Machine learning with remote sensing data to locate uncontacted indigenous villages in Amazonia

Robert S. Walker [1], Marcus J. Hamilton [2, 3]

[1] Department of Anthropology, University of Missouri, Columbia MO
[2] Department of Anthropology, University of Texas at San Antonio, San Antonio TX
[3] Santa Fe Institute, Santa Fe, NM

Corresponding Author:
Robert S. Walker [1]
Department of Anthropology, University of Missouri, Columbia MO USA 65203
Email address: walkerro@missouri.edu

Abstract

Background. The world's last uncontacted indigenous societies in Amazonia have only intermittent and often hostile interactions with the outside world. Knowledge of their locations is essential for urgent protection efforts, but their extreme isolation, small populations, and semi-nomadic lifestyles make this a challenging task.

Methods. Remote sensing technology with Landsat satellite sensors is a non-invasive methodology to track isolated indigenous populations through time. However, the small-scale nature of the deforestation signature left by uncontacted populations clearing villages and gardens has similarities to those made by contacted indigenous villages. Both contacted and uncontacted indigenous populations often live in proximity to one another making it difficult to distinguish the two in satellite imagery. Here we use machine learning techniques applied to remote sensing data with a training dataset of 500 contacted and 25 uncontacted villages.

Results. Uncontacted villages generally have smaller cleared areas, reside at higher elevations, and are farther from populated places and satellite-detected lights at night. A random forest algorithm with an optimally-tuned detection cutoff has a leave-one-out cross-validated sensitivity and specificity of over 98%. A grid search around known uncontacted villages led us to identify 3 previously-unknown villages using predictions from the random forest model. Our efforts can improve policies toward isolated populations by providing better near real-time knowledge of their locations and movements in relation to encroaching loggers, settlers, and other external threats to their survival.

Introduction

The ongoing colonization of Amazonia has brought waves of epidemics and violence for centuries with severe consequences for indigenous populations (Bodard, 1974; Hemming, 1978; Hurtado et al. 2001; Hamilton, Walker & Kesler, 2014). Amazingly, despite all the external pressures, remote areas in the upper Amazon watershed still support a number of remnant indigenous societies generally referred to as uncontacted or isolated populations (Vaz, 2001; Castillo, 2004; Ricardo & Ricardo 2011). Despite these labels, intermittent and often hostile interactions with the outside world are commonplace (Wallace, 2011). Most governmental and non-governmental organizations promote no-contact policies for these isolated indigenous populations with the belief that they are safest if left to themselves (Walker & Hill, 2015). However, encroachment from loggers, miners, settlers, and others is incessant and uncontacted societies represent the world's most critically endangered cultures (Walker, Kesler, & Hill, 2016). There is a need for good information on their locations and movements in hopes of improving their survival prospects moving forward.

Our project is part of a longitudinal remote surveillance program to conduct scientific studies of indigenous demography and spatial ecology to facilitate informed decisions by policy makers that will increase protection efforts for isolated indigenous populations (Walker & Hamilton, 2014; Walker, Hamilton & Groth, 2014). Our central goal is to gather as much information on isolated indigenous populations as possible without attempting any direct contact (Kesler & Walker, 2015). We maximize the use of available technologies to gather data remotely with no interference. Satellite imagery offers a safe, low-cost, and noninvasive method for studying population dynamics and spatial ecology of indigenous populations (Walker, Kesler, & Hill, 2016). Similarly important is the need to understand spatial resource needs of indigenous societies in a region heavily impacted by deforestation, as well as the potential importance of connections among subpopulations, known to contribute to population viability (Levins, 1969; Hanski, 1999). The irreversible threats from large-scale habitat loss via deforestation and conversion of land to agriculture and pasture paint a bleak future for uncontacted populations (Fagan & Shoobridge, 2005; Salisbury & Fagan, 2013; Walker, Kesler & Hill, 2016). The hope is that better data and methods can contribute improvements to this complex issue.

Applied machine learning is a vital tool for conservation work as a means to both collect and analyze more data at faster rates (Murray et al. 2018a). The growing use of machine learning methods to analyze large sets of biological, biophysical, spectral and climatological data has enabled accurate differentiation of the world's landscapes (Pettorelli et al. 2014). More germane to our work are forest classification projects (Hansen et al., 2013, Murray et al. 2018b). The Global Forest Change dataset was developed by classifying pixels using 15 or more high-resolution global composite images as predictors, each developed from over 500,000 Landsat images (Hansen et al., 2013).

80

81 The random forest algorithm is known to give excellent classification results and relatively quick
82 processing speed (Du et al., 2015, Pal, 2005, Rodriguez-Galiano et al., 2012). Random forests
83 (Breiman 2001) are an ensemble supervised learning method that builds multiple decisions trees
84 used here for the classification of village class (uncontacted versus contacted). Random forests
85 operate by constructing a multitude of decision trees. Some of the advantages of random forests
86 are that they are robust to inclusion of features that are irrelevant to classification, and they are
87 invariant to transformations of feature variables (Belgiu and Drăguţ 2016). For these reasons, the
88 random forest algorithm is popular for remote sensing data given its accuracy, speed, and ability
89 to handle high data dimensionality and multicollinearity.

90

Materials & Methods

92

93 **Data.** We combined the exact locations (centroids) of 25 uncontacted and 500 contacted
94 indigenous villages (Walker, Kesler & Hill, 2016). More information about our general project
95 along with high-resolution imagery for uncontacted villages is available at
96 https://isolatedtribes.missouri.edu. The locations of uncontacted villages were originally derived
97 from scouring high-resolution imagery using a combination of undergraduate helpers and various
98 maps made by governmental and non-governmental agencies in Colombia, Ecuador, Peru and
99 especially Brazil. Several additional locations have been pieced together from governmental
100 reports and news stories stemming from overflights. Contacted villages are from the Brazilian
101 government website (http://www.funai.gov.br/), and we included all of those that were in
102 western Amazonia (west of 60 degrees longitude, Figure 1).

103

104

105 Hansen and colleagues' (2013) Global Forest Change (GFC) project provides small-scale
106 deforestation at approximately 30 m resolution from Landsat sensors extending back to the year
107 2000. GFC version 1.5 goes up through the year 2017. We extracted the amount of detected
108 deforestation in 2x2 km squares surrounding each village's centroid and took the maximum area
109 cleared in any one particular year from across the 17-year period. We refer to this measure as
110 cleared area as it includes both the village and associated gardens but not those of neighboring
111 villages. In addition, our dataset has other features, including regional population density in the
112 nearest 100 square km (CIESIN, 2005), elevation at 30 m digital resolution from the Space
113 Shuttle Radar Topography Mission (Rabus et al. 2003), and distance to populated places at 10 m
114 resolution (Balk et al. 2006). We also included a local lights-at-night measure at 3 km resolution
115 (Pritchard, 2017, from https://earthobservatory.nasa.gov) using the distance from village centroid
116 to the nearest detected lights. Finally, distance to rivers of all the different Strahler stream orders
117 using the Global Self-consistent, Hierarchical, High-resolution Geography Database (Wessel &
118 Smith, 1996), along with the minimum distance to combined rivers of Strahler stream orders 1,
119 2, and 3, giving a total of 11 features used to train algorithms.

120

121   **Models.** Machine learning algorithms were performed with the R package caret. We found that
122   an untuned random forest algorithm had a fairly high combination of sensitivity (true positive
123   rate) and specificity (true negative rate) in the 0.8 to 0.9 range. As mentioned above, random
124   forest algorithm is an ensemble classifier that produces multiple decision trees, using a randomly
125   selected subset of training samples and variables. Other algorithms such as neural networks,
126   extreme gradient boosting tree, and lasso logistic regression were also relatively-high performing
127   but gave slightly lower values on one or the other metric.

128

129   The target classes in our sample are imbalanced with only 4.8% of villages in the sample being
130   uncontacted. During model training we noticed that varying the detection cutoff (also known as
131   the threshold) that classifies villages into one class or the other had large effects on the results
132   (the default cutoff is 0.5 majority rule). In addition, common loss metrics such as the area under
133   the ROC curve or the F1 score tended to give either high specificity or sensitivity with our data,
134   but not both.

135

136   To address the imbalanced data issue and improve model performance, we used a random forest
137   algorithm that iteratively tuned the cutoff value such as to simultaneously maximize both
138   specificity (true negative rate) and sensitivity (true positive rate). In other words, we instituted
139   cost-sensitive learning into the random forest (Elkan, 2001; Zadrozny et al. 2003; Khoshgoftaar
140   et al. 2007). The loss metric we used for training is the distance from a perfect model of
141   sensitivity of 1 and specificity of 1. We used 1,000 trees with 2 variables available for splitting at
142   each tree node. To evaluate models we used a leave-one-out cross-validation (non-nested) looped
143   over a range of cutoffs from 0.01 to 0.99 in increments of 0.01. Raising the cutoff value means a
144   higher level of evidence (i.e., more decision trees out of the total 1,000 trees that comprise the
145   random forest) is needed to assign the positive class (uncontacted) so it decreases sensitivity and
146   increases specificity. Here a sensitive cutoff of 0.2 yields a minimal distance metric and the
147   desired combination of high sensitivity and specificity metrics (Figure 2).

148

149   # Results

150

151   Our random forest algorithm, with an optimally-tuned cutoff of 0.2, yields a sensitivity of 1.0
152   and a specificity of 0.98 using leave-one-out cross-validation. This means that all uncontacted
153   villages are correctly classified and 98% of the contacted villages are correctly classified.
154   Therefore, our model has a strong ability to automatically distinguish between contacted and
155   uncontacted villages. In order of descending variable importance, uncontacted villages have 1)
156   smaller cleared areas, 2) longer distances from lights, 3) higher elevation, 4) longer distances to
157   populated places, 5) lower regional population density, 6) longer distances from rivers of all
158   Strahler stream orders up to and including 3, and 7) shorter distances to rivers of levels 4 and 5.
159   Figure 3 shows density plot comparisons for the top 4 features in terms of variable importance.

160

161 Given the success of our algorithm during cross-validation, we then moved to implement it for

162 predictive purposes. We did a grid search of all 2x2 km squares within a 100 km radius of the 5

163 clusters of known uncontacted villages (Figure 1). This approach does produce a high number of

164 false positives created by natural clearings (e.g., landslides, windfalls, etc.). Fortunately, most

165 natural clearings can be eliminated by simply removing all clearings that are less than 0.5 ha.

166 This left us with a sample of 20 clearings. Of these we were able to obtain high resolution

167 imagery for 8 of these and 3 contained newly-identified villages. One of these in Colombia

168 appears to be currently inhabited given that it has a single longhouse structure and shows

169 recently made clearings in Global Land Analysis and Discovery (GLAD, Tyukavina et al. 2016).

170 The GLAD alert system processes Landsat imagery as it becomes available to identify tree cover

171 change in near real-time. This is an invaluable system for monitoring both recent activity by

172 uncontacted villages, as well as encroaching deforestation from outsiders.

173

174 The other two newly-discovered sites are historical villages. One is from the uncontacted

175 Yanomami in northern Brazil inhabited from around year 2000 or earlier and until 2004. The

176 other is from Pano speakers on the border between Peru and Brazil and was probably inhabited

177 during a similar time period. The other 5 possible locations identified by the random forest

178 predictions with high resolution imagery available all appeared to be natural. Therefore, we

179 estimate our testing precision with this small sample as 0.375 (3 true positives divided by 8 total

180 cases).

181

182 ## Discussion

183

184 We used deforestation data from Landsat satellites to train algorithms to identify the locations of

185 uncontacted indigenous groups in Amazonia as part of an ongoing effort to better understand

186 their conservation status and threats. Our results show that uncontacted villages have smaller

187 cleared areas, reside at higher elevations, and are farther from populated places and satellite-

188 detected lights at night. Our random forest algorithm with an optimally-tuned cutoff has cross-

189 validated performance metrics of over 98%.

190

191 The case of the uncontacted Yanomami (also known as the Moxihatetea) is a good example of

192 the importance of a near real-time monitoring system. Their previous village was abandoned in

193 late 2014 and the Brazilian indigenous agency (FUNAI) and the Yanomami indigenous

194 association (Hutukara) were particularly worried that some disaster had befallen them since

195 much of the nearby area has seen invasions by gold miners. For a year and a half their

196 whereabouts were unknown. We began looking for them using Landsat data, but it was the

197 remote sensing fire alerts (FIRMS, Davies et al. 2009) that first alerted us to their exact location.

198 We tasked a DigitalGlobe satellite image on May 12, 2016 and were relieved to find out that they

199 were alive and well and clearing large gardens. The number of sections in their *shabono* village

200 structure had increased from 16 to 17. We relayed this information on to FUNAI and Hutukara
201 who then organized a flyover to officially confirm the location.
202
203 Remote sensing provides many advantages over flyovers, and we actually do not recommend
204 them. As we have shown, the information provided solely by remote sensing is sufficient to
205 identify uncontacted villages. Remote sensing is safe, low-cost, and noninvasive, while flyovers
206 are not. Population estimates are also crucial information for assessing trends in the demographic
207 health of isolated populations by measuring areas of fields, villages, and houses in satellite
208 imagery. Heads-up digitization of satellite imagery provides better population estimates than do
209 flyovers where most people are not visible because many hide or run away in fear. Remote
210 sensing offers the benefits of time-stamped evidence of occupation of areas inhabited by isolated
211 populations, along with movements through time (Walker, Kesler & Hill, 2016).
212
Conclusions
214
215 A dozen easily obtainable remote sensing measures allowed our random forest algorithm to
216 successfully classify uncontacted versus contacted villages. Extending the algorithm to make
217 predictions in a grid search greatly accelerates our ability to find and identify the locations of
218 uncontacted villages. Moving forward we anticipate using an even lower cutoff value because
219 the decreasing costs in satellite imagery make false positives from a more sensitive algorithm
220 relatively cheap to evaluate and discard. We anticipate that this method will become the primary
221 means by which to track and locate these same uncontacted villages, as well as undiscovered
222 locations of uncontacted villages.
223
224 One shortcoming of our classification model when applied to searching through unlabeled
225 satellite imagery is that it was not designed to classify natural landslides, windfalls, or riverbank
226 clearings. All of these natural processes also create deforestation signatures that further
227 complicate our searches. Future work could well include these, but in the meantime we filter our
228 predictions based on cleared area because natural clearings tend to be less than 0.5 ha while most
229 uncontacted villages have larger areas than that.
230
231 Our research is vital and timely as isolated groups are among the last remaining small-scale
232 subsistence populations living in a traditional lifestyle. The enormous and mounting pressure
233 from external threats create the possibility that isolated populations will disappear in the near
234 future. Better monitoring and tracking with remote sensing are tools that might provide more
235 informed conservation decisions concerning increased protection and land rights for the world's
236 most critically-endangered human cultures.
237
Acknowledgements
239

240   We thank Mark Flinn and the Comparative Methods course at the University of Missouri for
241   their help and suggestions.
242

References

244

245   Balk DL, Deichmann U, Yetman G, Pozzi F, Hay SI, Nelson A. 2006. Determining global
246   population distribution: methods, applications and data. *Advances in Parasitology* 62:119-156.
247

248   Belgiu M, Drăguţ L. 2016. Random forest in remote sensing: A review of applications and future
249   directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114:24-31.
250

251   Bodard L. 1974. *Green hell: massacre of the Brazilian Indians*. New York: Dutton.
252

253   Breiman L. 2001. Random forests. *Machine Learning* 45:5-32.
254

255   Castillo BH. 2004. *Indigenous peoples in isolation in the Peruvian Amazon*. Copenhagen:
256   International Work Group for Indigenous Affairs.
257

258   Center for International Earth Science Information Network (CIESIN). 2005. Gridded population
259   of the world: population density grid. Columbia University, Centro Internacional de Agricultura
260   Tropical.
261

262   Davies DK, Ilavajhala S, Wong MM, Justice CO. 2009. Fire information for resource
263   management system: archiving and distributing MODIS active fire data. *IEEE Transactions on*
264   *Geoscience and Remote Sensing* 47:72-79.
265

266   Du P, Samat A, Waske B, Liu S, Li Z. 2015. Random forest and rotation forest for fully
267   polarized SAR image classification using polarimetric and spatial features. *ISPRS Journal of*
268   *Photogrammetry and Remote Sensing* 105:38-53.
269

270   Elkan C. 2001. The foundations of cost-sensitive learning. *Proceedings of the IEEE*
271   *International Joint Conference on Artificial Intelligence* 17:973–978.
272

273   Fagan C, Shoobridge D. 2005. *An investigation of illegal mahogany logging in Peru's Alto Purús*
274   *National Park and its surroundings*. Durham NC: ParksWatch.
275

276   Hamilton MJ, Walker RS, Kesler D. 2014. Crash and rebound of indigenous populations in
277   lowland South America. *Sci. Rep.* 4, 4541.
278

279   Hanski I. 1999. *Metapopulation ecology*. Oxford: Oxford University Press.
280

281   Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D,
282   Stehman SV, Goetz SJ, Loveland TR, Kommareddy A. 2013. High-resolution global maps of
283   21st-century forest cover change. *Science* 342:850-853.

284

285 Hemming J. 1978. *Red gold: the conquest of the Brazilian Indians*. Cambridge: Harvard
286 University Press.

287

288 Hurtado AM, Hill KR, Kaplan H, Lancaster J. 2001. The epidemiology of infectious diseases
289 among South American Indians: a call for guidelines for ethical research. *Curr. Anth.* 42:425-
290 432.

291

292 Kesler DC, RS Walker. 2015. Geographic distribution of isolated indigenous societies in
293 Amazonia and the efficacy of indigenous territories. *PLoS ONE* 10:e0125113.

294

295 Khoshgoftaar TM, Golawala M, Van Hulse J. 2007. An empirical study of learning from
296 imbalanced data using random forest. IEEE Artificial Intelligence Tools 2:310-317.

297

298 Levins, R. 1969. Some demographic and genetic consequences of environmental heterogeneity
299 for biological control. *Bull. Entomol. Soc. Am.* 15:237–240.

300

301 Murray NJ, Keith DA, Bland LM, Ferrari R, Lyons MB, Lucas R, Pettorelli N, Nicholson E.
302 2018a. The role of satellite remote sensing in structured ecosystem risk assessments. *Science of*
303 *the Total Environment* 619:249-257.

304

305 Murray NJ, Keith DA, Simpson D, Wilshire JH, Lucas RM. 2018b. REMAP: An online remote
306 sensing application for land cover classification and monitoring. *Methods in Ecology and*
307 *Evolution*. 9:2019-2027.

308

309 Pal M. 2005. Random forest classifier for remote sensing classification. *International Journal of*
310 *Remote Sensing* 26:217-222.

311

312 Pappalardo SE, De Marchi M, Ferrarese F. 2013. Uncontacted Waorani in the Yasuní Biosphere
313 Reserve: geographical validation of the Zona Intangible Tagaeri Taromenane (ZITT). *PLoS ONE*
314 8:e66293.

315

316 Pettorelli N, Laurance WF, O'Brien TG, Wegmann M, Nagendra H, Turner W. 2014. Satellite
317 remote sensing for applied ecologists: opportunities and challenges. *Journal of Applied Ecology*
318 51:839-848.

319

320 Pritchard SB. 2017. The trouble with darkness: NASA's Suomi satellite images of earth at night.
321 *Environmental History* 22:312-330.

322

323 Rabus B, Eineder M, Roth A, Bamler R. 2003. The shuttle radar topography mission—a new
324 class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of*
325 *Photogrammetry and Remote Sensing* 57:241-262.

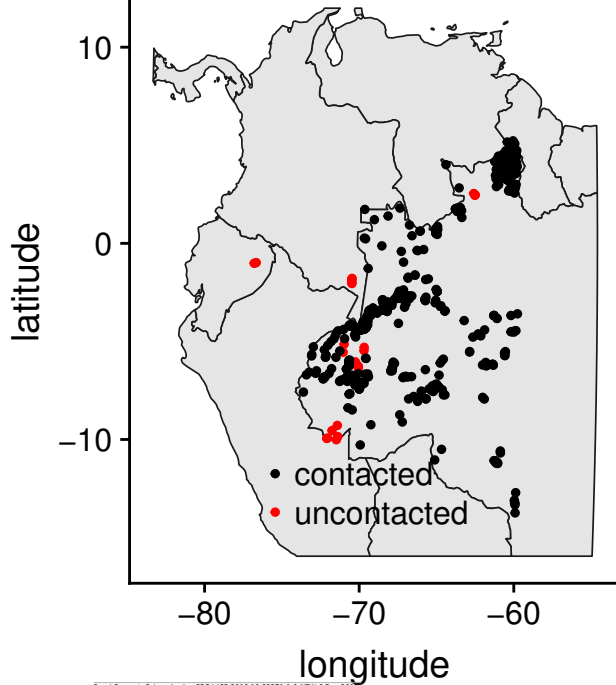
326

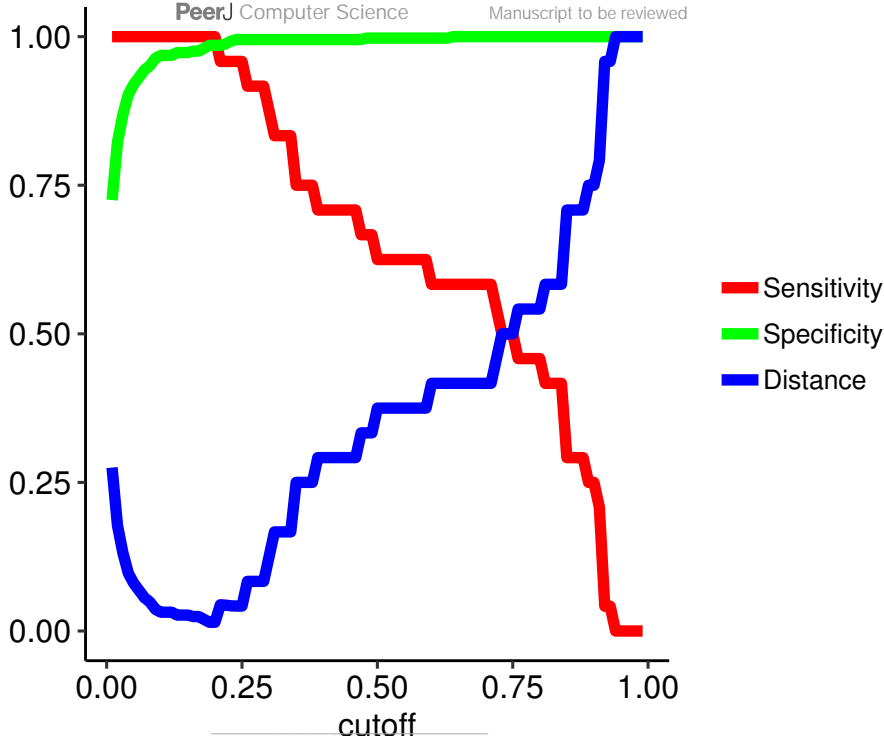
327 Ricardo B, Ricardo F. 2011. *Povos indígenas no Brasil*. São Paulo: Instituto Socioambiental.

328

329 Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M., Rigol-Sanchez JP. 2012. An
330 assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS*
331 *Journal of Photogrammetry and Remote Sensing* 67:93-104.
332
333 Salisbury DS, Fagan C. 2013. Coca and conservation: cultivation, eradication, and trafficking in
334 the Amazon borderlands. *GeoJ.* 78:41-60.
335
336 Tyukavina A, Hansen MC, Potapov PV, Krylov AM, Goetz SJ. 2016. Pan-tropical hinterland
337 forests: mapping minimally disturbed forests. *Global Ecology and Biogeography* 25:151-163.
338
339 Vaz A. 2011. *Isolados no Brasil. Política de estado: da tutela para a política de direitos – uma*
340 *questão resolvida?* Brasília: Estação Gráfica.
341
342 Walker RS, DC Kesler, KR Hill. 2016. Are isolated indigenous populations headed toward
343 extinction? *PLoS ONE* 11:e0150987.
344
345 Walker RS, Hill KR. 2015. Protecting isolated tribes. *Science* 348:1061.
346
347 Walker RS, Hamilton MJ. 2014 Amazonian societies on the brink of extinction. *Am. J. Hum.*
348 *Bio.* 26:570-572.
349
350 Walker RS, Hamilton MJ, Groth AA. 2014. Remote sensing and conservation of isolated
351 indigenous villages in Amazonia. *Royal Society Open Science* 1:140246.
352
353 Wallace S. 2011. *The Unconquered: In Search of the Amazon's Last Uncontacted Tribes*. New
354 York: Random House LLC.
355
356 Wessel P, Smith WHF. 1996. A global, self-consistent, hierarchical, high-resolution shoreline
357 database. *J. Geophys. Res.* 101:8741-8743, doi:10.1029/96JB00104.
358
359 Zadrozny B, Langford J, Abe N. 2003. Cost-sensitive learning by cost-proportionate example
360 weighting. *Proceedings of the IEEE International Conference on Data Mining* 3:435-442.
361

Figure 1. Map of study locations.

Map of 500 contacted indigenous villages in Brazil and 25 uncontacted indigenous villages in Brazil, Colombia, Ecuador, and Peru that were included in the study.

Model metrics obtained from training the random forest model across a range of cutoffs from 0.01 to 0.99 in increments of 0.01.

To train the random forest model we used leave-out-out cross-validation across a range of cutoffs from 0.01 to 0.99 in increments of 0.01. Raising the cutoff value means a higher level of evidence is needed to assign the positive class (uncontacted), which decreases sensitivity (true positive rate) and increases specificity (true negative rate). Here the optimal cutoff (0.2) gives a perfect cross-validated sensitivity of 1.0 and a specificity of 0.98. The distance is the distance from a perfect model which is minimized during training.

Smoothed kernel density plots comparing uncontacted to contacted indigenous villages.

The top 4 distinguishing features in terms of variable importance in the random forest model are uncontacted vilages have (A) smaller cleared areas, (B) farther distances to satellite-detected lights at night, (C) higher elevation, and (D) farther distances to populated places, on average. Plots A and C are best visualized on log scales.