Rangeland condition metrics for the Gobi Desert, derived from stakeholder evaluations

Otgonsuren Avirmed, Matt D. White, Khorloo Batpurev, Peter Griffioen, Canran Liu, Sergelenkhuu Jambal, Hayley Sime, Kirk Olson & Steve J. Sinclair

April 2018

Arthur Rylah Institute for environmental research

WCS

Оюу Толгой

SUSTAINABILITY EAST ASIA LLC

VICTORIA State Government | Environment, Land, Water and Planning

Photo credit
Desert vegetation, near Khanbogd, Mongolia (Steve Sinclair)

Disclaimer
This publication may be of assistance to you but the State of Victoria and its employees do not guarantee that the publication is without flaw of any kind or is wholly appropriate for your particular purposes and therefore disclaims all liability for any error, loss or other consequence which may arise from you relying on any information in this publication.

Accessibility

If you would like to receive this publication in an alternative format, please telephone the DELWP Customer Service Centre on 136186, email customer.service@delwp.vic.gov.au, or via the National Relay Service on 133 677 www.relayservice.com.au. This document is also available on the internet at www.delwp.vic.gov.au.

Rangeland condition metrics for the Gobi Desert, derived from stakeholder evaluations

ARI Technical Report 289

Otgonsuren Avirmed, Matt D. White, Khorloo Batpurev, Peter Griffioen, Canran Liu, Sergelenkhuu Jambal, Hayley Sime, Kirk Olson & Steve J. Sinclair.

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

Wildlife Conservation Society, Mongolia Country Program
Ulaanbaatar, Mongolia

April 2018

Authors

Otgonsuren Avirmed
Wildlife Conservation Society, Mongolia Country Program
POB 485, PO 38, Ulaanbaatar 15141, Mongolia
oavirmed@wcs.org

Matthew D. White
Arthur Rylah Institute for Environmental Research
123 Brown St Heidelberg, 3084, Victoria, Australia
Matt.white@delwp.vic.gov.au

Khorloo Batpurev
Arthur Rylah Institute for Environmental Research
123 Brown St Heidelberg, 3084, Victoria, Australia
khorloo.batpurev@delwp.vic.gov.au

Peter A. Griffioen
Arthur Rylah Institute for Environmental Research
123 Brown St Heidelberg, 3084, Victoria, Australia
Peter.Griffioen@delwp.vic.gov.au

Canran Liu
Arthur Rylah Institute for Environmental Research
123 Brown St Heidelberg, 3084, Victoria, Australia
Canran.Liu@delwp.vic.gov.au

Sergelenkhuu Jambal
Wildlife Conservation Society, Mongolia Country Program
POB 485, PO 38, Ulaanbaatar 15141, Mongolia
sjambal@wcs.org

Hayley Sime
LaTrobe University, Department of Ecology, Environment and Evolution
Bundoora, 3086, Victoria, Australia

Kirk Olson
Wildlife Conservation Society, Mongolia Country Program
POB 485, PO 38, Ulaanbaatar 15141, Mongolia
kolson@wcs.org

Steve J Sinclair
Arthur Rylah Institute for Environmental Research
123 Brown St Heidelberg, 3084, Victoria, Australia
Steve.sinclair@delwp.vic.gov.au

Contents

List of tables

List of figures

Summary

Context

Changes in the condition of rangelands in the Gobi Desert of Mongolia must be understood and quantified. This imperative is driven by land use intensification. There is now a need to account for expenditure on incentive schemes to manage grazing intensity, and to measure the impacts and mitigation measures associated with mining projects such as Oyu Tolgoi (OT), which has an offset program with a requirement to monitor rangeland condition. There is, however, currently no accepted means of quantifying rangeland condition. This project addressed the need for robust metrics to measure rangeland condition. The work was undertaken as a collaboration between Wildlife Conservation Society, Mongolia Country Program (WCS) and the Arthur Rylah Institute for Environmental Research (ARI). It was funded by Oyu Tolgoi.

Aims

We aimed to create condition metrics for five ecosystems in the Gobi Desert: True Desert, Desert Steppe, Semi Desert, Saxaul and Elm Forest. Such metrics must be able to distinguish sites of different condition, using only field-measured data for several (11-14) simple parameters (no stakeholder data is required for operation of the metrics).

Methods

We used a method adapted from previously-published work on Australian ecosystems (Sinclair *et al.* 2015, 2018). That method assumes that the concept of condition is inherently subjective, and thus the metric algorithm is derived from human opinion. 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. Ninety-four stakeholders contributed quantitative data that were used to derive the metrics. They represented four groups: nomadic pastoralists, specialists in botany, specialists in wildlife, and conservation practitioners and policymakers. They evaluated a set of hypothetical rangeland sites, providing each with a score between 100 (a desired state) and 0 (no values of the desired state retained). We used these evaluation data to train a model (an ensemble of 30 regression trees) capable of predicting the score based on the site variables. The models were converted directly into metrics for each system, and are presented as decision trees which can be implemented in a spreadsheet (as "if, then" statements).

Results

The primary test of the metrics' utility was to compare the metrics' scores derived from field-measured test sites with the median score assigned by a group of stakeholders for the same test sites. The test sites were not used to train the model. The metrics for three of five ecosystems were tested in this way (True Desert, Semi-Desert, Desert Steppe, the other systems remain untested). All tested metrics performed well (with r^2 values of 0.78, 0.82, 0.68 respectively). We also visualised the performance of the metrics using Multi Dimensional Scaling (MDS), where the model and each stakeholder was each represented as a point in a space defined by the evaluations. The metrics were positioned within the cluster defined by the stakeholders.

We confirmed that the selected variable sets adequately addressed stakeholder conceptions of condition, and that the field plot method adequately measured these variables, by demonstrating a close correlation between the scores provided by stakeholders in the field, and stakeholders assessing the same sites in a workshop context, where the sites were abstracted using only the selected variables to describe them. We performed this test for True Desert, Semi-Desert, Desert Steppe (with r^2 values of 0.81, 0.82, 0.53 respectively).

We also checked that each variable within the metrics related to the condition score in ways that would be generally expected in conservation biology, and that the relative importance of each variable in the regression tree models reflected their perceived importance in each ecosystem. We performed these checks for all five ecosystems.

Conclusions

We conclude that the data collection method and the metrics for deriving condition scores are robust and fit for purpose for the ecosystems we tested (True Desert, Semi-Desert, Desert Steppe). We suggest that the metrics for Saxaul and Elm Forest are likely to be useful, pending field testing for those metrics. We recommend that all metrics be mounted on a suitable web-based application, and be used for monitoring and reporting on rangeland condition between sites, over time and between ecosystems in the Gobi Desert.

1. Introduction

1.1 Background: The need to evaluate ecosystem condition in the Gobi

The Gobi Desert in central Asia is a vast arid region that lies in the rain-shadow of the Himalaya, in southern Mongolia and northern China. It experiences some of the most extreme weather conditions on the planet. Annual rainfall often falls short of 50 mm. Winter temperatures routinely drop below -30°C and daily maxima may not exceed -10°C for weeks. In summer, days above 30°C are common. Severe dust and sand storms can develop in the early afternoon as rapid increases in temperature result in powerful air movements.

Despite the extreme conditions, the Gobi has been populated for tens of thousands of years, and its human inhabitants have adapted to changes in climate over that period (Owen *et al.*, 1998). The ecosystems of the region have developed their characters under these shifting climatic and land use patterns, so that natural and anthropogenic influences on the environment are impossible to separate (Miehe *et al.* 2007).

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 then 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 being abolished, and livestock being 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 with hay and fodder 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, outside of accepted seasonal norms and chronic overgrazing leading to a perceived degradation in rangeland condition (Fernández-Giménez 1999). Livestock numbers, particularly goats, have more than doubled since the 1960s (Bedunah and Schmidt 2000, Tuvshintogtokh and Ariungerel 2013; Rao *et al.* 2015), although some authors have questioned the accuracy of the statistics, pointing out regional differences, and possible over- and under-reporting between different socio-political periods (Addison *et al.* 2012).

Increased grazing pressure is generally thought to cause ecological degradation and desertification, via the loss of palatable species, the increase of non-palatable species, the overall loss of vegetation cover, and soil erosion (Tserendash and Erdenebaatar 1993; Fernández-Giménez and Allen-Diaz 1999; Lkhagva *et al.* 2013), although the extent of degradation and its impacts are much-debated (Jamsranjav *et al.* 2018).

Concerns about environmental degradation have resulted in programs of research and monitoring of rangeland condition, supported by the Mongolian National Agency for Meteorology and Environmental Monitoring (NAMEM) and the Administration of Land Affairs, Geodesy and Cartography (ALAGaC), which maintain a nationwide spread of rangeland monitoring stations which collect a range of raw data.

Several international non-government organizations have supported herding communities to use rangelands sustainably. For example, the Swiss Agency for Development and Cooperation (SDC) has funded a project ("Green Gold") that engaged with ~53,000 herders to encourage their organization into Pasture User Groups (community based organizations of herders). This project encouraged new economic opportunities for herders, with the goal of enabling them to sustainably utilize rangeland resources. As a result of such programs, the resilience of herders to economic and climatic difficulties has apparently increased, although how this will affect the ecological condition of rangelands remains unclear.

Recently, mining has surpassed pastoralism as the major economic activity in the Gobi. In 2010, construction began on the Oyu Tolgoi (OT) mine project, which mainly extracts copper and gold bearing ores. 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 positive benefits over the lifetime of the mine, 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, these 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. To

date, there is no universally accepted conception of rangeland condition in the Gobi, nor any means to measure it. Different stakeholder groups perceive the environment differently, and have different priorities.

As described below in the project aims, this project seeks to address the need for tools to measure condition, and the need to link condition metrics to the views of stakeholder.

1.2 Evaluation of ecological condition

1.2.1 What is 'ecological condition'?

It is generally agreed that *ecological condition measures the retention (or loss) of the ecological attributes that characterise an ecosystem in its desired state.* Beyond this, however, there is much debate and controversy (Oliver *et al.* 2002, Buckland *et al.* 2005, Parkes & Lyon 2006, Gibbons & Freudenberger 2006, Stoddard *et al.* 2006, Sinclair *et al.* 2018). 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 or improvement?)
- Should condition measures allow direct comparison between ecosystems? (is there a 'common currency' that expresses condition in a steppe as well as a jungle?)

All these questions are controversial because they are ultimately subjective. Science or measurement cannot resolve them without interpretation by people. Ecological condition is unavoidably subjective (Daniel & Vining 1983, Keith & Gorrod 2006).

1.2.2 Subjectivity in ecological condition assessment

The subjectivity at the core of ecological condition assessment poses an apparent problem: Why is any evaluation credible, if it is merely an opinion or a value judgement? This problem is acute in cases where condition assessments are used for making decisions about land use, investments, project performance, or environmental regulation.

The issue of subjectivity is particularly complex in cases where there are multiple stakeholder groups with multiple viewpoints, and in cases where humans have interacted with the environment for so long that it is impossible to separate a pristine ecosystem from a managed ecosystem, and it may be desirable to maintain human land-use. Both situations apply in the Gobi Desert.

Despite these difficulties, 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 a standard set of measures which are combined to produce a score that reflects 'condition', in a way that conforms to stakeholders' views and the scientific literature (Parkes *et al.* 2003, Geneletti 2005, Sinclair *et al.* 2015, 2018). The resulting algorithm for assessment is commonly known as a "condition metric".

1.2.3 What does a condition metric do?

Condition cannot be measured directly (unlike length, weight, etc), because it is a composite, multi-variate concept (Schlacher *et al.* 2014; Sinclair *et al.* 2015, 2018, Venables and Boon 2016). An ecosystem condition metric is a formula for transforming multivariate information into a single number, reflecting the consensus opinion or values of stakeholders (Figure 1). It is thus an algorithm for reducing the dimensionality of data.
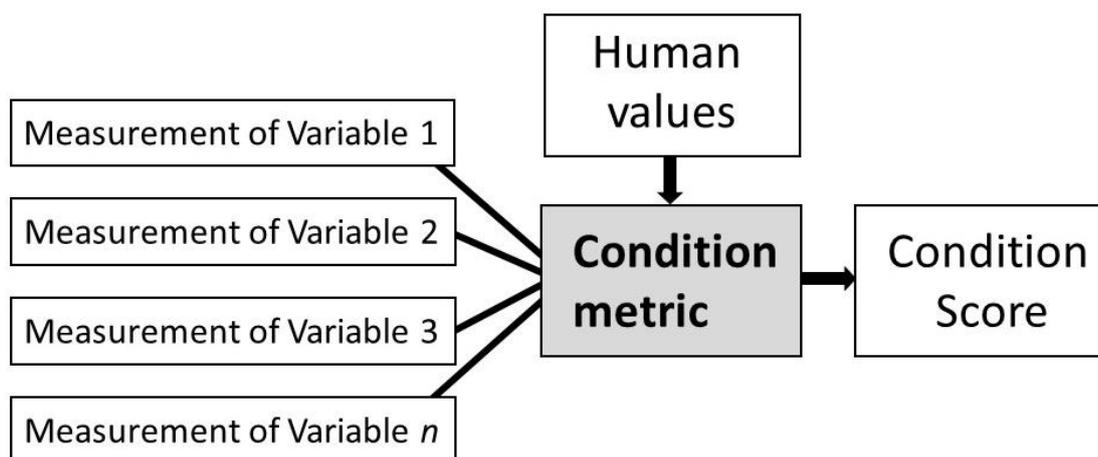
Figure 1. The basic function of a condition metric for a hypothetical ecosystem.

1.2.4 How can multiple variables be combined?

There are several ways in which a condition metric can transform multiple variables in to one. The most common approach is to assess them separately, and then add them together (or average them), sometimes with a weighting which emphasises some variables at the expense of others (Oliver *et al.* 2002; Parkes *et al.* 2003, Geneletti 2005, Reza *et al.* 2013, Schlacher *et al.* 2014). If the parameters are all the same type (e.g. the abundances of multiple related species), then they may be able to be combined with a good degree of mathematical rigour (Buckland *et al.* 2005).

Alternatives to the weighted addition of variables are available, but few scoring metrics use them. One is to take the highest value among the variables, and ignore the others (if the purpose of assessment is to seek outstanding attributes), or to take the lowest (if the purpose is to identify problems). Another is to use algorithms which combine the variables in ways that allow the value of one variable to influence the way in which another is used (Sinclair *et al.* 2015, 2018). This is a way of explicitly dealing with variable interactions.

Recent advances in machine learning have provided ways to derive such algorithms, including Regression Trees (Sinclair *et al.* 2015).

1.2.5 Distinguishing natural variation from changes in condition

All ecosystems vary over time. This is true from pristine sites to degraded sites. For example, Seasonal bursts of growth, flowering and decay cause changes in the abundances of plants. The movements of animals alter the faunal assemblage present at the site at any given moment, and may disturb the vegetation by consuming it or trampling it. Climatic variation between years means that each year differs from other years. Natural disturbances (e.g. flood, drought, heavy snowfall, sand movement) and seral changes in vegetation cause natural changes over years or decades. Each of these processes occur in the Gobi Desert.

Change presents problems for condition assessment. If measurements detect a change, how do we know whether the change represents normal fluctuation, or a meaningful change in condition? Put another way, if condition is assessed against a "desired state", how do we decide how much deviation from the desired state is normal, and how much represents degradation?

Variation between sites presents a similar problem. No two sites are identical. All sites vary from each other. Some of this variation is due to the inherent characteristics of the sites (e.g. some sites are sloping, others

are not). These inherent differences are not related to land use or degradation, and cannot be used to judge the condition of sites. However, other differences between sites may be related to ecological condition (e.g. one site has been bulldozed for a road, another has not). How do we decide how much difference is due to the inherent characteristics of sites, and how much represents differences of condition?

No solution to these problems exists, but several partial solutions may be employed:

- Condition assessments are confined to set times of year, to reduce temporal variation,
- "Desired states" are defined with sufficient tolerance to absorb natural spatial and temporal variation (however, increased tolerance may reduce the ability to resolve small scale or early-stage degradation),
- Multiple "desired states" are defined which represent different tolerable natural variants, and
- Monitoring is confined to the assessment of change over time at individual sites, but not designed to compare between sites (i.e. longitudinal studies).

1.3 Aims of the current work

The following specific aims guided the work presented here. They were formulated within the context described above, and the inherent limitations on the creation of condition metrics.

The work aimed to produce robust quality metrics for the target ecosystems that-

- can distinguish sites of different condition, including sites at the extreme ends of the condition spectrum,
- are based on data that is easily derived from field plots, which can be completed by any moderately skilled botanist within 1 hour, without follow-up laboratory analysis,
- can detect changes related to land-use change over multi-year periods,
- are not unduly influenced by natural and short-term fluctuations,
- are supported and justified by good data,
- are explicitly linked to the views of stakeholders,
- are tested on field data, and
- facilitate comparisons of condition both within and between ecosystems

The metric was 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 remote sensed data (although explicit links are made which will assist this in future).

1.4 Overview of the approach taken

The method used here was adapted from that published by Sinclair *et al.* (2015, 2018). The main components of the approach are described briefly here to orient the reader. More detail is provided in the body of the report.

1.4.1 Treatment of different ecosystems

We created a separate metric for each ecosystem. We considered making a combined, multi-ecosystem metric, but this approach was discarded based on preliminary work which showed little improvement in metric performance (not reported here).

1.4.2 Variable selection

We assumed up front that rangeland condition related to the vegetation and soil (not the animal community). This decision reflects that fact that vegetation and soil parameters are relatively easy to measure, respond directly to most degradation processes, and are relatively stable over the relevant time periods.

Variables were selected for each ecosystem based on stakeholder consultation. 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 (see Results 3.1).

1.4.3 Field plot design

A field plot was designed to measure the variables in each ecosystem. It is recommended that the plot is measured only in the season of peak growth (July - September), to reduce the influence of seasonal change on the monitoring data.

The efficacy of the plot was tested, by comparing stakeholder evaluations of real sites (without reference to the measured variable set) with evaluations of the same sites in a workshop context, where the sites were described from measurements taken using the field plot method (see Sections 3.2).

1.4.4 Stakeholder selection and description

The opinions given by stakeholders were used to create the metrics. Stakeholders were selected in consultation between OT and WCS. 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 stakeholders.

Stakeholders were grouped into four groups at the time of selection (Pastoralist, Specialist- Botany, Specialist- Wildlife, Conservation Practice and policy).

A self-assessed stakeholder questionnaire covering many different topics was then used to show the expertise that resided within these groups, and to show how discrete or mixed these groups were. It is essential that the stakeholder population is described, so that it is transparent which collective opinion is represented.

1.4.5 Metric creation

We sought a single metric for each ecosystem that spoke for the collective opinion of all stakeholders (i.e. we did not pursue multiple metrics representing different stakeholder segments). The opinions of stakeholders were explicitly used to create each metric.

Stakeholders were asked to evaluate and score a set of synthetic (i.e. fictional) sites, presented to them as site descriptions using relevant site variables. 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) 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 are summarised below:

- There is an explicit recognition in the method that the concept of ecological quality is subjective, and is derived from human preferences. A tool based on the evaluations of stakeholders can be said to directly represent or 'speak for' those stakeholders.

- The means of blending the multiple variables is driven by data, and is transparent and repeatable. Disagreement or criticism about the aggregation of the components in the metric could be settled by recourse to the data, or by the addition of new data.

- Regression trees can readily deal with multiple types of variables (categorical, binary, ordinal, continuous), and variables that interact (e.g. it is conceivable that the relationship between condition and forb cover may depend on shrub cover, if forb and shrub cover compensate each other with regard to important functions such as soil stability or cover for animals). They deal with these situations far more readily than weighted combinations (Kim and Park 2009).

- 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 (noted above).

All stakeholders' views were treated equally (i.e. they were unweighted), except for a small number of responses that were discarded because they were judged too aberrant from the shared opinion in an outlier detection process, or because the stakeholder did not follow the instructions and the responses could not be interpreted as required.

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.

1.4.6 Metric testing

The metrics were evaluated using the approaches published by Sinclair *et al*. (2018). Stakeholders were taken to a range of field sites, and asked to evaluate their condition. The sites were measured using the plot design, and the metric calculated a condition score for each site. We evaluated how well the metric performed in relation to the stakeholders, using regression and Multi-dimensional Scaling (MDS) approaches.

Due to the limitations of the field schedule, the metrics for three of five ecosystems were tested (Desert Steppe, True Desert, Semi-Desert). The other two (Saxaul, Elm Forest) remain untested.

1.5 The ecosystems covered by the metric

Five widespread ecosystems in the Gobi Desert are considered priorities for the development of condition metrics. These ecosystems have been delineated by previous studies (Jambal and Olson 2016). The definition or delineation of ecosystems was not part of the current project.

Three of the five of the ecosystems (Desert Steppe, Semi Desert and True Desert) are very broadly defined, being united by their basic physical structure (shrubs, grasses, etc.), but encompassing multiple plant communities and a wide range of landscape and soil types. Each of these systems may occur on sandy or stony soils, on valley floors, slopes and plateaux (Hilbig 1995, Radnaakhand 2016). In contrast, the other two ecosystems (Elm Forest and Saxaul) are more narrowly defined, characterised by a single dominant species in a relatively narrow landscape context. The five relevant ecosystems are described below, with an emphasis on their vegetation structure and composition.

1.5.1 Desert Steppe

Desert Steppe vegetation is dominated by perennial grasses and onions (Figure 2). It also supports a range of perennial forbs, shrubs and sub-shrubs. Annual grasses and forbs appear after rains. Desert Steppe occurs in a zone with annual average precipitation of 100-125 mm, and a growing season of 170-190 days, however rainfall may vary greatly between years. Within this climatic zone, Desert Steppe occurs across a range of geomorphic contexts, including sand plains, stony hills and valleys. In Mongolia, this ecosystem is generally found further north than True Desert or Semi Desert (Hilbig 1995, Radnaakhand 2016).

The most prominent grasses (Poaceae) are *Cleistogenes songorica, Stipa gobica*, *Stipa glareosa* and *Achnatherum splendens.* Common Onion (Alliaceae) species include *Allium polyrhizum* and *Allium mongolicum.* Prominent among the perennial shrubs and forbs are *Ajania achilleoides* (Asteraceae) and *Artemisia xerophytica* (Asteraceae). Annual species often include *Eragrostis minor* (Poaceae), *Aristida heymannii* (Poaceae) and *Bassia dasyphylla* (Chenopodiaceae), *Corispermum mongolicum* (Chenopodiaceae) and *Salsola collina* (Chenopodiaceae).

Desert Steppe is distinguished from the other Gobi ecosystems described here by the dominance of grasses and onions. It is distinguished from grassy steppes elsewhere in central Asia by the low rainfall, low biomass, drought-tolerant species and large inter-year variation in production and cover. In comparison to the other ecosystems noted below, Desert Steppe provides relatively reliable and nutritious fodder for livestock.
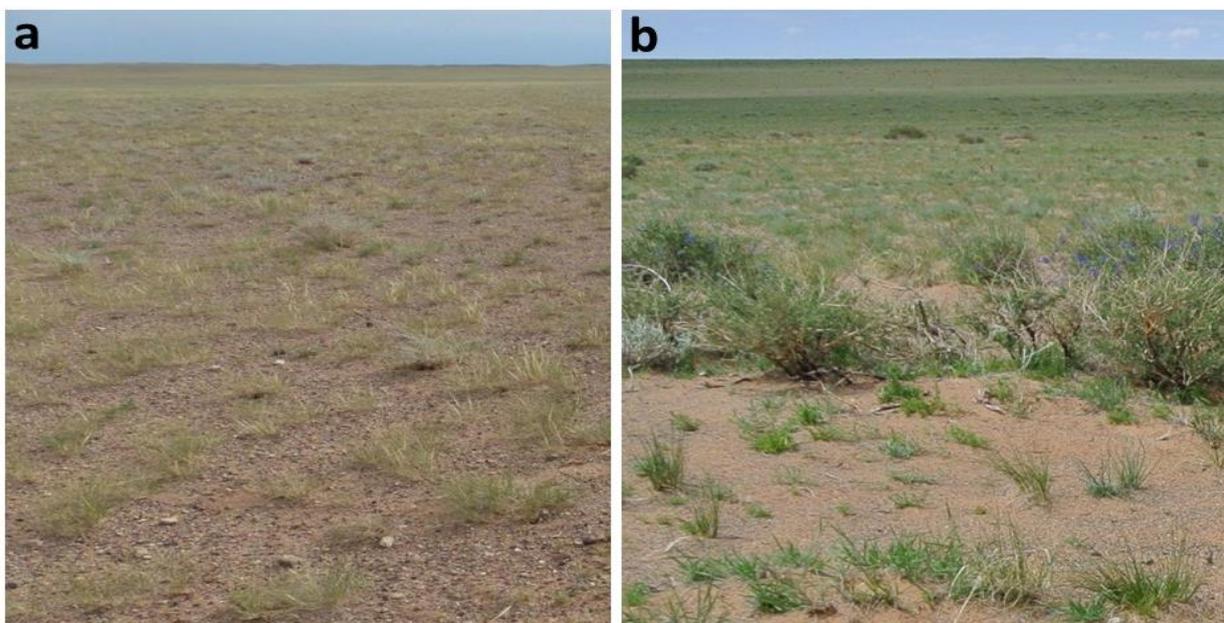
Figure 2. Two examples of Desert Steppe, showing the characteristic dominance by grasses.

Image a occurs on a stony plain, image b on a sandy slope.

1.5.2 True Desert

True Desert vegetation is dominated by low perennial shrubs which are tolerant of extreme drought (Figure 3). Grasses and forbs are usually sparse or absent. True Desert occurs in a zone with annual average precipitation of less than 100 mm. Rainfall varies greatly between years, with some years experiencing no precipitation at all (Hilbig 1995, Radnaakhand 2016). Within this low-rainfall zone, True Desert vegetation may be found across a range of geomorphic contexts, including sand plains, stony hills and valleys.

The shrub species vary from site to site depending on local conditions, but the most widespread and common are *Kalidium gracile* (Chenopodiaceae), *Nitraria sibirica* (Nitrariceae), *Reaumuria soongorica* (Tamaricaceae), *Salsola passerina* (Chenopodiaceae), and *Zygophyllum xantoxylon* (Zygophyllaceae). Saxaul (*Haloxylon ammodendron*, Chenopodiaceae) is often present, but places where Saxaul dominates to the exclusion of most other vegetation are defined as a separate ecosystem (see below).

Herders raise camels, goats and sheep in True Desert, but use of these areas is greatly limited by the lack of available water (Bedunah and Schmidt, 2000).
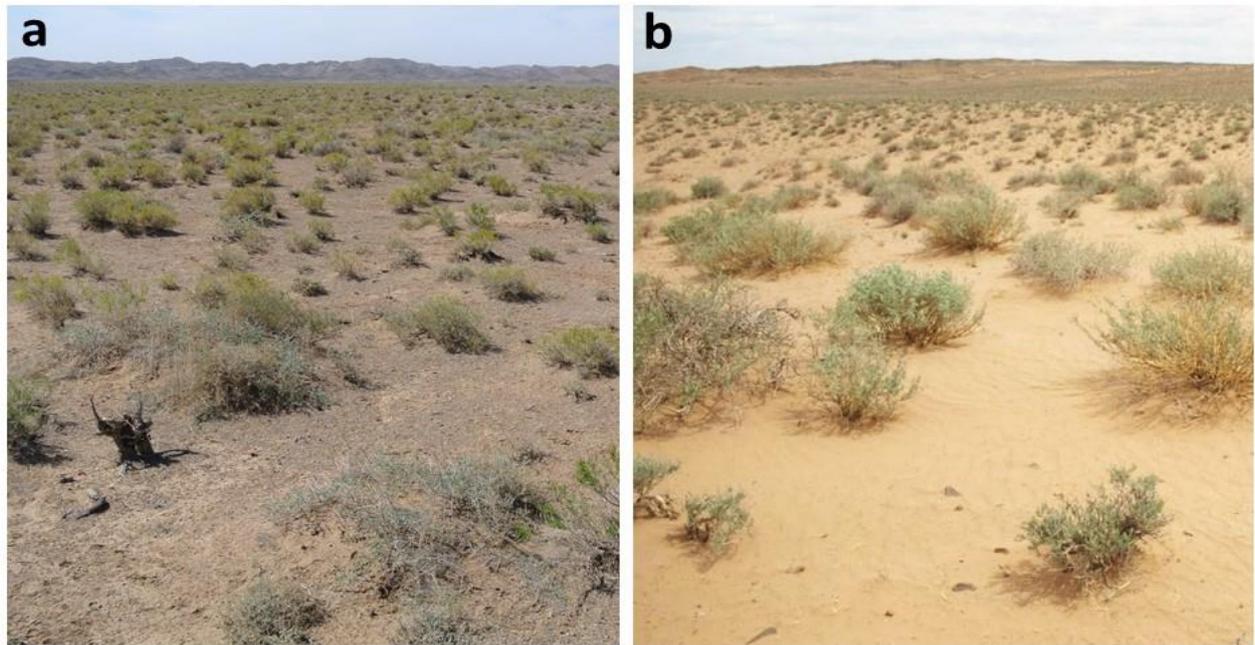
Figure 3. Two examples of True Desert, showing the characteristic dominance by low shrubs.

1.5.3 Semi Desert

Semi Desert vegetation is dominated by a mixture of grasses, shrubs and sub-shrubs (Figure 4). In this sense, it is midway between Desert Steppe (grassy) and True Desert (shrubby). Like those ecosystems, it may occur on a range of geomorphic contexts (Hilbig 1995).

The sub-shrub *Anabasis brevifolia* (Chenopodiaceae) is usually very abundant, and often strongly dominates the vegetation. The characteristic grasses and onions are the same as those noted above for Desert Steppe; although *Allium polyrhizum* is particularly prominent.

Productivity is relatively low in Semi-Desert, and drought is frequent, but this ecosystem supplies a significant amount of the forage for herders' livestock.

Figure 4. An example of Semi Desert, showing the mixture of grasses and low shrubs.

The dominant sub-shrub here is here *Anabasis brevifolia*.

1.5.4 Saxaul

The Saxaul ecosystem is defined by the dominance of a single species of shrub or small tree: Saxaul (*Haloxylon ammodendron*), which may grow to over 4 m in height (Figure 5). 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 central Asia tolerate these extreme conditions, so 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 with few other species (Zou *et al.* 2010), and because it provides important habitat for several wildlife species (Maclean 1996). Saxaul sometimes occurs in True Desert vegetation (above), but that ecosystem is distinguished by the high diversity and cover of other species.

Figure 5. An example of Saxaul, showing the dominance of Saxaul (*Haloxylon ammodendron*).

1.5.5 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 6) which form a patchy canopy (known locally as 'forest', although not meeting some global definitions of forest based on tree canopy). The ground-level vegetation is very sparse or almost absent, with occasional shrubs (e.g. *Nitraria sibicia* (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 sayr is former or potential Siberian Elm habitat, making the fine-scale delineation of this ecosystem difficult. For practical purposes, it is assumed that the occurrence (or definite evidence of past occurrence) of any Siberian Elms defines the Elm Forest ecosystem.

Siberian Elm trees provide an important 'drought-proof' food resource for camel herds.

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

1.5.6 Exclusions: places where the metrics will not apply

It is important to note that there are also places in the Gobi region where the metrics developed here are not intended to apply. These places include:

- Extremely rocky places (often ridgetops and peaks) where, even in a year and season of optimal rainfall, the total vegetation cover never exceeds 20%,

- Sand dunes,

- Granite outcrops,

- Rocky gorges,

- Narrow drainage lines vegetated by Almonds (*Amygdalus* species),

- Wetland ecosystems without Saxaul or Siberian Elms as their upper stratum (e.g. Saline lakes, Oases),

- Very heavily disturbed areas within 50 m of camps or wells.

2. Methods

2.1 Definition of condition

Our method for metric construction depended on stakeholders responding in a coherent manner to the evaluation exercises, with a shared concept of condition. Broadly, we defined condition as follows:

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

Each stakeholder was asked to bring their own personal idea of "desired state" to the exercise, within the following constraints.

Condition may include elements of "quality", "intactness", "health" or "conservation value". It may include consideration of any or all 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 were not to be included in the 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.

This conception of condition was explained to every stakeholder before they undertook the evaluation exercises.

2.2 Selection of variables to express and measure condition

2.2.1 Selection process

For each ecosystem, we sought the minimum set of measurable site variables that enable satisfactory evaluation of site condition. We attempted to select variables which-

- describe the main features of the vegetation of the ecosystem (i.e. dominant species and lifeforms),
- are likely to respond to the main expected pathways of degradation and recovery (e.g. grazing regimes, soil nitrification, soil disturbance),
- 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.

The following general considerations were taken into account when selecting the variables:

- The number of variables for each ecosystem cannot be too large, because stakeholders must be able to visualise and evaluate sites described using a list of the variables. The maximum appropriate number is not known in this context, but previous work has shown that 13 variables is tractable (Sinclair et al. 2018).

- Variables that describe the inherent characteristics of the site (i.e. which reveal only the type of site, rather than its condition) were excluded from the evaluation set. Such variables include ecosystem type, location (latitude, longitude), rainfall, soil type, etc.

- All variables were defined and scaled in relation to how they appear in August, when most site monitoring takes place in the Gobi region, to align the metric predictions with future data input.

- There are numerous plant species in the region; too many to include a variable for each species individually. To rationalise this richness, the plants were divided up into 'lifeform' groups. Lifeform groups are most informative when the species within them share similar structures, habitats, seasonal growth patterns, physiological tolerances, and responses to disturbance or management.

- Origin (native vs exotic) was not used to divide lifeforms because, for the Gobi region, there are very few exotic species, and no obvious binary distinction between natives and the few potentially invasive species (Radnaakhand 2016). This contrasts with the situation in the previous applications in Australia, where this variable was a strong driver of condition score (Sinclair et al. 2015, 2018).

Although the selection process was largely informal and qualitative, it was supported by a quantitative evaluation of the selected variables (see below Results 3.1).

The first stage involved an investigation of the English-language literature on ecological condition assessment and the ecology, degradation and recovery of ecosystems in the Gobi region, and other similar ecosystems in central Asia.

The second stage involved consultation with herders and scientists:

- The consultation with herders was facilitated by WCS, and involved unstructured conversations with two herder families. The herders were asked to "describe the features of the vegetation that indicate whether a site was in good condition or poor condition". Notes were taken in Mongolian, and translated into English by WCS.

- The consultation with scientists involved informal and unstructured conversations with ecologists in Mongolia and Australia.

The third stage was a structured survey of four WCS staff with extensive field experience in the Gobi region. This survey presented 22 possible variables, which were compiled after the first and second stages of consultation. The respondents were asked to vote for the top 10 and 15 most appropriate variables for each ecosystem, and provide comments on how the variables could be refined. These responses guided the final selection of variables. The variables that were considered on the WCS questionnaire, but ultimately not included for any ecosystem, were:

- Cover of *Cleistogenes* spp. (Poaceae)

- Cover of non-vascular plants

- Cover of Bare ground (soil, sand)

- Cover of exposed rock or pebbles

- Maximum depth of litter

- Density of holes created by animals

2.2.2 The selected variables

The final sets of variables for each ecosystem are shown in Table 1. The precise definitions of each variable are described below (Methods 2.2.3), along with the ecological rationale for the inclusion of each variable (Methods 2.2.4). Note that 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 do not present problems for the modelling approaches described below.

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

Variable	Desert Steppe	Semi Desert	True Desert	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 litter	✓	✓	✓		✓
Max height exposed roots pedestals	✓	✓	✓		✓
Cover annual forbs	✓	✓	✓		
Cover all annual grasses	✓	✓	✓		
Cover all succulent shrubs	✓	✓	✓		
Cover *Fabaceous* shrubs	✓	✓	✓		
Cover *Artemisia* species	✓	✓	✓		
Cover *Ulmus pumila*				✓	
Density adult *Ulmus pumila*				✓	
Density juvenile *Ulmus pumila* (suppressed)				✓	
Density juvenile *Ulmus pumila* (escaped)				✓	
Cover *Haloxylon*					✓
Density large *Haloxylon*					✓
Total number of variables	**14**	**14**	**14**	**11**	**11**

2.2.3 Definitions of terms used to express the variables

The terms used to define the site variables are explained below. It is important to note that these groups are not mutually exclusive (some species belong to multiple groups), and that some groups are nested within others.

- Shrubs: Dicotyledonous plants (of any family) which form perennial, above-ground woody stems. Such stems have secondary thickening, and can be "snapped". Common examples include *Artemisia xerophytica*, *Caragana leucophylla*, *Haloxylon ammodendron*, *Kallidium gracile, Nitraria sibirica, Oxytropis aciphylla* and *Reaumuria soongorica.*

- Forbs: Any species of angiosperm (monocot or dicot) that is not a shrub, and not a member of the Poaceae or Cyperaceae. Common examples include *Asparagus gobicus*, *Corispermum mongolicus*, *Ptilotrichum canescens*, *Rheum nanum* and *Scorzonera divaricata*. This group also includes sub-shrubs (or semi-shrubs) such as *Anabasis brevifolia*, *Peganum nigellastrum* and *Salsola collina*. It also includes the onions (*Allium* sp.).

- Grasses and sedges: Any species in the families Poaceae (grasses) or Cyperaceae (Sedges). Common examples include *Achnatherum splendens*, *Aristida heymannii*, *Carex duriuscula*, *Carex pediformis*, *Cleistogenes soongorica*, *Cleistogenes squarrosa*, *Stipa glareosa* and *Stipa gobica*.

- Annual: Any species which obligately complete its life-cycles in a single year. A common annual forb is *Corispermum mongolicus*. A common annual grass is *Eragrostis minor*.

- 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". Examples include several extremely common species such as *Anabasis brevifolia* and *Haloxylon ammodendron.*

- Fabaceous shrubs: Any shrub in the family Fabaceae (Peas). Prominent genera include *Caragana* and *Oxytropis*.

- *Artemisia* species: Any species in the genus *Artemisia*, whether a shrub (e.g. *A. xerophytica*) or a forb (e.g. *A. frigida*).

- Large *Haloxylon*. Any individual specimen of *Haloxylon ammodendron* that exceeds 1.5 m in total height.

- Adult *Ulmus pumila*. Any individual specimen of *Ulmus pumila* that exceeds 2.5 m in total height.

- Juvenile *Ulmus pumila.* Any individual specimen of *Ulmus pumila* that is between 0.5 m and 2.5 m in total height.

- Sapling *Ulmus pumila*. Any individual specimen of *Ulmus pumila* that is less than 0.5 m in total height.

- Juvenile *Ulmus pumila* (suppressed). Any juvenile *Ulmus pumila* that 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 juvenile *Ulmus pumila* that has one or a few extended recent branches that are likely to exceed 2.5 m and form a future trunk.

- Litter. Any plant material that is detached from the plant on which it formed (e.g. discarded leaves, twigs, etc.).

- 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).

2.2.4 Justification of variable inclusion and delineation

Total vegetation cover

Overall decreases in plant cover (or biomass) are generally interpreted as degradation (e.g. Fernández-Giménez & Allen-Diaz 2001, Yong-Zhong *et al.*, 2005; Pei *et al.* 2008, Tuvshintogtokh and Ariungerel 2013; Jamiyansharav *et al.*, 2018). There is, however, a great deal of debate about the relative contributions of climate and grazing to observed changes, and about precisely how decreased cover relates to condition and recoverability (Addison et al. 2012; Jamiyansharav *et al.*, 2018).

Cover of shrubs

In different systems, under different pressures, shrubs may be seen as invaders, or be highly valued where they stabilise soil and allow the accumulation of litter, carbon and nutrients, and in turn permit regeneration. For example, in grass-dominated steppes (here only 'Desert Steppe'), the literature suggests that, in general, overgrazing leads to increased shrub cover (e.g. Fernández-Giménez & Allen-Diaz 2001, Cheng *et al.* 2007). On the other hand, many systems are naturally shrub-dominated, and grazing sensitive shrubs may decrease with grazing (Stumpp *et al.* 2005; Pei *et al.* 2008). Taken together, it seems that shrub cover is likely to be related to perceptions of condition in very complex ways. This argues for the inclusion of shrub cover, but also suggests that shrubs should be split into multiple groups. Several groups of shrubs were singled out, as being likely to relate to cover in specific ways:

* Artemisia species appear to consistently increase with grazing (whether shrubs or forbs). This holds true for numerous species (*A. adamsii, A. frigida, A. glauca, A. laciniata, A. scoparia*) across numerous studies (e.g. Shiping & Yonghong 1999, Fernández-Giménez & Allen-Diaz 2001, Li *et al.* 2008, Pei *et al.* 2008, Yoshihara *et al.* 2010.). Given this genus is abundant and diverse in the Gobi region (Jambal & Olson 2016), it is an obvious choice to be singled out as a variable relevant to condition. It would be expected that low-moderate cover of Artemisia would be associated with higher site condition; however, some herders noted *A. frigida* as a component of healthy steppe, and this may influence this relationship.

* Leguminous shrubs (notably the numerous and abundant *Caragana* spp.) are distinguished from other shrubs because they increase soil nitrate via nitrogen fixation. In general, leguminous shrubs are palatable and often browsed, such that their depletion may be an indicator of overgrazing. Several studies have shown quantitatively that Fabaceous shrubs are associated with reduced grazing levels (e.g. *Caragana pygmea*, Fernández-Giménez & Allen-Diaz 2001; *Oxytropis glabra*, Pei *et al.* 2008).

* Succulent shrubs (almost exclusively Chenopodiaceae, which are abundant and diverse), are a prominent and obvious sub-group of shrubs in the Gobi region. Some are associated with high levels of disturbance (e.g. *Salsola passerina*, Fernández-Giménez & Allen-Diaz 2001), while others are associated with intact systems (e.g. Saxaul, below) (Note that forbs and sub-shrubs may also be succulent).

* *Saxaul (Haloxylon ammodendron)* is used to define an ecosystem (Saxaul), and is thus an obvious choice to be used as a variable relevant to condition. The abundance of Saxaul is clearly related to land use, degradation and perceptions of condition. In heavily utilised landscapes Saxaul may be depleted by collection for fuel. Saxaul also binds sand in places where few other species occur (Zou *et al.* 2010). Some herders suggested that livestock (especially camels) forced to consume too much Saxaul may become ill, suggesting that high Saxaul cover is not always desirable, especially without other species.

Cover of perennial grasses and sedges

Many studies have shown that grass cover is related to grazing, with palatable grasses being reduced under intense grazing pressure (Sasaki *et al.* 2005, Narantsetseg *et al.*,2015). The herders we interviewed identified many of the most common grasses and grass-like plants as being palatable and nutritious, including the grasses *Cleistogenes songarica*, *C. squarrosa*, *Stipa gobica*, *S.glareosa* and *S. grandis*, and the onions *Allium mongolicum* and *A. polyrhizum*.

Cover of all perennial forbs

In steppe ecosystems, Fernández-Giménez & Allen-Diaz (2001) found that forb biomass steadily decreased with increasing distance from waterpoints (i.e. lower forb biomass in less grazed plots). This relationship was not clear in desert steppe, with high forb biomass near waterpoints (driven by the ruderal *Chenopodium album*), low biomass at moderate distances, and increasing biomass at the greatest distances. It would seem

that forb biomass (or its surrogate, cover) is likely to be related to perceptions of condition, but in ways that may be complex.

Species richness (shrubs, grasses and sedges, forbs)

There is a general assumption in the global ecological literature that high species richness is desirable (Meir *et al*. 2004; Fleishman *et al*., 2006). Vascular plant species richness is related to grazing intensity in central Asian desert and steppe vegetation (Fernández-Giménez & Allen-Diaz 2001). Vascular species richness is thus very likely to relate to perceptions of condition.

Cover of litter

Litter is important for many ecological functions (Facelli *et al*. 1991), and it is known to vary significantly with different management regimes in desert steppe (most data from China), with litter decreasing under heavier grazing (Yong-Zhong *et al*. 2005; Li *et al*., 2008). It is thus reasonable to assume that litter cover is related to changes that represent degradation and recovery of ecosystem condition.

Maximum height of roots exposed by soil loss

Soil loss caused by wind erosion is generally considered a sign of degradation (Lal 1990; Batjargal 1997; Zhou *et al*. 2005, 2007), and the height of roots exposed by soil loss is a measure of the degree of recent soil loss.

Annuals vs perennials (forbs, grasses and sedges)

Grazing exclosure leads often to increases in the ratio of perennial to annual species, in both steppe and sand dunes (Katoh *et al*. 1998; Sasaki *et al*. 2005). This suggests that the lifespan of species (annual vs perennial) is a variable likely to be related to perceptions of condition.

2.3 Field sites

Field data were required for three purposes:

1. To gather stakeholder assessments of real sites, to test the appropriateness of the variables and the field sampling method (see Results 3.1, 3.2),

2. To allow real field sites to be incorporated into the set of sites assessed in the workshops (see Methods 2.7.4), to allow the model to be trained on realistic sites,

3. To gather stakeholder assessments of real sites, to enable the metric to be tested with field data (see results 3.3).

To achieve these purposes, we measured the selected variables across a set of field sites that were selected to cover the widest possible variation in the variables, and the widest possible condition spectrum.

Five days were allocated to field data collection. Given this time was constrained, field measurements were taken only for the desert ecosystems (Desert Steppe, Semi-Desert and True Desert), nominated as a priority by WCS, but not for Elm Forest or Saxaul. We sampled 28 sites in total, spread across Galba Gobi region, between Manlai and Gashuun Sukhait. The sites are described in detail in Appendix A. Their locations are shown in Figure 7.
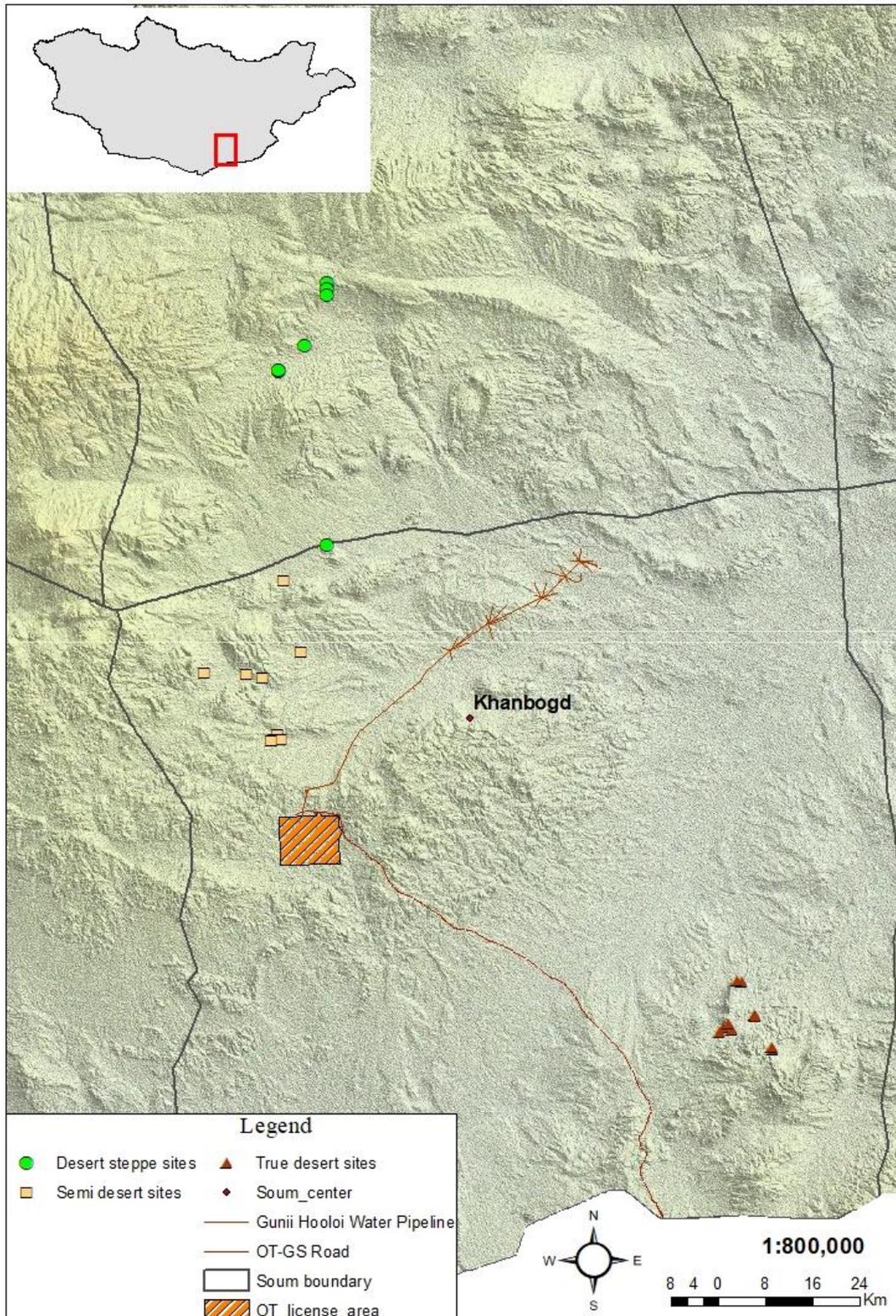
Figure 7. The locations of the field plots in Mongolia (inset) and the Khanbogd area.

2.4 Field sampling protocols for Desert systems

At each site, a 30 x 30 m (900 m²) 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.

The plot design described here is recommended for monitoring the following ecosystems:

- True Desert
- Semi Desert
- Desert Steppe

For Saxaul and Elm Forest, a different plot design is required, to capture variables specific to those ecosystems. The plot methods for these systems have not been trialled in the field. Draft (i.e. un-tested) methodologies for those ecosystems are presented in Appendix B.

2.4.1 Sampling vegetation 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 8.

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).
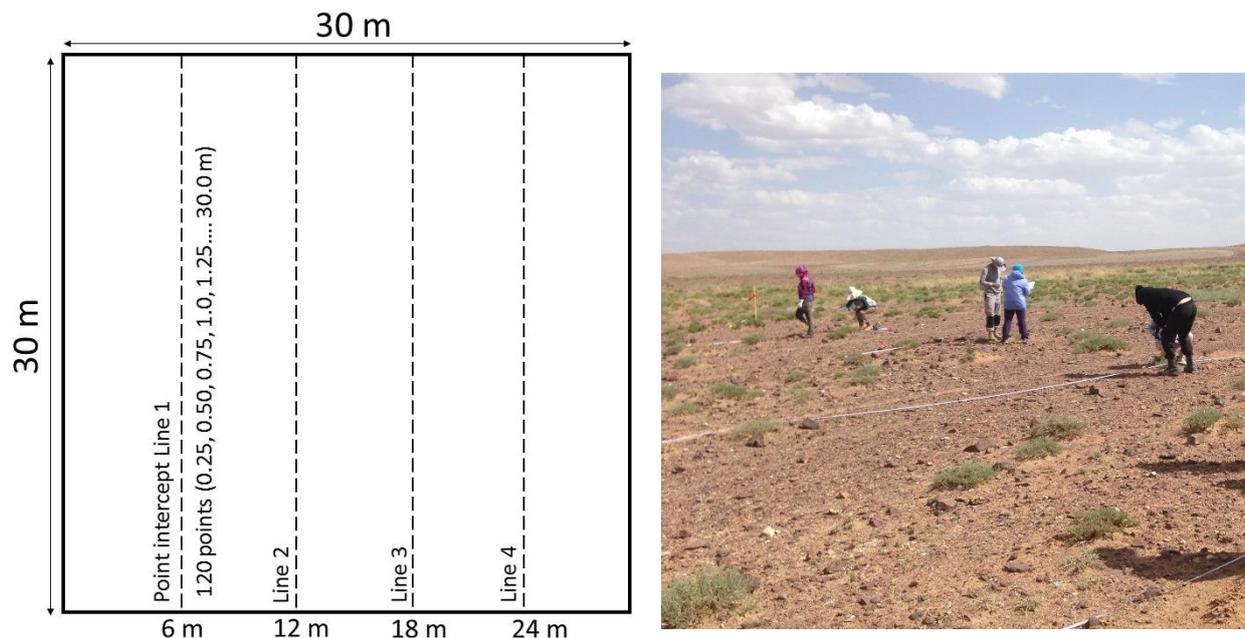
Figure 8. The plot method used to sample vegetation in the field.

Left: Diagram of the plot used to sample vegetation in the field. Right: Observers quantifying vegetation cover in plot, using tape measures to define the point intercept lines. The red flag marks the plot corner.

2.4.2 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.

2.4.3 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 9). 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.
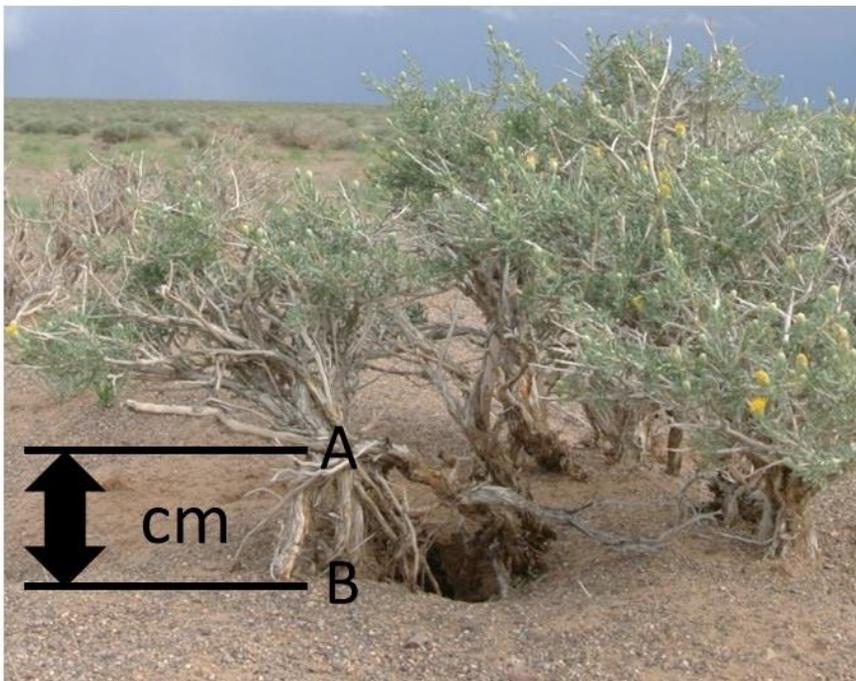
Figure 9. Measurement of roots exposed by soil loss.

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).

2.5 Expert stakeholders

Stakeholders with appropriate expertise (i.e. sufficient ecological knowledge to enable them to make evaluations based on simple vegetation data, regardless of their background or training) were selected by WCS with the intention of gathering opinions from a range of local stakeholders. A diverse group was selected (e.g. academics, botanists, zoologists, nomadic herders, land managers, consultant ecologists, amateur naturalists).

We interviewed 94 stakeholders in total, 52 male and 42 female. Eight stakeholders were non-Mongolian residents, the remainder were Mongolian. The stakeholders were offered compensation for their contribution.

Given the metric is intended to represent the consensus view of a stakeholder group, it is important to define the characteristics of this group (Sinclair *et al.* 2015). To assist in the description of the stakeholder group, each participant was asked to fill out a questionnaire describing their expertise, experience and affiliations. The questionnaire is included as Appendix C. These responses were used to show the representation of each organisation and skill set across the entire stakeholder group.

2.6 Stakeholder evaluation of field sites

We used the stakeholder evaluation of field sites for two purposes:

- To test the appropriateness of the selected variables for three of the ecosystems (Desert Steppe, Semi Desert, True Desert) (See results 3.1).
- To test the performance of the metric (See Results 3.3).

We took a group of expert stakeholders to all field sampling sites. Sixteen people participated in the field evaluations in total. Not all stakeholders were able to visit all sites, and the number who assessed any given field site varied between 9 and 15. At each site, we asked the stakeholders to do the following:

- Examine the plot (approx. 10 minutes was allowed).
- Evaluate the condition of the site, using a score between 0 and 100. The scores were recorded on a paper form.
- A score of 100 represents the highest ecological condition you could imagine for vegetation of this kind, at this site, in August, following a year of normal rainfall.
- A score of 0 represents the ecological condition of a site that has been degraded to the point where it retains none of the values associated with the ecosystem.

The evaluations were carried out independently. The stakeholders were asked not to communicate with each other prior to the submission of their assessments.

The participants were not provided with any instruction on which variables to consider, nor how to interpret or weight them. The stakeholders were aware that vegetation cover, species richness and soil loss were variables under consideration, given that these were measured at each plot. They remained unaware of how these basic concepts were expressed as variables (e.g. they remained unware of the distinction we made between life-forms).

2.7 Exemplar sites for workshop consultation

2.7.1 Rationale

In order to train models, we required evaluations of numerous sites, spanning a wide range of variation in each ecosystem. Given the unavoidable logistical constraints associated with field assessments in the vast Gobi region, most site evaluations were undertaken in the workshop context, using sites represented on cards. Each card described a single site for a single ecosystem, using the variables selected for that system.

The cover values on the cards were rounded to the nearest 5 for all covers above 10 (i.e. we used 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25…100 % cover).

The cards represented sites of three kinds, described below.

2.7.2 Card type 1: General synthetic sites

One hundred and twenty-one general synthetic sites were created for each ecosystem using the selected variables. These sites were designed to ensure that the dataset included sites covering the widest conceivable range of variation within each ecosystem, and including a wide range of permutations of values

for each variable. A wide-ranging site set in the training data reduces the need for the models to extrapolate when they encounter and assess a range of field-measured sites.

The cards were made by-

- consulting the literature and photographs of each ecosystem, and attempting to express many of the common variants of the ecosystem using the site variables, and

- systematically varying the site variables in combination and discarding implausible combinations.

2.7.3 Card type 2: Calibration sites

It is important that all stakeholders are evaluating sites on a common scale. To ensure this, we included a common set of pre-judged 'high' (3 cards) and 'low' (1 card) calibration sites within each set of synthetic sites. Every participant assessed these cards. The cards were not identified as calibration cards, and the participants were unaware of the calibration exercise. The calibration design was taken directly from previously published examples (Sinclair *et al.* 2015, 2018), as follows:

- The 3 'high' calibration cards were made by hand, subjectively, in consultation with WCS. They represented the most intact, highest condition sites that were considered possible for each system (generally very high species richness, high cover). The information used to construct these sites was taken from a pilot questionnaire of WCS staff.

- The 'low' calibration card was created for each ecosystem to represent a site that had no vegetation cover (all cover and richness variables set to 0), and some erosion (the maximum height of roots exposed by erosion set to 20 cm).

2.7.4 Card type 3: Real sites

The real sites we sampled in the field (10 True Desert, 10 Semi-Desert, 8 Desert Steppe) were converted into site cards, using the site measurements of the variables. These cards were included among the card set, but were not revealed to be real sites. Given the field work was only completed after the first workshop, these cards were only available for workshops 2 and 3, and the remote individual consultation (see Methods 2.8).

2.7.5 Card set composition

Each participant assessed 15 or 17 cards for each ecosystem. The composition of their set was as follows:

- 4 calibration cards (all participants encountered the same cards for each ecosystem).

- 11 general synthetic sites (each participant had a random selection of 11 cards taken from a pool of 121 cards for the ecosystem).

- 2 real sites; where available. These cards were only available for the ecosystems we sampled in the field, after workshop 1.

The different types of cards were not marked, and participants were unaware of the different cards in their set.

2.7.6 Presentation of consultation materials

The sites were presented to the stakeholders as printed cards, in A5 format. The cards included the random number, the variables representing that site, and a box for the stakeholder to write the score. The card sets were produced in both Mongolian and English, allowing each stakeholder to select which language they felt comfortable with. An example card is shown in Figure 10.

We automated the production of the cards (as pdf files ready for use) from the data describing the sites. We did this by creating images of each Mongolian and English phrase, and creating a script to call up the correct image and place it in the correct position on the card, based on the data describing the sites. This was processed with three packages jpeg (Urbanek 2014), plotrix (Lemon 2006) and grDevices (R Core Team 2016) in R (R Core Team 2016).

Some stakeholders may not be used to visualising covers from numerical values. We assisted them by presenting cover diagrams. Each cover value was represented by three cover images accurately representing that cover: One strongly clumped, one more dispersed, and one randomly dispersed. These were produced by colouring cells on a grid either black (cover) or white (non cover) using ArcGIS 10.3 (ESRI). The pixel counts confirm that the cover represented in each image is correct. The concept of cover, and the idea that cover could be more-or-less dispersed was introduced in the workshop introduction.

Desert steppe

Random site number   128

Cover all shrubs	5%			
Cover all succulents (shrubs or forbs)	2%			
Cover Fabaceous shrubs	3%			
Cover *Artemisia* species	1%			
Cover all perennial grasses / sedges	20%			
Cover all annual grasses	0%			
Cover perennial forbs and sub-shrubs	5%			
Cover annual forbs	5%			
Total vegetation cover	35%			
Number of shrub species	6			
Number of grass / sedge species	3			
Number of forb species	13			
Max. height exposed roots	2 cm			
Cover litter	3%			

Your name _____

Your condition score (0 – 100) [ ]

Figure 10. An example of a site card used for consultation.

This site, expressed in English, represents a real, field-measured site in the Desert Steppe ecosystem. Note the variables names are shortened for clarity.

2.8 Data elicitation workshops

2.8.1 Basic structure

The stakeholder evaluation data were collected from a series of workshops, detailed in Table 2. Figure 11 shows one of these workshops in progress.

Table 2: Summary of the expert consultation campaign.

Workshop	Date	Location	Number of participants
Workshop 1	4th August 2017	Khanbogd	35
Workshop 2	9th August 2017	Ulaanbaatar	21
Workshop 3	8th September 2017	Ulaanbaatar	11
Remote individual consultation	20th September – 31st October 2017	-	27
Total			**94**

The structure of each workshop was the same, and proceeded as follows:

- WCS presented an introductory talk on the need to develop a rangeland metric.
- The participants were provided with an audio-visual presentation that covered:
  - The delineation of the five relevant ecosystems.
  - The definitions of all variables, including the categorisation of plant species in to lifeforms. At the workshops, all groups were introduced in detail, with photographs and diagrams where necessary. The relationships between these botanical classes and commonly understood pastoralist terms were discussed (e.g. camel plants vs sheep plants).
  - The concept of condition.
  - The exercises they were expected to complete.
- Questions were taken, and open discussion was permitted among the participants.

All workshops were facilitated by between 3 – 4 facilitators, who were able to answer technical and ecological questions. The workshops were conducted in a mixture of Mongolian and English, with questions and answers translated between these languages as appropriate.

Clarifying questions were allowed throughout, but no communication was permitted between participants during the quantitative elicitation exercises.

2.8.2 Remote individual consultation

Many stakeholders could not attend any of the three scheduled workshops. To ensure that they contributed data, we allowed them to undertake the exercises on their own, supported by instructions provided by a 16 minute video covering the introductory material provided in the workshops. The card sets were sent to the stakeholders via email. The stakeholders were asked to print out the cards, fill them in, and return the filled cards by scanning and emailing them.

2.8.3 Elicitation exercise

At each workshop, the following exercise was undertaken five times, once for each of the ecosystems, in the following order: True Desert, Semi-Desert, Desert Steppe, Saxaul, Elm Forest. The design of the exercise was taken directly from that tested and published by Sinclair *et al.* (2015, 2018).

Each of the participants was given a set of site cards for one ecosystem.

The stakeholders were asked to rank the sites in order of the relative "condition" of each site. Tied ranks were permitted.

The stakeholders were then asked to quantitatively evaluate the condition site by writing a score on each card reflecting the quality of the sites. Again, ties were permitted. It was explained that scores need not be evenly distributed across the range of possible scores.

All participants were required to mark one of their cards 0 and one card 100. If any stakeholders felt that their concept of 0 or 100 was not represented in their card set, they were allowed to create their own new card that represented 0 and/or 100. Only one person (of 94) chose to use this option, for a single ecosystem. As described in Methods 2.7.3, we ensured that all stakeholders encountered a common set of expected very high condition and very low condition sites in their set.

Figure 11. Stakeholders evaluating site cards. Workshop 1, Khanbogd.

2.9 Data cleaning and outlier removal

2.9.1 Treatment of non-compliant card sets

All stakeholders assessed the calibration cards, and all were asked to score their card set to include scores of 0 and 100 (see section 2.4, above). This is important because it provides confidence that all evaluations across multiple stakeholders are scored on a comparable scale. Some stakeholders did not follow this instruction, and provided card sets that did not include scores of 0 and 100. We tolerated card sets that spanned a score range of at least 85 (e.g. 5-90, or 0-85), but discarded all sets that spanned less than this range.

All stakeholders were asked to score every card. However, some left cards blank. We tolerated card sets that included scores for more than half of the cards, but discarded those sets that had fewer than half of the cards scored.

On these critera, the cards from 2 of 94 participants were discarded entirely (from all ecosystems). The remainder (92) had card sets retained in at least one ecosystem. The data from a further 19 individuals were discarded for at least one ecosystem. The remaining 73 individuals contributed data to all ecosystems for which they filled out a card set.

2.9.2 Removal of outlying card sets using an aberrance measure

It is assumed that successful condition metrics will represent the consensus opinion of informed stakeholders. Some stakeholders provided evaluations that were very far removed from the consensus. Such outlying opinions may occur for three possible reasons:

1. Stakeholders may have valid and well-considered opinions that differ from the consensus view.

2. Some stakeholders may lack the requisite knowledge to make a sensible judgement, and their response represents damaging noise in the dataset.

3. Some stakeholders may have provided scores which do not reflect their true opinion because they mis-interpreted the task, or simply wrote the wrong score on their card.

Ideally, outliers stemming from 2 and 3 should be removed. Arguably, those stemming from 1 should be incorporated into the metric. There is no way of distinguishing these possibilities from the data alone.

We decided to remove whole card sets that retained high numbers of outlying scores (we did not retain any cards within these sets). Each ecosystem was treated separately: an individual participant who's card set was removed from the data for one ecosystem was not (necessarily) removed from the others. This reflects the fact that stakeholders may be more or less experienced with different systems.

We detected outliers by assigning every observation an aberrance score, using the method of Liu *et al.* (2017). To do this we built a preliminary regression tree using the R package 'rpart' (Therneau *et al.* 2017), using the site variables to predict the score. We applied this model to the same data used to train the model, to obtain a predicted score for all the stakeholder evaluations. Then, we calculated the absolute difference between the observed and predicted scores, which indicated the aberrance.

In order to understand the relative degree of aberrance, we compared our real stakeholders to dummy stakeholders. We made these dummies by copying the full evaluation dataset, creating new stakeholder ID numbers, and replacing all the scores with randomly generated scores between 0 and 100. This set of fake stakeholders thus retained the same over-all card set as the real stakeholders, and retained the calibration cards.

Each model calculation of aberrance included every real evaluation, and a single dummy. We repeated this process 90 times, each time with a different dummy. Only one dummy can be used at a time, because the inclusion of too many would perturb the model (by diluting the real opinion with noise), and make the assignment of aberrance scores less meaningful. Once the individual observations were assigned an aberrance, we assigned each card set (i.e each stakeholder) an aberrance score for the ecosystem in question, using the median aberrance score of the set.

To decide which card sets to remove, an aberrance threshold is required. This threshold can only be chosen subjectively. We discarded all card sets that had a median aberrance score greater than the 3rd quartile of the random dummies. This threshold can be seen in Figure 12. We believe that this removes most of the stakeholders who did not understand the task or know the ecosystems sufficiently well, but leave sufficient signal. Their removal does not imply their evaluations are "wrong", just un-helpful for defining a consensus.
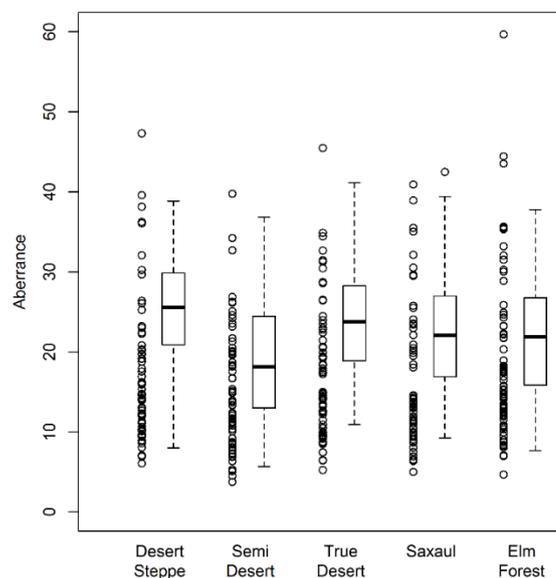
Figure 12. The aberrance scores for all card sets, compared to dummy sets.

Each open circle represents one card set, the dummy card sets are represented by the box plots, where the central bar represents the median, the ends of the boxes represent the first and third quartiles, and the whiskers extend to 1.5 times the inter-quartile range. Note the dummy box plots are generally more aberrant than the real stakeholders, as expected. We discarded all card sets that were more aberrant than the 3rd quartile of the dummies (i.e. above the top of the box).

2.9.3 Detecting and repairing key-stroke errors

Key-stroke errors (i.e. mis-entry of the raw data from the cards) are most damaging when they differ greatly from the correct value. We screened the data for keystroke errors, and repaired the errors we found, by targeting the most aberrant observations.

Starting from the most aberrant evaluations, we checked the entire card set to which the observation belonged, and corrected any key-stroke errors. We did this until ~10% of the cards (732) had been checked and repaired. We also assigned an aberrance value to each card set, by assigning it the median aberrance score from all cards in that set. We checked the most aberrant sets until we had checked 20% of all data (1494 cards).

We found a key-stroke error rate of 1.4%. We corrected these errors, and assumed that the remaining errors would have an acceptably low impact.

2.9.4 Overall impact of data cleaning and outlier removal

Table 3 shows the numbers of card sets which were removed during the data cleaning and outlier removal stages.

The final datasets included between 70 and 79 individual stakeholders. Previous work has examined how many stakeholders are required to support a stable model in this context. Studies on Australian grassland (Sinclair *et al*. 2015) and woodland (Sinclair *et al*. 2018) have both suggested that the use of more than approximately 20 stakeholders, each assessing 15 sites, is adequate. On this basis, we are confident that the use of 70 - 79 (per ecosystem) was more than adequate.

Table 3: Summary of the card sets accepted into the final dataset.

	Desert Steppe	Semi Desert	True Desert	Saxaul	Elm Forest
Number of stakeholder card sets	92	91	91	91	91
Non-compliant card sets	8	12	7	9	7
Card sets removed as outliers	9	9	5	8	7
Card sets remaining in final dataset	**75**	**70**	**79**	**78**	**74**

2.10 Modelling to predict condition score from the site variables

2.10.1 Adjusting the prevalence of the calibration cards

All stakeholders were given four calibration sites in their set (see Methods 2.7.3). As a result, the calibration sites were evaluated many more times than any other sites, and are thus over-represented in the dataset. To counter this, the calibration sites were culled in all models, such that 80% of all calibration observations, selected at random, were discarded (having served their purpose as calibration sites).

2.10.2 Regression tree ensembles

The condition metrics presented here are algorithms that transform multiple site variables into a single condition score. Each metric is a group (ensemble) of regression trees.

Regression trees operate by partitioning the data into subsets based on the predictors, such that each tree is a series of tests that eventually lead to a prediction. The regression trees were constructed from the training data using the open-access platform CLUS (https://dtai.cs.kuleuven.be/clus) (Blockeel *et al*. 1999). Regression trees were considered appropriate because they are capable of handling multiple variables of different kinds (categorical, ordinal, continuous), and variables that are nonlinear and interacting (Breiman *et al*. 1984; Kim & Park 2009).

The 'score' was set as the modelling target, and the site variables as the predictors.

The single score predicted by each of our final metrics is the result of model "ensembling", where the central tendency among many trees is used as a single consensus prediction. We constructed 30 trees for each ecosystem, each of which predicts different targets using different paths through the data. The single score is the median of the 30 trees. We used the median, rather than the arithmetic mean (as used previously in this

context (Sinclair *et al.* 2015, 2018), and for most regression tree ensembles) because the 30 predictions were found to be often skewed, with some outliers that would unduly influence the mean.

We proceeded through several stages of modelling, as follows:

- For all five ecosystems, a 10-fold cross validation process was used first, where a random selection of 90% of the data are used to train the models, and these are evaluated against the remaining 10%. This process is repeated 10 times, until all the data has been used as training and test data. The resultant model statistics provide some indication of how stable the models are likely to be (all models were found to be suitably stable). We trialled many different combinations of model settings, and selected those which produced the best model statistics (as described for the models below).

- For the three desert ecosystems, we built models for the purpose of testing the metrics against field data, as shown in Results 3.3, using the model settings we judged to be most appropriate: we used all of the synthetic sites to train the models (but excluded the field sites, which were held out for testing). For the models we used for testing, each tree was constructed from a different subset of the data, a process designed to increase accuracy but prevent overfitting (Breiman 1996). The data we selected was stratified to include 15 exemplar sites from each of 10 classes defined by total vegetation cover. We selected this stratification approach after discovering that total vegetation cover was the single variable most able to explain condition (See results 3.4). Thus, each tree was built using 150 observations (unless fewer than 150 were available in the stratum).

- For all five ecosystems, we took the model settings which produced the best models (above) and included all of the data in the training set (including, for the desert ecosystems, the field sites assessed as cards and as assessed in the field). These "all data" models remain untested, but are expected to be as good or better than the models produced with test data withheld. These "all data" models were used to create the final metric supplied to the client.

2.11 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.* 2015, 2017). There are three factors expected to contribute to this:

- Regression trees partition the data into bins, and return a prediction representing the centre of the bin. Thus, they never predict the outer edge of the bin, and extreme high or low scores are excluded.

- The ensembling process uses a measure of central tendency among many models (whether the arithmetic mean, the median, etc), such that the extremes of the score range are never returned by the ensemble.

- The calibration cards are set to represent unrealistically high condition sites, such that the stakeholders may reserve their 100 score for sites that are rarely (if ever) encountered in nature. This reduces the scores available to most real sites.

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 4 shows the highest and lowest scores used for each ecosystem, and the re-scaling function derived from these scores. We acknowledge that there are other re-scaling options available, the two most obvious being: 1) fitting to the card set evaluations, 2) fitting to the field evaluations.

Table 4: The parameters and functions used to re-scale the raw ensemble predictions.

Ecosystem	Lowest score returned by raw ensemble	Highest score returned by raw ensemble	Function used to re-scale raw ensemble median predictions
Desert Steppe	4.3	96.8	= (Raw-4.3) / 0.925
Semi Desert	4.1	92.3	= (Raw-4.1) / 0.882
True Desert	4.3	89.7	= (Raw-4.3) / 0.854
Saxaul	4.6	87.2	= (Raw-4.6) / 0.820
Elm Forest	8.3	100	= (Raw-8.3) / 0.917

2.12 Exploring the relationship between condition and each variable in isolation

To visualize the relationships between score (as described by the raw evaluation data, not the models) and each of the variables, we fitted a non-parametric local polynomial response function using locally weighted scatterplot smoothing (LOESS smoothing). In all cases we applied a two-degree polynomial with the span (proportion of the full data set used to calculate the mean response as a function of the predictor) set to 0.8. Confidence intervals were generated using the standard errors and degrees of freedom from the same models, implemented in R (R Development Core Team 2014).

2.13 Quantifying the relative importance of the variables

To illustrate the contribution of each variable to model performance, we compared the final model for each ecosystem to 1) a set of models with each variable withheld from the model, and 2) a set of models using only each single variable in isolation. We judged the contribution of each variable by the change in r^2, compared to the final model.

Although r^2 is a poor means of objectively assessing model performance *per se*, as argued above, it is adequate for comparing the relative differences between a single base model and comparable models with reduced input variable sets.

It must be remembered that some variables are nested within other variables (e.g. the cover of Fabaceous shrubs is a subset of the cover of all shrubs), making the relevant contributions of the variables difficult to disaggregate. Such disaggregation is required to show how important some basic components of the vegetation data are, relative to each other (richness vs cover; annuals vs perennials). These differences may be important when trying to understand the relationships between human concepts of condition and the data that can be remotely sensed. To facilitate this interpretation, Appendix E explores the relative importance of some simplified and non-nested variables which have been derived from the full variable set.

2.14 Presenting the metrics

All five final metrics are encoded as text strings, which may be used as formulae in a Microsoft Excel spreadsheet (once the relevant cell references are defined). These strings refer to the value of each variable that is measured, and use a series of "IF, THEN" statements to define the structure of the 30 regression trees which make up each model, and return a predicted score based on the values of the variables. The final score is the median of these 30 predictions, re-scaled using the functions shown in Table 4. A spreadsheet with all cell references defined, and ready to be used as a metric for each of the five ecosystems, has been supplied to the client.

3. Results

3.1 Testing the utility of the selected site variables

The variables selected to quantify condition must relate to site condition as perceived by the relevant stakeholders. To test this, we compared the site score assigned to a site by stakeholders in the field, with the score assigned to the same site by stakeholders, in an indoor workshop context, where the site was only represented by a card that described the site with the selected variables. No photograph of the site was provided.

A close correlation between these scores implies that:

- The measured variables capture and express something relevant to the concept of condition that the stakeholders applied to their field assessment, and therefore that the variables are appropriate for condition assessment.
- The plot method adequately measured the variables.

The logic of this test is shown in Figure 13.

The results of this test are shown for Desert Steppe, Semi-Desert and True Desert in Figure 14. For all three ecosystems there is a clear positive relationship between the scores assigned in the field, to the scores assigned to the corresponding cards in the workshops (r^2 = 0.53, 0.82, 0.81 respectively). This confirms that the selected variables are appropriate for the assessment of condition, and fit for purpose.

We did not perform this test for Elm Forest or Saxaul.
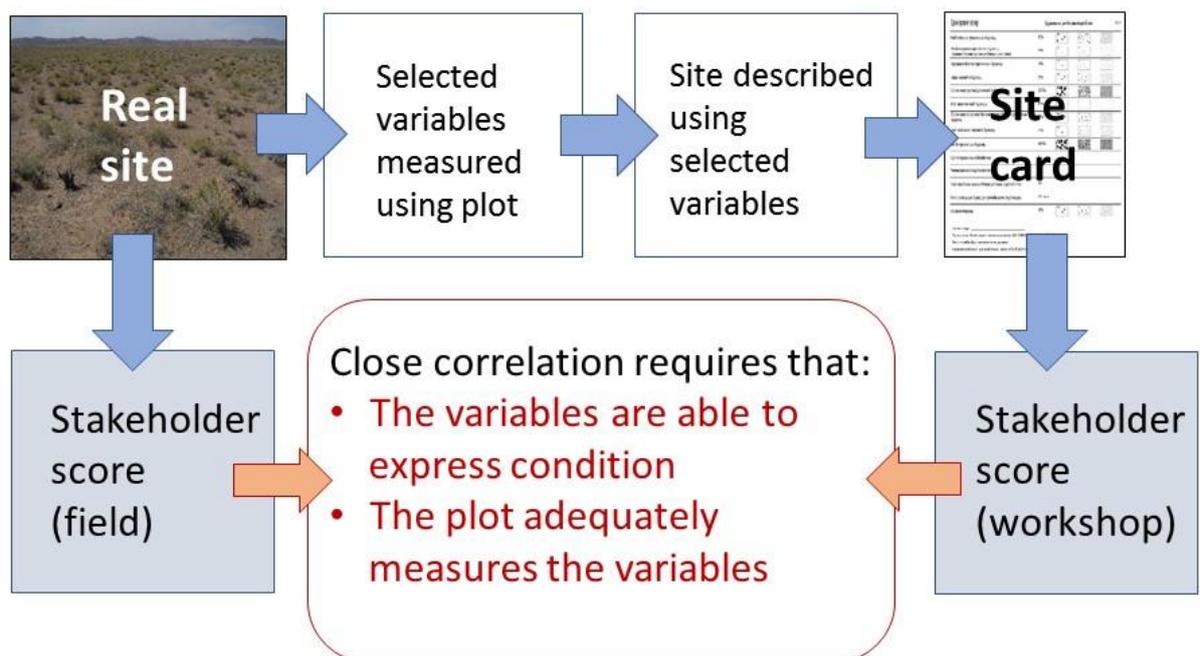
Figure 13. The logic of the test used to assess the suitability of the variables.
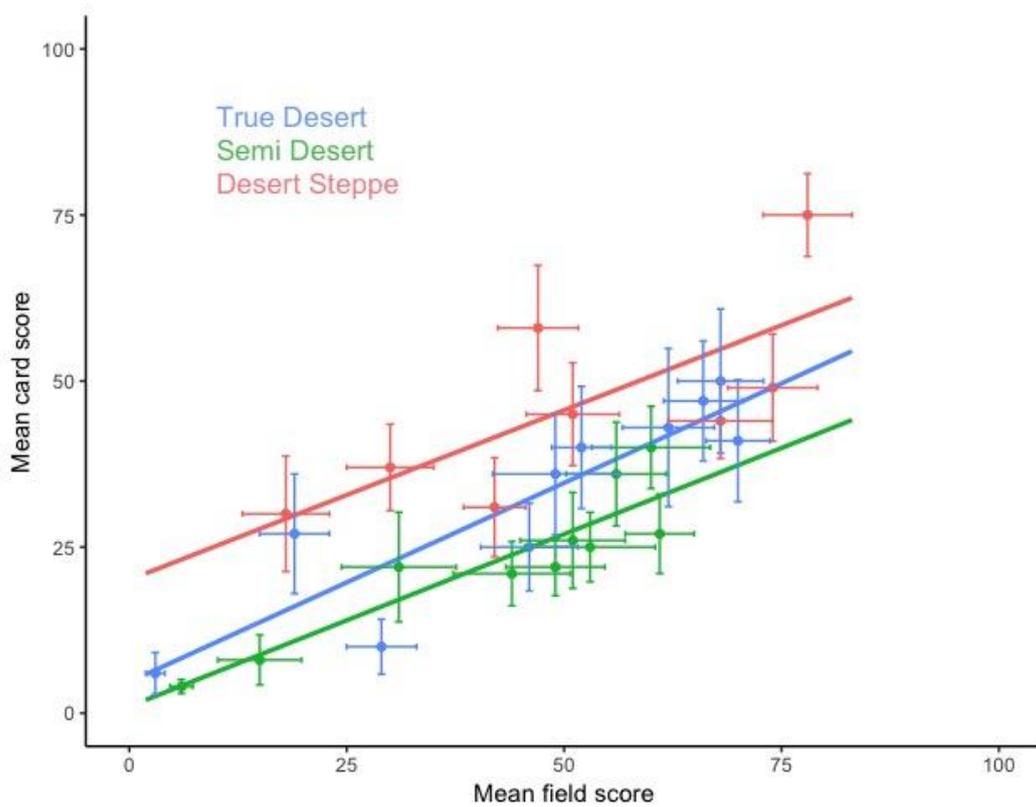
Figure 14. Relationship between field- and card-based assessments of the same sites.

The mean values of several assessments are reported (n= 9-14 for field scores, n=7-13 for card scores; see text), with error bars representing 1 Standard Error. The regression expressions that describe the relationships: True Desert: card = 0.599 x field + 4.689; Semi-Desert: card = 0.519 x field + 0.992; Desert Steppe: card = 0.511 x field + 20.063. These equations could be used to tilt the metric to the field scores, if desired (See Methods 2.11). Note that the relationships for all systems show slopes that are substantially less than 1. This is explored in Discussion 4.2.

3.2 The utility of the field plot method

It is important that the selected variables can be measured in the field using a method that is relatively rapid, repeatable, and able to measure the variables with a degree of precision that adequately captures site condition.

The test described above (Figure 14, Results 3.1) confirms that the point intercept plot method used here is adequate. If it was not, the translation of the real site's attributes to the site card would fail, and there would be little correlation between the stakeholders' field scores and the workshop scores.

These variables were measured in the field with an investment of <1 hour per site (2 people required), which is considered acceptable. We conclude that the plot method is fit for purpose.

3.3 The performance of the condition metrics

3.3.1 Tests against field evaluations

We assessed the ability of the metric to predict the consensus score among a group of expert stakeholders for a set of field sites. We performed this test for the desert ecosystems, True Desert, Desert Steppe and Semi Desert, as shown in Figure 15. Due to the limitations of field work, we did not perform this test for Saxaul or Elm Forest. Due to the limited availability of stakeholders, our test group did not include any herders.

We asked between 9 and 15 stakeholders to evaluate the condition of a range of field sites (each a 900 m^2 plot; the number of stakeholders varied between the sites as described in Methods 2.6). Although we provided guidance on how to conceive condition, we did not tell them which variables to incorporate, nor how to integrate them into a score. All stakeholders provided their own scores without consultation with others. We then measured the attributes of each site using a plot (see Methods 2.4), and derived a condition score for each plot using the appropriate metric.

In order to assess its performance, we treated the metric as another stakeholder 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. A successful metric would be expected to cluster with the group of human stakeholders, and close to the centre of this cluster.

The plots in Figures 15 span the full possible evaluation space. 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.

For all three ecosystems tested, the metric plots with the human evaluators. This suggests that the metrics are evaluating condition in a manner comparable to the stakeholders, in line with the project objectives. When compared to individual stakeholders, the metric performed better than some humans at finding the consensus (e.g. stakeholders 9 and 10 for True Desert, see Figure 15), but less well than others (e.g. stakeholder 2 for Desert Steppe).

We used the same data to demonstrate metric performance in a second way, for the three desert ecosystems. 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 close to the consensus view would be expected to plot a line intersecting (0,0) with a slope of 1.0 (i.e. 45°). Figure 16 shows the relationships for each stakeholder (grey lines) and the metric (black line). The metric performs well, with each line lying close to those of the human stakeholders, and each line correlating closely with the median values from the human stakeholders (True Desert r^2 = 0.78, Semi Desert r^2 = 0.82, Desert Steppe r^2 = 0.68). The metric provides a narrower score range for the real sites than the stakeholders (note the steep slope of the lines in Figure 16). For example for Semi Desert, the median stakeholder scores for the field sites ranged between 4 and 69, while for the same sites the metric predicted between 4 and 42.

The metric we used for the tests described above was not exposed to the field sites previously, and was not trained using the evaluations of these sites in the field or the workshop (Thus the test set is a true 'hold out'). The metric was re-scaled as described in Methods 2.11.

After testing, we made a new metric which did include the field sites. This 'all-data' model is the final model provided and recommended for use. It is untested, but, given it was trained on a slightly more extensive dataset including real sites, it is assumed to perform no less well than the version of the metric tested here.
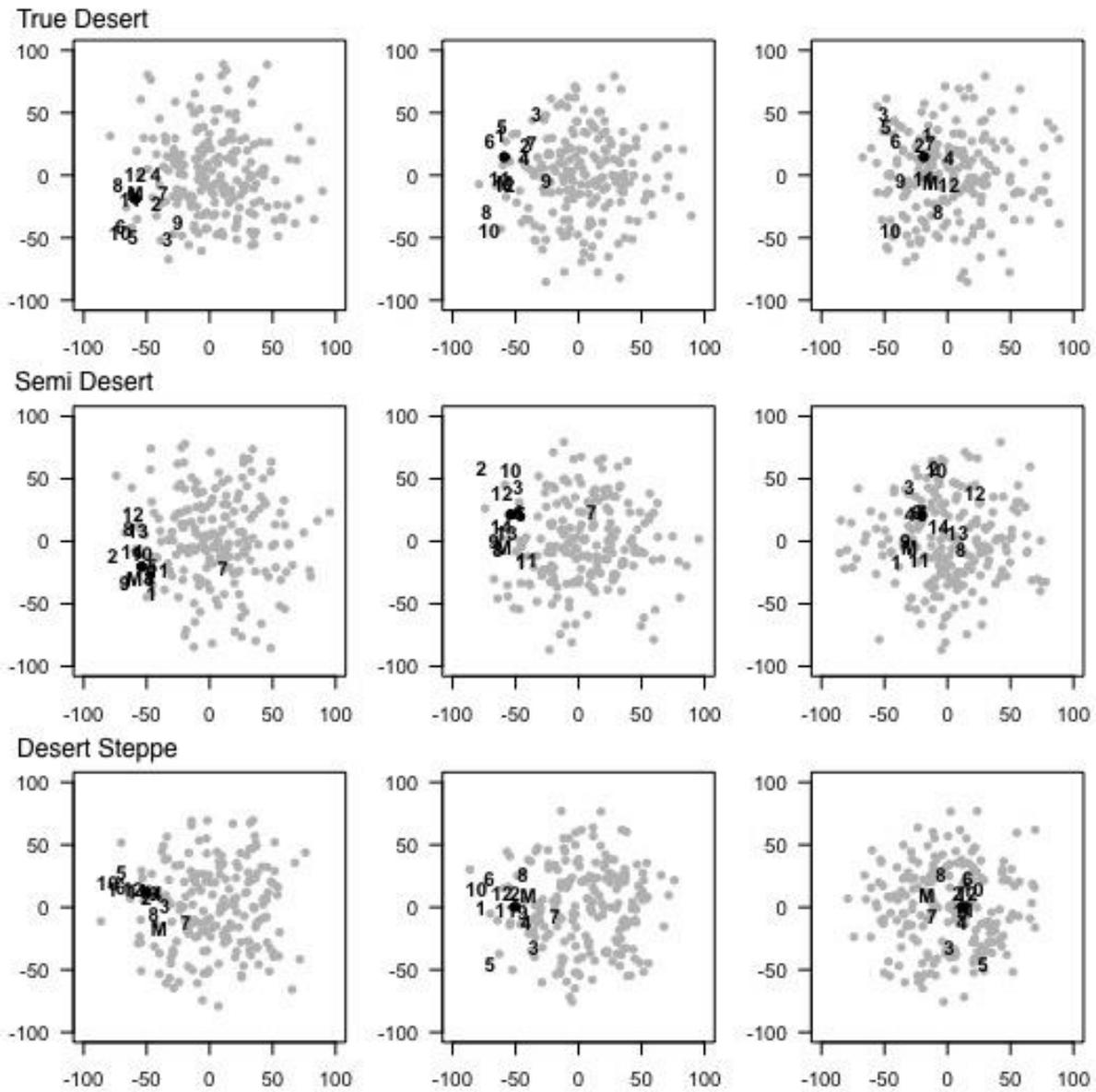
Figure 15. Evaluation of the metrics for the desert ecosystems, by comparison to stakeholders.

Multi-dimensional scaling (MDS) is used to determine whether the metrics return scores within the sensible evaluation space defined by real human evaluators. The space represents the evaluation space, defined from the field assessments. The numbers are human stakeholders, with each person designated by a number. M is the metric. 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. Left panel: MDS dimension 1 (horizontal), MDS dimension 2 (vertical); Middle panel: MDS dimension 1 (horizontal), MDS dimension 3 (vertical); Right panel: MDS dimension 2 (horizontal), MDS dimension 3 (vertical).
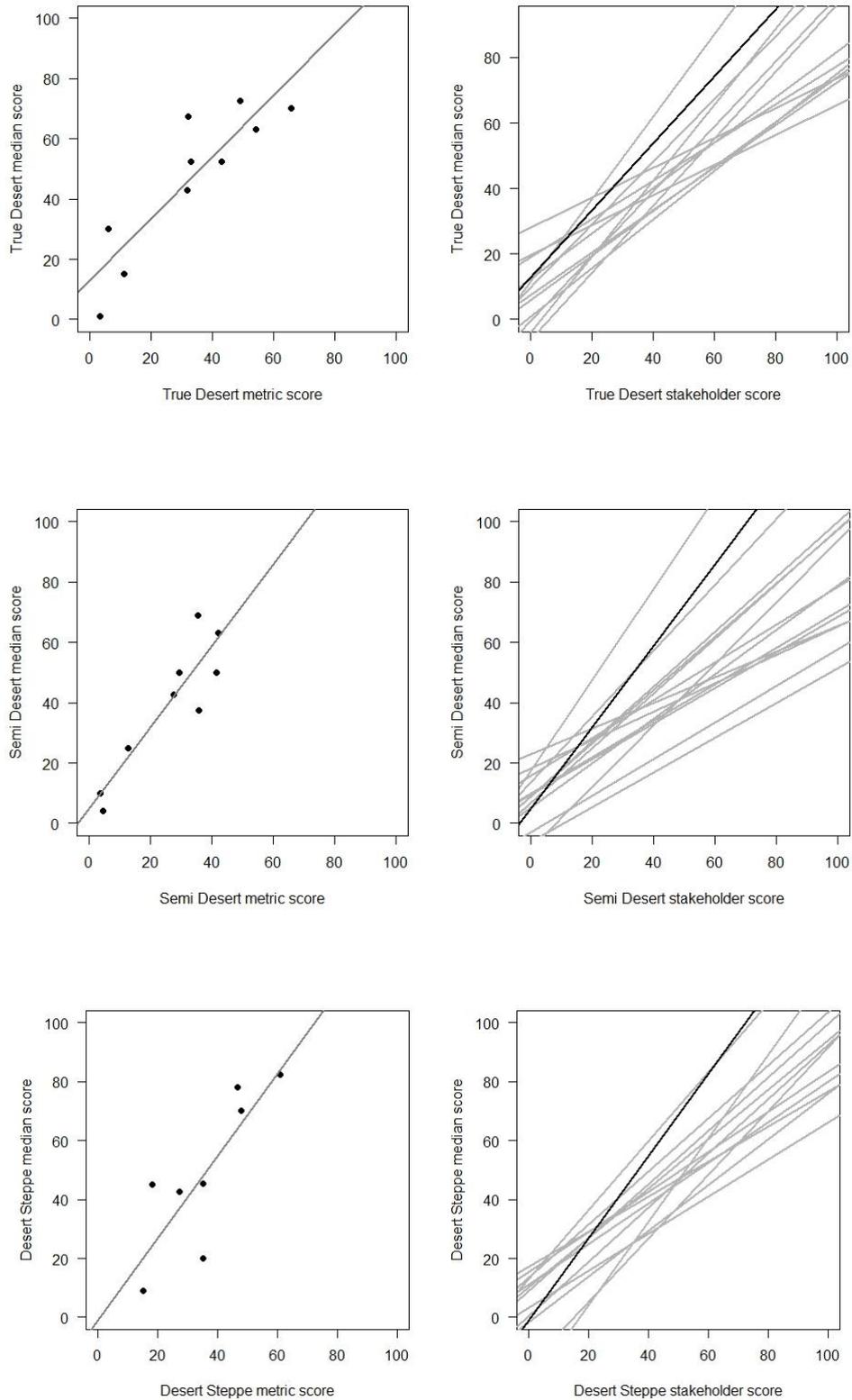
Figure 16. Relationships between metrics and human stakeholders, for the desert ecosystems.

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).

3.3.2 Performance statistics

We believe that the field-based tests presented above are the most valuable forms of metric evaluation, however additional information can be gleaned from some basic model performance statistics. The r^2 values for the following four scenarios are shown in table 5:

- The 10-fold cross validation exercise, where 90% of the data are used to train a model to predict the remaining 10% of data, ten times over until all the data has been used to train and to predict. The value presented is the mean r^2 from the ten tests.
- The final post-testing "all data" model (The version of the metric recommended for use), where all synthetic and field data are used to build a model. This model is then used to back-predict all of the workshop card-based assessments.
- The metric tested in Results 3.3 (a model produced from all of the synthetic sites), used to predict the stakeholder field-based assessments (i.e. multiple values for each field site).
- The metric tested in Results 3.3 (a model produced from all of the synthetic sites), used to predict the median of the stakeholder field-based assessments (i.e. a single value for each field site). This situation is identical to that shown in Figure 16.

These data are particularly useful in allowing a preliminary assessment of the Saxaul and Elm Forest metrics, which were not tested using field data. In the cases where all five ecosystem models can be compared, the Saxaul and Elm Forest metrics perform comparably well to the three desert ecosystems. Thus, there is no reason to believe that these untested metrics are inferior to the metrics for the desert ecosystems.

Table 5: r^2 values for several model scenarios.

The top line summarises the data used to train the model, the next line summarises the target against which the model predictions were tested.

Training data	Card-based (10Xval)	All data	Synthetic	Synthetic
	▼	▼	▼	▼
Test target	Card based	Card-based assessments	Stakeholder Field (all assessments)	Stakeholder Field (median)
Desert Steppe	0.53	0.49	0.17	0.68
Semi Desert	0.62	0.62	0.29	0.82
True Desert	0.40	0.41	0.22	0.78
Saxaul	0.64	0.62	NA	NA
Elm Forest	0.53	0.53	NA	NA

3.4 The roles of the individual variables

The method we used to create the metrics does not explicitly quantify the role of each variable. Nonetheless, it is useful to understand the role and influence of each variable as a means of checking that the data driving the model follows some expected ("common sense") patterns. These patterns include:

- Species richness should be positively related to condition score in virtually all systems, although the strength of this relationship may vary between systems. It is a generally accepted tenet of conservation biology that each species holds intrinsic value, and that high species richness is valuable (Meir *et al.* 2004; Fleishman *et al.*, 2006).

- The abundance (cover, density, etc) and health of so-called 'foundational' species should be strongly related to condition, in the ecosystems in which they occur. 'Foundational species' are those which shape ecosystem processes and on which many other species depend (Ellison *et al.* 2005, Ellison and Degrassi 2017). Clearly, Siberian Elm and Saxaul are such species in the ecosystems they define.

- The species (or species groups) which respond most strongly to the mechanisms of degradation (e.g. are killed by trampling, are palatable) should be strongly related to condition.

- The abundance of exotic species should be negatively related to condition score in virtually all systems, although the strength of this relationship would be expected to vary (D'antonio *et al.* 2002). It was determined up front that exotic species are barely relevant in the Gobi Desert, and this consideration does not apply here.

We used two ways to explore the role of individual variables.

- The first is to quantify their relative influence in prediction. This can be done by 1) systematically withholding each variable from the model, 2) using each variable in isolation; and then assessing the relative impacts by examining the change in r^2.

- The second is to examine the nature (shape) of the relationship between condition score and each variable in isolation in the raw opinion dataset. Such relationships were not sought explicitly from the stakeholders, but can be extracted from the multi-variate data-set.

We discuss the findings below, for the different systems. In general, the expected patterns are indeed evident in the data and the metrics, providing confidence that they reflect the generally accepted concepts of condition and degradation.

3.4.1 Desert ecosystems (True Desert, Desert Steppe, Semi Desert)

Several general patterns are obvious for the desert systems, as shown in Figures 17 to 22. First, 'Total vegetation cover' is easily the most informative single variable for the 3 desert ecosystems (i.e. when used alone to predict the score). Conversely, when it is removed from the model, it has very little impact (compare the left and right panels of Figures 17, 19 and 21). This is because the useful information it contains is also present among the other variables (which is not surprising, given it is derived from them). This means that, in general, stakeholders score sites with higher vegetation cover more highly than those with lower vegetation cover. This is consistent with prior views of condition in desert systems, where higher cover is generally assumed to signal intact vegetation (e.g. Addison *et al.*, 2012).

The importance of the plant lifeforms reflects the structure of the communities: In True Desert (defined by its shrub dominance), the 'Cover of all shrubs' is the most important lifeform variable; in Semi Desert, the 'Cover of all shrubs' and the 'Cover of all perennial grasses and sedges' are of roughly the same importance; while in Desert Steppe (defined by its grass dominance) the 'Cover of all perennial grasses and sedges' is far more important than 'Cover of all shrubs'. The relationships between the influential cover variables, and the score assigned by the stakeholders also show patterns consistent with expectations. For example, Shrub cover yields a peak condition score at a relatively high cover in True Desert (~40% cover), and a much lower cover in Desert Steppe, where grasses dominate, and shrubs are relatively inconsequential (~20% cover).

Interestingly, species richness proved to be generally less informative in all the desert ecosystems than the cover variables. However, the relationship between the richness of all groups and the score assigned by stakeholders was consistently positive, for all desert ecosystems, as expected (Meir *et al.* 2004; Fleishman *et al.*, 2006) (Figures 18, 20 and 22).

In general, the degree of redundancy among the variables is very high. All variables can be removed from the model without reducing the r^2 substantially, meaning that the 'condition' information contained in most variables is also captured by the remaining variables.

3.4.2 Saxaul

Unlike the desert systems, the single most informative variable for predicting condition in Saxaul is not 'Total vegetation cover', but the cover of Saxaul (*Haloxylon*) – the foundational species of this ecosystem. The importance of this variable reflects the pre-eminent importance of this species in the Saxaul ecosystem. Consistent with this, the next most important variable in isolation is the 'Density of large *Haloxylon*'. This variable has less relative impact when removed, suggesting that its information content is somewhat nested within Cover *Haloxylon*. The next most important variables in isolation are 'Cover all shrubs' and 'Cover all succulent shrubs'. These also relate to the cover of *Haloxylon*.

'Cover all perennial grasses and sedges' and 'Cover perennial forbs' are comparatively unimportant. The richness variables are comparatively unimportant in Saxaul, even less so than litter. Despite their relative unimportance, the richness variables all show a clear positive relationship with score, as expected (Figure 24).

The centrality of *Haloxylon* in this community, and the stakeholders' conception of its condition, is also emphasised by the fact that it is possible to maximise the score (i.e. score 100) with a site that only supports Saxaul (See Appendix D). No species (i.e. no cover or richness) of any forbs, grasses or sedges are required to reach this score.

Like the desert ecosystems, the degree of redundancy among the variables is very high.

3.4.3 Elm Forest

In Elm Forest, the single most important variable is again 'Total vegetation cover'. This is followed closely by several variables relating to the foundational species in this community, Siberian Elm. The second most important variable is 'Cover of *Ulmus*', which is positively related to condition, and which increases steadily from zero to approximately 40%, and then continues to increase until 75% is reached (which we acknowledge is excessively high for real sites). The third most important variable is 'Density of adult *Ulmus*', which is again positively related to condition, and which increases steadily from zero until it peaks close to 25 trees per 900m^2 plot (This represents a stand of trees with the average trunk separated from its nearest neighbour by approximately 5m, which is physically possible but exceptionally dense).

The variables related to other species (shrubs, forbs grasses and sedges) are relatively unimportant, but more important than the variables that deal with the density of juvenile and sapling *Ulmus*.

Like the other ecosystems, the degree of redundancy among the variables is very high.

Figure 17. The relative importance of the variables in the metric for True Desert.

Left panel: Shows the ability of each variable to predict score (points showing r^2), compared to the full 30-tree ensemble model with all variables (dotted line). Right panel: Shows the impact on r^2 of omitting each variable from the full model (dotted line).

Figure 18. Relationships between selected variables and workshop evaluations, True Desert.

The lines are derived from LOESS smoothing. Each grey point represents a stakeholder evaluation.

Figure 19. The relative importance of the variables in the metric for Semi Desert.

Left panel: Shows the ability of each variable to predict score (points showing r^2), compared to the full 30-tree ensemble model with all variables (dotted line). Right panel: Shows the impact on r^2 of omitting each variable from the full model (dotted line).

Figure 20. Relationships between selected variables and workshop evaluations, Semi Desert.

The lines are derived from LOESS smoothing. Each grey point represents a stakeholder evaluation.
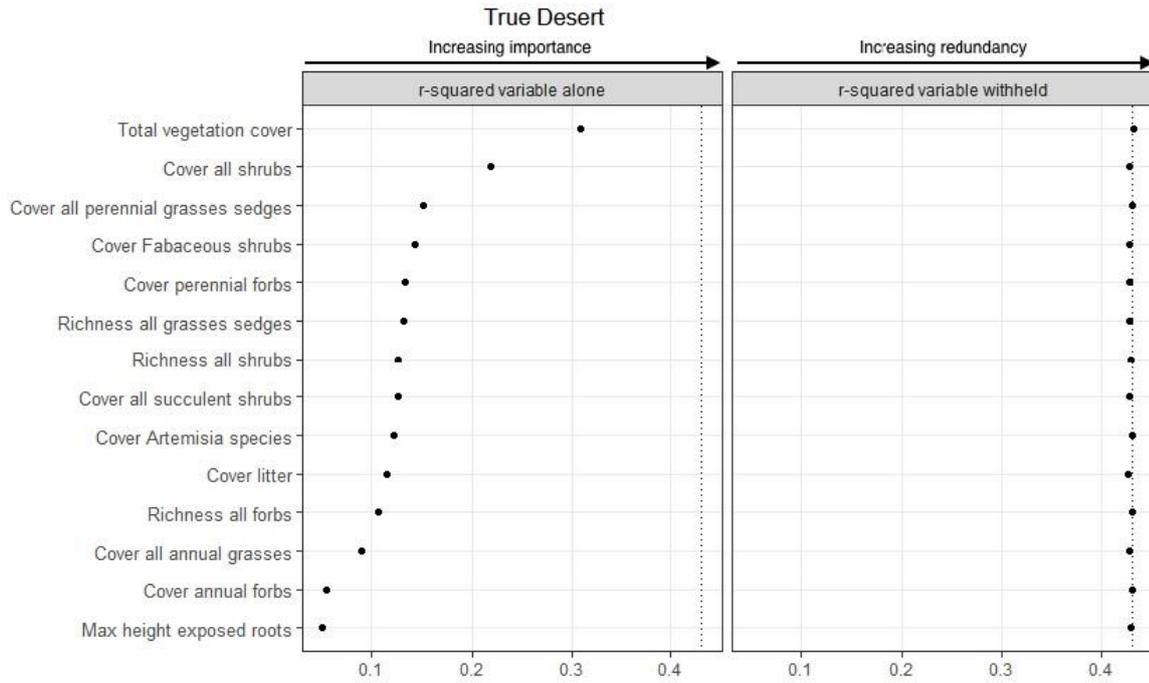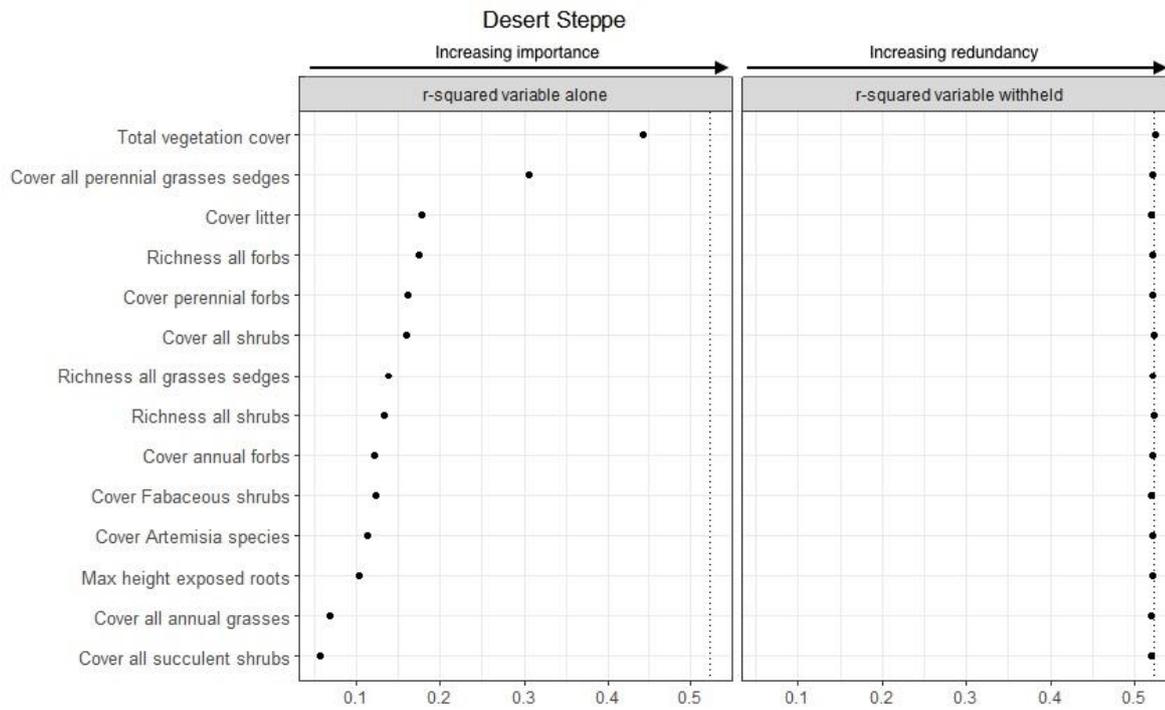
Figure 21. The relative importance of the variables in the metric for Desert Steppe.

Left panel: Shows the ability of each variable to predict score (points showing r^2), compared to the full 30-tree ensemble model with all variables (dotted line). Right panel: Shows the impact on r^2 of omitting each variable from the full model (dotted line).
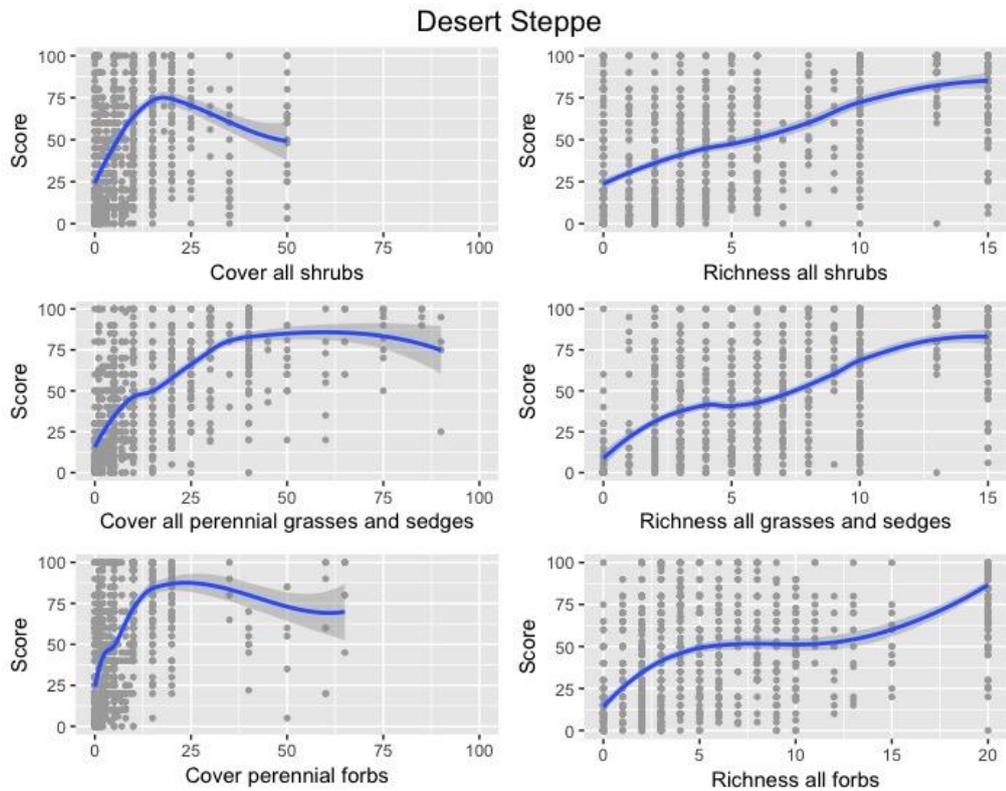
Figure 22. Relationships between selected variables and workshop evaluations, Desert Steppe.

The lines are derived from LOESS smoothing. Each grey point represents a stakeholder evaluation.
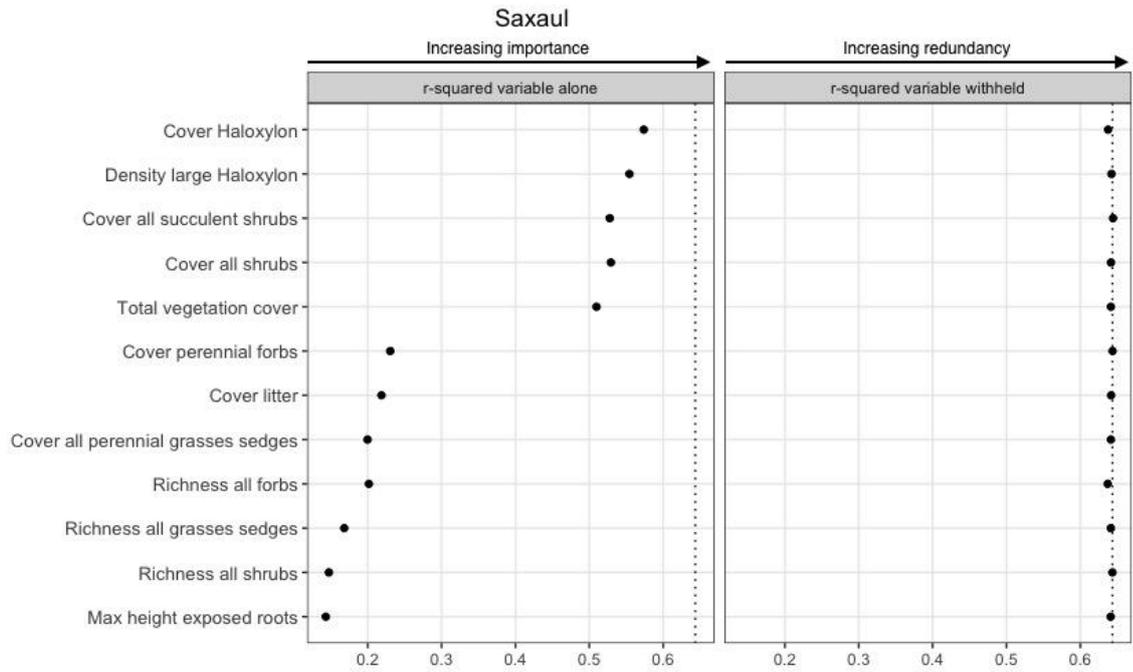
Figure 23. The relative importance of the variables in the metric for Saxaul.

Left panel: Shows the ability of each variable to predict score (points showing r^2), compared to the full 30-tree ensemble model with all variables (dotted line). Right panel: Shows the impact on r^2 of omitting each variable from the full model (dotted line).
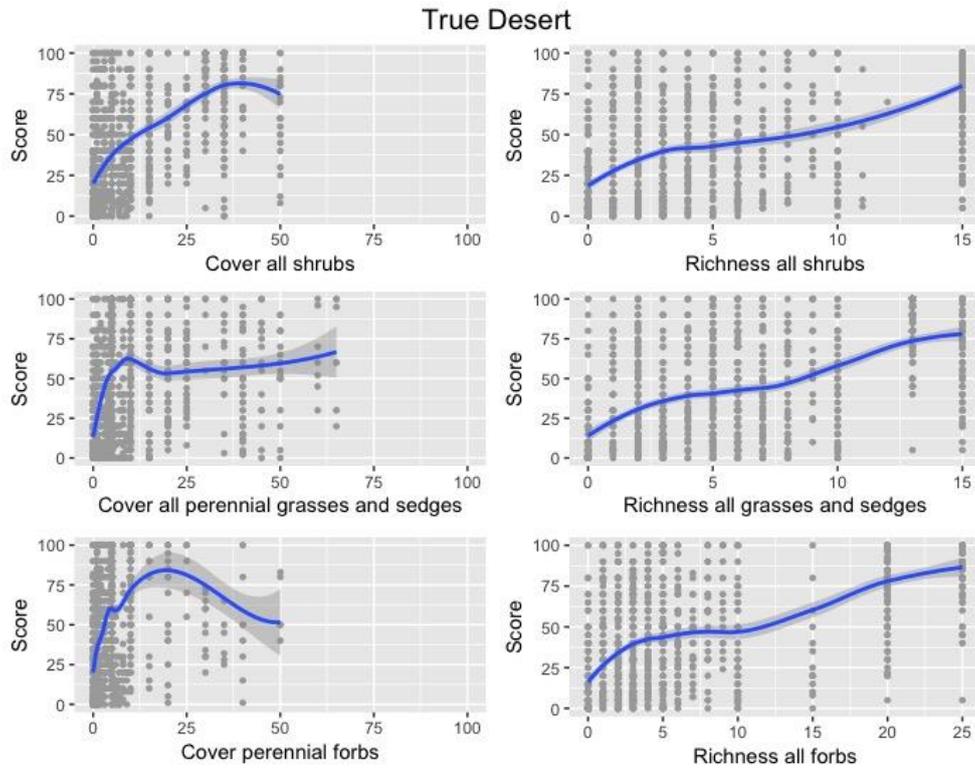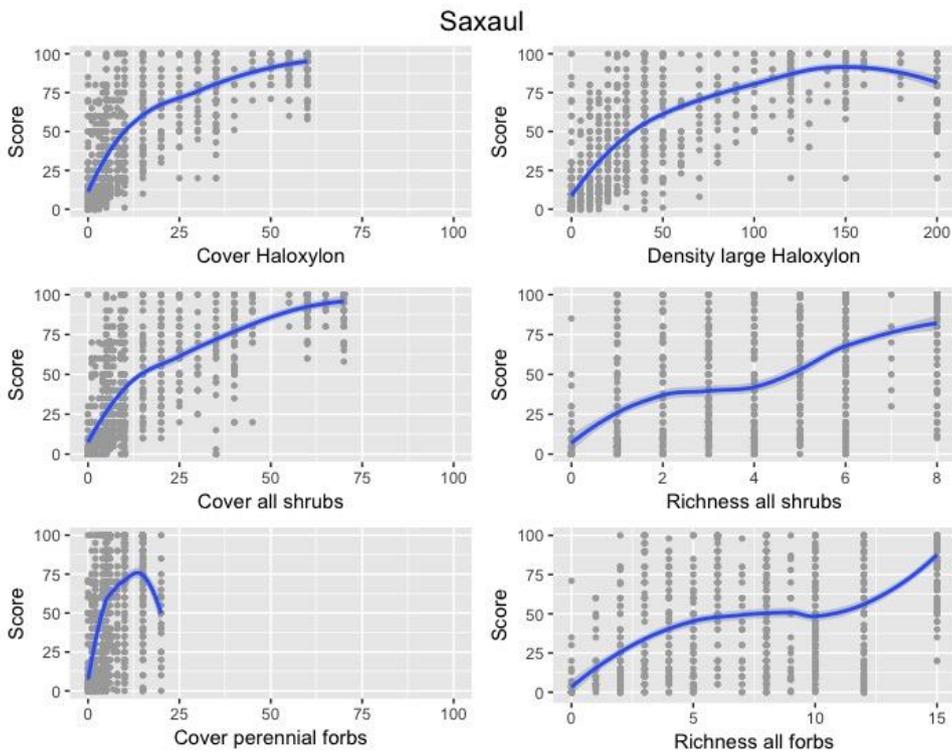

Figure 24. Relationships between selected variables and workshop evaluations, Saxaul.

The lines are derived from LOESS smoothing. Each grey point represents a stakeholder evaluation.
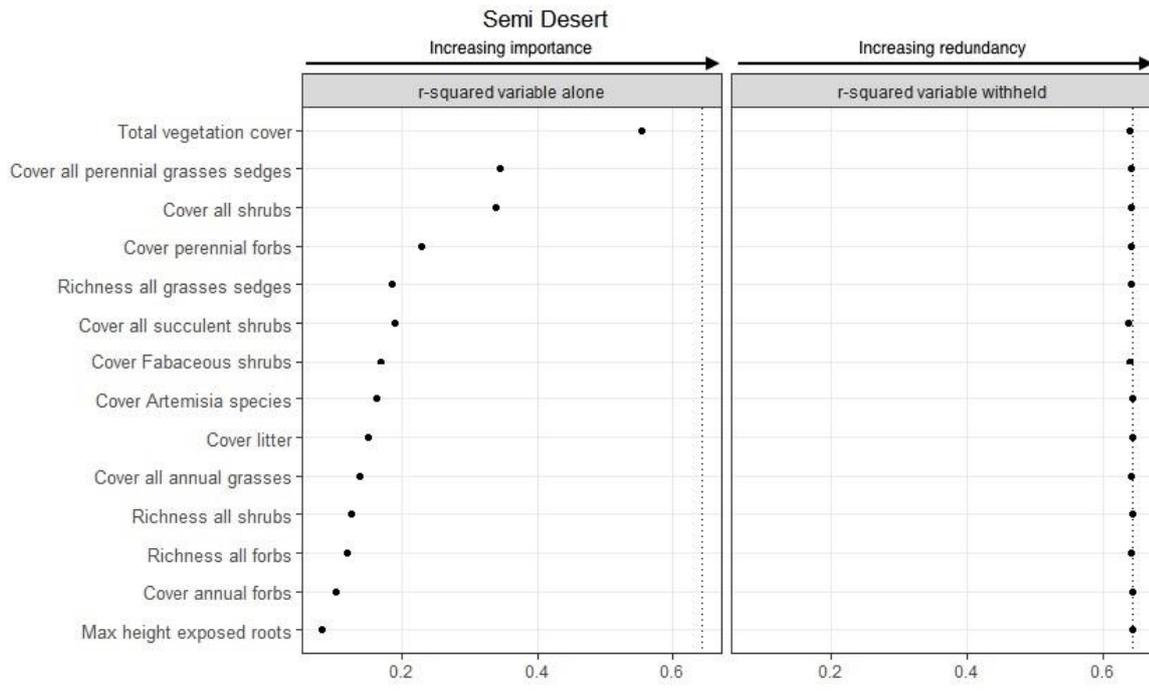
Figure 25. The relative importance of the variables in the metric for Elm Forest.

Left panel: Shows the ability of each variable to predict score (points showing r^2), compared to the full 30-tree ensemble model with all variables (dotted line). Right panel: Shows the impact on r^2 of omitting each variable from the full model (dotted line).
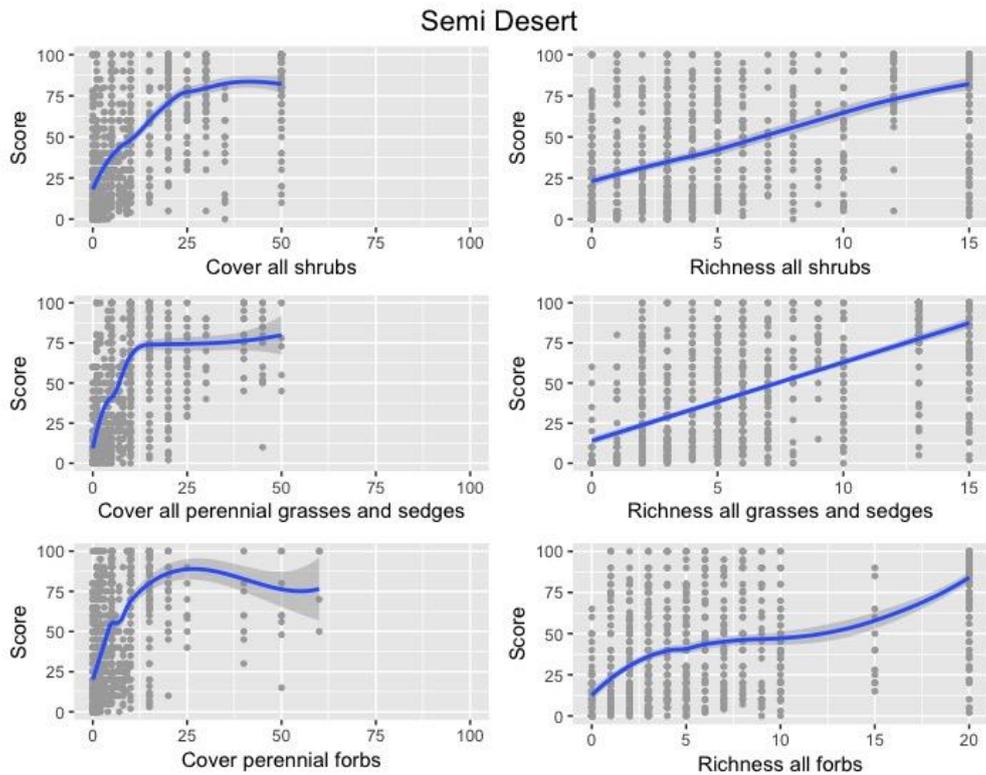

Figure 26. Relationships between selected variables and workshop evaluations, Elm Forest.

The lines are derived from LOESS smoothing. Each grey point represents an stakeholder evaluation.
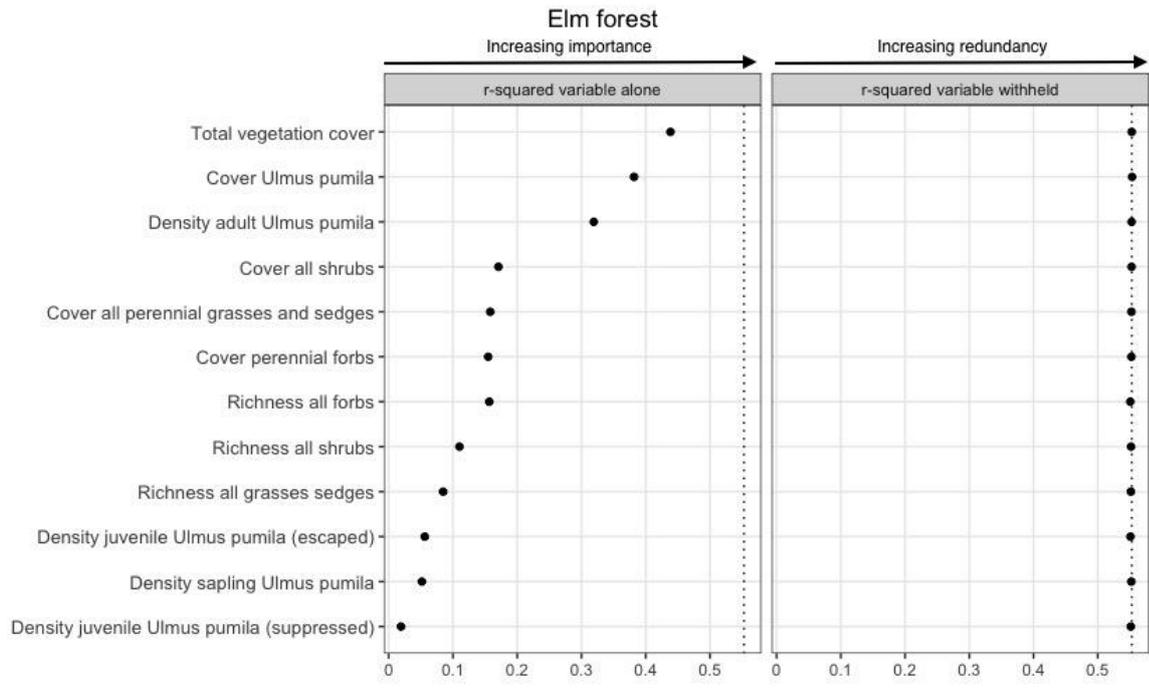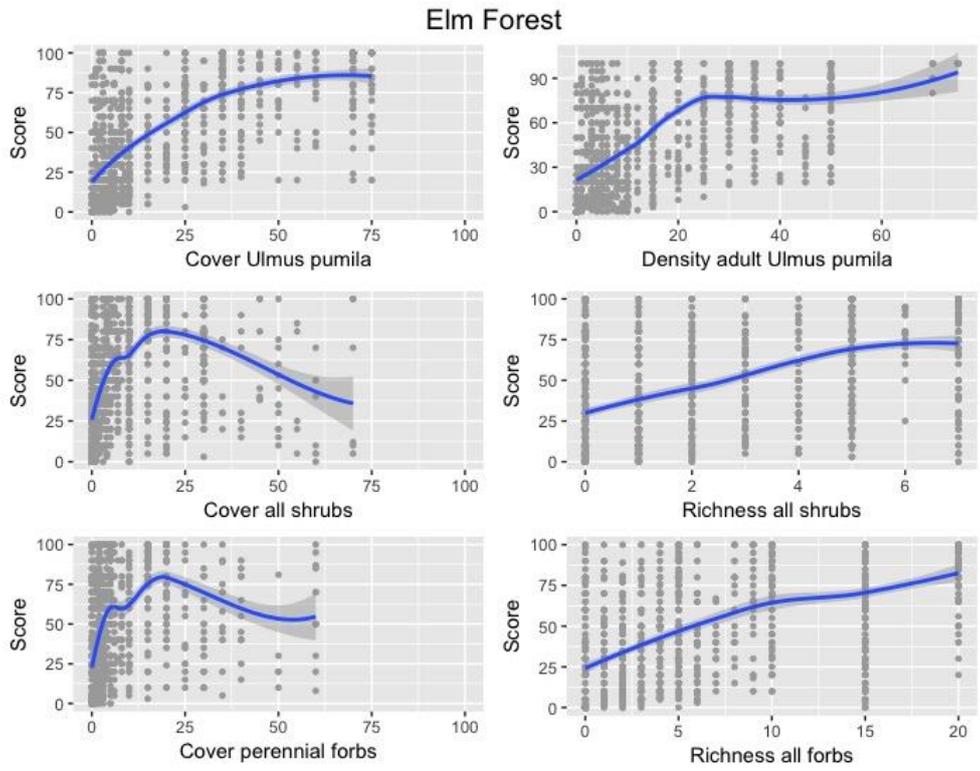
3.5 Application of the metrics to field-based monitoring data

The plot method described in Methods 2.4 was used by WCS during the field monitoring campaign for 2017, covering the desert ecosystems (not under the project reported in this report). As shown in table 6, the metric returned scores spanning a relatively wide range. Given that the monitoring sites were not chosen with the aim of sampling the absolute 'best' or 'worst' of the ecosystems, such that score of 0 or 100 would not be expected, we conclude that the scaling of the metric is sufficient to provide an acceptable spread among real monitoring sites.

Table 6: Score ranges from WCS 2017 monitoring data

Ecosystem	Number of sites in 2017	Minimum Score	Maximum score
Desert Steppe	17	15	78
Semi Desert	29	2	83
True Desert	29	13	57

3.6 Describing the characteristics of the stakeholders

It is important to describe the collective knowledge and experience of the stakeholders that defined the metric. The composition of our stakeholder group by gender and by primary area of expertise ('group') is summarised in table 7. This shows a strong representation from all four groups: conservation policymakers and practitioners, plant specialists and herders, and a relatively lower representation from wildlife specialists. This is reasonable, given the focus of the metrics on vegetation structure and composition.

It is important to recognise, however, that each individual brings a range of expertise, and is more than a representative of a single group. To summarise this diversity of knowledge within each individual and within each group, we apportioned the expertise of each individual into the four primary areas of expertise based on their self-assessment in a range of topic areas (that were themselves assigned to the four primary areas, Appendix C) (Figure 27). Each group contains a well-rounded mix of expertise, but with a clear focus on the topics that define the group. The Conservation policymakers and practitioners have the most even spread of expertise, while the pastoralists are the most narrowly focussed, with 72% of their expertise in pastoralism.

Table 7: Summary of stakeholders by gender and primary area of expertise.

	Pastoralist	Specialist- Botany	Specialist- Wildlife	Conservation policy and practice	Total
Female	6	15	4	16	**41**
Male	19	9	5	18	**51**
Total	**25**	**24**	**9**	**34**	**92**
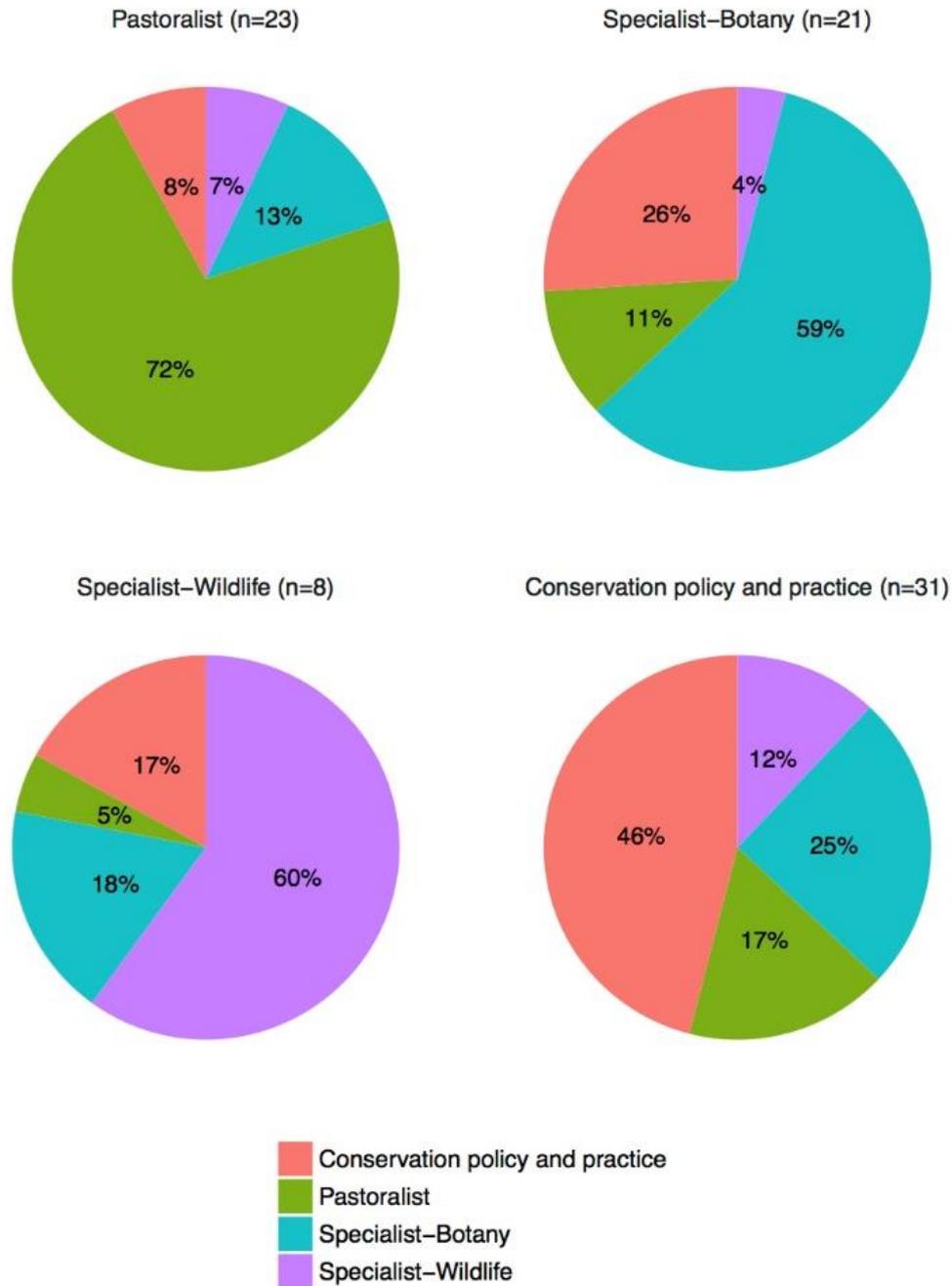
Figure 27. Summary of the diversity of expertise within each group of stakeholders.

Each pie chart represents one of the four stakeholder groups. The divisions within the chart show how the expertise within that group is distribution among its members; making the point that every primary group contains a range of skills, expertise and viewpoints. The number of stakeholders (n=83) is reduced from the total set because some people did not complete their "about you" evaluation forms.

4. Discussion

4.1 Variable selection

The unstructured and iterative variable selection process resulted in robust and useful variable sets for each ecosystem. This was directly demonstrated for the desert ecosystems (True Desert, Semi Desert and Desert Steppe). For the Saxaul and Elm Forest ecosystems this can be inferred from the model performance, the relative importance of the key variables, and the sensible relationships between condition score and each variable.

The high degree of redundancy among the variables for every ecosystem (as revealed in Figures 17, 19, 21, 23 and 25) suggests that fewer variables could have been used to produce robust metrics. This would have saved effort in the stakeholder workshops. However, there is no ongoing disadvantage or cost to the inclusion of all the variables, given that the point intercept plot measures almost all of them at once without additional effort. Variables can be measured within an hour at each plot, and require no further laboratory analysis.

We conclude that the variable sets are fit for purpose.

4.2 Metric performance

The metrics performed well compared to human stakeholders, generally predicting scores close to the stakeholder scores (Figures 15 and 16). When compared to individual stakeholders evaluating field sites, the metrics for the desert ecosystems performed better than some humans at finding the consensus (median), but less well than others. The metric, however, brings benefits that even the best-performing observer cannot bring: The metric provides consistent and repeatable results, is always freely available, is incorruptible, and transparent and defensible with regard to method. The metrics which were not tested with field data (Saxaul and Elm Forest) performed as well as the desert ecosystem metrics when model performance statistics are considered, suggesting that these metrics are also robust.

The consensus among the stakeholders exists among a very wide range of responses. Some cards received scores from different stakeholders varying over a full 90 points of score range (see the raw evaluation data presented in Appendix F). This means that any single stakeholder may occasionally find that their own evaluation differs from the metric score (and the true median response of other stakeholders) by over 40 points. This is an inevitable consequence of the evaluation data, which are highly variable.

The metric provided a narrower score range for the real sites than the stakeholders. For example, in Semi Desert, the median stakeholder scores for the field sites ranged between 4 and 69, while for the same sites the metric predicted between 4 and 42. There may be several reasons for this:

- The stakeholders who assessed the sites in the field did not undertake the calibration exercise, where they would have been systematically reminded what a pristine (100) and a valueless (0) site were like. The uncalibrated stakeholders may have expanded their score range to better differentiate among the real sites they were asked to assess. The model, in comparison, was calibrated.

- The metric may tend to score real sites lower, because its expectation of what a high-condition site may be inflated by the use of extremely high-cover and high-richness (hence high condition) sites in the workshop, which may be never found in nature.

- The pool of stakeholders who assessed the field sites was smaller than the total stakeholder set used to create the metric (This is essentially inevitable, given the practical constraints of working in the Gobi Desert). The smaller test set may be more influenced by differing or outlying opinions compared to the metric.

The score range could be adjusted by applying a different re-scaling function (See Methods 2.11), but we suggest this is unnecessary, particularly given the degree of resolution found when the metrics were applied to the 2017 field data.

We conclude that the metrics for the desert ecosystems are fit for their intended purposes. These are discussed below.

4.3 Application to monitoring programs

4.3.1 Assessing change over time

The metric is expected to perform well as a means of measuring change (degradation or improvement) over time, at a given site. Yearly differences in precipitation will lead to fluctuations in cover and the detectability of species, but it is expected that multiple years of data will enable longer-term averages to reveal any meaningful changes in condition.

The ability of metric to separate the real 2017 monitoring sites from each other suggests that the desert sampling plots and metrics have a degree of resolution that is adequate for real applications (where resolution refers to the ability of the model and the plot method to register and differentiate relevant changes to the variables). Informal tests of all five metrics (by changing the parameters and observing the score change) also confirm that small, realistic changes in cover or richness are, in general, appropriately reflected in a small score change. We encountered one exception to this: the Elm Forest metric was unable to distinguish between sites that had no adult *Ulmus* trees and sites that had a low density of trees. Testing of this metric will reveal whether this is problematic at real sites, and whether a post-model adjustment is required for such sites.

Grazing exclosure plots paired with adjacent plots offer perhaps the best means of demonstrating the performance of the metric in this task (assuming, as most authors agree, that grazing can lead to degradation). We should expect to see a gradual separation in the condition score of the fenced plot from its pair, although this change may take years, and may be punctuated by fluctuations in the interim. There will also be some cases where this separation is not seen- for example, if the plots were not degraded by grazing at the time of fencing, if the plots are too inherently dissimilar, or if the plots were so degraded that no recovery is possible.

4.3.2 Comparing between sites and ecosystems

The 2017 monitoring results summarised in Table 6 show that the desert metrics possess sufficient resolution to distinguish between real sites. While we suggest that the metrics are fit for this purpose, some caution in interpreting the results is required, due to the inherent variation between sites (as discussed in Introduction 1.2.5). We do not know the degree to which inherent inter-site variation (as opposed to variation in condition) will be reflected in their scores.

It is worth noting that the metrics are limited by the resolution of the ecosystem typology used to direct them (i.e. the distinction between one ecosystem type and another). The metrics require the ecosystem to be defined: if site data from a True Desert site is measured with the Semi Desert metric, a different score will be produced as compared to the True Desert metric. If sites are mis-classified or mis-mapped, they may receive a misleading condition score. The relative coarseness of the ecosystem typology (only 3 desert ecosystems, across a range of landscape positions and species compositions) means that much inter-site variation is hidden within each type, and that this variation is liable to be recorded as differences in condition, even if it represents inherent differences in site type. A finer-level ecosystem typology would allow more precise and meaningful distinctions to be drawn between sites.

The way in which condition was conceived was not system-specific. Thus, a score of 100 in one system should be equivalent to a 100 score in another system.

4.3.3 Extrapolation

As presented here, the metrics require site-based measurements, and produces site-based condition scores. Ideally, managers and ecologists would like to learn about the extent of changes in condition across large areas. Remotely-sensed variables (such as the satellite data that are regularly collected across the Gobi region) can potentially facilitate landscape-wide assessments, because they cover every site (every pixel). Ideally, sites with measured condition could be used to train models that can use the remotely-sensed data to extrapolate, and provide complete coverage of condition information.

There are two problems with the use of remotely-sensed data to extrapolate. The first is related to ecosystem circumscription (see Discussion 4.3.2, above). Since each metric is ecosystem-specific, it is necessary to know the ecosystem type that occurs at every site (every pixel). Currently, the ecosystem mapping for the Gobi is unlikely to be sufficiently resolved, and will cause errors if the wrong metric is modelled at the wrong location.

The second problem is that some of the variables that define condition in our metrics (e.g. species richness) cannot currently be measured from remotely-sensed data. This means that the condition metrics do not have

reliable data with which to extrapolate. Both of these issues can be solved to some degree: possibly by considering the ecosystem type as probabilistic (i.e. with uncertainty), and accounting for the degree of error introduced into the score due to the lack of knowledge of some variables (such as species richness). It is fortunate that much of the condition score is driven by variables that can be remotely sensed (e.g. total vegetation cover, see Appendix E).

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 metrics for Desert Steppe, Semi Desert and True Desert be used for monitoring in the Gobi Desert. Such monitoring may include comparisons between sites, between years and between ecosystems.

- The metrics for Saxaul and Elm Forest be used as draft metrics, with the knowledge that any minor changes made after testing (see below) can be back cast onto the raw data.

- 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 2.6). These assessments should be recorded. They will permit ongoing comparisons to be made between the metrics and the expectations of stakeholders.

5.2 Communications

- The metrics be published in peer-reviewed format.

- A short summary document is produced, which describes the final metric product and its capabilities, without the degree of detail and discussion included here.

- Any promotion of the metric stresses that stakeholder opinion varies widely, that no metric could perfectly reproduce a single stakeholder's evaluation of a site, and that any given stakeholder may find that some of their evaluations differ markedly from the metric.

5.3 Further work

- The metrics for Saxaul and Elm Forest be tested by comparing the median scores from a group of suitable expert stakeholders to the scores provided by the metric, using the approach presented in Results 3.3.1

- The field plot method for Elm Forest be trialled, and its utility and practicality be assessed.

- Opportunities for extrapolating condition across the landscape using remotely sensed data be explored. This will enable reporting on landscape-scale changes, not just plot-based changes. Given that not all of the metric variables are readily detectable from remotely sensed data, this work may involve deriving approximations of the metrics that are able to be sensed, and quantifying their relationship to the metrics.

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.

Batjargal, Z. (1997). Desertification in Mongolia. RALA report, 200, 107-113.

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.

Cheng, X., An, S., Chen, J., Li, B., Liu, Y. and Liu, S. (2007) Spatial relationships among species, above-ground biomass, N, and P in degraded grasslands in Ordos Plateau, north western China. *Journal of Arid Environments* **68,** 652-667.

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.

D'antonio, C A.R.L.A. and Meyerson, L. A. (2002). Exotic plant species as problems and solutions in ecological restoration: a synthesis. *Restoration Ecology* **10,** 703-713.

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.

Ellison, A.M., Bank, M.S., Clinton, B.D., Colburn, E.A., Elliott, K., Ford, C.R., Foster, D.R., Kloeppel, B.D., Knoepp, J.D., Lovett, G.M., Mohan, J., Orwig, D.A., Rodenhouse, D.A., Sobzak, N.L., Stinson, W.V., Stone, K.A., Swan, J.K., Thompson, C.M., Holle, J., Von, B. and Webster, J.R. (2005). Loss of foundation species: consequences for the structure and dynamics of forested ecosystems. *Frontiers in Ecology and the Environment* **3,** 479–486.

Ellison, A.M. and Degrassi, A.L. (2017). All species are important, but some species are more important than others. *Journal of Vegetation Science* **28,** 669–671.

Facelli, J. M., and Pickett, S. T. (1991). Plant litter: its dynamics and effects on plant community structure. *The Botanical Review* **57,** 1-32.

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. 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.

Fleishman, E., Noss, R. F., and Noon, B. R. (2006). Utility and limitations of species richness metrics for conservation planning. *Ecological Indicators* **6,** 543-553.

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.

Hilbig, W. (1995). Vegetation of Mongolia. SPB Academic Publishing.

Jambal, S. and Olson, K. (2016). Rangeland condition in the South Gobi. Appendix E; Annual monitoring report, Wildlife Conservation Society, Mongolia.

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.

Katoh, K., Takeuchi, K., Jiang, D., Nan, Y., and Kou, Z. (1998). Vegetation restoration by seasonal exclosure in the Kerqin Sandy Land, Inner Mongolia. *Plant Ecology* **139,** 133-144.

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

Lal, R. (1990). Soil erosion and land degradation: the global risks. In Advances in soil science (pp. 129-172). Springer, New York, NY.

Lemon, J. (2006). Plotrix: a package in the red light district of R. *R-News* **6,** 8-12.

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.

Li, Y.H. (1989) Impact of grazing on *Aneurolepidium chinense* steppe and *Stipa grandis* steppe. *Acta Oecologica* **10,** 31–46.

Li, C., Hao, X., Zhao, M., Han, G. and Williams, W. D. (2008). Influence of historic sheep grazing on vegetation and soil properties of a Desert Steppe in Inner Mongolia. *Agriculture, Ecosystems & Environment* **128,** 109-116.

Liang, C., Michalk, D.L. and Millar, G.D. (2002). The ecology and growth patterns of *Cleistogenes* species in degraded grasslands of eastern Inner Mongolia, China. *Journal of Applied Ecology* **39,** 584-594.

Liu, J., Zhu, X., Yu, F., Dong, M., Zhang, S. and Wang, R. (2003). Spatial heterogeneity of *Ulmus pumila* open forest ecosystem in Otindag sandy land. *Huan Jing Ke Xue* [环境科学, Environmental Science] **24,** 29-34.

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.

McIntyre, S. and Lavorel, S. (2007). A conceptual model of land use effects on the structure and function of herbaceous vegetation. *Agriculture, Ecosystems & Environment* **119,** 11-21.

Meir, E., Andelman, S. and Possingham, H.P. (2004). Does conservation-planning matter in a dynamic and uncertain world? **Ecology Letters 7,** 615-622.

Miehe, G.I, Schlütz, F., Miehe, S., Opgenoorth, L., Cermak, J., Samiya, R., Jäger, E.J. and Wesche, K. (2007) Mountain forest islands and Holocene environmental changes in Central Asia: a case study from the southern Gobi Altay, Mongolia. *Palaeogeography, Palaeoclimatology, Palaeoecology* **250,** 150-166.

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.

Olson, K., Avirmed, O., Jambal, S. and Strindberg, S. (2016). Monitoring the health of Elm trees (*Ulmus pumila*) in Khanbogd area; Appendix D: Annual monitoring report. Wildlife Conservation Society, Mongolia Program.

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.

Pei, S., Fu, H. and Wan, C. (2008). Changes in soil properties and vegetation following exclosure and grazing in degraded Alxa desert steppe of Inner Mongolia, China. *Agriculture, Ecosystems & Environment* **124,** 33-39.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Radnaakand, T. (2016). The flowers of the Mongolian Gobi Desert. Admon Publishing, Ulaanbaatar, Mongolia.

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

Sasaki, T., Okayasu, T., Takeuchi, K., Jamsran, U. and Jadambaa, S. (2005). Patterns of floristic composition under different grazing intensities in Bulgan, South Gobi, Mongolia. *Grassland Science* **51,** 235-242.

Schlacher, T.A., Schoeman, D.S., Jones, A.R., Dugan, J.E., Hubbard, D.M., Defeo, O., Peterson, C.H. Weston, M.A., Maslo, B., Olds, A.D., Scapini, F., Nel, R., Harrise, L.R., Lucrez, S., Lastra, M., Huijbers, C.M. and Connolly, R.M. (2014). Metrics to assess ecological condition, change, and impacts in sandy beach ecosystems. *Journal of Environmental Management* **144,** 322-335.

Shiping, W., Yonghong, L., Yanfen, W. and Yuanhong, H. (1998). The Succession of *Artemisia frigida* Rangeland and Multivariation Analysis Under Different Stocking Rates in Inner Mongolia. *Acta Agrestia Sinica* **4,** 8.

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.

Urbanek, S. (2014). jpeg: Read and write JPEG images. R package version 0.1-8. https://CRAN.R-project.org/package=jpeg

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.

Wang, Y.S. (1992). Vegetation dynamics of grazing succession in the *Stipa baicalensis* steppe in Northeastern China. *Vegetatio* **98,** 83–95.

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.

Yoshihara, Y., Chimeddorj, B., Buuveibaatar, B., Lhagvasuren, B. and Takatsuki, S. (2008). Effects of livestock grazing on pollination on a steppe in eastern Mongolia. *Biological Conservation* **141,** 2376-2386.

Zemmrich, A., Manthey, M., Zerbe, S. and Oyunchimeg, D. (2010). Driving environmental factors and the role of grazing in grassland communities: a comparative study along an altitudinal gradient in Western Mongolia. *Journal of Arid Environments* **74,** 1271-1280.

Zhang, P., Yang, J., Zhao, L., Bao, S. and Song, B. (2011). Effect of *Caragana tibetica nebkhas* on sand entrapment and fertile islands in steppe–desert ecotones on the Inner Mongolia Plateau, China. *Plant and Soil* **347,** 79-90.

Zhao, H. L., Zhao, X. Y., Zhou, R. L., Zhang, T. H. and Drake, S. (2005). Desertification processes due to heavy grazing in sandy rangeland, Inner Mongolia. *Journal of Arid Environments* **62,** 309-319.

Zhao, H.L., Zhou, R.L., Su, Y.Z., Zhang, H., Zhao, L.Y. and Drake, S. (2007). Shrub facilitation of desert land restoration in the Horqin Sand Land of Inner Mongolia. *Ecological Engineering* **31,** 1-8.

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 A Field site descriptions

Summary of the DESERT STEPPE field sites. All sites were assessed on the 7th August 2017. Species contributing >0.5% absolute cover are listed, in descending order of cover.

Site No.	Latitude	Longitude	Grazing excl.?	Summary description
DS 126	43.46877	106.90668	No	Undulating lower slopes. Moderately high grass cover. *Cleistogenes squarrosa* (Poaceae) *Stipa glareosa* (Poaceae), *Ajania achilleoides* (Asteraceae)
DS 127	43.87444	106.91991	No	Undulating plain. Targeted by WCS as a very intact example of Desert Steppe. Very high cover and richness of grasses, forbs and shrubs. *Stipa glareosa*, *Allium mongolicum* (Alliaceae), *Artemisia xerophytica* (Asteraceae), *Cleistogenes squarrosa*, *Salsola collina* (Chenopodiaceae), *Ajania achilleoides*, *Caragana korshinskii* (Fabaceae), *Asparagus gobicus* (Asparagaceae), *Corispermum mongolicum* (Amaranthaceae), *Convolvulus ammanii* (Convolvulaceae), *Krascheninnikovia ceratoides* (Chenopodiaceae).
DS 128	43.86565	106.91982	No	Undulating plain, sandy soil. Targeted by WCS as an intact example of Desert Steppe. High cover and richness of grasses, forbs and shrubs. *Stipa glareosa*, *Cleistogenes squarrosa*, *Allium mongolicum*, *Corispermum mongolicum*, *Salsola collina*, *Caragana korshinskii*, *Caragana leucophylla*, *Artemisia xerophytica*, *Caryopteris mongolica (Verbenaceae)*.
DS 129	43.86576	106.91919	No	Undulating plain. Targeted by WCS as an intact example of Desert Steppe. High cover and richness of grasses, forbs and shrubs. *Stipa glareosa*, *Cleistogenes squarrosa*, *Artemisia xerophytica*, *Allium mongolicum*, *Ajania achilleoides*, *Asparagus gobicus*, *Salsola collina*.
DS 130	43.85629	106.91959	No	Undulating plain. Disturbed by numerous parallel vehicle tracks. Very low cover of grasses, substantial bare ground. *Ajania achilleoides, Artemisia xerophytica, Stipa glareosa, Cleistogenes squarrosa Carex pediformis* (Cyperaceae), *Salsola collina, Corispermum mongolicum.*
DS 131	43.77819	106.86852	No	Undulating plain near stony rise. Immediately adjacent to winter camp. Heavily disturbed, largely bare ground, animal dung and disturbance and nutrient-loving annual and perennial forbs. *Atriplex sibirica* (Chenopodiaceae), *Halogeton glomeratus* (Chenopodiaceae), *Peganum nigellastrum* (Zygophyllaceae)..
DS 132	43.73945	106.81374	No	Undulating plain, sandy soil. An unusual example of Desert Steppe with substantial shrub cover, along with moderate grass and forb cover. Low richness. *Brachanthemum gobicum* (Asteraceae), *Stipa glareosa, Allium mongolicum.*
DS 133	43.74069	106.81457	No	Undulating plain, sandy soil. Moderate grass, shrub and forb cover. *Stipa glareosa, Allium polyrhizum, Ajana achilleoides, Allium mongolicum.*

Summary of the SEMI-DESERT field sites. Sites 126 – 134 were assessed on the 6th August 2017, site SD 135 was assessed on the 7th August 2017. Species contributing >0.5% absolute cover are listed, in descending order of cover.

Site No.	Latitude	Longitude	Grazing excl.?	Summary description
SD 126	43.17705	106.79336	Yes	Low sandy plain, with sand hummocks and bare ground. 100 m from a well, but inside the WCS F4-100 exclosure fence since 2015. Shrub dominated. *Nitraria sibirica* (Nitrariaceae), *Peganum nigellastrum* (Zygophyllaceae), *Reaumuria soongorica* (Tamaricaceae), *Salsola passerina* (Chenopodiaceae), *Allium polyrhizum* (Alliaceae), *Aristida hegmannii* (Poaceae).
SD 127	43.17711	106.79328	No	Low sandy plain, with sand hummocks and bare ground. 100 m from a well, and unprotected from grazing. Immediately adjacent to SD 126. *Nitraria sibirica*, *Peganum nigellastrum*, *Allium polyrhizum*, *Allium mongolicum*.
SD 128	43.16888	106.78502	Yes	Slight rise on a stony plain. 1000 m from a well, but inside the F4-1000 exclosure fence since 2015. Relatively high cover of grasses and forbs. *Salsola passerina*, *Allium polyrhizum*, *Anabasis brevifolia*, *Allium mongolicum*, *Potaninia mongolica* (Rosaceae).
SD 129	43.169081	106.78475	No	Slight rise on a stony plain. 1000 m from a well, and unprotected from grazing. Immediately adjacent to SD 128. Moderate cover of forbs. *Allium polyrhizum*, *Salsola passerina*.
SD 130	43.16806	106.78154	No	Undulating plain. Substantially disturbed by recent powerline construction, with disturbed soil, wheel tracks and very low vegetation cover. *Allium polyrhizum*.
SD 131	43.27392	106.64359	No	Undulating plain. High cover and richness of grasses, shrubs and forbs. *Zygophyllum xantoxylon*, *Stipa glareosa*, *Potaninia mongolica*, *Allium polyrhizum*, *Allium mongolicum*, *Cleistogenes squarrosa*, *Anabasis brevifolia*.
SD 132	43.27179	106.73057	No	Low sandy plain, with sand hummocks and bare ground, but moerate grass cover. *Nitraria sibirica*, *Stipa glareosa*, *Allium mongolicum*, *Cleistogenes squarrosa*
SD 133	43.26588	106.76445	No	Low sandy plain, with very high forb (Allium) cover. *Allium polyrrhizum*, *Anabsis brevifolia*, *Allium mongolicum*, *Aristida hegmannii*, *Reaumuria soongorica*.
SD 134	43.30394	106.84744	No	Sandy plain with high levels of bare sand and apparent erosion. Low vegetation cover. *Nitraria sibirica*, *Aristida hegmannii*.
SD 135	43.41522	106.81415	Yes	Lower valley floor. 1000 m from well, but inside the WCS F5-1000 exclosure fence since 2015. High cover of grasses and shrubs. *Potaninia mongolica*, *Cleistogenes squarrosa*, *Allium mongolicum*, *Anabsis brevifolia*.

Summary of the TRUE DESERT field sites. Sites TD 127 – TD 129 were assessed on the 3rd August 2017, the remainder on the 4th August 2017. Species contributing >0.5% absolute cover are listed, in descending order of cover.

Site No.	Latitude	Longitude	Grazing excl.?	Summary description
TD 127	42.705778	107.708889	Yes	Lower valley floor, 100 m from a well, but within the WCS F3-100 exclosure fence since 2015. Sparse, low shrubland. *Peganum nigellastrum* (Zygophyllaceae).
128	42.698694	107.70109	Yes	Lower valley floor, 1000 m from a well, but within the WCS F3-1000 exclosure fence since 2015. Moderately sparse shrubland. *Haloxylon ammodendron* (Chenopodiaceae), *Salsola passerina* (Chenopodiaceae) and *Reaumuria soongorica* (Tamaricaceae).
129	42.701767	107.72556	No	Lower valley floor, immediately adjacent to river bed. Moderately dense and diverse shrubland. *Zygophyllum xantoxylon* (Zygophyllaceae), *Artemisia xanthrochroa* (Asteraceae) and *Haloxylon ammodendron*.
130	42.711009	107.7183	No	Rocky mid-slope, Low but moderately dense shrub- and forb-land. *Anabasis brevifolia* (Chenopodiaceae), *Salsola passerina*, *S. laricifolia* and *Reaumuria soongorica*, with a small amount of grass cover (*Cleistogenes squarrosa*).
131	42.72256	107.77454	No	Lower valley floor, immediately adjacent to river bed. Moderately dense and diverse shrubland. *Asterothamnus centrali-asiaticus* (Asteraceae)*, Reaumuria soongorica, Zygophyllum xantoxylon, Caragana korshinskii (Fabaceae) and Artemisia xanthrochroa.*
132	42.72162	107.77512	No	Lower valley floor, slightly further from river bed than site 131. Moderately dense and diverse shrubland. *Kalidium gracile* (Chenopodiaceae), *Salsola passerina*, *Reaumuria soongorica*, *Nitraria sibirica*.
133	42.66991	107.80882	No	Undulating valley slope. Nominated by WCS as an exemplar of intact True Desert. Diverse shrubland. *Reaumuria soongorica*, *Zygophyllum xantoxylon*, *Salsola passerina*, *S. laricifolia*, *Anabsis brevifolia*, *Sympegma regelii* (Chenopodiaceae) and *Potaninia mongolica* (Rosaceae).
134	42.67288	107.80758	No	Undulating valley slope, adjacent to river bed and protected area well. Substantially disturbed by wheel tracks. Sparse shrubland. *Salsola passerina, Reaumuria soongorica, Nitraria sibirica.*
135	42.77596	107.74918	No	Mid-slope. Immediately next to a winter camp. Heavily disturbed, largely bare ground and animal dung. No species contributed >0.5% cover, only two species present (*Reaumuria soongorica*, *Nitraria sibirica*).
136	42.77766	107.74074	No	Mid-slope, near a dis-used well. Sparse shrubland with low diversity. *Reaumuria soongorica*, *Nitraria sibirica*, *Kalidium gracile*.

Appendix B Draft field sampling protocols for Saxaul and Elm Forest

Saxaul

The Saxaul ecosystem is not adequately addressed by the method described and tested in this report (Methods 2.4), because it is addressed by variables that only apply to Saxaul. The following adapted method is recommended for trialling in Saxaul.

Sampling plot design

The sampling plot should be set up as described in Methods 2.4.

Sampling vegetation and litter cover

Sampling of cover should be undertaken as described in Methods 2.4, using 4 point intercept lines and 480 sampling points. The 'Cover of Haloxylon' can be derived from the pointing data.

Sampling species richness

Sampling species richness should be undertaken as described in methods 2.4, using a ten-minute timed search of the plot.

Sampling the maximum height of roots exposed by soil loss

Sampling the maximum height of roots exposed by soil loss should be undertaken as described in methods 2.4.

Measuring the density of 'Large Haloxylon'

The 10-minute search of the plot already required to quantify richness should also include a count of all *Haloxylon* plants >1.5 m tall. This count provides the measurement for 'Density of large *Haloxylon* plants'.

Elm Forest

There are three reasons why the plot method described and tested in this report (Methods 2.4) is not suited to use in Elm Forest:

- The Elm Forest ecosystem requires the measurement of several additional variables,
- The Elm Forest ecosystem often occurs in narrow linear strips which may be less than 30m across, such that the recommended 30 x 30 m plot does not fit within the ecosystem,
- The Siberian Elm trees occur patchily, at a scale larger than the patterning at the 900 m² plot.

Sampling plot

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 m²) 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 m²) 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.

Measuring the density of adult Ulmus *(per 900m²)*

The following method should be used to calculate the density of adult *Ulmus*.

- The number of adult *Ulmus* should be counted within the 900 m² plot (regardless of its shape).
- If there are 5 or more adult *Ulmus* counted, no further counting is required, and the 'density' is simply the number counted.
- If fewer than 5 adult *Ulmus* are found in the plot (which is very frequently the case), then an ever-expanding radius around the plot (centred on 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:

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

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

- 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:

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

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

◻ **E** (Area of Elm Forest ecosystem within radius R)

◻ Not Elm Forest

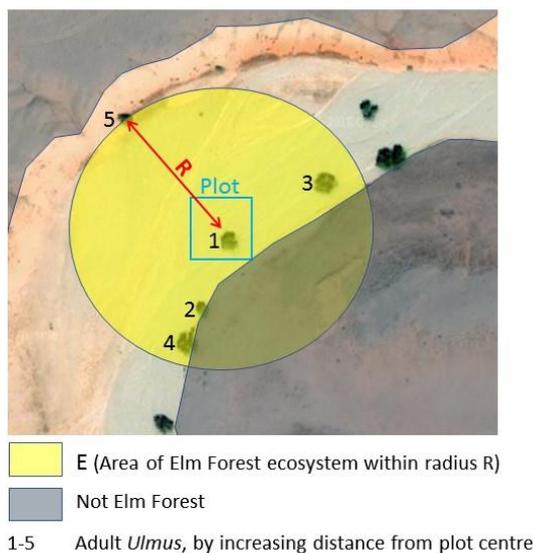1-5   Adult *Ulmus*, by increasing distance from plot centre

Figure 1. Diagram showing the method for calculating *Ulmus* density around the plot.

Measuring the density of juvenile and sapling Ulmus

The densities of 'Juvenile *Ulmus* (suppressed)' and 'Juvenile *Ulmus* (escaped)' should be measured as for adult *Ulmus*.

Sampling vegetation cover

- If a 30 x 30 m plot was used, cover should be measured exactly as described in Methods 2.4.
- 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' and must be included in the point data.

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.

Appendix C Questionnaire used to characterise stakeholders

Introduction

The questionnaire used to quantify the expertise and affiliations of the stakeholders is reproduced below. It has been modified in format to fit the style of this report. The categories used to sort the questions have also been added; these were not included on the questionnaire presented to the stakeholders.

About you

Name: _____

Gender: _____

Age: _____

Please complete the following table. You may fill in multiple boxes.

Does this describe you….	Currently…? (Y/ N)	For how long during your life? (years)
Mongolian resident		
Deriving primary income from herding livestock		
University academic		
Government employee		
Employee of environmental NGO		
Employee in mining industry		
Member of a natural history club or society		
Consultant ecologist		

If your income is largely from herding livestock, how many of the following animals would you have managed, in a typical year over the last ten years?

Livestock type	Number
Sheep	
Goats	
Camels	
Cattle	

Your knowledge

Please complete the table below. You may fill in multiple boxes.
- Write 1 if you consider yourself an <u>expert</u> / <u>very experienced</u> in the topic area described,
- Write 2 if you consider yourself not an expert, but still quite knowledgeable,
- Leave blank if you have little or no knowledge or expertise.

Question		Question category (not included on questionnaire as presented to stakeholders)
Raising sheep		Pastoralism
Raising goats		Pastoralism
Raising cattle		Pastoralism
Raising camels		Pastoralism
Hunting wild game		Pastoralism
Horse husbandry		Pastoralism
Animal illness and veterinary care		Pastoralism
Management of vegetation with grazing		Pastoralism
The ecology or natural history of vascular plants		Specialist - Botany
The ecology or natural history of non-vascular plants		Specialist - Botany
Scientific sampling of vegetation		Specialist - Botany
Invasive / pest plant species		Specialist - Botany
The use of plants for medicine		Specialist - Botany
Ecology or natural history of Birds		Specialist - Wildlife
Ecology or natural history of Mammals		Specialist - Wildlife
Ecology or natural history of Reptiles & amphibians		Specialist - Wildlife
Ecology or natural history of Invertebrates		Specialist - Wildlife
Invasive / pest animal species		Specialist - Wildlife
Environmental Policy		Conservation policy and practice
Long term environmental change (1000s of years)		Conservation policy and practice
Mathematics and statistics		Conservation policy and practice
Management of vegetation with fire		Conservation policy and practice
Geological processes		Conservation policy and practice
Soil processes (including erosion)		Conservation policy and practice
Navigation across the landscape		Conservation policy and practice

Appendix D Variable combinations which result in the highest scores

The table below shows examples of variable combinations which represent the maximum condition score (100, after re-scaling) for each ecosystem.

For all systems, multiple combinations will return the highest score (i.e. there is not a single benchmark). Table 1 shows only the site with the minimum cover and minimum richness values which returns a 100 score (with the constraint that the internal logic of the nested and related variables is upheld). May sites with higher cover or richness values will also return a score of 100.

Variable	True Desert	Semi Desert	Desert Steppe	Saxaul	Elm Forest
Cover of all vegetation	61	95	76	61	91
Cover of all shrubs	48	56	11	61	6
Cover of all succulent shrubs	16	56	0	61	NA
Cover of all Fabaceous shrubs	6	4	3	NA	NA
Cover of all Artemisia species	3	1	1	NA	NA
Cover of Haloxylon	NA	NA	NA	61	NA
Cover of all perennial grasses and sedges	1	16	44	0	11
Cover of all annual grasses	4	4	0	NA	NA
Cover of all perennial forbs	6	16	16	0	0
Cover of all annual forbs	2	3	5	NA	NA
Cover of litter	6	8	4	0	NA
Species richness of all shrubs	10	4	11	1	4
Species richness of all grasses and sedges	7	10	9	0	0
Species richness of all forbs	9	7	14	0	5
Density of large Haloxylon	NA	NA	NA	4	NA
Max height of roots exposed by soil loss	0	0	0	0	NA
Cover of Ulmus	NA	NA	NA	NA	74
Density of adult Ulmus	NA	NA	NA	NA	26
Density of juvenile Ulmus (escaped)	NA	NA	NA	NA	5
Density of juvenile Ulmus (suppressed)	NA	NA	NA	NA	5
Density of sapling Ulmus	NA	NA	NA	NA	0

Appendix E The relative importance of simplified variables

The relative importance of each of the variables was explored in Results 3.4, by comparing the final model for each ecosystem to 1) a set of models with each variable withheld from the model, and 2) a set of models using only each single variable in isolation. We judged the contribution of each variable by the change in r^2, compared to the final model.

It was noted that many of the variables are nested or overlapping, making the interpretation of some basic patterns difficult. Here, we provide a clearer illustration of the contribution of different ecosystem elements to the concept of condition.

We created a simplified set of variables which were not nested. These variables were simply derived from the existing variables (i.e. It can be assumed that total richness of non-woody species = Richness of forbs + Richness of Grasses and sedges). The simplified variable set was created to match the variables that are likely to be available via remote sensing. Table 1 shows the simplified, derived variable set.

We then repeated the analysis with these simplified variables. Figures 1 to 5 show the results.

Table 1. The simplified variable sets used to explore the contribution of different elements to the concept of condition.

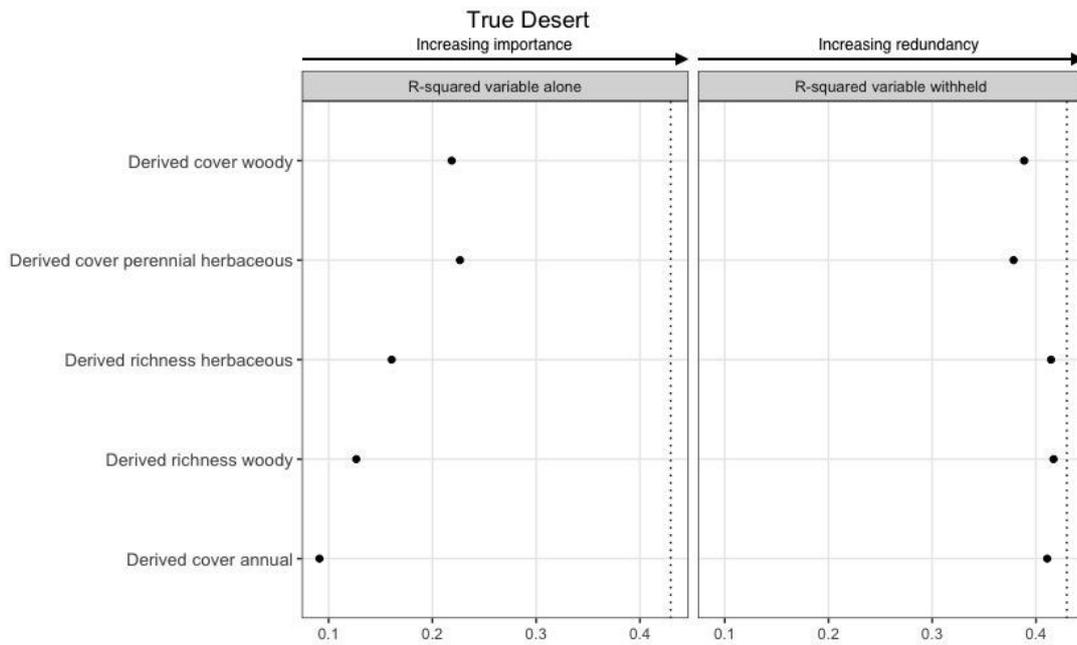| Derived variable	Derivation
Cover of woody perennials	Cover of all shrubs + Cover of *Ulmus pumila* (where relevant)
Cover of herbaceous perennials	Cover of all perennial forbs + Cover of all perennial grasses and sedges
Cover of annual vegetation	Cover of all annual forbs + Cover of all annual grasses
Species richness of woody species	Species richness of all shrubs + One additional species where *Ulmus* is present (where relevant)
Species richness of herbaceous species	Species richness of all forbs + Species richness of all grasses and sedges
Cover of litter	As in full model

Figure 1. The relative importance of the derived (simplified) variables in the metric for True Desert, as judged by their ability to predict score alone, and the impact of omitting them from the metric (30-tree ensemble model). The left panel shows the ability of each variable to predict score (points showing r^2), compared to the full model with all variables (dotted line). The right panel shows the impact on r^2 of omitting each variable from the full model (dotted line).
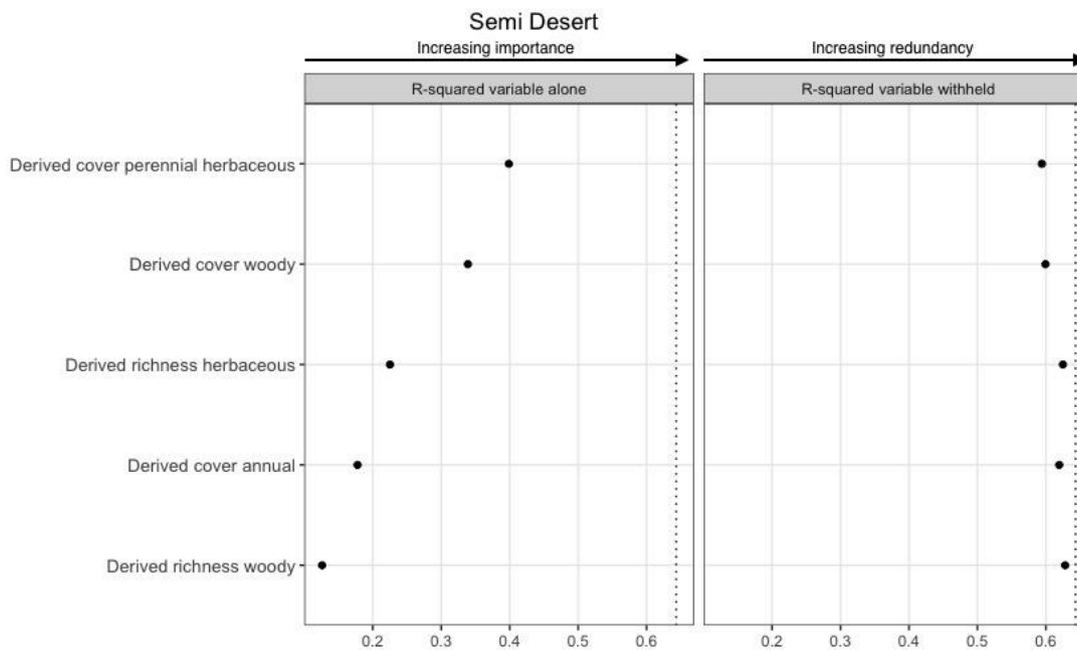

Figure 2. The relative importance of the derived (simplified) variables in the metric for Semi Desert, as judged by their ability to predict score alone, and the impact of omitting them from the metric (30-tree ensemble model). The left panel shows the ability of each variable to predict score (points showing r^2), compared to the full model with all variables (dotted line). The right panel shows the impact on r^2 of omitting each variable from the full model (dotted line).
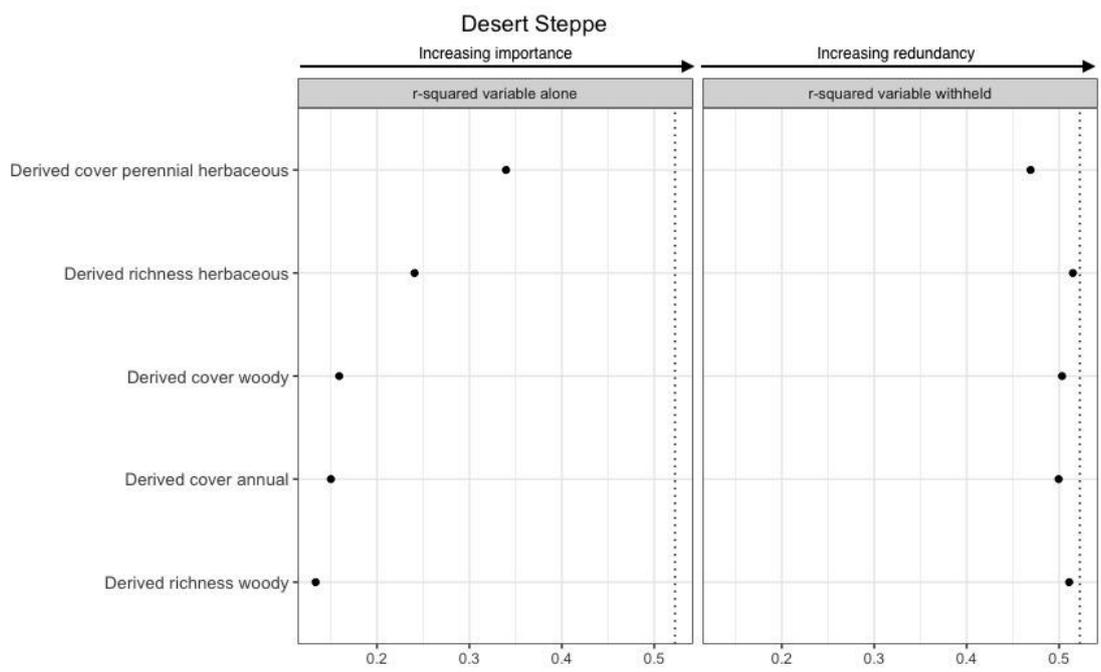
Figure 3. The relative importance of the derived (simplified) variables in the metric for Desert Steppe, as judged by their ability to predict score alone, and the impact of omitting them from the metric (30-tree ensemble model). The left panel shows the ability of each variable to predict score (points showing r^2), compared to the full model with all variables (dotted line). The right panel shows the impact on r^2 of omitting each variable from the full model (dotted line).
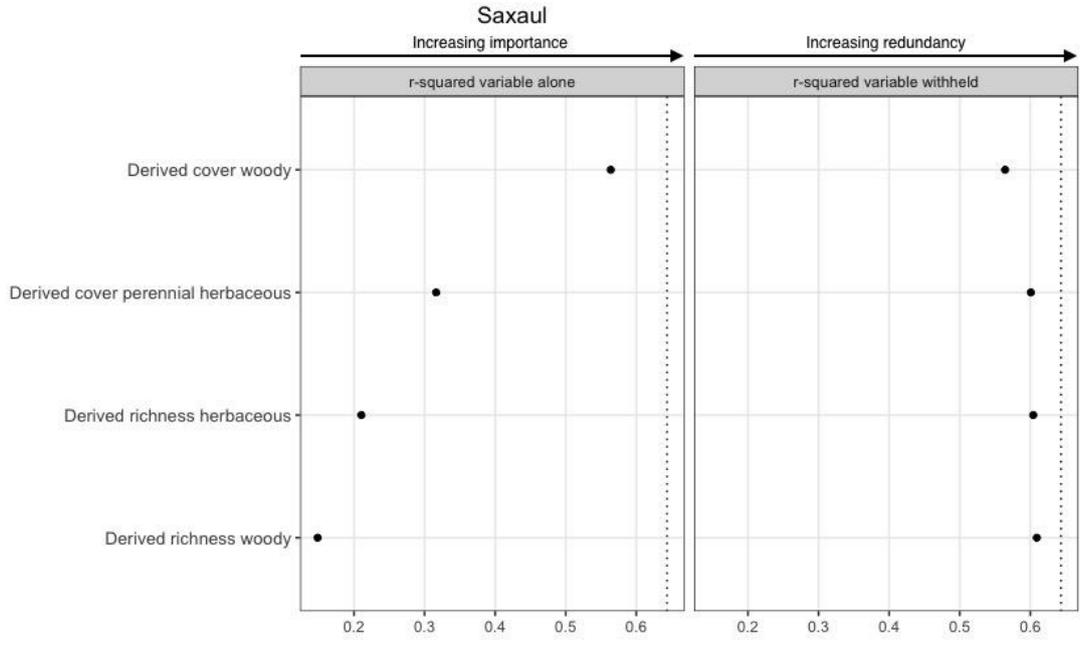

Figure 4. The relative importance of the derived (simplified) variables in the metric for Saxaul, as judged by their ability to predict score alone, and the impact of omitting them from the metric (30-tree ensemble model). The left panel shows the ability of each variable to predict score (points showing r^2), compared to the full model with all variables (dotted line). The right panel shows the impact on r^2 of omitting each variable from the full model (dotted line).
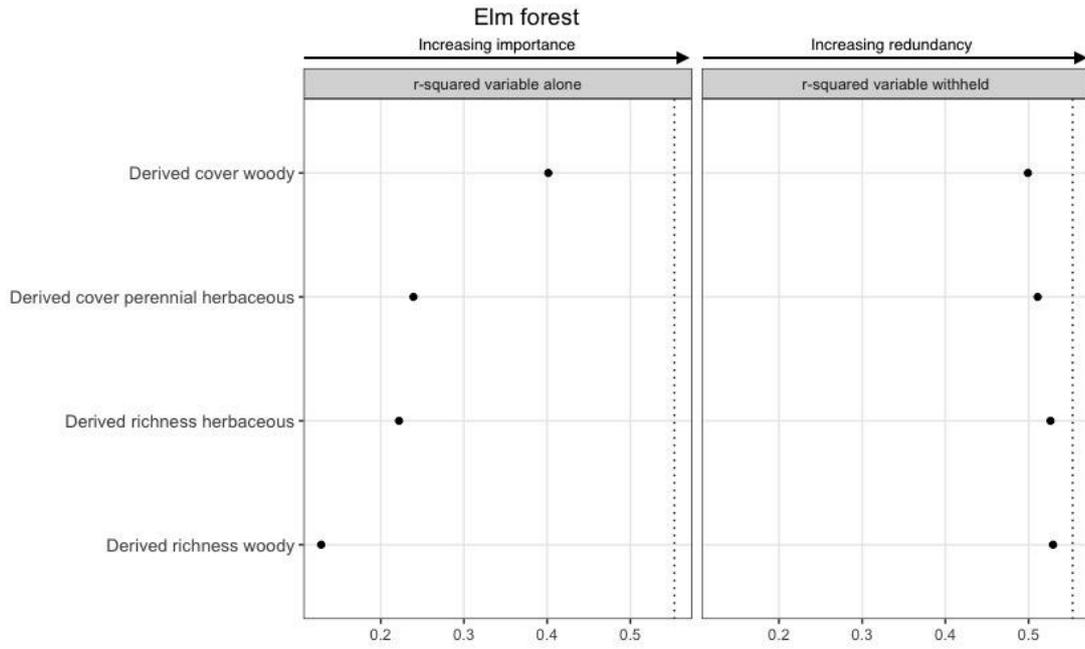
Figure 5. The relative importance of the derived (simplified) variables in the metric for Elm Forest, as judged by their ability to predict score alone, and the impact of omitting them from the metric (30-tree ensemble model). The left panel shows the ability of each variable to predict score (points showing r^2), compared to the full model with all variables (dotted line). The right panel shows the impact on r^2 of omitting each variable from the full model (dotted line).

Appendix F Raw evaluation data

The figures below summarise the raw evaluation data used to create the final metrics. They include data from all workshops, but do not include the data removed during the outlier treatment process.

Each ecosystem is represented by a separate panel.

Each card is arranged along the horizontal axis by its median evaluation score. The vertical axis shows the individual evaluations made by stakeholders for each site, so that each site is represented by a vertical set of points.

The hollow circles represent stakeholder evaluations. The black points represent the predictions of the final metrics.