1   **Predictive analysis across spatial scales links zoonotic malaria to deforestation**

2

3   Patrick M. Brock[1], Kimberly M. Fornace*[2], Matthew J. Grigg[3], Nicholas M. Anstey[3], Timothy

4   William[4,5], Jon Cox[2], Chris J. Drakeley[2], Heather M. Ferguson[1], Rowland R. Kao[1, 6]

5

6

7   [1]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical,

8   Veterinary and Life Sciences, University of Glasgow, Glasgow, G61 1QH, UK

9   [2]London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

10   [3]Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin

11   University, Darwin, NT 0810, Australia

12   [4]Infectious Diseases Unit, Clinical Research Centre, Queen Elizabeth Hospital, Kota Kinabalu

13   88560, Sabah, Malaysia

14   [5]Infectious Diseases Society Sabah-Menzies School of Health Research Clinical Research

15   Unit, Kota Kinabalu 88560, Sabah, Malaysia

16   [6]Royal (Dick) School of Veterinary Studies and Roslin Institute, Easter Bush Campus,

17   University of Edinburgh, EH25 9RG, UK

18

19

20

21   * Corresponding author: Kimberly.Fornace@lshtm.ac.uk

22   **ABSTRACT (max 200 words)**

23

24   The complex transmission ecologies of vector-borne and zoonotic diseases pose challenges

25   to their control, especially in changing landscapes. Human incidence of zoonotic malaria

26   (*Plasmodium knowlesi*) is associated with deforestation although mechanisms are unknown.

27   Here, a novel application of a method for predicting disease occurrence that combines

28   machine learning and statistics is used to identify the key spatial scales that define the

29   relationship between zoonotic malaria cases and environmental change. Using data from

30   satellite imagery, a case control study, and a cross-sectional survey, predictive models of

31   household-level occurrence of *P. knowlesi* were fitted with 16 variables summarised at 11

32   spatial scales simultaneously. The method identified a strong and well-defined peak of

33   predictive influence of the proportion of cleared land within 1 km of households on *P.*

34   *knowlesi* occurrence. Aspect (1 and 2km), slope (0.5km) and canopy regrowth (0.5km) were

35   important at small scales. In contrast, fragmentation of deforested areas influenced *P.*

36   *knowlesi* occurrence probability most strongly at large scales (4 and 5 km). The identification

37   of these spatial scales narrows the field of plausible mechanisms that connect land use

38   change and *P. knowlesi*, allowing for the refinement of disease occurrence predictions and

39   the design of spatially-targeted interventions.

40

41

42   **Key words (3-6 only):** disease ecology, zoonoses, malaria, *Plasmodium knowlesi*, boosted

43   regression trees, disease occurrence prediction

44

45

INTRODUCTION (4367 words)

48 Infectious disease mapping plays a vital role in guiding public health policy and practice [1].

49 For diseases with environmental drivers, such as malaria, mapping has supported the

50 ongoing and successful drive to reduce the number of infections worldwide and has been

51 pivotal to understanding the effectiveness and progress of this effort [1-4]. As control

52 reduces incidence, the geographical distribution of infection becomes more heterogeneous

53 [5]. In situations where few data are available, predicted probability of disease occurrence

54 can be mapped in place of measures such as incidence or prevalence. This approach has

55 been applied to a variety of infectious disease systems using methods that combine the

56 strengths of machine learning and statistics, originally developed to more accurately map

57 species distributions in ecology (e.g. [6-8]). In addition to geostatistical mapping, disease

58 occurrence mapping has helped describe the spatial distribution of infectious diseases

59 worldwide, and provided information relevant to the design and execution of disease

60 control programmes (e.g. [9-11]).

61

62 Ensemble boosted regression tree (BRT) analysis is one such method that is now widely

63 used for disease occurrence mapping [6, 11, 12]. BRT analysis is increasingly used to identify

64 patterns in large infectious disease datasets, building on analytical developments in

65 macroecology [12-15], and has been used to generate hypotheses from these patterns [15].

66 BRT analysis combines decision trees, in which trees are grown with binary splits of

67 predictor values to minimise prediction errors, and boosting, in which a collection of models

68 are combined [16]. It allows for the uneven distribution of variation in predictor variables

69 without the need for transformation, is not biased by correlation between predictors, can

70 incorporate complex interactions and fit non-linear functions [16].

71

72 A disadvantage of disease occurrence mapping is the difficulty identifying how different

73 factors contribute to models that generate their spatial predictions; predictions may be

74 sufficiently reliable, but it may not be clear why [14]. This is particularly problematic in

75 relation to the scale of processes that could give rise to spatial heterogeneity of disease, as

76 the environmental data used to predict occurrence are usually aggregated on a single

77 spatial scale (e.g. square grid cells of 5 km x 5 km). This may be unavoidable if, for example,

78 satellite data are only available at a fixed resolution, or census data are pre-aggregated over

79 administrative units. However, even when disaggregated data are available at high

80 resolution, there is often no evidence-based methodological recourse to guide decisions on

81 the appropriate spatial scale for inclusion in models. Ecological processes occur at different

82 spatial scales and the scale at which analyses of disease distributions are conducted

83 influences the inferred contribution of the determinants of those distributions [17-19].

84

85 Differences between the spatial scales of the underlying biological processes that drive

86 disease transmission and the scale imposed on models by the aggregation of predictor

87 variables (such as into raster grid cells) is likely to be particularly influential in models of

88 zoonoses and vector-borne diseases. Transmission dynamics of these diseases arise from

89 the interaction of multiple species and the environment, likely occurring over a variety of

90 spatial scales, which makes it less likely that predictors aggregated at a single spatial scale

91 will capture important variation, especially if the influences of multiple scales are

92 dependent on one another, and when few data are available [20].

93

94 *Plasmodium knowlesi* malaria is a vector-borne zoonosis in South East Asia, which usually

95 infects long-tailed (*Macaca fascicularis*) and pig-tailed macaques (*Macaca nemestrina*) [21].

96 Transmitted by the *Anopheles leucosphyrus* group of mosquitoes, changes in forest cover

97 impact vector habitats as well as macaque and human distributions [22]. Identified as a

98 potentially lethal infection in humans and a major public health concern in 2004 [23], *P.*

99 *knowlesi* is now the most common cause of malaria in Malaysia and parts of Indonesia,

100 global hotspots of tropical deforestation [24-26]. It may be misdiagnosed or undiagnosed

101 across South East Asia, and the World Health Organisation has advised it be incorporated

102 into ongoing malaria elimination programmes [27]. Due to this increasing public health

103 concern, *P. knowlesi* was proposed as a global priority for disease mapping [4] and has since

104 been mapped by BRT analysis, using historical data to highlight priority areas for

105 surveillance [6].

106

107 This study introduces a novel approach to spatial scale in disease occurrence prediction as a

108 tool to identify the key scales that define the relationship between a zoonosis of serious

109 public health concern (*Plasmodium knowlesi* malaria) and the rapidly changing landscape

110 implicated in its spillover from macaques to humans in South East Asia. Where the highest

111 numbers of cases have been reported (Malaysian Borneo), *P. knowlesi* incidence has been

112 positively associated both with forest cover and historical forest loss [28]. However, the

113 mechanisms of the proposed influence of deforestation on *P. knowlesi* transmission are

114 unknown; for example, this could be due to changes in macaque densities, vector bionomics

115 or human behaviour. For the purposes of control, this precludes the assessment of which

116 part(s) of the transmission cycle to target and which kind of interventions are most likely to

117 be effective at which spatial scales. For example, if regulating land use change to reduce the

118 proximity of macaque to humans, how far should regulated zones extend from planned or

119 existing settlements? The spatial scales that define *P. knowlesi* occurrence identified by this

120 study provide important hitherto missing information to inform such spatially targeted

121 control measures.

122

123 **METHODS**

124

125 *Ethics approval and informed consent*

126 This study was approved by the Medical Research Sub-Committee of the Malaysian Ministry

127 of Health and the Research Ethics Committee of the London School of Hygiene and Tropical

128 Medicine. Written informed consent was obtained from all participants.

129

130 *Case and household data*

131 Data on household locations of consenting PCR-confirmed *P. knowlesi* cases (n=206) were

132 obtained from a case control study carried out between 2012 and 2014 in Kudat and Kota

133 Marudu districts, Northern Sabah, Malaysian Borneo [29] and used as presence points. In

134 this study, control households were selected in the vicinity of cases households, making

135 them unsuitable for use as absence points due to spatial sampling bias. Instead, absence

136 households were identified from the sampling frame of a cross-sectional survey geo-locating

137 all households within 180 randomly selected villages in four districts in Northern Sabah

138 (Fornace et al, in prep). Absence points were identified from households not reporting

139 clinical knowlesi cases within the two districts included in the case control study. These

140 absence points were filtered so that there were no more than 5 per village, with the first

141   absence point in each village sampled randomly, and the remainder chosen to maximise the

142   total distance between absence points within that village to ensure spatial

143   representativeness. Absence points were excluded if they were further than 5 km from a

144   presence point (to prevent large areas being covered only by absences), nearer than 0.2 km

145   to a presence point or did not have permanent residents. Presence and absence points were

146   excluded if they were located within an urban area, determined using administrative

147   boundaries, as travel histories suggest cases reported in urban areas are unlikely to have

148   been contracted in urban areas [29]. These filters resulted in a dataset including 206

149   presence points, 43 of which were located on the island of Banggi, and 1324 absence points,

150   105 of which were located on the island of Banggi. All household locations were visited and

151   geolocated using a handheld GPS (Garmin, USA).

152

153   *Landscape variables*

154

155   Data on forest cover at 30m resolution was obtained from Hansen et. al, [26], with annual

156   forest cover defined categorically as over 50% canopy cover based on data derived from

157   Landsat imagery. Although this definition of forest may not differentiate between forest and

158   plantations, canopy cover has previously been associated with *P. knowlesi* incidence [28].

159   Cases were approximately evenly divided between 2013 (n = 101) and 2014 (n = 105), and

160   as the annual classified satellite data composition method tracks back in time as far as

161   necessary to find cloud-free imagery covering all locations, a frequent issue in Borneo [26],

162   forest data was extracted from the 2014 annual composite as it was most likely to represent

163   the environment contemporaneous with case reporting.

164

165   Scalable variables were extracted from forest cover data, including proportions of recent

166   (previous year) and historical (previous 5 years) forest loss and cleared areas (Table 1). Data

167   on forest gain was only available aggregated over the period 2000-2012 and was included to

168   represent types of land use distinct from straightforward forest persistence or clearance,

169   such as agroforestry. Perimeter area ratio (P:A) was used as a proxy for fragmentation of

170   these land cover categories, as variation in P:A was more evenly distributed across variables

171   than other fragmentation measures.

172

173   Other environmental variables previously associated with malaria [30] were included as

174   predictors in BRT models, including elevation, aspect and slope [31]. Average annual

175   normalized difference vegetation index (NDVI), which quantifies the greenness of

176   vegetation, was calculated from the Landsat imagery used as input for the Hansen et al. [26]

177   2014 classification. Additionally, the standard deviation of NDVI was also included, as

178   variance in NDVI values in space may identify habitat type contrasts and boundaries. To

179   address the possibility of reporting bias, the distance to the nearest clinic and the minimum

180   distance to any road were included in a subset of BRT models. A list of clinics in the study

181   area was obtained from the Ministry of Health, Malaysia, and all clinics and roads were geo-

182   located using a hand-held GPS (Garmin 62s, Schaffhausen, Switzerland). All variables were

183   extracted at 30m resolution.

184

185   *Spatial scales*

186   16 scalable variables (Table 1) were summarised over buffer areas determined by a

187   maximum overland distance of 0.1, 0.2, 0.5, 1, 2, 3, 4, 5, 7.5, 10 and 20 km ('spatial scales')

188   from each household. Maximum overland distances (i.e. areas containing all grid cells less

189 than the threshold overland distance from the focal household) were used rather than

190 circular buffers to exclude parts of the landscape separated from focal households by water.

191

192 *Ensemble boosted regression tree analysis*

193 To balance the influence of presence and absence points [32] and quantify uncertainty [8],

194 models were run on 100 datasets, each including all presence points (n = 206) and an equal

195 number of randomly sampled (without replacement) absence points. To describe variation

196 in the contribution of variables to predictive ability across scales, a model was fitted with all

197 scalable variables included at all spatial scales (11 spatial scales and 16 variables giving 176

198 predictors). An additional model was fitted in which two non-scalable variables (shortest

199 distance to clinic and road) were added (178 predictors). To compare overall predictive

200 ability across scales, eleven ensemble models were fitted, one for each spatial scale (16

201 predictors each). A version of all models was fitted to data from the mainland only,

202 excluding cases not on the main island of Borneo (e.g. on Banggi island) to examine whether

203 these associations were impacted by the inclusion of households within smaller land areas.

204

205 Models were fitted by 10-fold cross-validation, dividing the dataset into 10 training sets with

206 each comprising a unique combination of 9 subsets of the data with the remaining subset

207 withheld for independent validation [16]. Model predictive ability was assessed using area

208 under the receiver operator curve (AUC). The tree complexity parameter of the boosted

209 regression tree analysis was set at 5, so that each decision tree built as part of the model

210 included five nodes, allowing for complex interactions between predictor variables. The

211 learning rate, which determines the contribution of each decision tree to a BRT model, was

212 tuned to between 0.0001 and 0.002 to minimise prediction error during cross-validation

213   (23). Marginal effect curves, the effect of the change in one unit of the predictor on the

214   probability of disease occurrence, were plotted for all predictors by scale.

215

216   *Relative variable importance*

217

218   Profiles of relative variable importance (RVI) for landscape variables across spatial scales

219   were derived from models that included all scales simultaneously so that the importance of

220   scale variable-combinations could be assessed while accounting for the contributions of all

221   other variable-scale combinations and interactions between them. RVI measures the

222   number of times a variable is selected for splitting during the construction of a BRT model,

223   weighted by the squared improvement of the model due to the split, averaged over all trees

224   in the model [16]. To aid the interpretation of RVI across scales within variables, Spearman

225   rank correlation matrices comparing values between all pairwise combinations of scales

226   were plotted for each variable

227

228   To test whether peaks of RVI were driven by changes in variance available to BRT models

229   across scales, variance was superimposed on RVI profiles. This is a necessary check, as if RVI

230   tracked variance across correlated scales within variables, we could not preclude differences

231   in RVI across scales arising due to an artefact of available variance alone. To aid

232   interpretation, variances were plotted as proportions of maximum variance across scales for

233   each landscape variable. Relative variance was compared with median RVI using Spearman

234   rank correlation tests across the whole study site.

235

236   *Case clusters*

237   To investigate whether analysis across spatial scales could be used to distinguish different

238   sets of epidemiological circumstances driving *P. knowlesi* spillover, a cluster analysis was

239   performed on the model fitted (whole-study-site, scalable variables only) marginal

240   probabilities of occurrence for each scalable variable (n = 176) for all cases (n = 206). Cases

241   were clustered into two groups using Ward's minimum variance method [33].

242

243   *Data availability*

244   All analyses were performed in R and code and sample environmental data are available at:

245   https://github.com/kfornace/monkeybar. Due to data confidentiality, human disease and

246   household data are available through contacting relevant ethics committees as described in

247   [29, 34].

248

249   **RESULTS**

250

251   *Relative variable importance across scales*

252   RVI was extracted from an ensemble BRT model of *P. knowlesi* occurrence in Sabah,

253   Malaysian Borneo, including 176 predictors and 16 scalable landscape variables (Table 1.)

254   summarised at 11 spatial scales (Fig. 1). The emergent peaks in RVI profiles show that the

255   influence of several variables on *P. knowlesi* occurrence prediction is strongly dependent on

256   the spatial scale of their aggregation. The median relative importance of the proportion of

257   cleared land was more than threefold higher when aggregated over a radius of 1 km from

258   households than at any other scale in the mainland-only model, and more than twofold

259   higher in the whole-study-site model (Fig. 1c). This was also the variable-scale combination

260   with the highest RVI of the 176 predictors included in the whole-study-site model (Fig. S1a).

261   The corresponding marginal effect curve shows that probability of *P. knowlesi* occurrence

262   was greater at lower proportions of cleared land within 1 km of households (Fig. 2).

263

264   The RVI profiles of five other variables included peaks at similar scales (Fig. 1 & Table 1):

265   mean aspect (1 and 2 km), mean slope (0.5 km), gain all years (0.5 km), population density

266   (2 km) and loss previous year (0.5 km). The probability of *P. knowlesi* occurrence was

267   predicted to be highest on west-facing slopes (higher aspect values, averaged over 1 and 2

268   km), which were relatively steep (averaged over 0.5 km), that both gained a relatively high

269   proportion of canopy cover between 2000 and 2012 and lost a relatively high proportion of

270   canopy in 2014 (both averaged over 0.5 km), and where (averaged over 2 km) few people

271   lived (Fig. 2).

272

273   The fragmentation of forest loss was also an important predictor of *P. knowlesi* occurrence

274   but only at relatively large spatial scales (e.g. 4-5km, Fig. 1f and 1h). A similar pattern was

275   observed both for the fragmentation of forest loss in the previous year (peak at 5 km) and in

276   the previous five years (peaks at 4 km and 5 km), with the highest probability of *P. knowlesi*

277   occurrence predicted when the landscape distribution of forest loss was most fragmented

278   on these scales (Fig. 2).

279

280   The fragmentation of cleared land (as distinct from forest loss, see Table 1) in the previous

281   year was important at 5 km (Fig. 1d), as well as at three other scales (0.1, 0.2 and 0.5 km).

282   The importance of three consecutive scales for one variable is likely to be due to correlation

283   across scales, and correlations were high in this case (Fig. S3d). However, the correlation

284   between small (0.1, 0.2 and 0.5 km) and large scale (5 km) aggregations was substantially

285   lower (Fig. S3d), which might suggest a real biological influence of this variable on two

286   scales simultaneously. However, as the variance in this predictor variable was correlated

287   with RVI (Fig. S4) at small spatial scales, the possibility of their importance being artefactual

288   at these scales cannot be ruled out, as higher variance is likely to lead to more frequent

289   inclusion of variables in the decision trees that make up BRT models. The same

290   interpretational caveat applies to the standard deviation of NDVI at 0.1km (Fig. S4).

291

292   *Variance across scales*

293   In general, the peaks of RVI (Figure 1) do not arise from an artefact of correlation with

294   variance (Fig. S4 and Table S1). However, in the case of the fragmentation of cleared land in

295   the previous year, some caution is required in the interpretation of the importance of the

296   smaller spatial scales. First, the comparison of variance with RVI across scales (Figure S4d)

297   and their correlation (Table S1) suggest that RVI may be influenced by variance available to

298   the model. Second, as the grid cells that make up the landscape variable layers are square,

299   the perimeter length of patches will be overestimated at small scales [35]. In addition, the

300   marginal effect curve for cleared P:A (previous year) at 5 km covers a greater range of

301   predicted probability than those at the smaller scales of 0.1, 0.2 and 0.5 km (Fig. 2).

302

303   Although the standard deviation of NDVI at 0.1 km appears in the top 16 variable-scale

304   combinations, the same caveat relating to changing variance across scales applies as above

305   because RVI tracks variance (Fig. S4). Therefore, it is possible that 0.1 km emerges as the

306   most important scale due to an artefact of variance available to the model, rather than due

307   to the influence of an underlying biological process on this scale. In addition, the marginal

308   effect curve for SD NDVI 0.1 km does not suggest a strong influence on *P. knowlesi*

309 occurrence probability (Fig. 2). The same applies to the importance of cover P:A at 0.1 km,

310 as RVI tracks variance across scales (Fig. 2 and Table S1), and perimeters will be over-

311 estimated at small scales.

312

313 *Non-scaled variables*

314 The median prediction accuracy (area under the receiver operator curve, AUC) of *P.*

315 *knowlesi* occurrence across the whole study site was 0.76. The inclusion of two non-scalable

316 variables, the shortest distance from households to the nearest clinic and road were

317 included, increased this to 0.78. The shortest distance to road had the highest RVI in this

318 model (Fig. S1b), with the probability of *P. knowlesi* occurrence predicted to be highest at

319 households furthest from roads (Fig. S2). The addition of the two non-scalable variables only

320 increased median AUC by 0.02, and gave rise to only minor changes in the most important

321 variable-scale combinations (Fig. S1) and negligible differences in their marginal effect

322 curves (Fig. 2 and S2). This suggests much of the variation explained by distance to roads

323 and clinics is explained by included landscape factors; for example, distance to roads is likely

324 highly correlated with population density and forest cover. This model was used to generate

325 *P. knowlesi* human case occurrence predictions for all the households (Fig. 3a). The

326 corresponding plot of prediction error by household shows there is little clustering of

327 prediction error in space, and therefore that the model is not overly influenced by

328 households in one area (Fig. 3b).

329

330 *Case clusters*

331 The division of case locations only (n = 206) by the marginal occurrence probabilities of the

332 whole-study-site model into two clusters produced one cluster of 93 cases (cluster A) and

333   another of 113 cases (cluster B). The two clusters appear to be spatially distinct, with cluster

334   A mainly occurring on the mainland of the district of Kudat, and cluster B occurring on the

335   island of Banggi and in the south of the Kudat peninsula (Fig. 2c). Exploration of the

336   differences between clusters by examination of the 15 variable-scale combinations with the

337   highest median marginal probability differences between clusters showed that cases in

338   cluster A were characterised by low canopy cover, high proportion of cleared land and high

339   population density at large spatial scales (Fig. S5).

340

341   *Prediction accuracy across scales*

342   The ability of single-scale BRT models to predict *P. knowlesi* occurrence varied from an AUC

343   of 0.55 (little better than a random model) to a maximum of 0.82. Models fitted to the

344   smallest spatial scales had the lowest predictive power, those fitted to intermediate scales

345   had the highest predictive power, and models that included all scales simultaneously

346   performed better on average than all single-scale models (Fig. S6).

347

348   **DISCUSSION**

349

350   A key unanswered question about *P. knowlesi* transmission is what mechanism(s) give rise

351   to the observed association between deforestation and human *P. knowlesi* incidence [28].

352   This study examines the influence of the absence of forest (cleared land), the process of

353   forest loss, and the landscape distribution of forest loss (fragmentation) by spatial scale.

354   This not only provides evidence that landscape fragmentation influences *P. knowlesi*

355   spillover into humans, as it is thought to for other zoonoses such as Lyme disease [36] and

356   Ebola [37], but also identifies the spatial scale of the influence of fragmentation on knowlesi

357   transmission (within 4 and 5 km of households).

358

359   Consideration of the multiple spatial scales identified by this new analytical approach with

360   corresponding marginal effect curves can suggest drivers of the observed patterns of

361   disease occurrence. The effects of human, macaque and vector movement and density likely

362   contribute to the spatial scale at which different landscape factors are predictive. For

363   example, if individuals are exposed outside the house, the large-scale influence of the

364   fragmentation of deforested areas (4-5 km) could emerge as a property of *P. knowlesi*

365   spillover if humans commuted to fragmented deforested areas over distances of up to 5 km,

366   and/or were at risk while there because of the nature of their work. This is consistent with

367   the findings of a case-control study undertaken in the same area, including an increased risk

368   of knowlesi (but not non-knowlesi) malaria in those walking to or from work or school [29].

369   Alternatively, macaque troops may respond to deforestation on this emergent scale,

370   because they move distances of up to 5 km in response to fragmentation beyond a

371   threshold, exposing households in sink areas to an increase in macaque density, which

372   would be consistent with what estimates there are of *M. fascicularis* home ranges [38]. The

373   step-like marginal effect curve of the fragmentation of deforestation on the probability of *P.*

374   *knowlesi* occurrence suggests such a threshold effect. In addition, increasing values of the

375   fragmentation of cleared land at 5 km predicted a similar step-like increase in occurrence

376   probability. This suggests that the deforestation fragmentation result is not only an effect of

377   the immediate disturbance of forest removal on *P. knowlesi* transmission, but one that is

378   rather (or also) influenced by the habitat geometry it leaves behind [39]. Although 5km was

379   chosen as the maximum distance due to village distribution and the small spatial scale of

380 this study site (including islands), future work could explore whether landscape variables

381 influence transmission at larger distances or explore the mechanisms behind these

382 associations.

383

384 The probability of *P. knowlesi* occurrence was highest when the proportion of cleared land

385 within 1 km of households was low. This suggests that households isolated in patches of

386 forest or plantation (with less than 10 % of the area within 1 km cleared) may be at the

387 highest *P. knowlesi* exposure risk. This is in line with the traditional man-in-the-forest

388 human *P. knowlesi* risk profile, which suggests that individuals who work on clearing forest

389 or on plantations (usually adult men) are at highest risk of *P. knowlesi* infection, and

390 additionally consistent with studies describing high vector densities in forest areas [22, 40].

391 When averaged over this same scale, aspect also had an important influence on predicted *P.*

392 *knowlesi* occurrence. Aspect is associated with *P. falciparum* infection in humans [30] but is

393 identified here as a potential determinant of *P. knowlesi* human infection risk for the first

394 time. As households situated on west-facing slopes had the highest probabilities of disease,

395 this may plausibly be because these households receive more sunlight in the afternoon,

396 resulting in higher temperatures. For *P. falciparum,* increased temperature has been shown

397 to shorten the duration of the incubation period in the mosquito or the length of the

398 gonotrophic cycle, or speed up the development or increase the survival probability [41,

399 42]. Alternatively, this association could arise through correlation between aspect and

400 agricultural practice, with the peak of aspect RVI at 1 km arising from the way people

401 modify (and the way both people and macaques use) agricultural land near households. *P.*

402 *knowlesi* occurrence was also predicted to be higher at households on relatively steep

403 slopes, which, as for aspect discussed above, could be a result of the influence of

404   temperature on mosquito life history and infection dynamics, and/or the way that humans

405   and macaques respond to slope. For example, if relatively steep slopes are uncultivatable,

406   they may provide refuge from disturbance for macaques. That canopy regrowth (gain all

407   years, Table 1) had high RVI at the same scale as slope, suggests that peridomestic land use

408   has an important influence over this scale, and therefore that the latter interpretation is

409   more likely. Although this study has not equivocally identified mechanisms by which land

410   use change influences human *P. knowlesi* infection risk, by mining the extra information

411   contained within the spatial scale signatures of associations it has pared down the many

412   plausible possibilities to a manageable number for further investigation. Future studies

413   could additionally expand this analysis to evaluate the impact of different land use or forest

414   types.

415

416   A challenge to a synthesis of *P. knowlesi* epidemiology across South East Asia is the

417   considerable regional variation in infection patterns and risk profiles. The degree to which

418   infection risk is concentrated in men who work in forests or plantations, the extent to which

419   peridomestic transmission occurs, and whether human-vector-human transmission occurs

420   under natural conditions are open questions [29, 43, 44]. Cluster analysis partitioned cases

421   occurring in this part of Malaysian Borneo into two geographical groups, each with distinct

422   risk profiles. Cluster A cases occurred at households around which where there was

423   relatively low forest cover, relatively high proportions of cleared land, relatively high

424   population density, and that were immediately surrounded by fragmented forest cover

425   compared with cluster B cases. These differences may reflect regional variation in the

426   history of land use – the conversion of forest on the island of Banggi from the coast inwards,

427   for example – and therefore the distinction between two sets of drivers of *P. knowlesi*

428 spillover from macaques to humans. This novel approach to identifying transmission

429 heterogeneities in disease occurrence datasets could be refined through integration with

430 other sources of data, such as travel histories and human GPS tracking data, and developed

431 into an effective tool for the surveillance of epidemiological transitions [45].

432

433 **CONCLUSION**

434

435 The consideration of multiple spatial scales can add value to analysis of disease occurrence

436 by delivering more accurate spatial predictions, and identifying the key spatial scales of

437 transmission. In the case of *P. knowlesi*, the application of a data mining approach has

438 teased apart the potentially conflicting influences of forest cover and forest loss [28] on

439 disease occurrence, identifying the latter as an effect of fragmentation on relatively large

440 spatial scales and the former as an effect of the proportion of cleared land nearer to

441 households. This could provide the key to the prediction of disease risk under models of

442 future land use, and the design of spatially-targeted disease interventions. This new scale-

443 focussed approach could be widely applied to other zoonoses and vector-borne diseases of

444 public health concern.

FIGURE & TABLE LEGENDS

454

455 Table 1. The ten scalable landscape variables classified from Landsat satellite imagery used

456 in the analysis [26]. Grid cells estimated as > 50 % tree crown cover density by were defined

457 as forested. Perimeter area ratio (P:A) was used as a proxy for fragmentation as variation in

458 P:A was more evenly distributed across variables than any other measure.

459

460 Figure 1. Relative variable importance (RVI) of all variable-scale combinations from BRT

461 models of *P. knowlesi* occurrence (176 predictors). See Table 1 for variable definitions.

462 Green points represent the whole-study-site, blue points the mainland-only model. Purple

463 boxes indicate the 16 variable-scale combinations with the highest RVIs, detail of which is

464 shown in Figure S1a.

465

466 Figure 2. Marginal effect curves of the 16 variable-scale combinations with the highest

467 relative variable importance across the whole study site (176 predictors)

468

469 Figure 3. The locations of all households included in the study, showing a) occurrence

470 probability predictions from the whole-study-site model (176 predictors); b) the prediction

471 error from the same model; and c) the location of the two clusters of case households.

Table 1.

Variable name	Details	Composite year
Cover (previous year)	*Proportion of forested grid cells*	*2014*
Cover P:A (previous year)	*Perimeter area ratio of forested grid cells*	*2014*
Cleared (previous year)	*Proportion of non-forested grid cells*	*2014*
Cleared P:A (previous year)	*Perimeter area ratio of non-forested grid cells*	*2014*
Loss (previous year)	*Proportion of grid cells that changed from forested to non-forested*	*2014*
Loss P:A (previous year)	*Perimeter area ratio of grid cells that changed from forested to non-forested*	*2014*
Loss (previous 5 years)	*Proportion of grid cells that changed from forested to non-forested*	*2010-2014*
Loss P:A (previous 5 years)	*Perimeter area ratio of grid cells that changed from forested to non-forested*	*2010-2014*
Gain (all years)	*Proportion of grid cells that changed from non-forested to forested*	*2000-2012*
Gain P:A (all years)	*Perimeter area ratio of grid cells that changed from forested to non-forested*	*2000-2012*
NDVI	*Normalised difference vegetation index, calculated from composite Landsat image*	*2014*
NDVI SD	*Standard deviation of normalised difference vegetation index, calculated from composite Landsat image*	*2014*
Elevation	*Metres above sea level (ASTER Global Digital Elevation Model)*	*2014*
Slope	*Maximum rate of change in elevation, calculated from ASTER GDEM*	*2014*
Population density	*Population density estimates*	*2010*

Aspect	*Direction of the steepest down slope (in degrees), calculated from ASTER DGEM*	*2014*

474

475 1. Bhatt, S., et al., *The effect of malaria control on Plasmodium falciparum in Africa*
476   *between 2000 and 2015.* Nature, 2015. **526**(7572): p. 207-211.
477 2. Gething, P.W., et al., *Declining malaria in Africa: improving the measurement of*
478   *progress.* Malar J, 2014. **13**: p. 39.
479 3. Hay, S.I., et al., *Global mapping of infectious disease.* Philos Trans R Soc Lond B Biol
480   Sci, 2013. **368**(1614): p. 20120250.
481 4. Pigott, D.M., et al., *Prioritising Infectious Disease Mapping.* PLoS Negl Trop Dis, 2015.
482   **9**(6): p. e0003756.
483 5. Sturrock, H.J.W., et al., *Mapping Malaria Risk in Low Transmission Settings:*
484   *Challenges and Opportunities.* Trends Parasitol, 2016. **32**(8): p. 635-645.
485 6. Shearer, F.M., et al., *Estimating Geographical Variation in the Risk of Zoonotic*
486   *Plasmodium knowlesi Infection in Countries Eliminating Malaria.* PLoS Negl Trop Dis,
487   2016. **10**(8): p. e0004915.
488 7. Alegana, V.A., et al., *Advances in mapping malaria for elimination: fine resolution*
489   *modelling of Plasmodium falciparum incidence.* Sci Rep, 2016. **6**: p. 29628.
490 8. Bhatt, S., et al., *The global distribution and burden of dengue.* Nature, 2013.
491   **496**(7446): p. 504-7.
492 9. Messina, J.P., et al., *The global distribution of Crimean-Congo hemorrhagic fever.*
493   Trans R Soc Trop Med Hyg, 2015. **109**(8): p. 503-13.
494 10. Messina, J.P., et al., *Mapping global environmental suitability for Zika virus.* Elife,
495   2016. **5**.
496 11. Pigott, D.M., et al., *Updates to the zoonotic niche map of Ebola virus disease in*
497   *Africa.* Elife, 2016. **5**.
498 12. Han, B.A., et al., *Rodent reservoirs of future zoonotic diseases.* Proc Natl Acad Sci U S
499   A, 2015. **112**(22): p. 7039-44.
500 13. Evans, M.V., et al., *Data-driven identification of potential Zika virus vectors.* Elife,
501   2017. **6**.
502 14. Escobar, L.E. and M.E. Craft, *Advances and Limitations of Disease Biogeography*
503   *Using Ecological Niche Modeling.* Front Microbiol, 2016. **7**: p. 1174.
504 15. Stephens, P.R., et al., *The macroecology of infectious diseases: a new perspective on*
505   *global-scale drivers of pathogen distributions and impacts.* Ecol Lett, 2016. **19**(9): p.
506   1159-71.
507 16. Elith, J., J.R. Leathwick, and T. Hastie, *A working guide to boosted regression trees.* J
508   Anim Ecol, 2008. **77**(4): p. 802-13.
509 17. Brady, O.J., et al., *Vectorial capacity and vector control: reconsidering sensitivity to*
510   *parameters for malaria elimination.* Trans R Soc Trop Med Hyg, 2016. **110**(2): p. 107-
511   17.
512 18. Parratt, S.A., E. Numminen, and A. Laine, *Infectious disease dynamics in*
513   *heterogeneous landscapes.* Annual Review of Ecology, Evolution, and Systematics,
514   2016. **47**: p. 283-306.
515 19. Stefani, A., et al., *Studying relationships between environment and malaria incidence*
516   *in Camopi (French Guiana) through the objective selection of buffer-based landscape*
517   *characterisations.* International Journal of Health Geographics, 2011. **10**(65).
518 20. Wardrop, N.A., et al., *Interpreting predictive maps of disease: highlighting the pitfalls*
519   *of distribution models in epidemiology.* Geospat Health, 2014. **9**(1): p. 237-46.
520 21. Daneshvar, C., et al., *Clinical and laboratory features of human Plasmodium knowlesi*
521   *infection.* Clin Infect Dis, 2009. **49**(6): p. 852-60.

522   22.   *Wong, M.L., et al., Seasonal and Spatial Dynamics of the Primary Vector of*
523   *Plasmodium knowlesi within a Major Transmission Focus in Sabah, Malaysia. PLoS*
524   *Negl Trop Dis, 2015. **9**(10): p. e0004135.*
525   23.   *Singh, B., et al., A large focus of naturally acquired Plasmodium knowlesi infections in*
526   *human beings. Lancet, 2004. **363**(9414): p. 1017-24.*
527   24.   *Lubis, I.N., et al., Contribution of Plasmodium knowlesi to multi-species human*
528   *malaria infections in North Sumatera, Indonesia. J Infect Dis, 2017.*
529   25.   *Moyes, C.L., et al., Defining the geographical range of the Plasmodium knowlesi*
530   *reservoir. PLoS Negl Trop Dis, 2014. **8**(3): p. e2780.*
531   26.   *Hansen, M.C., et al., High-resolution global maps of 21st-century forest cover*
532   *change. Science, 2013. **342**(6160): p. 850-3.*
533   27.   *World Health Organisation Regional Office for Western Pacific, Expert consultation*
534   *on Plasmodium knowlesi malaria to guide malaria elimination strategies. 2017,*
535   *World Health Organization: Manila, Philippines.*
536   28.   *Fornace, K.M., et al., Association between Landscape Factors and Spatial Patterns of*
537   *Plasmodium knowlesi Infections in Sabah, Malaysia. Emerg Infect Dis, 2016. **22**(2): p.*
538   *201-8.*
539   29.   *Grigg, M.J., et al., Individual-level factors associated with the risk of acquiring human*
540   *Plasmodium knowlesi malaria in Malaysia: a case control study. Lancet Planetary*
541   *Health, 2017. **1**: p. e97-104.*
542   30.   *Weiss, D.J., et al., Re-examining environmental correlates of Plasmodium falciparum*
543   *malaria endemicity: a data-intensive variable selection approach. Malar J, 2015. **14**:*
544   *p. 68.*
545   31.   *Land Processes Distributed Active Archive Center (LP DAAC), Advanced Spaceborne*
546   *Thermal Emission and Reflection Radiometer Global Digital Elevation Model (ASTER*
547   *GDEM) Version 2. 2015, NASA EOSDIS Land Processes DAAC, USGS Earth Resources*
548   *Observatoin and Science (EROS) Center: Sioux Falls, South Dakota.*
549   32.   *Barbet-Massin, M., et al., Selecting pseudo-absences for species distribution models:*
550   *how, where and how many? Methods in Ecology and Evolution, 2012. **3**: p. 327-338.*
551   33.   *Ward, J.H., Hierarchical grouping to optimize an objective function. Journal of the*
552   *American Statistical Association, 1963. **58**(301): p. 236-244.*
553   34.   *Fornace, K.M., et al., Exposure and infection to Plasmodium knowlesi in case study*
554   *communities in Northern Sabah, Malaysia and Palawan, The Philippines. PLoS Negl*
555   *Trop Dis, 2018. **12**(6): p. e0006432.*
556   35.   *Hargis, C.D., J.A. Bissonette, and J.L. David, The behaviour of landscape metrics*
557   *commonly used in the study of habitat fragmentation. Landscape Ecology, 1998. **13**:*
558   *p. 167-186.*
559   36.   *Allan, B.F., F. Keesing, and R.S. Ostfeld, Effect of forest fragmentation on Lyme*
560   *Disease risk. Conservation Biology, 2003. **17**(1): p. 267-272.*
561   37.   *Rulli, M.C., et al., The nexus between forest fragmentation in Africa and Ebola virus*
562   *disease outbreaks. Sci Rep, 2017. **7**: p. 41613.*
563   38.   *Fooden, J., Systematic Review of Southeast Asian Longtail Macaques Macaca*
564   *fascicularis. Fieldiana, 1995. **81**.*
565   39.   *Tucker Lima, J.M., et al., Does deforestation promote or inhibit malaria transmission*
566   *in the Amazon? A systematic literature review and critical appraisal of current*
567   *evidence. Philos Trans R Soc Lond B Biol Sci, 2017. **372**(1722).*

568   40.   Barber, B.E., et al., A prospective comparative study of knowlesi, falciparum, and
569   vivax malaria in Sabah, Malaysia: high proportion with severe disease from
570   Plasmodium knowlesi and Plasmodium vivax but no mortality with early referral and
571   artesunate therapy. Clin Infect Dis, 2013. **56**(3): p. 383-97.
572   41.   Weiss, D.J., et al., Air temperature suitability for Plasmodium falciparum malaria
573   transmission in Africa 2000-2012: a high-resolution spatiotemporal prediction. Malar
574   J, 2014. **13**: p. 171.
575   42.   Mordecai, E.A., et al., Optimal temperature for malaria transmission is dramatically
576   lower than previously predicted. Ecol Lett, 2013. **16**(1): p. 22-30.
577   43.   Manin, B.O., et al., Investigating the Contribution of Peri-domestic Transmission to
578   Risk of Zoonotic Malaria Infection in Humans. PLoS Negl Trop Dis, 2016. **10**(10): p.
579   e0005064.
580   44.   Brock, P.M., et al., Plasmodium knowlesi transmission: integrating quantitative
581   approaches from epidemiology and ecology to understand malaria as a zoonosis.
582   Parasitology, 2016. **143**(4): p. 389-400.
583   45.   Han, B.A. and J.M. Drake, Future directions in analytics for infectious disease
584   intelligence: Toward an integrated warning system for emerging pathogens. EMBO
585   Rep, 2016. **17**(6): p. 785-9.
586