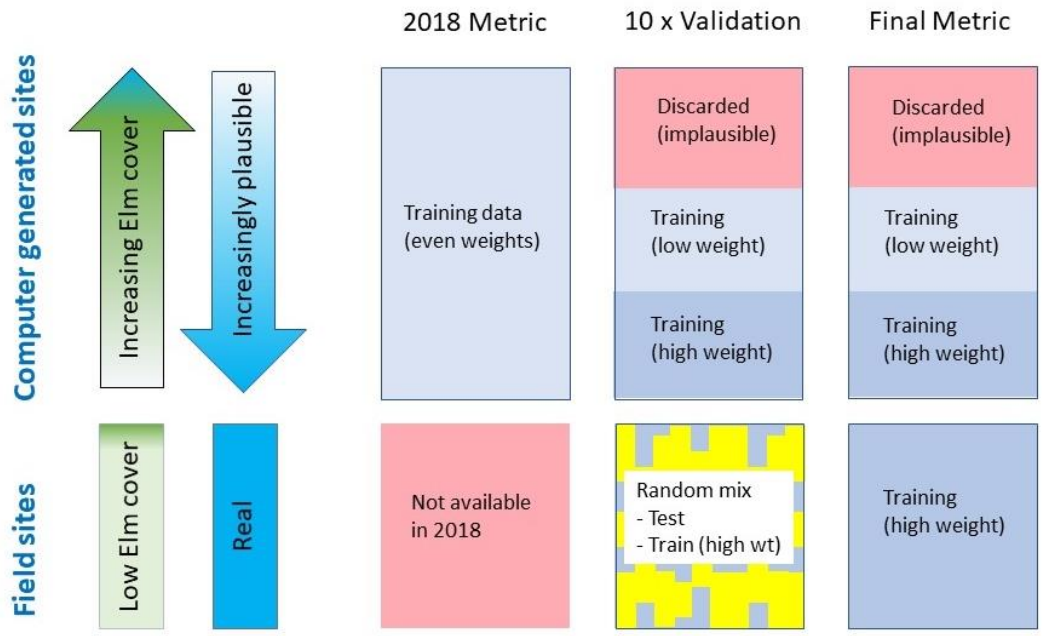


Figure 8. Training data strategy in this report, vs the strategy in 2018.

The Elm Forest system is used here as an example.

2.9 Scaling the model predictions

The Regression Tree model ensembles are expected to predict across a contracted score range when applied to real data (i.e. not spanning 0 to 100) (Sinclair et al. 2017; Avirmed et al. 2018).

The contracted score range may be perceived as problematic, if it does not match the expectations of stakeholders. This can be rectified by rescaling (stretching) the predictions. This can be done without changing the relative scores or the rank order of the sites (Sinclair et al. 2015).

We rescaled the predictions as follows:

- We found the “highest” and “lowest” scores able to be predicted by the ensemble.

- The lowest was found by calculating the score for a site that had no vegetation cover for any species, zero species richness, and 'height of roots exposed by erosion' set to 50 cm.
- The highest was found by calculating scores for sites that maximised the scores for each variable, with reference to the plots presented in Results 3.4. The values used to define the highest score are recorded in Appendix D.
- We applied the following formula:

Re-scaled prediction = (raw prediction – lowest) / ((highest – lowest)/100)

Table 5 shows the highest and lowest scores used for each ecosystem, and the re-scaling function derived from these scores. It is notable that the Saxaul metric required almost no rescaling, given the raw predictions encompass the range 0-99.

Table 5. The parameters and functions used to re-scale the predictions

Ecosystem	Lowest score returned by raw ensemble	Highest score returned by raw ensemble	Function used to re-scale raw ensemble median predictions
Elm Forest	0	81	=Raw / 0.81
Saxaul	0	99	=Raw / 0.99

2.10 Presentation of the metrics

For ease of use, the metrics that result from the models described above are presented in Microsoft Excel format. The user needs only to enter or copy the values of the measured sites into the appropriate cells, and the spreadsheet calculates a score.

This is achieved by encoding the regression trees as "IF, THEN" statements, which refer to the input cells. Each tree is represented by a separate statement, and returns its own score prediction. The final score returned by the spreadsheet is the median of these 30 predictions.

3 Results

3.1 Testing the performance of the existing metrics

3.1.1 Tests against field evaluations

We assessed the ability of the metrics to predict the consensus score among a group of expert stakeholders for a set of field sites. The field sites were selected to sample a broad range of the condition spectrum, and were not used to create the metrics.

We treated each metrics as another evaluator among the human evaluators. We plotted the position of each evaluator (human and model) in evaluation space, defined by the scores given to all sites (Figure 9). To provide context, we also plotted 100 random (uninformed) evaluators. The meaningful evaluations would be expected to occur in a small subset of the possible evaluation space, in an area distinct and more compact than the random evaluation space. A successful metric would be expected to cluster with the group of human stakeholders, and close to the centre of this cluster (Sinclair et al. 2018).

The ordinations in Figure 9 show the test results for the Elm Forest and Saxaul metrics, as well as the Desert Steppe metric reported in Avirmed et al. (2018) to provide a comparison. Several observations are important:

- All metrics produce scores that are to some degree aligned with the consensus of the expert evaluations (i.e. In Figure 9, 'M' clusters with the numbered experts, reasonably close to the point marking the median),
- The metrics for Elm Forest and Saxaul are clearly less aligned to the consensus than the Desert Steppe metric (i.e. In Figure 9, the model (M) is generally on the outside of the cluster of evaluators, and further from the median than in Desert Steppe),
- The human evaluators cluster more tightly for Desert Steppe, than for Elm Forest or Saxaul. This indicates that the experts are less in agreement about these latter ecosystems. Their consensus is poorly defined. This may be due, by chance, to the different set of stakeholders in this study, compared to Avirmed et al. (2018). However, we believe it is very likely due to the inherent nature of the ecosystems, as anticipated (See Introduction 1.5.3). This result equates to less signal and more noise, meaning that the prospect of creating a consensus metric is materially more difficult for Elm Forest and Saxaul.

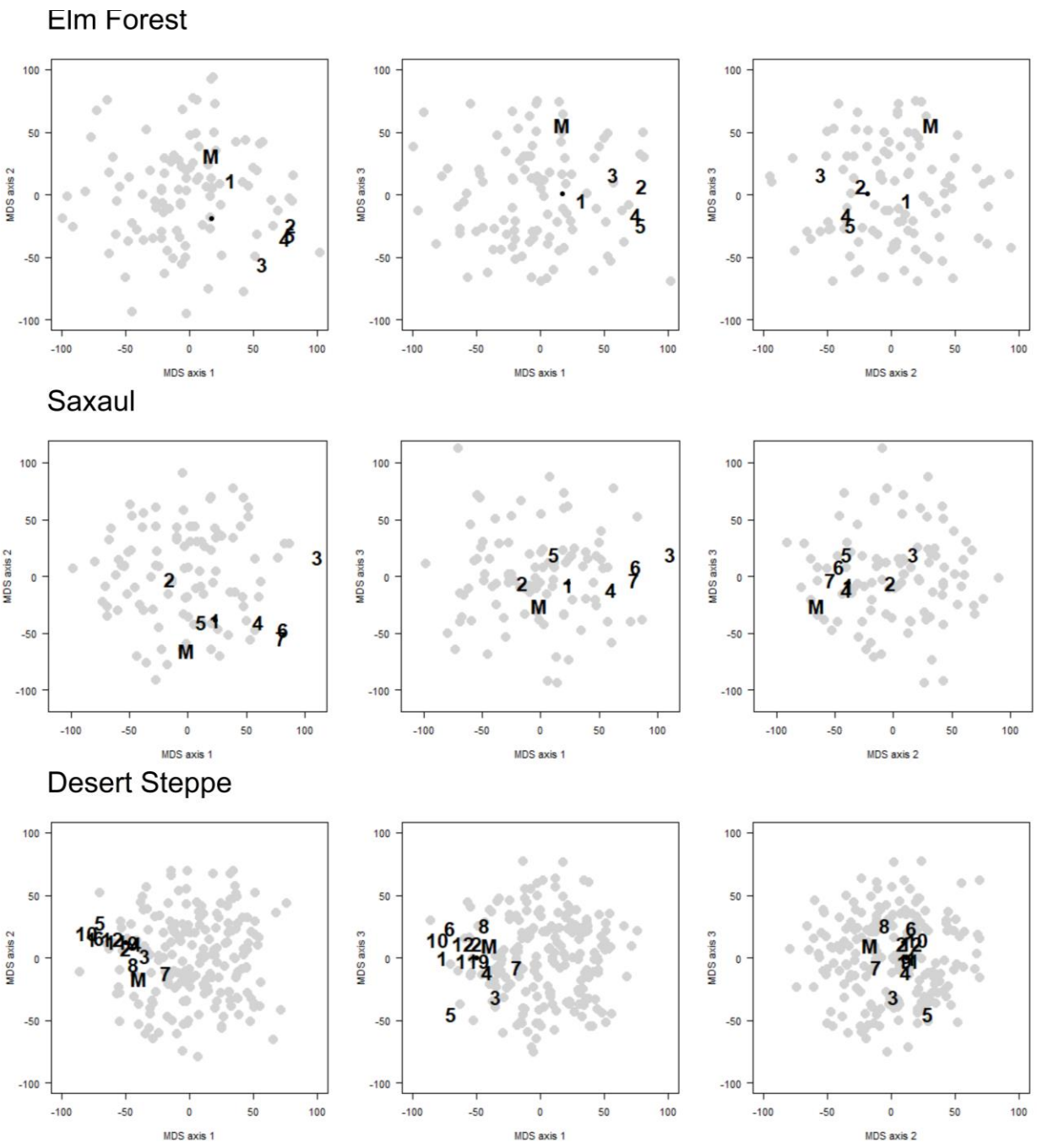


Figure 9. Evaluation of the 2018 metrics by comparison to stakeholders.

Multi-dimensional scaling (MDS) is used to determine whether the metrics reported in Avirmed et al. (2018) return scores within the evaluation space defined by real human evaluators. The graph space represents the possible evaluation space. The numbers are human stakeholders, with each person designated by a number. Note that the numbers are used consistently within an ecosystem, but do not represent the same people across different ecosystems. M is the metric for each ecosystem. The black dot is the median of the evaluators, considered the consensus evaluation, and the target for the metric. The grey points represent 100 dummy evaluators. The data presented here for Elm Forest do not include the outlying observer who was removed.

We used the same data to demonstrate metric performance in a second way (Figure 10). We took a single stakeholder, and regressed their site evaluations with the median site evaluation from all other evaluators (whether human or metric). We did this for each stakeholder in turn, including the metric (which was compared to the humans). An evaluator that was perfectly aligned to the consensus view would be expected to plot a line intersecting (0,0) with a slope of 1.0 (i.e. 45^{o}).

Figure 10 shows the relationships for each stakeholder (grey lines) and the metrics (black lines). Again, several observations are important:

- All metrics are positively correlated with the expert observers, indicating that they produce scores that, in general, correctly distinguish poorer sites from better sites, in line with the stakeholder consensus,
- The Elm Forest and Saxaul metrics produce lines with steeper slopes than the stakeholders (compare the black line to the grey lines in Figure 10). This indicates that the metrics tend to provide scores within a narrower range than the stakeholders, and
- The Elm and Saxaul metrics produce a line that does not meet the origin (0,0), but rather crosses at approximately 20 points on the stakeholder axis (compared to the Desert Steppe line, which passes close to the origin (0,0)). This probably indicates that the Elm Forest and Saxaul metrics lack resolution in the lower part of the score range. They tend to collapse to their lowest scores in situations where the stakeholders are still able to provide useful discrimination.

Taken together, these results suggest that the Elm Forest and Saxaul metrics do represent the stakeholder consensus view of condition, but more poorly than the Desert ecosystems tested in Avirmed et al. (2018). Given this, and in line with the project aims, it was deemed necessary to refine the metrics.

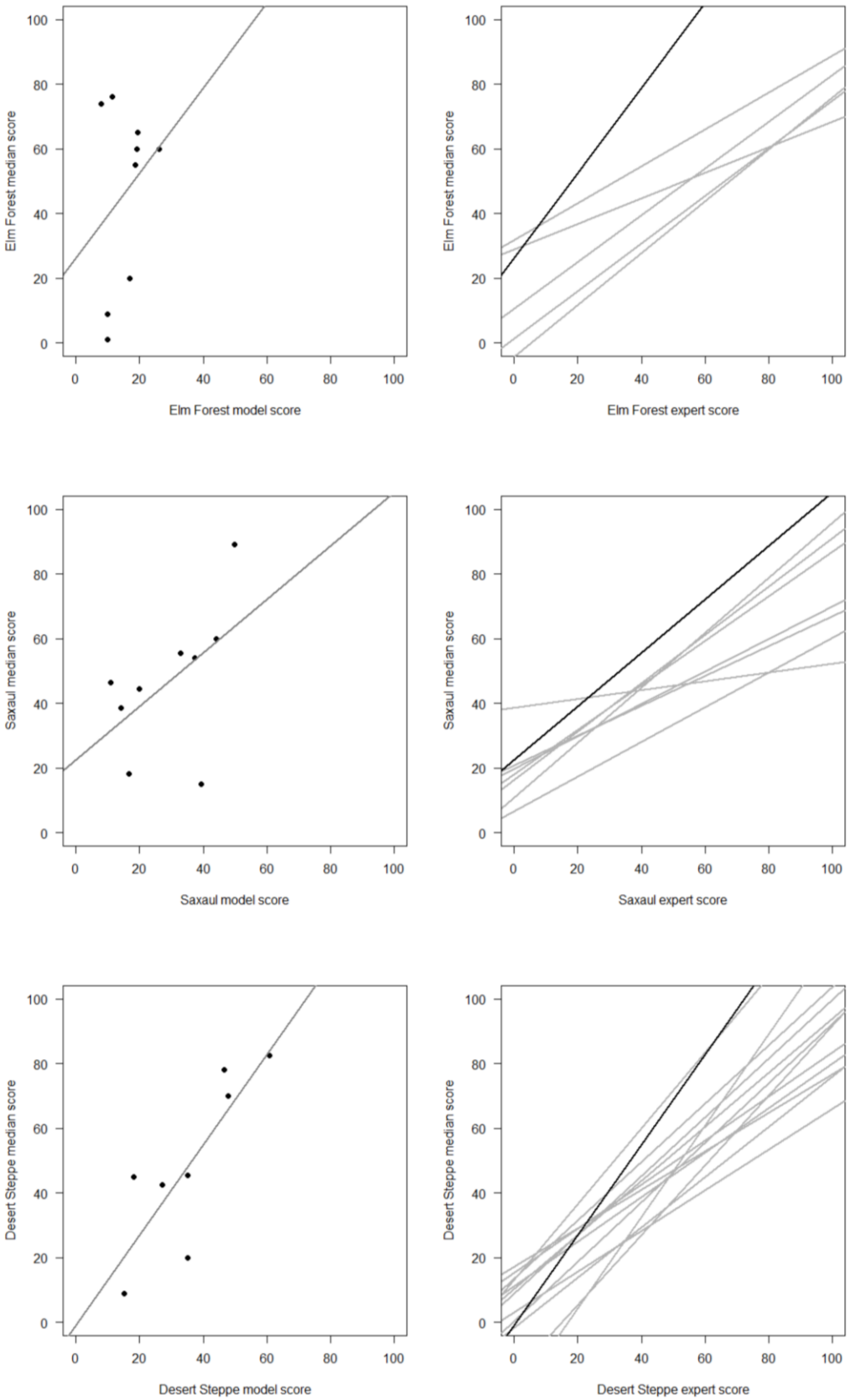


Figure 10. Relationships between metrics and human stakeholders.

Left panels: The horizontal axis shows the metric score for the field sites. The vertical axis shows the median score of the stakeholders.

Right panels: The black line represents the metrics compared to the median of the human stakeholders. Each of the grey lines represents a human stakeholder (each compared to the remainder of the human stakeholders and the model).

3.2 Creation and testing of refined metrics

3.2.1 Diagnosis of the problems with the existing metrics

We believe that the relatively poor performance of the Elm Forest and Saxaul metrics was largely caused by-

1) A lack of consensus among stakeholders for these ecosystems, as suggested in the Introduction (1.5.3), and
2) bias in the original training data. The computer-generated sites were skewed towards sites with unrealistically high cover and richness values. Real field sites resemble only a small range of the data used in training, at the lower end of the score range. This resulted in the models returning relatively low scores and having low resolution when dealing with the attributes of real sites.

We are able to address the second point. It appears that the computer-generated sets included far too many sites with implausibly high vegetation cover and species richness. This occurred due to the lack of field data and experience available to guide the creation of realistic computer-generated sites.

We believe that the stakeholders, when confronted with such sites, evaluated them as being exceptionally high condition, despite being unlikely to occur in the field, and scored them very highly. They then adjusted the scores on all other sites downwards to be consistent with these unrealistic sites. This caused a number of interrelated problems:

- The models were trained with data that was heavily focussed on sites with characteristics that are rare or impossible, such that many of the decisions in the regression tree models are 'wasted' on situations that never occur,
- The models received insufficient training on sites that resemble real field sites, causing them to discriminate poorly among real sites,
- Real field sites are scored poorly by the metrics and pushed to the lower portion of the score range, further reducing the discrimination of the metrics.

Figures 11 and 12 provide quantitative evidence to support these diagnoses. They show the field measured and computer-generated sites in ordination space defined by the site attributes. The field sites, which were deliberately selected to cover a wide range of variation in vegetation, occupy a small portion of the ordination space (the portion known to be realistic), while the computer-generated sites occupy a far larger portion of ordination space (which we suggest extends into implausible regions).

3.2.2 Creating improved training sets

In order to improve the models, new training sets were required which emphasised sites with realistic characteristics. As described in the methods, one key step was to cull the set of training sites, to focus on those with more realistic attributes. To do this, we used two criteria (as described in detail in Methods 2.8.1):

- The opinions of stakeholders, who were asked to rate whether each site was likely to be found in reality, or was implausible.
- The site attributes; removing sites with very high vegetation cover.

Figures 14 and 15 suggest that those criteria we used to cull the sites did indeed remove unrealistic sites, given that the discarded sites generally fell further from the real sites in ordination space (compare the large space occupied by the discarded sites (x), compared to the narrower overlapping space occupied by both the retained sites (O) and the real field sites (●)). By culling the dataset, we have demonstrably focussed the attention of the model on the types of sites we know are realistic.

We also improved the training sets by adding the newly-acquired field-based sites. This seems to be particularly important for the Saxaul ecosystem, where Figure 11 shows that our training set did not previously cover some of the actual characteristics of real Saxaul sites (i.e. there are 2 field sites which lie quite outside of the computer-generated sites in ordination space).

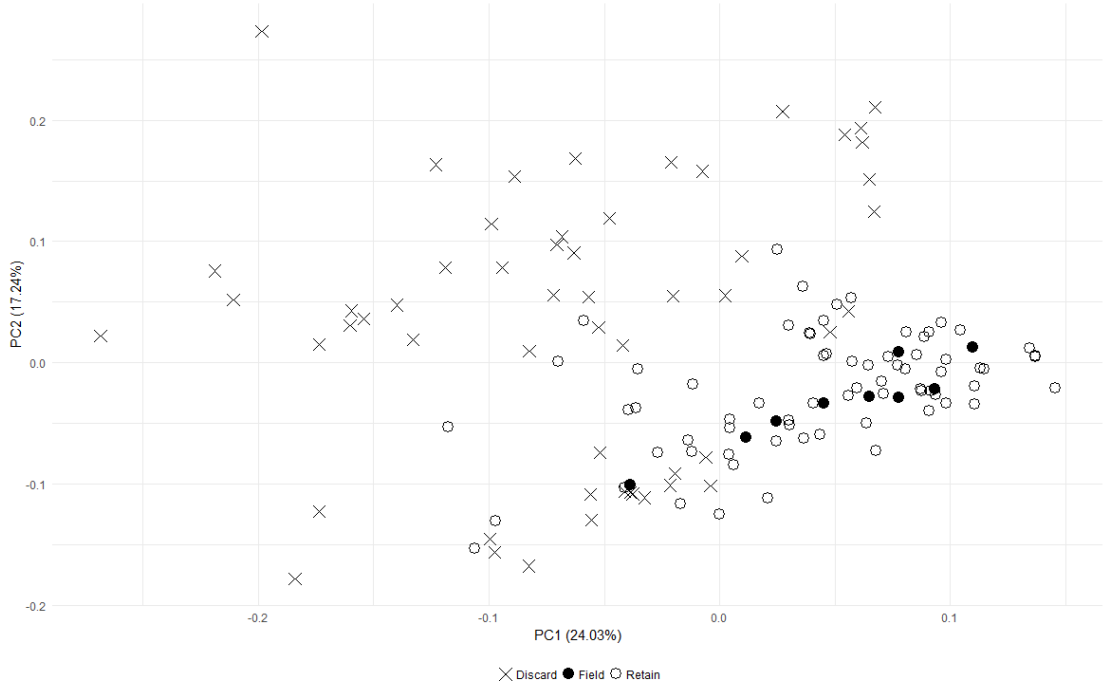


Figure 11. An ordination showing the training data and test data for the Elm Forest ecosystem

The ordination space is defined by the site variables. The sites are coloured by whether they are field sites, which define a portion of the ordination space known to be occupied by real sites in nature; computer-generated sites that we culled from the dataset (as described in Methods 2.8.1), or Computer-generated sites we retained on the basis that they were plausible.

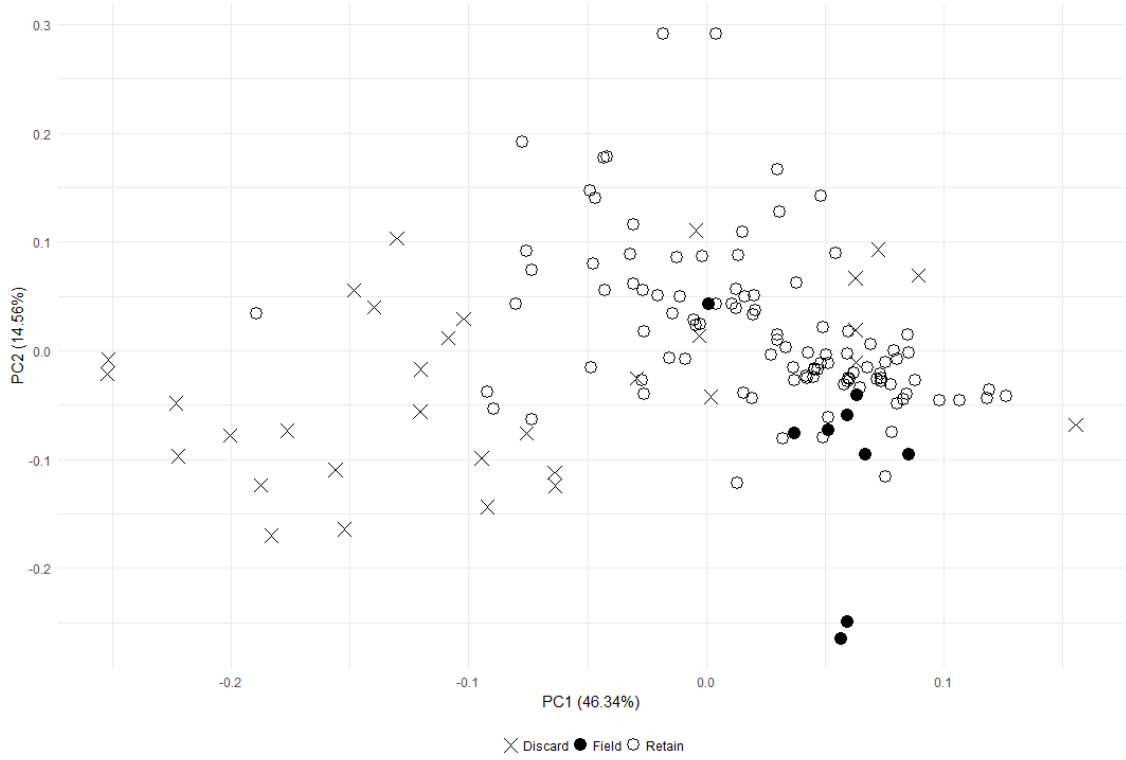


Figure 12. An ordination showing the training data and test data for the Saxaul ecosystem.

The ordination space is defined by the site variables. The sites are coloured by whether they are field sites, which define a portion of the ordination space known to be occupied by real sites in nature; computer-generated sites that we culled from the dataset (as described in Methods 2.8.1), or Computer-generated sites we retained on the basis that they were plausible.

3.2.3 Testing the performance of new models

We used the new training data and the approaches described in the methods to create new models. First, we tested these new models in exactly the same way as the old models. The results of these tests are shown in Figures 13 and 14, which recapitulate the tests described above, shown in Figures 9 and 10. The new Elm Forest and Saxaul metrics are clearly better able to predict the expert consensus in the test set.

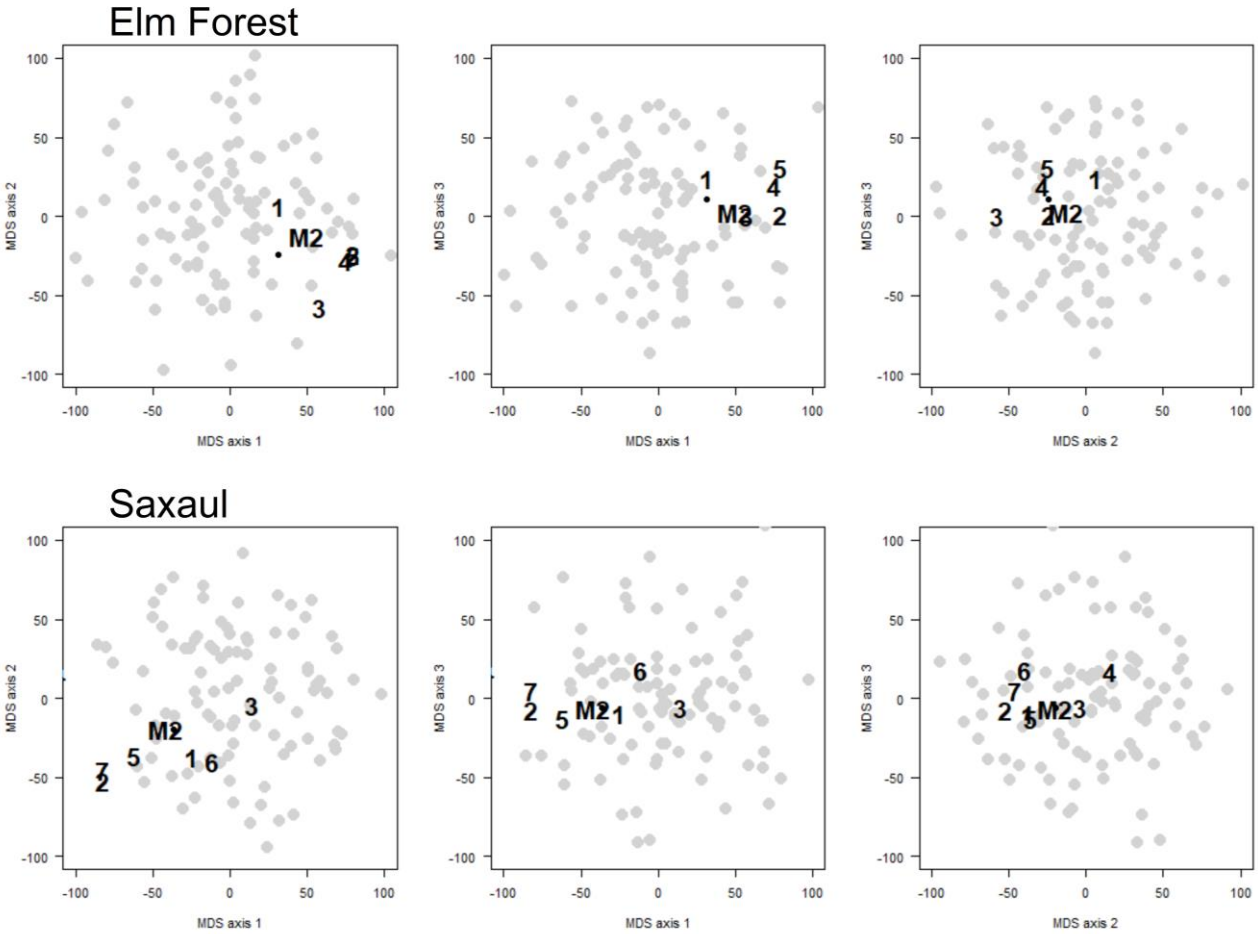


Figure 13. Evaluation of the NEW Elm Forest and Saxaul metrics by comparison to stakeholders.

This figure should be compared to Figure 9. Multi-dimensional scaling (MDS) is used to determine whether the metrics return scores within the evaluation space defined by real human evaluators. The graph space represents the possible evaluation space. The numbers are human stakeholders, with each person designated by a number. Note that the numbers are used consistently within an ecosystem, but do not represent the same people across different ecosystems. M2 represents the New Elm Forest and Saxaul metrics. The black dot is the median of the evaluators, considered the consensus evaluation, and the target for the metric. The grey points represent 100 dummy evaluators. The data presented here for Elm Forest do not include the outlying observer who was removed (See Appendix 1).

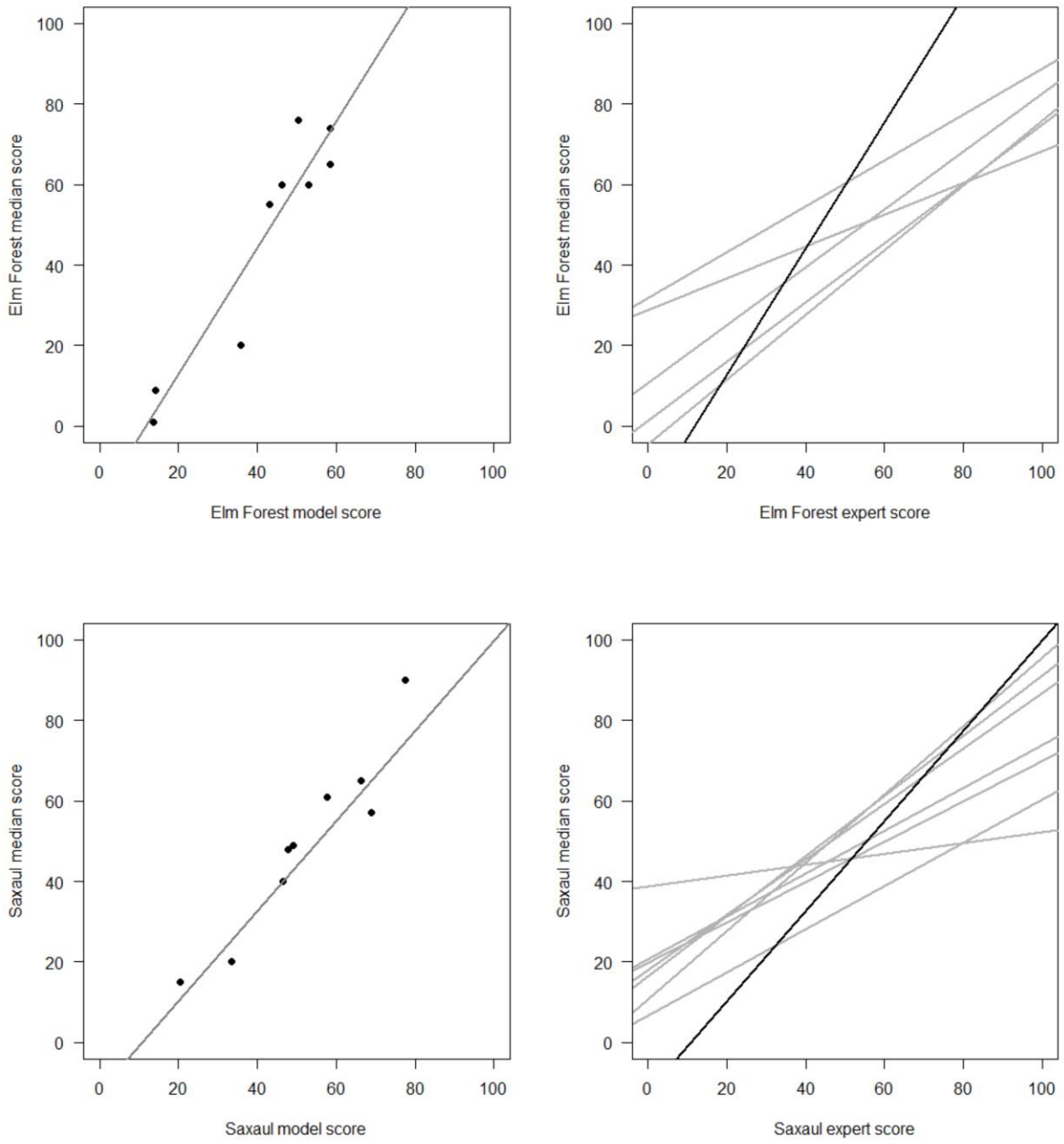


Figure 14. Relationships between the NEW metrics and human stakeholders.

This figure should be compared to Figure 10. Left panels: The horizontal axis shows the metric score for the field sites. The vertical axis shows the median score of the stakeholders. Right panels: The black line represents the metrics compared to the median of the human stakeholders. Each of the grey lines represents a human stakeholder (each compared to the remainder of the human stakeholders and the model).

The tests shown in Figures 13 and 14 are not as stringent as the initial tests, because the test data have been used to train the model. Because of this, there is a significant risk that the models are over-fit, such that their apparently excellent performance (Figures 13 and 14) is narrowly focussed on the particular field sites used for testing, and that this performance does not translate into performance that is transferrable to all sites. Overfitting is troubling because it gives a false impression of how capable a model is of extrapolation.

To test how much of the improved model performance seen in Figures 13 and 14 is due to over-fitting, we performed a cross-validation exercise. As described in the methods, we compared the coefficients of determination (r^2) for three model test scenarios:

- The original metric (Avirmed et al. 2018) against ten sub-sets of test data (each set being the median of the evaluations from a randomly selected set of 5 field sites),
- Ten cross validation models which used the improvement strategy described above (culling, stratification, introduction of field sites) against the same 10 sets of test data as above (where each cross-validation model was trained on a different set of field sites (n=4) to those used in the test (n=5)).
- The final models presented in Figures 13 and 14, against the same 10 test sets described above.

If the improvement in model performance seen in Figures 13 and 14 was entirely due to overfitting, we would expect to see the cross-validation models perform no better than the original metric. If, on the other hand, the improved performance was due to the data improvement strategy, we would expect to see the cross validation models (which employed this strategy) perform better than the original metric.

This test is, however, hampered by the very small field set available (n=9), in two ways:

- Having only 5 field sites in the test set means that variation in the field evaluation data (see Introduction 1.5.3 and Results 3.1.1) would be expected to cause wide variation in r^2 between the 10 trials, adding 'noise' to the r^2 comparison.
- Using only 4 field sites to train the cross-validation models gives them less chance to improve than the final models which could learn from all 9. This means that the improvement strategy being tested could not be fully implemented in the test.

The results shown in Figure 15 reveal different results for the two ecosystems.

- For the Elm Forest metric the improvement strategy (culling, rescaling and adding other field sites) seems to have improved model performance. This is shown by the slight increase in r^2 in the cross-validation exercise. Even so, the r^2 value for the final model (with all data used for training) was higher still, suggesting that the model is somewhat over-fit to the field data.
- In contrast, we found no evidence that the strategy improved the Saxaul metric, because there was no increase in r^2 over the old Saxaul model. In fact, the r^2 declined in this test. It is highly likely that the Saxaul model is substantially over-fit to the particular field data collected in 2019. The degree to which this is a problem is explored in the discussion.

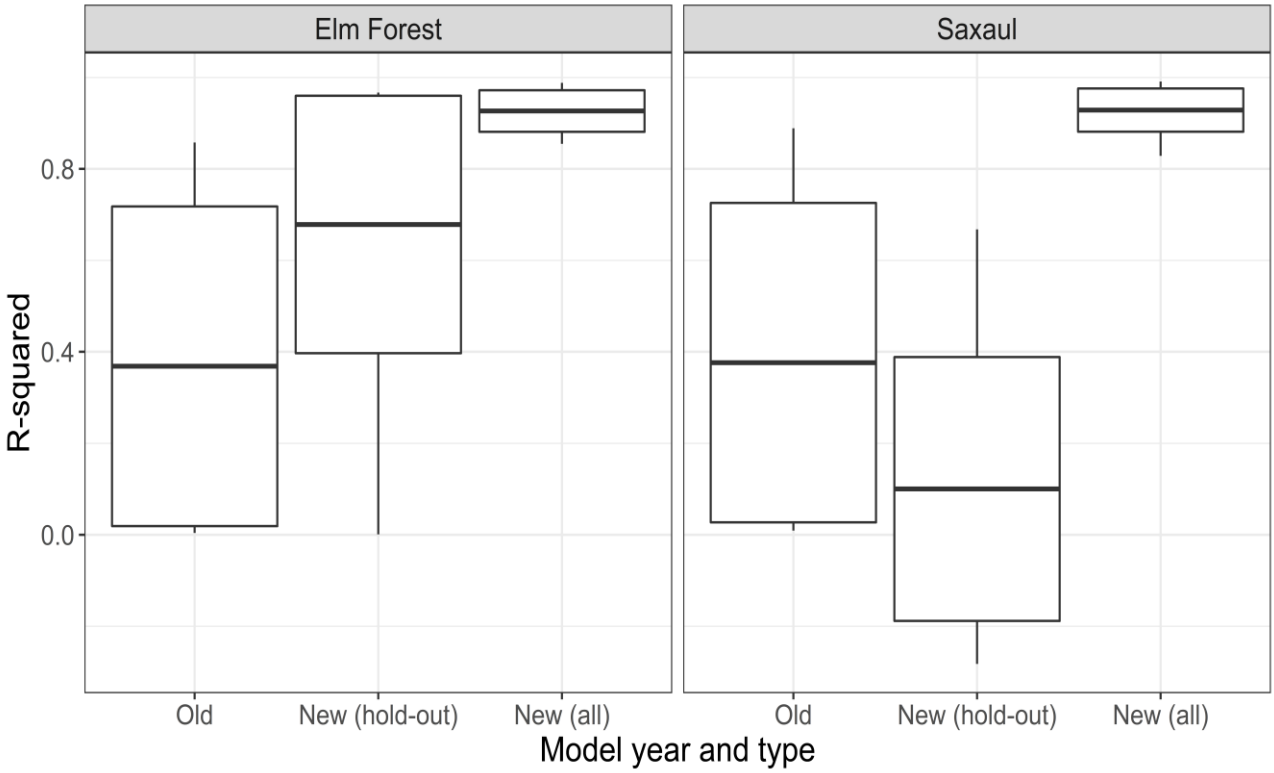


Figure 15. The coefficient of determination (r^2) for three modelling strategies.

Each strategy is compared to a set of 10 hold out test sets, each composed of randomly-selected sites evaluated in the field. The boxplots show the full range of variation in r^2 across the 10 tests (whiskers, along with the mean (horizontal line) and the 1st and 3rd quartiles.

4 Discussion

4.1 The final models are fit for purpose

The final metrics presented here show a positive relationship with the consensus view of the stakeholders, demonstrating that they are capable of measuring meaningful differences in ecological condition. Judged across all tests, they performed about as well as most of the human evaluators (Figures 9, 10, 13 and 14), in systems shown to be the subjects of only weak stakeholder consensus (Figure 9). A metric that performs about as well as human evaluators is useful, with distinct advantages over any human stakeholder: It provides consistent and repeatable results, is always freely available, is incorruptible, and transparent and defensible with regard to method.

We regard the metrics presented here as being fit for use in field monitoring programs.

4.2 Model refinement and over-fitting

Our attempts to improve the existing metrics had mixed success. While we succeeded with the Elm Forest metric (Figures 13, 14 and 15), the Saxaul Metric is more problematic, because we could not demonstrate that its apparent improvement (Figures 13 and 14) is due to anything beyond overfitting to the test dataset (Figure 15). Overfitting is generally troubling because it gives a false impression of how capable a model is of extrapolation. Despite this, we still chose to present the possibly-overfit model as the final model for use. We judged that the benefits of the re-modelling strategy are likely to outweigh any disadvantages from overfitting, for several reasons:

- There is no evidence that the new model using all data performs less well than the original model.
- The field test sites were specifically selected to represent a wide range of field conditions. If the model is overfit to these sites, it is at least overfit to a large and relevant portion of reality.
- The test set we used was very small (5 sites per test in the cross validation sets), meaning that r^2 values are volatile (see the spread of r^2 values in Figure 15). There may be improvements in the new model that we could not detect. This may be the case given the similarity in the diagnosis of problems for the two models, and the fact that the improvement strategy worked well for the Elm Forest metric.

4.3 Limitations to use and interpretation

No condition metric is perfectly capable of representing all aspects of ecological condition and degradation. There are several important aspects where the metrics presented here are limited, including the following two areas which are relevant to Elm Forest and Saxaul.

4.3.1 There may be cases where the variables fail to capture a relevant phenomenon

There may be phenomena that are not well captured by the variables but may be perceived as being relevant on site. One case may be the physiological health of individual plants at a site. The metrics presented here do not use this information, and the metric score does not change if plants are suffering from stress, or are defoliated, discoloured or wilted. This decision was made early on in the project, given how difficult it is to rapidly quantify plant physiological stress, and how this can fluctuate rapidly.

Despite this, physiological stress is sometimes conspicuous, and stakeholders in the field may give it significant weight in their field evaluations. This was highlighted by one Saxaul site (Site 144, see Appendix 1). Here, the density of large Saxaul plants was measured as being relatively high (22 in the 900m2 plot), but the large plants were defoliated (for an unknown reason). The stakeholders who inspected this site presumably took this into account and provided a relatively low score (median score 15). The original metric, without any information about defoliation, provided a far higher score (36). Although the final metric provided a score closer to the expert median (20; probably due in part to overfitting, as described above), it remains true that the metric is simply ‘unaware’ of the issue of physiological health and plant decline. There may be other such variables that the metric does not consider.

Given the performance of the metrics in the tests, this problem does not hamper model performance in general; however it may return results that vary greatly from stakeholder opinions at a small minority of sites.

4.3.2 The score range may be miscalibrated

The score range from 0 to 100 is intended to cover the full range of condition states that an ecosystem experiences, so that any process of degradation and recovery can be quantified. In our approach, the upper score range is set by the stakeholders' implicit consensus that some sites are sufficiently desirable (or intact) that they deserve a score of 100. An issue arises if the stakeholders are not familiar with how the system appears when it is intact, or if the concept of 'desirableness or intactness' is unclear. This may be an issue for the Elm Forest ecosystem.

In historic times, Elm Forests have existed as scattered clumps or individual Elm trees, with Siberian Elm cover rarely exceeding ~10% over significant areas. The system is apparently perceived as a savannah woodland where tree cover and recruitment is kept inherently low by ecological processes. It is clear that stakeholders value sites with higher Elm cover, rewarding higher scores to sites with higher cover and density of Elm trees. Stakeholders familiar with the system generally seem to perceive that sites deserve scores in excess of 50 even when Elm cover is lower than ~5%, and in excess of 85 for sites where Elm cover reaches 25%.

In the last couple of years, evidence has emerged from exclusion plots that when all browsing animals are excluded, Siberian Elms grow and recruit far more densely rapidly than previously expected (WCS, anecdotal information). This suggests that the Elm Forest Ecosystem may be easily capable of supporting Elms at covers far higher than appreciated by most stakeholders; possibly approaching an extensive closed canopy (well in excess of 25% cover). Whether such dense stands have ever actually existed before is not known.

This raises difficult questions about the idea of what is desirable, and how to conceive of an intact Elm Forest: Is the state of the Elm Forest ecosystem in the absence of all browsers a 'desirable' state? Can the desirable state be one that may never have existed before, but may exist in future? Is such a state relevant to management and restoration? Has the presence of browsers kept this system in universally low condition for millennia, or is exclosure a situation that is so unusual that it is mere hypothetical curiosity? These questions are very difficult to answer in a landscape that has been exposed to grazing and browsing livestock for millennia.

These issues have implications for the metric's score range: Under the revised metric, sites with 25% Elm cover may achieve scores exceeding 85. This means that if such a site was managed to encourage Elm growth (e.g. by grazing exclosure), there is little room for improvement in the score range, and many gains in Elm canopy will not translate into score gains (the score will simply hit a ceiling at or near 100). This may suggest that the metric should be re-calibrated, to allow such sites to increase their scores. On the other hand, if the metric was stretched in this way, it would mean that most site in the current landscape would receive a low score, and that the ability of the metric to discriminate between them would be hampered. Indeed, this is the very problem that we set out to solve in this report (Results 3.2.1).

There is no easy or correct answer to these questions. Here, we have taken the view that the metrics must perform a) to discriminate between real sites, as they appear today, b) in line with stakeholder expectations. This is a conscious decision, made in full awareness of the difficulties in setting a score range.

We acknowledge that some Elm Forest sites under grazing exclusion may move "off the scale" in terms of both metric score and stakeholder expectations. If this occurs, it will be necessary to collectively re-learn how this system functions and adjust the metrics in accordance with revised stakeholder understanding. Provided raw data are retained from all assessments, this is an exercise that can be achieved with full transparency, and the ability to back-cast scores onto any previously assessed sites.

4.4 Conclusion

All of the issues and limitations discussed here underscore the complexity of using quantitative techniques in the realms of subjective value judgement and imperfect understanding. Despite these issues, and the complex ecological processes which occur in all the ecosystems we have investigated, it remains clear that there is a signal amongst the noise: there is a collective agreement among stakeholders that ecosystems are degraded or improved by specific changes in form and composition, and these shifts can be reliably quantified.

5 Recommendations

Based on the work presented here, and an understanding of the project context, we make the following recommendations.

5.1 Application of the metrics

- The new metrics for Elm Forest and Saxaul be used for monitoring in the Gobi Desert. Such monitoring may include comparisons between sites, between years and between ecosystems.
- All metrics be mounted on a secure platform (e.g. a web-based application) where the metric structure cannot be corrupted, and the metrics are easily accessible to the relevant stakeholders.
- All raw point intercept and species richness data be retained securely for all sites. This allows other research and monitoring projects to use the data.
- All field workers who implement field plots in future years be asked to make a subjective assessment of the sites' condition (using the approach described in Methods 2.4). These assessments should be recorded. They will permit ongoing comparisons to be made between the metrics and the expectations of stakeholders.

5.2 Implications for future work

- Future projects of this kind would be improved if field data was collected at the commencement of the project and used to guide the creation of the computer-generated sites. This would avoid the situation where too many of the sites used to train the model are unrealistic.
- Future projects would benefit from the collection of a larger number of field validation sites. The fluctuation in r^2 shown in Figure 15 suggests that the training data used here were barely adequate.

6 References

Addison, J., Friedel, M., Brown, C., Davies, J., and Waldron, S. (2012). A critical review of degradation assumptions applied to Mongolia's Gobi Desert. *The Rangeland Journal* **34,** 125-137.

Avirmed, O., White, M.D., Batpurev, K., Griffioen, P.A., Liu, C., Jambal, S., Sime H, Olson, K. and Sinclair, S.J. (2018). Rangeland condition metrics for the Gobi Desert, derived from stakeholder evaluations. Arthur Rylan Institute Technical Report 289. Arthur Rylah Institute and Wildlife Conservation Society Mongolia Country Program, report for Oyu Tolgoi.

Bedunah, D.J. and Schmidt, S.M. (2000). Rangelands of Gobi Gurvan Saikhan National Conservation Park, Mongolia. *Rangelands* **22,** 18-24.

Blockeel, H., Dzeroski, S. and Grbovic, J. (1999). Simultaneous prediction of multiple chemical parameters of river water quality with TILDE. Pages 32–40 in: Zytkow, J.W. and Rauch, J. (Eds) 'Principles of data mining and knowledge discovery'. Springer, Berlin.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and regression trees. Chapman & Hall, New York.

Buckland, S.T., Magurran, A.E., Green, R.E. and Fewster, R.M. (2005). Monitoring change in biodiversity through composite indices. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **360,** 243–254.

Daniel, T.C. and Vining, J. (1983) Methodological issues in the assessment of landscape quality. Pages 39-84 in: Altman *et al*. (Eds) 'Behavior and the natural environment'. Springer, New York.

Gibbons, P. and Freudenberger, D. (2006). An overview of methods used to assess vegetation condition at the scale of the site. *Ecological Management & Restoration* **7,** S10-S17.

Fa-min, L., Yan-qing, W., Jian-ping, S. and Ming-wu, D. (2003). Effects of water stress on *Haloxylon ammodendron* seedlings in the desert region of Heihe inland river watershed, Gansu Province, China. *Journal of Forestry Research* **14,** 197-201.

Fernández-Giménez, M. E. (1999). Sustaining the steppes: a geographical history of pastoral land use in Mongolia. Geographical Review, 89(3), 315-342.

Fernández-Giménez, M.E. and Allen-Diaz, B. (2000). The role of Mongolian nomadic pastoralists' ecological knowledge in rangeland management. *Ecological Applications* **10,** 1318–1326.

Fernández-Giménez, M.E., and Allen-Diaz, B. (2001). Vegetation change along gradients from water sources in three grazed Mongolian ecosystems. *Plant Ecology* **157,** 101-118.

Godínez-Alvarez, H., Herrick, J. E., Mattocks, M., Toledo, D., and Van Zee, J. (2009). Comparison of three vegetation monitoring methods: their relative utility for ecological assessment and monitoring. *Ecological indicators* **9,** 1001-1008.

Jamiyansharav, K., Fernández-Giménez, M.E., Angerer, J.P., Yadamsuren, B. and Dash, Z. (2018). Plant community change in three Mongolian steppe ecosystems 1994–2013: applications to state-and-transition models. *Ecosphere* **9,** 1-26, Article e02145.

Jamsranjav, C., Reid, R.S., Fernández-Giménez, M.E., Tsevlee, A., Yadamsuren, B. and Heiner, M. (2018). Applying a dryland degradation framework for rangelands: the case of Mongolia. *Ecological Applications* [in press] 2018; DOI: 10.1002/eap.1684.

Keith, D. and Gorrod, E. (2006) The meanings of vegetation condition. *Ecological Management & Restoration* **7:** S7-S9.

Lkhagva, A., Boldgiv, B. Goulden, C. Yadamsuren, O. and Lauenroth, W. (2013). Effects of grazing on plant community structure and aboveground net primary production of semiarid boreal steppe of northern Mongolia. *Grassland Science* **59,** 135-145.

Maclean, G. (1996). Avian Adaptations to Deserts of the Northern and Southern Hemispheres: a Comparison. Curtin University of Technology, School of Environmental Biology, Bulletin No. 17.

Narantsetseg, A., Kang, S., and Ko, D. (2015). Distance-to-well effects on plant community based on palatability and grazing tolerance in the desert-steppe of Mongolia. In Proceedings of Building Resilience of Mongolian Rangelands: A Trans-disciplinary Research Conference, June 9-10, 2015. Colorado State University. Libraries.

Oliver, I., Smith, P.L., Lunt, I. and Parkes, D. (2002). Pre-1750 vegetation, naturalness and vegetation condition: What are the implications for biodiversity conservation? *Ecological Management and Restoration* **3:** 176-178.

Oliver, I., Jones, H. and Schmoldt, D.L. (2007). Expert panel assessment of attributes for natural variability benchmarks for biodiversity. *Austral Ecology* **32:** 453-475.

Owen, L. A., Richards, B., Rhodes, E. J., Cunningham, W. D., Windley, B. F., Badamgaray, J. and Dorjnamjaa, D. (1998). Relic permafrost structures in the Gobi of Mongolia: age and significance. *Journal of Quaternary Science* **13,** 539-547.

O.T. (2012) Comprehensive Environmental and Social Impact Assessment. Oyu Tolgoi, Chinggis Avenue 15, Sukhbaatar District, Ulaanbaatar, Mongolia

Parkes, D. and Lyon, P. (2006). Towards a national approach to vegetation condition assessment that meets government investors' needs: A policy perspective. *Ecological Management and Restoration* **7,** S3-S5.

Rao, M.P., Davi, N.K., D'Arrigo, R.D., Skees, J., Nachin, B., Leland, C., Lyon, B., Wang, S-Y. and Byambasuren, O. (2015). Dzuds, droughts, and livestock mortality in Mongolia. *Environmental Research Letters* **10,** 0740012. doi:10.1088/1748-9326/10/7/074012

Shiping, W. and Yonghong, L. (1999). Degradation mechanism of typical grassland in Inner Mongolia. *Chinese Journal of Applied Ecology* **4**.

Sinclair, S. J., Griffioen, P., Duncan, D. H., Millett-Riley, J. E. and White, M. D. (2015). Quantifying ecosystem quality by modeling multi-attribute expert opinion. *Ecological Applications* **25,** 1463-1477.

Sinclair, S. J., Bruce, M. J., Griffioen, P., Dodd, A. and White, M. D. (2018). A condition metric for *Eucalyptus* woodland derived from expert evaluations. *Conservation Biology* **32,** 195-204.

Stumpp, M., Wesche, K., Retzer, V. and Miehe, G. (2005). Impact of grazing livestock and distance from water source on soil fertility in southern Mongolia. *Mountain Research and Development* **25,** 244-251.

Stoddard, J.L., Larsen, D.P., Hawkins, C.P., Johnson, R.K. and Norris, R.H. (2006). Setting expectations for the ecological condition of streams: the concept of reference condition. *Ecological Applications* **16,** 1267-1276.

Tuvshintogtokh, I. and Ariungerel, D. (2013). Degradation of Mongolian grassland vegetation under overgrazing by livestock and its recovery by protection from livestock grazing. In The Mongolian Ecosystem Network (pp. 115-130). Springer Japan.

Venables, A. and Boon, P.I. (2016). What environmental, social or economic factors identify high-value wetlands? Data-mining a wetlands database from south-eastern Australia. *Pacific Conservation Biology* **22,** 312-337.

Wesche, K., Walther, D., Von Wehrden, H. and Hensen, I. (2011). Trees in the desert: Reproduction and genetic structure of fragmented *Ulmus pumila* forests in Mongolian drylands. *Flora* **206,** 91-99.

Wood, N. and Lavery, P. (2000). Monitoring seagrass ecosystem health – the role of perception in defining health and indicators. *Ecosystem Health* **6,** 134−148.

Xu, G., Yu, D., Xie, J., Tang, L. and Li, Y. (2014). What makes *Haloxylon persicum* grow on sand dunes while *H. ammodendron* grows on interdune lowlands: a proof from reciprocal transplant experiments. *Journal of Arid Land* **6,** 581-591.

Zou, T., Li, Y., Xu, H. and Xu, G.Q. (2010). Responses to precipitation treatment for *Haloxylon ammodendron* growing on contrasting textured soils. *Ecological Research*, **25,** 185-194.

Appendix 1. Site details

Site	Latitude	Longitude	WCS field site code
Elm Forest			
126	42.78704	108.199	EPO1
127	42.79238	108.1731	EPO2
128	42.7825	108.0991	EPO3
129	42.50076	107.5551	EPO4
130	42.599	107.5507	EPO5
131	42.56472	107.0341	EPO6
132	42.7027	106.9677	EPO7
133	42.91882	106.9359	EPO8
134	43.195	107.1189	EPO9
Saxaul			
137	43.28493	107.2704	HA1
138	43.31856	107.3743	HA2
139	43.40388	107.4283	HA3
140	43.40165	107.4285	HA4
141	43.35236	107.5897	HA5
142	43.11248	107.7279	HA7
143	42.68401	107.899	HA8
144	42.54868	107.1683	HA9
145	42.53989	106.9247	HA10

Appendix 2. Outlier removal

Avirmed et al. (2018) removed experts from the training set if they were shown to have provided evaluations that were contrary to the consensus. The consensus was defined by a preliminary model representing the pool of all observers. Such observers were removed because they cannot improve a model of the consensus opinion, only add noise to the modelling process. This pragmatic approach to outlier removal does not necessarily imply that these observers provided "poor" evaluations, only evaluations that were not helpful (They may hold legitimate minority opinions, or they may have misunderstood the task, or they may lack the requisite knowledge of the ecosystem). Observers were judged on an ecosystem-by-ecosystem basis, such that an individual could be removed from the dataset for one ecosystem but retained for another.

In the current exercise, where new evaluations were made in the field, we also checked for outliers. Here, we defined an outlier as a person who's set of evaluations were negatively correlated to the consensus view of the other stakeholders; revealed when this observer was plotted against the median of all remaining observers (i.e. they were judged against others on a comparable task). Again, we made this judgement ecosystem-by-ecosystem (assuming that a person may provide a useful set of evaluations for one system, and not for another).

We found only one observer in the Elm Forest evaluations with a set of evaluations that were negatively correlated to the consensus. This observer was removed from the evaluation dataset at the beginning, and no data or tests described in the body of the report includes their evaluations.

Figure A1 shows the same data as plotted in Figure 10 in the main report, with this observer included: they are the negatively-sloped line in the Elm Forest plot.

No other outlier removal was done in this project.

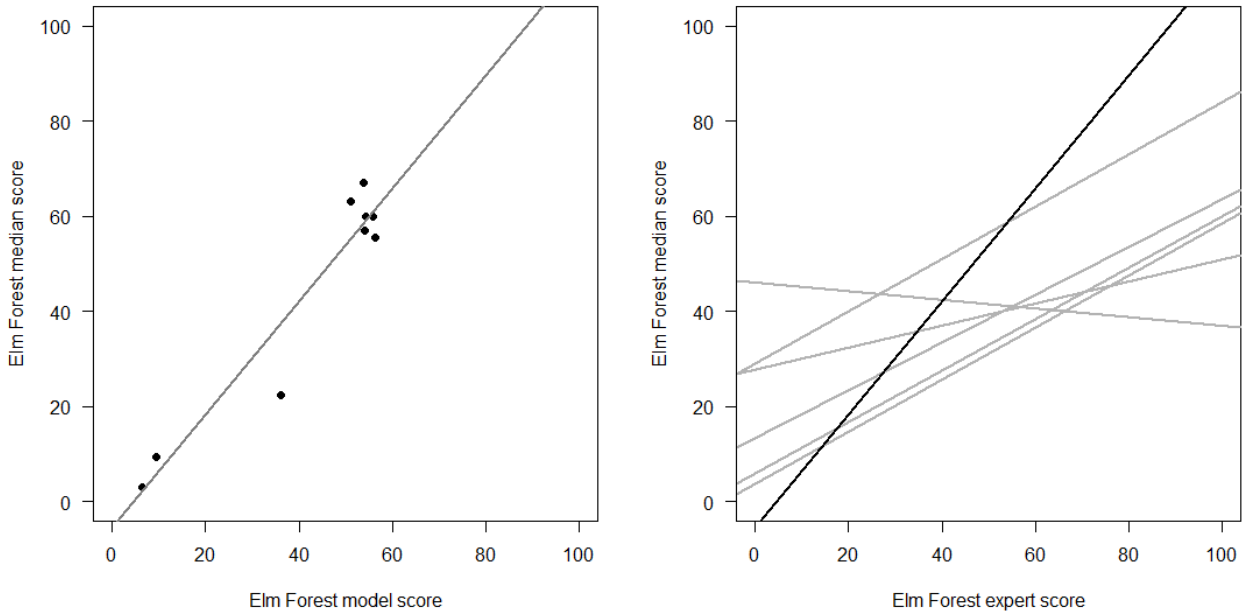


Figure A1. Metric performance assessed by correlation to the pool of other observers, with outlying stakeholder retained.

Appendix 3. Further information on the attributes of the sites

Figures A2 and A3 show the same ordinations as Figures 11 and 12, with vectors added to show the direction of change in the key site variables. It can be seen that the real field sites reside in the portion of the plots with generally low richness and cover.

Figure A2. An ordination showing the site attributes of the Elm Forest sites used in this project.

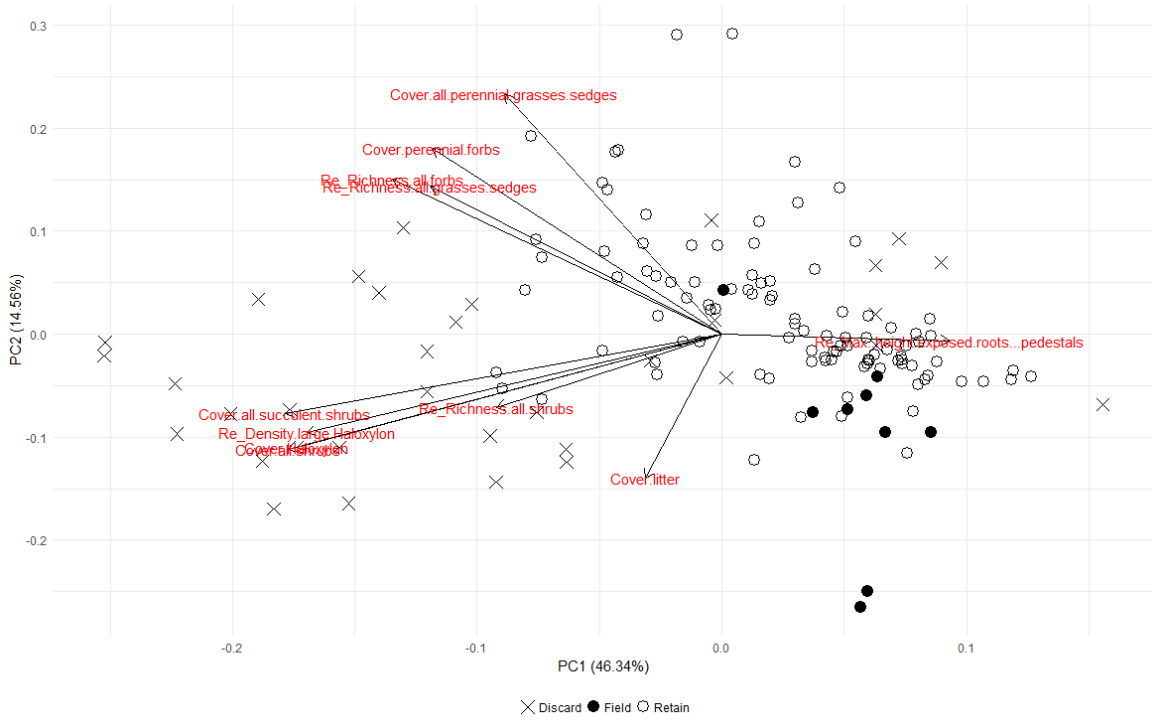


Figure A3. An ordination showing the site attributes of the Saxaul sites used in this project.

www.delwp.vic.gov.au

www.ari.vic.gov.au