A random forest approach for predicting the presence of *Echinococcus multilocularis* intermediate host *Ochotona spp.* presence in relation to landscape characteristics in western China

Christopher G. Marston[a]*, F. Mark Danson[b], Richard P. Armitage[b], Patrick Giraudoux[c], David R.J. Pleydell[d], Qian Wang[e], Jiamin Qui[e], and Philip S. Craig[b]

[a] School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool L3 3AF, UK.

[b] School of Environment and Life Sciences, University of Salford, Manchester M5 4WT, UK.

[c] Department of Chrono-environment and Institut Universitaire de France, University of Franche-Comt_e, Place Leclerc, 25030 Besançon Cedex, France.

[d] INRA, UMR-1351 CMAEE, Domaine Duclos, Prise D'eau, 97122 Petit Bourg, Guadeloupe.

[e] Sichuan Centers for Disease Control and Prevention, Chengdu 610041, Sichuan, China.

* Corresponding author

Abstract

Understanding distribution patterns of hosts implicated in the transmission of zoonotic disease remains a key goal of parasitology. Here, random forests are employed to model spatial patterns of the presence of the plateau pika (*Ochotona spp.*) small mammal intermediate host for the parasitic tapeworm *Echinococcus multilocularis* which is responsible for a significant burden of human zoonoses in western China. Landsat ETM+ satellite imagery and digital elevation model data were utilized to generate quantified measures of environmental characteristics across a study area in Sichuan Province, China. Land cover maps were generated identifying the distribution of specific land cover types, with landscape metrics employed to describe the spatial organisation of land cover patches. Random forests were used to model spatial patterns of *Ochotona spp.* presence, enabling the relative importance of the environmental characteristics in relation to *Ochotona spp.* presence to be ranked. An index of habitat aggregation was identified as the most important variable in influencing *Ochotona spp.* presence, with area of degraded grassland the most important land cover class variable. 71% of the variance in *Ochotona spp.* presence was explained, with a 90.98% accuracy rate as determined by 'out-of-bag' error assessment. Identification of the environmental characteristics influencing *Ochotona spp.* presence enables us to better understand distribution patterns of hosts implicated in the transmission of Em. The predictive mapping of this Em host enables the

identification of human populations at increased risk of infection, enabling preventative strategies to be adopted.

Keywords: *Echinococcus multilocularis*, *Ochotona*, remote sensing, random forests, landscape metrics, classification.

1. Introduction

Human Alveolar Echinococcosis (HAE), caused by the parasitic tapeworm *Echinococcus multilocularis* (Em), is an emerging pathogen for which increased prevalence and range expansion is documented in many regions of the northern hemisphere (Eckert, 1996; Eckert *et al.*, 2001). It is a highly pathogenic zoonosis with over 94% mortality in untreated patients ten years after diagnosis (Wang *et al.*, 2010), and is increasingly recognised as a major population health problem (Zhang *et al.*, 2014). The known Em range includes Europe, North America, Japan, the former USSR, Central Asia and China where new foci are being discovered (Wang *et al.*, 2001; Giraudoux *et al.*, 2013a), with prevalence rates of greater than 10% observed in Gansu and Sichuan provinces, China (Craig *et al.*, 1992; Li *et al.*, 2010). The spatial distribution of Em is highly variable, with significant regional and local differences in parasite prevalence resulting in patchy distributions generally not reflected in Em and HAE distribution maps (Eckert *et al.*, 2001; Giraudoux *et al.*, 2006; 2013a).

The Em transmission cycle is based on the predator-prey relationships between canid definitive hosts such as fox, coyote and wolf and small mammal intermediate hosts (Rausch, 1995; Eckert *et al.*, 2001). Within a definitive host adult tapeworms produce eggs at regular intervals which are shed in faeces, contaminating the environment (Raoul *et al.*, 2001). The parasite lifecycle then undergoes a free-egg stage, with intermediate hosts infected through oral ingestion of eggs when feeding (Eckert, 1996). The transmission cycle is completed when definitive hosts are infected by predating infected intermediate hosts. Em exploits a large number of intermediate host species (>40) (Eckert *et al.*, 2001; Giraudoux *et al.*, 2013b), however the epidemiological importance of these hosts varies (Rausch, 1995).

Domestic dogs can also be infected and, due to their close contact with human populations, are a significant infection risk to humans (Rausch, 1995; Moss *et al.*, 2013; Zhang *et al.*, 2014) via accidental ingestion of Em eggs. Prevalence rates of Em infection in domestic dogs of up to 33% are recorded in Tibetan communities of western Sichuan Province, China (Budke *et al.*, 2005), with Craig *et al.* (2000) and Wang *et al.* (2001) identifying owned dogs as a major transmission source to humans in Gansu Province, and the eastern Tibetan plateau, China, respectively (Wang *et al.,* 2010).

Dog re-infection studies in Sichuan Province, China, suggest that domestic dog populations are quickly re-infected by Em, and may contribute to an active peri-domestic transmission cycle (Giraudoux *et al.,* 2013a; Moss *et al.,* 2013). Wang *et al.* (2010) also found that Em worm burden in dogs exhibited a statistically significant relationship to maximum burrow densities of a key Em intermediate host, the plateau pika (*Ochotona spp.*) in the surrounding landscape in Shiqu County, Ganze Tibetan Autonomous Prefecture, China. This study failed to identify significant relationships between dog worm burden and burrow density of another potential Em small mammal intermediate host present in this region, *Microtus spp.*, thus suggesting that the rapid Em re-infection rates in domestic dogs, shown by Moss *et al.* (2013), is probably linked to surrounding high densities of *Ochotona spp.*

Small mammal species often exhibit specific preferences for optimal habitats, with species distributions influenced by the locations of these key habitats (Raoul *et al.,* 2008). Small mammal populations are shown to respond to optimal habitat availability, particularly the ratio of optimal habitat to total land area (Giraudoux *et al.,* 2003; Pleydell *et al.,* 2008). Consequently, landscape change is known to affect the population dynamics of wild mammals (Lidicker, 1995), with increases in the optimal habitat proportions correlated with population outbreaks of *Microtus arvalis* and *Arvicola terrestris* in France (Giraudoux *et al.,* 1997), and *M. limnophilus* and *Cricetulus longicaudatus* in south Gansu, China (Giraudoux *et al.,* 1998; Craig *et al.,* 2000). This process is hypothesised to be significant for Em transmission (Giraudoux *et al.,* 1997), so that pathogen transmission may vary through time and space due to landscape modification. Elsewhere in China, small mammal spatial distributions are shown to be modified by landscape disturbances such as deforestation in Gansu (Giraudoux *et al.,* 1998), afforestation in Ningxia (Raoul *et al.,* 2008), and overgrazing and fencing practices on the Tibetan plateau (Wang *et al.,* 2004; Raoul *et al.,* 2006).

Pastureland degradation due to overgrazing has also been linked to increased small mammal densities, for example *Ochotona spp.*, *Microtus spp.*, *Cricetulus kamensis* and *Myospalax baileyi* (Raoul *et al.,* 2006) on the eastern Tibetan plateau, China, where HAE is endemic (Wang *et al.,* 2004; Li *et al.,* 2010). In Shiqu county, China, grass height was negatively related to *Ochotona curzoniae* burrow abundance suggesting that overgrazing in this area increased abundance of this species (Wang *et al.,* 2010). With high *Ochotona spp.* densities significantly associated with infection of domestic dogs (Wang *et al.,* 2010), foxes and humans (Craig *et al.,* 2000), pastureland degradation resulting from overgrazing could prove a significant driver of increased human Em incidence in this region.

Previous studies of Em and landscape using remote sensing techniques in southern Gansu Province, China, identified strong links between landscape composition and HAE prevalence (Craig *et al.,* 2000; Giraudoux *et al.,* 2003; Danson *et al.,* 2004). This suggested that grassland and tree/shrub habitats capable of sustaining cyclically high populations of susceptible intermediate hosts were key spatial determinants of Em transmission (Danson *et al.,* 2003), and indicated that landscape composition could provide a useful predictor of Em and HAE (Pleydell *et al.,* 2008; Giraudoux *et al.,* 2013b).

On the Tibetan plateau the black-lipped pika or plateau pika (*Ochotona curzoniae*) is thought to be one of the principal intermediate hosts in the Em transmission cycle (Giraudoux *et al.,* 2006; Zhang *et al.,* 2014). Pika are social mammals that tend to be spatially clumped (Arthur *et al.,* 2008), with average individual home range sizes for *Ochotona curzoniae* of 1,375 ± 206m^2 (Smith & Gao, 1991) and population densities ranging from 100 to 400 pikas ha^{-1} on the Tibetan plateau (Jiapeng *et al.,* 2013). Given the contrast between the biomass of *Ochotona spp*. (high) to *Microtus spp.* (low) in Shiqu county (Wang *et al.,* 2010), the role of *Ochotona spp*. in transmission to dogs may be highly significant (Giraudoux *et al.,* 2013a).

The research presented here builds on this previous work and investigates a critical phase of the Em transmission cycle, where the parasite is carried by small mammal intermediate hosts. Satellite remote sensing and *in-situ* ecological datasets are used to investigate the spatial relationship between *Ochotona spp.* presence and specific landscape characteristics to identify and better understand these links using random forests. Key landscape variables hypothesised to influence *Ochotona spp.* presence, and their relative importance, are determined and used to map *Ochotona spp.* presence over a broader geographical area. The hypotheses addressed are: (1) *Ochotona spp*. presence is statistically related to key environmental variables which can be used to predict species presence over larger areas; and (2) In the geographical area of interest, *Ochotona spp.* presence is specifically linked to areas of degraded grassland.

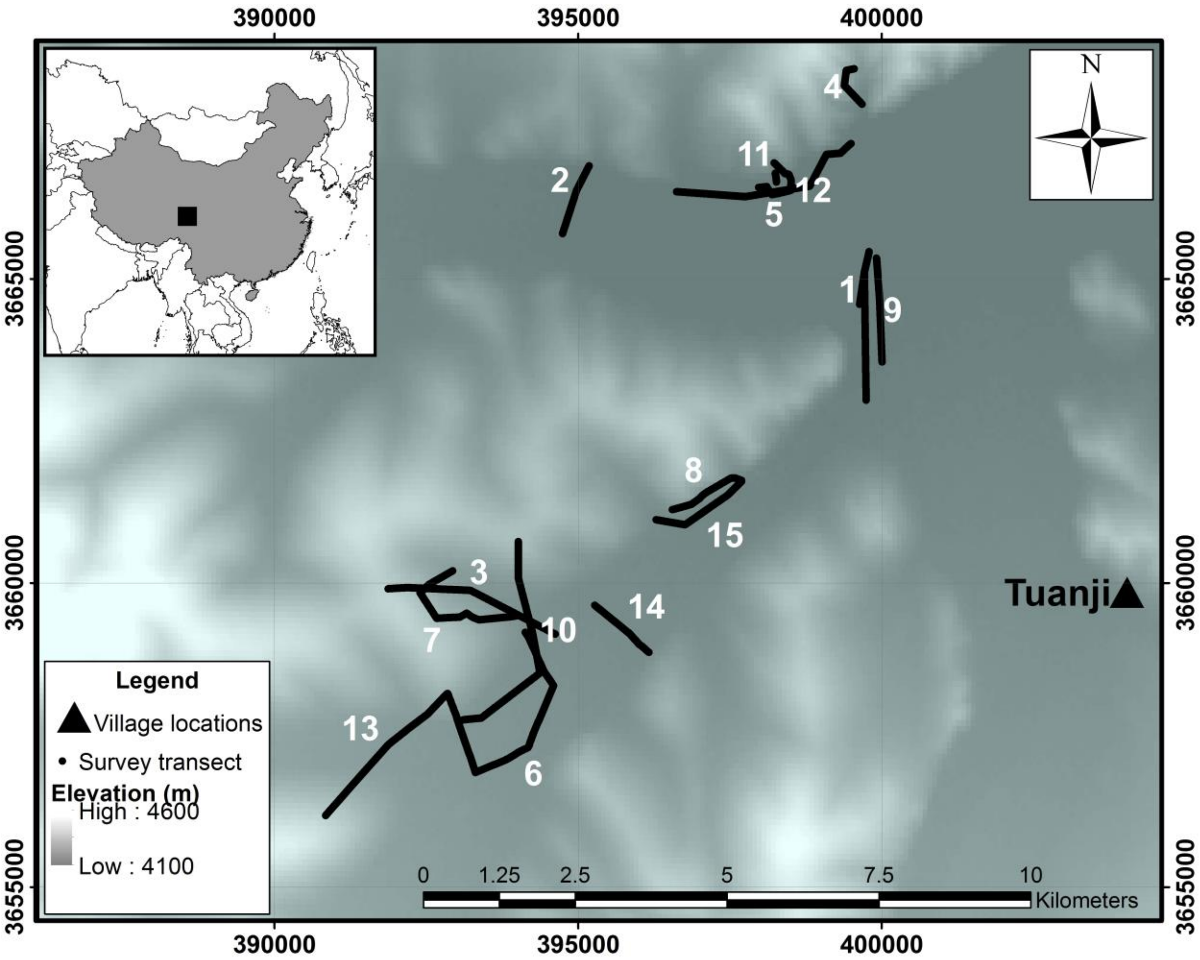
To identify the key landscape features influencing *Ochotona spp.* presence, random forest (RF) analysis methods are highly appropriate. RF are an ensemble learning technique developed by Breiman (2001) based on a combination of a large set of classification and regression trees. They are well-suited to handling large datasets with correlated predictor variables (Svetnik *et al.*, 2003), handle a variety of data types (Duro *et al.*, 2012), are non-parametric (Strobl *et al.*, 2008), make no assumption of independence concerning the data being analysed (Perdiguero-Alonso *et al.*, 2008), and are robust to outliers, noise and over-fitting (Breiman, 2001). They have been used as analytical tools for a variety of applications (Svetnik *et al.*, 2003) including remote sensing analysis (Duro *et al.*, 2012; Abdel-Rahman *et al.*, 2013) and parasitological studies (Perdiguero-Alonso *et al.*, 2008).

Random forest algorithms employ recursive partitioning to generate multiple decision trees and average individual tree predictions across the entire forest (Duro *et al.*, 2012; Abdel-Rahman *et al.*, 2013). Each iteration uses two-thirds of the data to train the RF while the remaining third, the 'out of bag' (OOB) samples, are retained for testing the prediction error of the RF (Duro *et al.*, 2012). The OOB error estimate also generates variable importance measures by comparing increases in OOB error when that variable is randomly permuted while all others are left unchanged, enabling ranking of the importance of individual variables (Abdel-Rahman *et al.*, 2013). The OOB error estimate removes the need for cross-validation via a set-aside test dataset (Perdiguero-Alonso *et al.*, 2008).

2. Materials and methods

The research focused on a study area near the town of Tuanji, Shiqu county, Ganze Tibetan Autonomous Prefecture, Sichuan Province, China (Fig 1). This is located on the eastern edge of the Tibetan plateau (Lat 33.04° Lon 97.97°) at altitudes between 4000-4300 metres, and dominated by semi-natural grassland. Although above the tree line, variation in herb and shrub vegetation produces a variety of land cover types. Heavy grazing by yak in this region has resulted in extensive areas of degraded grassland. Within Shiqu county, at least three townships have been found to be local *foci* for HAE, showing that a transmission cycle is, or has been active here (Wang *et al.*, 2001).

Figure 1. Study site map with numbered survey transects and SRTM DEM (USGS, 2006) site elevation and UTM WGS84 zone 47N grid displayed. [SINGLE COLUMN FIGURE]



2.1 Study design

Fifteen transects of varying length (220-4750m) totaling approximately 35 km and comprising 3481 transect points were surveyed in July 2001 (Table 1), with transect routes pre-selected to sample the maximum number of land cover types. At ten meter intervals along the transects small mammal activity indicators were recorded. Visual sightings of small mammals and species-specific indicators including foraging corridors, ground holes, and small mammal faeces, all identifiable to species or genus level (Raoul *et al.,* 2006; Wang *et al.,* 2010), were used as evidence of small mammal presence using methods established by Giraudoux *et al.* (1998). Transects were mapped using a GPS with an accuracy of approximately 15 m.

At this study site the small mammal community predominantly comprised two *Ochotona* species both known to be Em intermediate hosts, *Ochotona curzoniae* (black-lipped pika), and *Ochotona cansus* (Gansu pika), the latter recorded sporadically compared to the former. Due to similarities between the two species resulting in identification difficulties, they were grouped together to form a generic *Ochotona spp.* group. *Microtus irene, M. oeconomus, M. leucurus* and *Cricetulus kamensis* small mammals were also observed but, given the very extensive *Ochotona spp.* colonies in the study area in comparison to the sparse records of these other species, and the established links between *Ochotona spp.* and Em infection in dogs (Wang *et al.*, 2010), our investigation focused exclusively on *Ochotona spp.*

Altitude, slope and aspect values for each transect point were extracted from 90m resolution Shuttle Radar Topographic Mission (SRTM) digital elevation models. A Landsat ETM+ satellite image (3 July 2001) was acquired (path 134 row 37), geometrically corrected, with snow and cloud masks created to exclude these areas of the image from further analysis. ERDAS IMAGINE was used to perform a maximum likelihood supervised classification on the image using nine land cover classes: village, road, long grass, water, short grass, upper *Potentilla* shrubland, bare ground, degraded grassland, and wet grassland. Classification accuracy assessment was performed using 365 reference points collected from high-resolution imagery of the survey area using established techniques (e.g. Duro *et al.*, 2012). Reference points exhibiting temporal change in land cover type between Landsat ETM+ image and reference high resolution imagery acquisition dates were disregarded to minimise potential error.

When investigating the relationships between landscape and *Ochotona spp.* issues of scale and the spatial arrangement of different land cover class patches within the landscape should be considered (Pleydell *et al.,* 2008; Pleydell & Chrétien, 2010). A common approach is to quantify landscape characteristics around a point of interest using a circular buffer centred at the observation (Pleydell & Chrétien, 2010). However, as the optimal buffer size cannot be known *apriori*, multiple nested buffers with radius increments between 100m and 500m in 100m increments were generated for each transect point, enabling landscape influence over multiple ranges to be investigated. Within each nested buffer, the area of each land cover class was recorded. To minimise collinearity between these nested land cover area measurements (variables calculated using smaller buffers partly measures the same area as the larger buffers), but to retain the nested spatial structure, a new set of variables Z100m, Z200m, Z300m, Z400m and Z500m were created following the methodology of Rhodes *et al.* (2009) such that:

Z100m = X100m.
Z200m = X200m - X100m.
Z300m = X300m - X200m.
Z400m = X400m - X300m.
Z500m = X500m - X400m.

where X100m,…,X500m are the land cover class coverage data for the 100m,…,500m buffer sizes respectively, and the Z200,…,Z500m provide the difference between the original variables and the variable nested within it (Rhodes *et al.*, 2009).

Landscape structure and composition are important determinants of species distributions and population viability (Rhodes *et al.*, 2009), with the amount of suitable habitat present and the level of landscape fragmentation both important factors for biological population abundance and distribution (Fahrig, 2003). Here, the aggregate properties of the spatial organisation of land cover patches within a 500m radius buffer surrounding each transect point are examined using landscape metric methods within FRAGSTATS (McGarigal *et al.*, 2002). Eighteen landscape level metrics were generated (see Table 1). Pairwise correlation was performed between metrics values, with all correlations exhibiting an r^2 value of <0.5 indicating that the landscape metrics variables were not highly correlated.

Table 1. Landscape metrics included in the analysis (McGarigal *et al.*, 2002).

Metric Type	Metric	Acronym
Area and edge metrics	Total Area	TA
	Largest Patch Index	LPI
	Patch Area Distribution	AREA_AM
Shape metrics	Perimeter-Area Ratio Distribution	PARA_AM
	Fractal Index Distribution	FRAC_AM
	Contiguity Index Distribution	CONTIG_AM
Aggregation metrics	Aggregation Index	AI
	Patch Cohesion Index	COHESION
	Landscape Division Index	DIVISION
	Splitting Index	SPLIT
	Euclidean Nearest Neighbor Distance Distribution	ENN_AM
	Connectance	CONNECT
Diversity metrics	Patch Richness	PR
	Shannon's Diversity Index	SHDI
	Simpson's Diversity Index	SIDI
	Shannon's Evenness Index	SHEI
	Simpson's Evenness Index	SIEI

Random forest (RF) analysis was performed to identify potential causal linkages between *Ochotona spp.* presence and the environmental variables of nested land cover class areas, the landscape metrics, and topographical variables of elevation, slope and aspect (ntrees = 10000, number of variables tried at each split = 21). The OOB data samples generated importance measures for each variable, and tested the prediction error of the generated RF.

Random Forest analysis was performed in the R statistical environment using the randomForest package (Liaw & Wiener, 2002). The RF was then used to produce a predicted *Ochotona spp.* distribution map. A point grid was generated for a 45km x 45km area surrounding the survey transect locations with 30m point spacing. Data values for each explanatory variable included in the RF were calculated for each vector grid point. The RF was applied in a predictive classifier capacity with the vector grid datasets as input variables and predicted *Ochotona spp.* presence or absence as the output. Predicted values were converted from vector to raster format using ArcMap 10.1.

3. Results

The overall land cover classification accuracy using 365 reference locations was 83.84% (Table 2). Of the 3481 sample points sampled along 15 transects, *Ochotona spp.* were present at 1246 points (35.8%). For individual transects the rate of *Ochotona spp.* presence ranged from 0% (transects 1, 11 and 15) to 88% (transect 2) indicating a patchy distribution across the study area (Table 3).

Table 2. Supervised classification confusion matrix and accuracy assessment. Overall Kappa statistic = 0.816

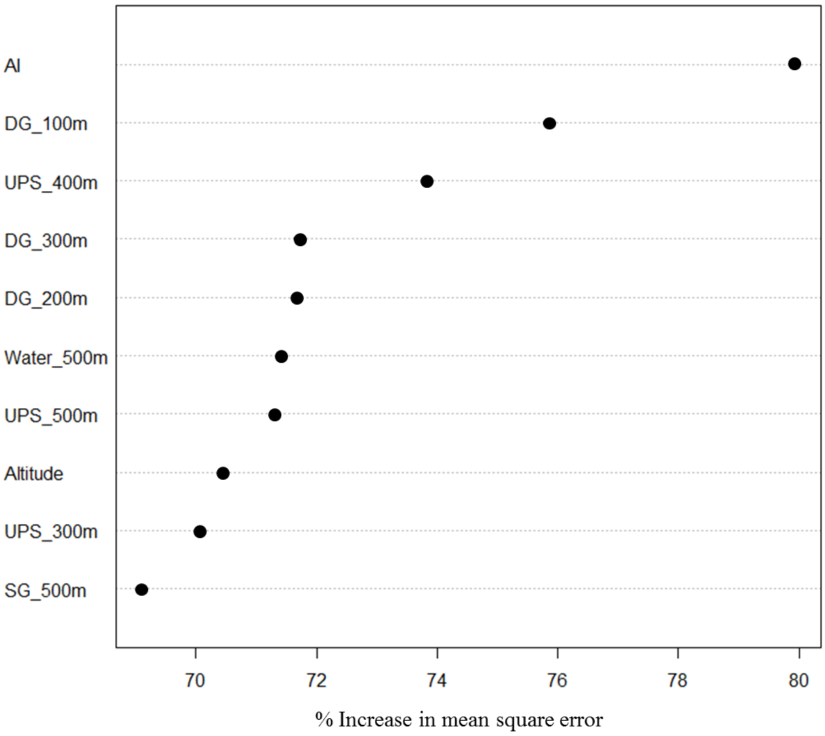
	Reference										
Classified	Village	Road	Long grass	Water	Short grass	Upper *potentilla* shrubland	Bare ground	Degraded grassland	Wet grassland	Sum of row	User's accuracy (%)
Village	22	0	0	0	0	0	0	0	0	22	100.00
Road	0	41	0	0	3	0	0	0	0	44	93.18
Long grass	0	0	18	0	0	0	1	0	0	19	94.74
Water	0	1	2	44	1	0	0	1	4	53	83.02
Short grass	0	0	0	0	31	2	0	0	0	33	93.94
Upper *potentilla* shrubland	0	1	2	0	5	20	0	2	0	30	66.67
Bare ground	0	1	0	0	0	0	44	2	0	47	93.62
Degraded grassland	1	2	2	3	7	1	4	45	0	65	69.23
Wet grassland	0	4	4	3	0	0	0	0	41	52	78.85
Sum of column	23	50	28	50	47	23	49	50	45	365	
Producers accuracy (%)	95.65	82.00	64.29	88.00	65.96	86.96	89.80	90.00	91.11		Overall accuracy = 83.84

Table 3. Survey transect *Ochotona spp*. presence and elevation ranges.

Transect	Number of survey points along transect	Number of points with *Ochotona spp.* present	Number of points with *Ochotona spp.* absent	*Ochotona spp*. presence (%)	Elevation range of transect (m)
1	276	0	276	0.0	4280-4480
2	133	117	16	88.0	4290-4334
3	320	89	231	27.8	4294-4350
4	94	1	93	1.1	4299-4360
5	346	28	318	8.1	4287-4350
6	475	363	112	76.4	4285-4501
7	274	129	145	47.1	4387-4532
8	137	61	76	44.5	4309-4484
9	182	10	172	5.5	4299-4366
10	424	242	182	57.1	4160-4348
11	22	0	22	0.0	4160-4160
12	172	1	171	0.6	4160-4259
13	339	204	135	60.2	4177-4262
14	109	1	108	0.9	4182-4300
15	178	0	178	0.0	4190-4492
Total	3481	1246	2235	35.8	4160-4532

RF analysis explained 70.78% of the variance in *Ochotona spp*. presence or absence. Fig 2 shows the ten environmental variables determined as most important by the RF in relation to *Ochotona spp*. presence. Aggregation Index (AI) was identified as the single most important variable, however it was the only landscape metric in the top ten ranked variables. Three of the top five variables were degraded grassland (DG), with DG at the 100m buffer size second, at the 300m buffer size fourth, and at the 200m buffer size fifth. Upper *Potentilla* shrubland (UPS) was also important but at the larger buffer sizes of 400m (third ranked importance), 500m (seventh) and 300m (ninth). Water at 500m was sixth highest ranked, with altitude eighth, and short grass (SG) at the 500m buffer tenth.

Figure 2. Variable importance scores for the top ten variables as identified by the RF, with corresponding % increase in mean square error when that variable is randomly permuted. Percent variance explained = 70.78%, number of trees = 10000, mean square of residuals = 0.07, number of variables tried at each split = 21. AI = Aggregation Index; DG = degraded grassland; UPS = upper *Potentilla* shrubland; SG = short grass. [SINGLE COLUMN FIGURE]



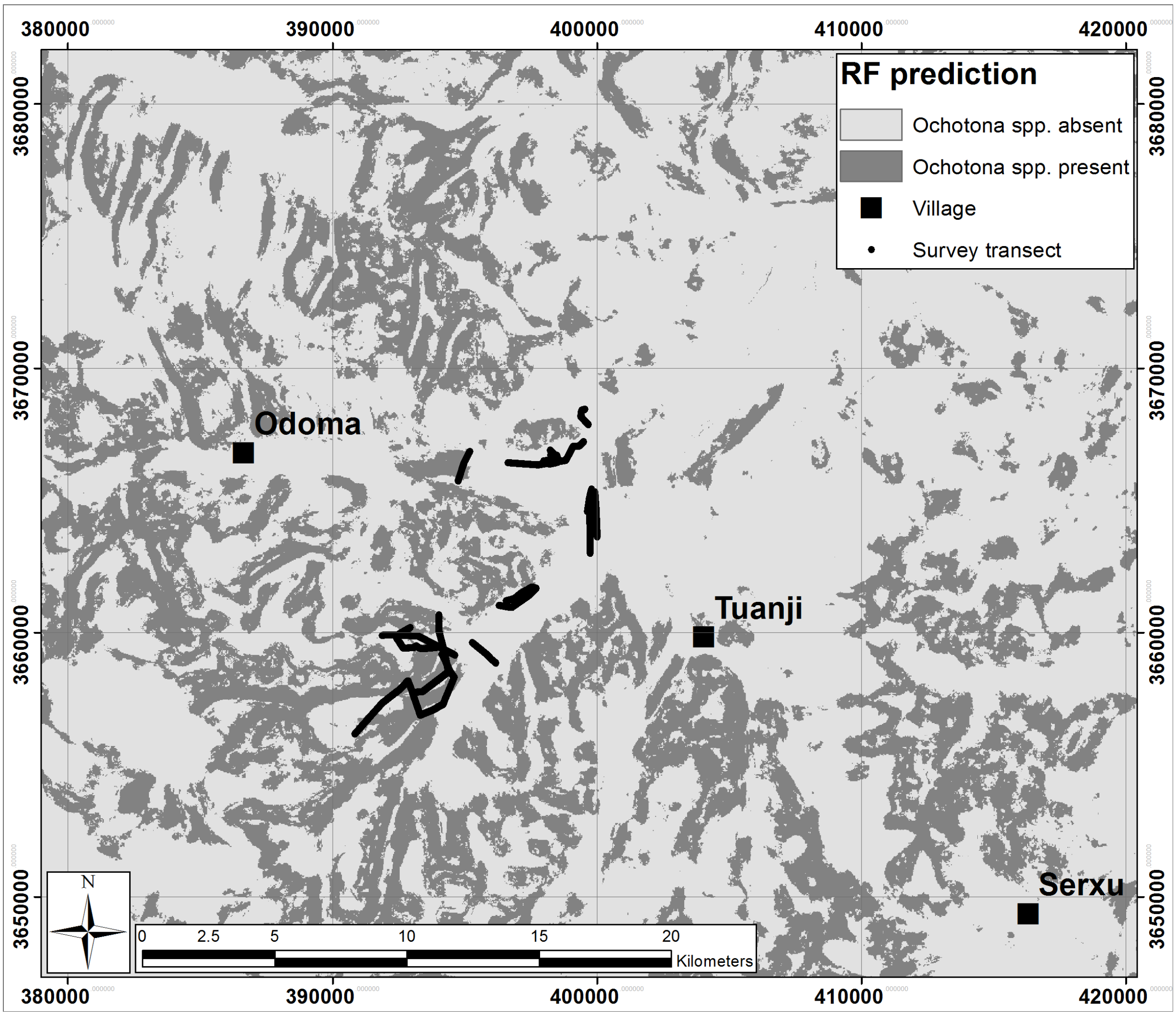
A confusion matrix of the predicted values was generated using the OOB data samples to assess the RF predictive accuracy (Table 4). Results indicate that the RF performed with a high level of accuracy, with a 90.98% accuracy rate. Of the incorrectly predicted samples, the false positives (150) and false negatives (164) were similar in magnitude.

Table 4. RF confusion matrix of predicted versus observed *Ochotona spp*. presence (1) and absence (0). Total correct = 3167, total incorrect = 314, percentage of survey points predicted correctly = 90.98%

Observed value	Predicted value		Total
	0	1	
0	2085	150	2235
1	164	1082	1246
Total	2249	1232	3481

The map produced (Fig 3) shows the predicted areas of *Ochotona spp*. presence with patchiness in these areas observed at the local scale. Areas of predicted presence occur across the area, but are more extensive to the south, west, and north-west of the original survey transects, with sparser areas of predicted presence to the east and north-east.

Figure 3. Predicted *Ochotona spp*. presence (red) or absence (blue) with original survey transects overlaid and UTM WGS84 zone 47N grid displayed for context. [SINGLE COLUMN FIGURE]



4. Discussion

This research examined a critical phase of the *Echinococcus multilocularis* (Em) transmission cycle, and adopted an analytical approach using random forests (RF) to model and predict *Ochotona spp*. presence in relation to landscape characteristics within a highly endemic area of the Tibetan plateau for Em. We found that the environmental variables analysed explained 70.78% of the variance in *Ochotona spp*. presence. It is argued thus that (1) *Ochotona spp.*

presence is statistically related to key environmental variables which can be used to predict species presence over large areas; and (2) in the geographical area of interest *Ochotona spp.* presence is specifically linked to areas of degraded grassland.

The application of RF for predictive modelling of *Ochotona spp.* presence, based on landscape characteristics has provided a clearer understanding of the influence of key landscape variables in this region. The environmental variables analysed explained 70.78% of the variance in *Ochotona spp.* presence, with a 90.98% accuracy rate indicating that the RF methods employed enabled accurate modelling of *Ochotona spp.* presence. Given these encouraging results, we then generated predictive maps of *Ochotona spp.* presence across a larger spatial extent within the same bio-geographical area to identify potential hot-spots of presence meriting further investigation as reservoir zones of the zoonotic parasite *Echinococcus multilocularis*.

This analysis enabled comparison of the relative importance of the environmental predictors, with the aggregation index (AI) landscape metric ranked with the highest importance. AI is computed where each land cover class is weighted by its area in the landscape, scaled to account for the maximum possible number of like adjacencies given any landscape composition (McGarigal *et al.*, 2002). The interpretation is that buffered areas containing larger aggregations, or clusters of land cover patches of the same type, are of importance in influencing *Ochotona spp.* presence. However, eight of the ten highest ranked variables are particular land cover class variables suggesting that the presence of specific land cover classes was, with the exception of AI, of greater importance in influencing *Ochotona spp.* presence than land cover patch spatial arrangement.

RF assessment indicated that degraded grassland (DG) at the 100m buffer size was the most important land cover class variable. At the 200m and 300m buffer sizes DG was again the highest ranked land cover variable. Although UPS (400m) and water (500m) were the highest ranked land cover variables at those respective buffer sizes, the ranking of DG as second, fourth and fifth most important variables overall, and highest at the three buffer sizes closest to the survey transect points, indicates that DG could be considered the most important land cover variable of influence. Smith & Gao, (1991) determined that the average home range for *Ochotona curzoniae* is 1,375 ± 206m^2, placing the principle area of activity of an individual *Ochotona spp.* within the 100m buffer area, supporting the RF result that DG at the 100m buffer size is the most important land cover variable influencing *Ochotona spp.* presence. This reinforces previous studies that have sought to understand the drivers of *Ochotona spp.* presence in the study region such as Raoul *et al.* (2006), and visual field observations, indicating that higher *Ochotona spp.* densities were more commonly present in areas with low vegetation cover. It should be noted, however, that in some areas of degraded grassland where transects were surveyed *Ochotona spp.* were not present. This may be due to patchy local-scale extinctions during *Ochotona spp.* population cycles in this area.

Of particular concern in the study area is the impact of heavy grazing by yak resulting in large areas of degraded grassland. Past studies have shown that land cover changes and grazing practices can increase the likelihood of small mammal population outbreaks that are suggested to play a significant role in Em transmission (Wang *et al.*, 2004). If this heavy grazing results in larger *Ochotona spp.* populations and more frequent population outbreaks due to increased optimal habitat availability, this could potentially contribute to increasing levels of Em transmission, resulting in greater risk to human populations.

4.1 Conclusions

We have used random forests (RF) to successfully model the environmental variables influencing spatial patterns in the presence of the *E. multilocularis* intermediate host *Ochotona spp.* in western China. The predictive use of random forests to indicate likely areas of *Ochotona spp.* presence could form a valuable contribution to systematic modelling describing the broader *E. multilocularis* transmission pathways between *Ochotona spp.* small mammal intermediate hosts, both sylvatic (fox) and domestic (dog) definitive hosts, and susceptible human populations. Given the relationships established previously by Wang *et al.* (2010) correlating density of *Ochotona spp.* burrows with domestic dog infection rates, this methodology could enable identification of domestic dog populations at risk of continual re-infection through predation of *Ochotona spp.* and thus help identify areas of active *E. multilocularis* transmission. In conjunction with the possibility of applying these techniques over larger geographical regions utilizing the extensive coverage of satellite imagery, such information could facilitate the design of pre-emptive disease control measures including targeted treatment of dogs with antihelminthic drugs to disrupt the Em transmission cycle in that region, thus reducing Em infection risk in local human populations.

Acknowledgements

Special thanks to F. Raoul, JP Quéré, D. Rieffel, N. Bernard, R. Scheifler, A. Vaniscotte and Alastair Graham for their valuable assistance. This research has been co-funded by the US National Institutes of Health and National Science Foundation (EID TW001565-01 & 05) from the Fogarty International Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Fogarty International Center or the National Institutes of Health. This is an article of the GDRI (International research network) "Ecosystem health and environmental disease ecology".

References

Abdel-Rahmana, E.M., Ahmed, F.B., & Ismail, R. (2013) Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. Int. J. Remote Sens. 34, 712–728.

Arthur, A.D., Pech, R.P., Davey, C., Jiebu, Yanming, Z., & Hui, L. (2008) Livestock grazing, plateau pikas and the conservation of avian biodiversity on the Tibetan plateau. Biol. Conserv. 141, 1972-1981.

Breiman, L. (2001) Random forests. Mach. Learn. 45, 5–32.

Budke, C.M., Campos-Ponce M., Wang Q. & Torgerson P.R. (2005) A canine purgation study and risk factor analysis for echinococcosis in a high endemic region of the Tibetan plateau. Vet. Parasitol. 127, 43-49.

Craig, P.S., Liu, D., Shi, D., Macpherson, C.N.L., Barnish, G., Reynolds, D., Gottstein, B. & Wang, Z. (1992) A large focus of alveolar echinococcosis in central China. Lancet, 340, 826-831.

Craig, P.S., Giraudoux, P., Shi, D., Bartholomot, B., Barnish, G., Delattre, P., Quéré, J.P., Harraga, S., Bao, G., Wang, Y., Lu, F., Ito, A. & Vuitton, D.A. (2000) An epidemiological and ecological study of human alveolar echinococcosis transmission in south Gansu, China. Acta Trop. 77, 167-177.

Danson, F.M., Graham, A.J., Pleydell, D.R.J., Campos-Ponce, M., Giraudoux, P. & Craig, P.S. (2003) Multi-scale spatial analysis of human alveolar echinococcosis risk in China. Parasitology, 127, S133-S141.

Danson, F.M., Craig, P.S., Man, W., Shi, D.Z., & Giraudoux, P. (2004) Landscape dynamics and risk modelling of human alveolar echinococcosis, Photogramm. Eng. Rem. S. 70, 359-366.

Duro, D.C, Franklin, S.E., & Dube, M.G. (2012) Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests. Int. J. Remote Sens. 33, 4502–4526.

Eckert, J. (1996) Echinococcus multilocularis and alveolar echinococcosis in Europe (except parts of eastern Europe). Alveolar Echinococcosis. Strategy for eradication of alveolar echinococcosis of the liver (ed. by J. Uchino and N. Sato), pp 27-43. Fuji Shoin, Sapporo.

Eckert, J., Gemmell, M.A., Meslin, F.X., & Pawlowski, Z.S. (2001) WHO/IOE manual on Echinococcosis in humans and animals: a public health problem of global concern. OIE/WHO, Paris.

Fahrig, L. (2003) Effects of habitat fragmentation on biodiversity. Annu. Rev. Ecol. Syst. 34, 487-515.

Giraudoux, P., Delattre, P., Habert, M., Quéré J.P., Deblay, S., Defaut, R., Duhamel, R., Moissenet, M.F., Salvi, D. & Truchetet, D. (1997) Population dynamics of fossorial water vole (Arvicola terrestris scherman): a land use and landscape perspective. Agr. Ecosyst. Environ. 66, 47-60.

Giraudoux, P., Quéré, J.P., Delattre, P., Bao, G., Wang, X., Shi, D., Vuitton, D. & Craig, P.S. (1998) Distribution of small mammals along a deforestation gradient in south Gansu, China. Acta Theriol. 43, 349-362.

Giraudoux, P., Craig, P.S., Delattre, P., Bao, G., Bartholomot, B., Harraga, S., Quéré, J.P., Raoul, F., Wang, Y., Shi, D, and Vuitton, D.A. (2003) Interactions between landscape changes and host communities can regulate Echinococcus multilocularis transmission. Parasitology, 127, S121-S131.

Giraudoux, P., Pleydell, D.R.J., Raoul, F., Quéré, J.P., Wang, Q., Yang, Y., Vuitton, D.A., Qiu, J., Yang, W. & Craig, P.S. (2006) Transmission ecology of Echinococcus multilocularis: What are the ranges of parasite stability among various host communities in China. Parasitol. Int. 55, S237-S246.

Giraudoux, P., Raoul, F., Afonso, E., Ziadinov, I., Yang, Y., Li, L., Li, T.Y., Quéré, J.P., Feng, X.H., Wang, Q., Wen, H., Ito, A. & Craig, P.S. (2013a) Transmission ecosystems of Echinococcus multilocularis in China and Central Asia. Parasitology, 140, 1655-1666.

Giraudoux, P., Raoul, F., Pleydell, D.R.J., Li, T., Han, X., Qui, J., Xie, Y., Wang, H., Ito, A. & Craig, P.S. (2013b) Drivers of Echinococcus multilocularis transmission in China: small mammal diversity, landscape or climate? PLOS Neglect. Trop. D. 7, 1-12.

Jiapeng, Q., Wenjing, L., Min, Y., Weihong, J., & Yanming, Z. (2013) Life history of the plateau pika (Ochotona curzoniae) in alpine meadows of the Tibetan Plateau, Mamm. Biol. 78, 68-72.

Li, T., Chen, X., Zhen, R., Qiu, J., Qiu, D., Xiao, N., Ito, A., Wang, H., Giraudoux, P., Sako, Y., Nakao, M. & Craig, P.S. (2010) Widespread co-endemicity of human cystic and alveolar echinococcosis on the eastern Tibetan Plateau, northwest Sichuan/southeast Qinghai, China. Acta Trop. 113, 248-256.

Liaw, A. & Wiener, M. (2002) Classification and regression by randomForest. R News, 2, 18–22.

Lidicker, W.Z. (1995), Landscape Approaches in Mammalian Ecology and Conservation. University of Minnesota Press, Minneapolis.

McGarigal, K., Cushman, S.A., Neel, M.C., & Ene, E. (2002) FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. Available at the following web site: www.umass.edu/landeco/research/fragstats/fragstats.html

Moss, J.E., Chen, X., Li, T., Qiu, J., Wang, Q., Giraudoux, P., Ito, A., Torgerson, P.R. & Craig, P.S. (2013) Reinfection studies of canine echinococcosis and role of dogs in transmission of Echinococcus multilocularis in Tibetan communities, Sichuan, China. Parasitology, 28, 1-8.

Perdiguero-Alonso, D., Montero, F.E., Kostadinova, A., Raga, J.A., & Barrett, J. (2008) Random forests, a novel approach for discrimination of fish populations using parasites as biological tags. Int. J. Parasitol. 38, 1425–1434.

Pleydell, D.R.J., Yang, Y.R., Danson, F.M., Raoul, F., Craig, P.S., McManus, D.P., Vuitton, D.A., Wang, Q. & Giraudoux, P. (2008) Landscape Composition and Spatial Prediction of Alveolar Echinococcosis in Southern Ningxia, China. PLOS Neglect. Trop. D. 2, e287.

Pleydell, D.R.J. & Chrétien, S. (2010) Mixtures of GAMs for habitat suitability analysis with overdispersed presence / absence data. Comput. Stat. Data An. 54, 1405-1418.

Raoul, F., Deplazes, P., Nonaka, N., Piarroux, R., Vuitton, D.A. & Giraudoux, P. (2001) Assessment of the epidemiological status of Echinococcus multilocularis in foxes in France using ELISA coprotests on fox faeces collected in the field. Int. J. Parasitol. 31, 1579-1588.

Raoul, F., Quéré, J.P., Rieffel, D., Bernard, N., Takahashi, K., Scheifler, R., Wang, Q., Qiu, J., Yang, W., Craig, P.S., Ito, A. & Giraudoux, P. (2006) Distribution of small mammals along a grazing gradient on the Tibetan plateau of western Sichuan, China. Mammalia, 42, 214-225.

Raoul, F., Pleydell, D.R.J., Quéré, J.P., Vaniscotte, A., Rieffel, D., Takahashi, K., Bernard, N., Wang, J.L., Dobigny, T., Galbreath, K.E. & Giraudoux, P. (2008) Small-mammal assemblage response to deforestation and afforestation in central China. Mammalia, 72, 320–332.

Rausch, R.L. (1995) Life cycle patterns and geographic distribution of Echinococcus species. Echinococcus and Hydatid Disease (ed. by R.C.A. Thompson & A.J. Lymbery), pp. 89-134. Cab International, Wallingford,

Rhodes, J.R., McAlpine, C.A., Zuur, A.F., Smith, G.M. & Ieno, E.N. (2009) GLMM Applied on the Spatial Distribution of Koalas in a Fragmented Landscape. Mixed Effects Models and

Extensions in Ecology with R (ed. by A.F. Zuur, E.N. Ieno, N.J. Walker, A.A. Saveliev & G.M. Smith), pp. 469-492. Springer, New York.

Smith, A.T., & Gao, W.X. (1991) Social Relationships of Adult Black-Lipped Pikas (Ochotona curzoniae). J. Mammal. 72, 231-247.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. & Zeileis, A. (2008) Conditional variable importance for random forests. BMC Bioinformatics, 9, 307.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. & Feuston, B.P. (2003) Random Forests: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Model. 43, 1947–1958.

USGS (2004) Shuttle Radar Topography Mission, 1 Arc Second scene SRTM_ n33_e097, Unfilled Unfinished 2.0, Global Land Cover Facility, University of Maryland, College Park, Maryland, February 2000.

Veit, P., Bilger, B., Schad, V., Schafer, J., Frank, W. & Lucius, R. (1995) Influence of environmental factors on the infectivity of Echinococcus multilocularis eggs. Parasitology, 110, 79-86.

Wang, Q., Qiu, J.M., Schantz, P.M., He, J.G., Ito, A. & Liu, F.J. (2001) Risk factors for development of human hydatidosis among households raising livestock in Tibetan areas of western Sichuan Province. Chin.J. Parasit. Dis. Parasitol. 19, 289-293.

Wang, Q., Vuitton, D.A., Qui, J., Giraudoux, P., Xiao, Y., Schantz, P.M., Raoul, F., Li, T., Yang, W & Craig, P.S. (2004) Fenced pasture: a possible risk factor for human alveolar echinococcosis in Tibetan pastoralist communities of Sichuan, China. Acta Trop. 90, 285-293.

Wang, Q., Raoul, F., Budke, C., Craig, P.S., Yong-fu, X., Vuitton, D.A., Campos-Ponce, M., Qiu, D.C., Pleydell, D.R.J., & Giraudoux, P. (2010) Grass height and transmission ecology of Echinococcus multilocularis in Tibetan communities, China, Chinese Med. J-Peking, 123, 61-67.

Zhang, W., Zhang, Z., Wu, W., Shi, B., Li, J., Zhou, X., Wen, H., and McManus, D.P. (in press) Epidemiology and control of echinococcosis in central Asia, with particular reference to the People’s Republic of China, Acta Trop.
http://dx.doi.org/10.1016/j.actatropica.2014.03.014