Manuscript Click here to view linked References

Predictors of bovine *Schistosoma japonicum* infection in rural Sichuan, China

Elise Grover[1], Sara Paull[1], Katerina Kechris[2], Andrea Buchwald[1,3], Katherine James[1], Yang Liu[4], Elizabeth J. Carlton[1*]

[1] Department of Environmental and Occupational Health, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, USA

[2] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, USA

[3] Center for Vaccine Development and Global Health, University of Maryland School of Medicine, Baltimore, USA

[4] Institute of Parasitic Diseases, Sichuan Center for Disease Control and Prevention, Chengdu, China

* Corresponding author

Email: elizabeth.carlton@cuanschutz.edu

Note: Supplementary data is associated with this article.

Abstract

In China, bovines are believed to be the most common animal source of human schistosomiasis infections, though little is known about what factors promote bovine infections. The current body of literature features inconsistent, and sometimes contradictory results, and to date, few studies have looked beyond physical characteristics to identify the broader environmental conditions that predict bovine schistosomiasis. Because schistosomiasis is a sanitation-related, water-borne disease transmitted by many animals, we hypothesized that several environmental factors – such as the lack of improved sanitation systems, or participation in agricultural production that is water-intensive – could promote schistosomiasis infection in bovines. Using data collected as a part of a repeat cross-sectional study conducted in rural villages in Sichuan, China from 2007 to 2016, we used a Random Forests, machine learning approach to identify the best physical and environmental predictors of bovine *S. japonicum* infection. Candidate predictors included: 1) physical/biological characteristics of bovines, 2) human sources of environmental schistosomes, 3) socio-economic indicators, 4) animal reservoirs, and 5) agricultural practices. The density of bovines in a village and agricultural practices such as the area of rice and dry summer crops planted and the use of night soil as an agricultural fertilizer, were among the top predictors of bovine *S. japonicum* infection in all collection years. Additionally, human infection prevalence, pig ownership and bovine age were found to be strong predictors of bovine infection in at least one year. Our findings highlight that presumptively treating bovines in villages with high bovine density or human infection prevalence may help to interrupt transmission. Furthermore, village-level predictors were stronger predictors of bovine infection than household-level predictors, suggesting future investigations may need to apply a broad ecological lens to identify potential underlying sources of persistent transmission.

Keywords: schistosomiasis, buffalo, cattle, machine learning, prevention and control, China

1. Introduction

Bovines have long-been recognized as a key contributor to human schistosomiasis in Asia (Gray et al., 2007; Guo et al., 2006), and emerging evidence suggests they may be contributing to human schistosomiasis in Africa (Colley and Loker, 2018).Schistosomiasis is among the most burdensome helminth infections worldwide, with transmission being reported in a total of 78 countries in 2018 and approximately 230 million people in need of preventative treatment (World Health Organization, 2020). Although great strides have been made in the last few decades in global schistosomiasis control (World Health Organization, 2020), pockets of reemergent or persistent transmission highlight the need for careful consideration of possible local drivers of transmission (Kittur et al., 2019; Song et al., 2016). A poignant example is found in *Schistosoma japonicum* transmission in China, where despite well-established control programs and considerable progress towards elimination since the mid-1950s (Xu et al., 2016), a national report highlighted that there remained 450 endemic counties in 2020, a quarter of which had not achieved elimination criteria by year's end (Zhang et al., 2021). *S. japonicum* is transmitted by at least 40 species of wild and domestic mammals (Gray et al., 2009b), and animal activities near likely transmission sites may be important sources of reemergence and persistence. Although there are several domesticated and wild animals that can carry and transmit *S. japonicum*, (Li et al., 2015; Van Dorssen et al., 2017), estimates from Eastern China suggest that bovines may be responsible for as much as 75% of human transmission (Guo et al., 2006). This substantial contribution is thought to be related to the high degree of environmental overlap between humans and bovines during agricultural production, as well as the large amount of fecal output of bovines, which is estimated to be as high as 100 times that of human fecal production each day (Gray et al., 2009b; He et al., 2001; Ross et al., 2001). Additionally, the high frequency of livestock movement via the livestock trade within mountainous regions of China may facilitate the spread of *S. japonicum* infections across the region (Zheng et al., 2000; Zhou et al., 2012).

Bovine parasitic flatworm infections also have important veterinary impacts. The consequences of *Schistosoma bovis* and other parasitic flatworms can be quite severe on the animal, including: mortality, extreme morbidity and organ damage, emaciation, reduced milk yields and other production issues, and greater susceptibility to other pathogens (Al-Gaabary et al., 2009; Alemneh et al., 2015; McCauley et al., 1983).

Despite strong evidence that bovines are an important driver of human *S. japonicum* infection (Colley and Loker, 2018; Gray et al., 2007; Guo et al., 2006; Rudge et al., 2013) and of human schistosomiasis in Africa, there

are major gaps in our understanding of the factors that contribute to bovine *Schistosoma* infection. Most studies assessing bovine infection risks have focused primarily on veterinary schistosome species, the results of which have been inconsistent. For example, several recent assessments of *Schistosoma bovis* infection risk in Eastern Africa assessed bovine sex, age, breed and body condition as predictors of bovine infection status with contradictory results (Chanie et al., 2012; Defersha and Belete, 2018; Gebremeskel et al., 2017; Kebede et al., 2018; Lulie and Guadu, 2014; Tsega and Derso, 2015; Yihunie et al., 2019). Whereas assessments from 2012 and 2017 in Northwest Ethiopia found that neither age, sex nor the bovine's body condition was associated with bovine schistosomiasis infection (Chanie et al., 2012; Gebremeskel et al., 2017), a 2018 assessment found no significant effect of sex or breed, but suggested that bovines with poor body condition and bovines aged 2-5 had the highest infection risk (Defersha and Belete, 2018). By contrast, a 2019 study from northwestern Ethiopia recently highlighted that female sex, breed and poor body condition were all associated with bovine infection, while age was not found to be associated (Yihunie et al., 2019). Reasons for such discrepancies have not been fully elucidated, though Defersha & Belete (2018) hypothesize that it may be related to variations in management practices for different bovine groups (e.g. separation of sexes or of age groups) and different grazing ranges and patterns allowed on different farms (e.g. smaller grazing area of very young and very old bovines) (Defersha and Belete, 2018).

Outside of eastern Africa, few studies have characterized predictors of bovine schistosomiasis infection. One study from Malaysia found that older age, low weight and male sex, were all risk factors for *Schistosoma spindale* infection in a range of different cattle species (Tan et al., 2015). By comparison, a study conducted primarily among water buffaloes (96.2% water buffalos, 3.8% cattle) in Southern China found that *S. japonicum* infection intensity was highest in young bovines (<2 years) (Gray et al., 2007). These seemingly contradictory results may potentially be explained by isolation and limited grazing for calves in Malaysia (Tan et al., 2015), as well as potential genus-related differences in acquired immunity and self-cure rates (Li et al., 2014), as studies assessing worm establishment success in the two genera have found that cattle are more susceptible to infection than water buffalo (He et al., 1992; Xu et al., 1993). Nevertheless, water buffaloes may still act as important hosts in marshland areas of China, as they are more likely to spend time in water, and therefore more likely to be involved in the *S. japonicum* transmission cycle (He et al., 2001).

Not only is there a great deal of disagreement in the current body of literature over the key risk factors for bovine schistosomiasis infection, but studies to date have almost exclusively focused on physical characteristics rather than broader environmental conditions. There is a considerable literature that documents the role of social and environmental conditions in human schistosomiasis. For example, water, sanitation and hygiene (WASH) infrastructure, agriculture and fishing production, irrigation and night soil use – that is, the collection of either treated or untreated human and/or animal waste for use as fertilizer – have all been implicated as significant risk factors for human *Schistosoma* infection (Carlton et al., 2015; Grimes et al., 2015; Southgate, 1997; Zhou et al., 2018). While it is plausible that many of these factors pose similar risks to bovines, this has not yet been tested. Given that bovine *parasitic flatworm* infections can yield significant morbidity and economic losses, and contribute substantially to human *S. japonicum* infection risk, studies aimed at assessing potential predictors of *S. japonicum* infection in bovines are of paramount importance.

To address this gap, we set out to assess a range of potential physical and environmental predictors of bovine *S. japonicum* infection in 2007, 2010 and 2016 at the individual, household and village-levels in a region where schistosomiasis persistence has been demonstrated to exist in both humans and bovine populations. We draw from the literature on human and bovine schistosomiasis to define a broad set of candidate predictors that describe individual bovine characteristics. In so doing, our study helps to fill the gaps in knowledge about the conditions that predict *S. japonicum* infection in bovines, as well as the relevance of two additional assessment scales (household and village-level) that have rarely been considered in the studies conducted to date. The results of this study will ultimately serve as a critical stepping stone toward developing appropriately-scaled and targeted schistosomiasis intervention and control activities.

2. Materials and Methods

2.1 Village selection

This study was conducted in two rural counties in Sichuan, China where schistosomiasis reemerged and persisted despite aggressive control efforts. Surveys of environmental and social risk factors, as well as human and bovine infection were conducted in 2007, 2010 and 2016. In 2007, three of eight Sichuan counties where schistosomiasis had reemerged (Liang et al., 2006), were selected for inclusion in our study based on surveillance record availability and the local control stations' willingness and capacity to collaborate, and a set of 17 to 19 villages was selected in each county, as described previously (Carlton et al., 2011). However, in May

139 2008, a 7.9 magnitude earthquake severely impacted one of the study counties (Carlton et al., 2015), such that, in
140 2010, follow-up surveys were conducted in 2 counties (36 villages). In an effort to hone in on the highest risk
141 locations within the two study counties in a region where infections were declining, surveillance records were
142 reexamined in 2016, from which, a total of ten villages were selected for inclusion in the 2016 data collection. Of
143 these, seven were villages that were also included in 2007 and 2010, while three were new. We restrict this
144 analysis to the two counties surveyed in all three years, noting that our villages are not a representative sample of
145 Sichuan, but rather a survey of high-risk locations where schistosomiasis has persisted despite control efforts. All
146 villages are located in the hilly regions of rural Sichuan and ranged from approximately 20 to 150 households and
147 50 to 200 residents. Finally, the analysis presented here was restricted to villages where bovines were both
148 present and tested for *S. japonicum* infection: 35 villages in 2007, 30 villages in 2010 and 8 villages in 2016.
149 Details about the total number of included villages, households and bovines are provided in Supplementary Table
150 S1. *2.2 Census, household questionnaires and infection surveys*

151 A village census was conducted in each collection year and all residents over the age of five were invited
152 to participate in surveys for *S. japonicum* infection. In addition, attempts were made to survey all bovines in the
153 village for *S. japonicum* infection. In the summers of 2007, 2010 and 2016, the head of each household was
154 asked to complete a household survey that contained closed-ended questions related to socioeconomic status,
155 domestic and farm animal ownership, sanitation and water access and agricultural practices. Bovine age, type
156 and sex were collected at the time of the bovine infection surveys in 2007 and 2010 (these data were not
157 collected in 2016). Trained staff from the Sichuan Center for Disease Control and Prevention and the county
158 Schistosomiasis Control Stations piloted and conducted all surveys in the local Sichuan dialect. Only bovines that
159 had both household survey data and *S. japonicum* infection data were included in this assessment
160 (Supplementary Table S1).

161 *S. japonicum* infection surveys were conducted by attempting to test three stool samples on three
162 consecutive days from eligible humans and all bovines in the village. Infection surveys were conducted in
163 November and December of 2007 and 2010, and July 2016. Bovines were isolated in a pen or tied up until a stool
164 was produced on three separate days (consecutive, when possible). All stool samples were transported to the
165 central laboratory soon after collection to be examined using the miracidium hatching test, following standard
166 protocols (Department of Disease Control, 2000). To account for the short survival and rapid hatching of bovine

167 miracidia, the bovine samples were examined for miracidia at one, three and five hours after preparation for at
168 least two minutes each time, whereas human samples were assessed at two, five and eight hours after
169 preparation. One sample from each human was also examined using the Kato Katz thick smear procedure in
170 2007 and 2010 (Katz et al., 1972). A bovine was considered positive for *S. japonicum* if any miracidium hatching
171 test was positive. A human was considered positive for *S. japonicum* if any miracidium hatching test or the Kato-
172 Katz test was positive. Each person who tested positive was notified and referred to the local anti-schistosomiasis
173 control station for treatment. When bovines tested positive, owners were notified and the animal was referred to
174 the county veterinary station for treatment.

175 *2.3 Ethics statement*

176 This study was approved by the Sichuan Institutional Review Board, the University of California, Berkeley,
177 Committee for the Protection of Human Subjects, and the Colorado Multiple Institutional Review Board. All
178 participants provided written, informed consent. The collection of bovine samples was determined to be exempt
179 from review by the Animal Care and Use Committee at the University of California, Berkeley and the Institutional
180 Animal Care and Use Committee at the University of Colorado Anschutz.

181 *2.4 Predictor selection and definitions*

182 The primary outcome of interest in this analysis was bovine *S. japonicum* infection in 2007, 2010 and
183 2016. All candidate predictors were defined using either the household surveys or the human infection surveys
184 and were divided into five categories: 1) biological/physical characteristics; 2) potential human sources of
185 environmental schistosomes; 3) socio-economic indicators; 4) potential animal reservoirs/sources of infection; 5)
186 agricultural risk factors (Table 1). We included agriculture as its own category because bovines are frequently
187 employed in agricultural work in China (Zheng et al., 2000), and because different crop types and agricultural
188 practices have their own inherent exposure risks (e.g. planting wet crops like rice may increase the likelihood of
189 contact with snail habitat and exposure to cercariae (Gordon et al., 2019)). Variables with hypothesized similar
190 mechanisms of transmission risk were aggregated where possible. Three crop type categories were created:
191 winter crops (primarily rapeseed and wheat), summer dry crops (primarily corn, peanuts and vegetables) and
192 summer wet crops (rice). Night soil use was also included as an agricultural risk factor and divided into three
193 categories: night soil use on winter crops, dry summer crops and wet summer crops.

There were minor variations in the household survey content and question formulation across the study period (e.g., pig ownership was not assessed in 2007). Where possible, continuous/discrete predictors were included over binary measures. We excluded binary variables from the analysis if they represented very rare (<10%) or very common conditions (>90%). Continuous variables were excluded when >90% of the observations took a single value. For example, household dog and pig ownership were both excluded in 2016 because >90% of the households owned dogs while >90% did not own any pigs. A composite household asset score (0-9) was developed for use in this assessment, which included eight household assets assessed in all three collection years (washing machine, television, air conditioner, refrigerator, computer, car, motorcycle), as well as a binary measure indicating that the home was constructed of concrete, wood or bricks (vs. adobe).

Because prior work has demonstrated that group-level measures can serve as important predictors of schistosomiasis infection in humans (Carlton et al., 2015), we also generated village-level predictors from the household survey data. Village-level variables represent all households that participated in the household survey from a given village, even if they didn't own bovines. Village-level variables were either the village-average value of continuous household measures, or for binary variables, the proportion of the village population reporting the condition. Notably, the village-level variables excluded all observations from the bovine's own household, and instead used only the data from the other households in the village that participated in the household survey. This allowed for an assessment of how the surrounding village environment impacts individual bovine infection risk, independent of the home environment, whereas the household-level variables aim to unpack the influence of the unique household environment on bovine infection status. The aforementioned criteria led to a total of 31 predictors.

2.5 Analysis

Among the bovines with *S. japonicum* infection data (i.e. the outcome variable), those that lacked all household survey data (i.e. the predictor variables) were excluded from the analysis (30/503 bovines in 2007; 36/233 bovines in 2010; and 1/72 in 2016). Infection prevalence was similar among the excluded bovines (11/67, 16.4%) as compared to those included in this analysis (111/741, 15.0%). Among included bovines, missingness was generally low: the variable with the most missing data was bovine sex in 2007 (21.6% missing). Missing values were imputed separately for each collection year using the rfImpute function from the "randomForest" package in R (Breiman, 2018; RColorBrewer and Liaw).

222 Two preliminary descriptive analyses were carried out. First, spatial patterns of bovine infection
223 prevalence by village were visualized using ESRI's ArcGIS ArcMap software release 10.5.1 (Environmental
224 Systems Research Institute (ESRI), 2017). Second, categorical versions of the candidate predictors were
225 generated in order to evaluate changes in the distribution of predictors over time as well as to allow a qualitative
226 examination of the distribution of predictors by bovine infection status.
227 To determine which of our candidate predictors serve as the best predictors of bovine schistosomiasis
228 infection in 2007, 2010 and 2016, a random forests (RF) machine learning approach was used. For each year,
229 25% of the data was reserved for validation, while the remaining 75% was used for model construction. To
230 address class imbalance in our outcome of interest (13.3%, 17.3%, and 19.7% of bovines were *S. japonicum*
231 positive in 2007, 2010 and 2016, respectively), over sampling of the minority class was conducted. For model
232 tuning, 10-fold cross validation was performed using the Caret package in R to help select the optimal maximum
233 node size and the number of variables to try at each branch. Once the optimal value of each of these parameters
234 was determined, a final model was run using 5000 trees per forest (Breiman and Cutler, 2011).
235 For each collection year, we conducted a total of ten rebalancing and model tuning iterations to assess
236 the degree of stability in our variable importance rankings. The mean decrease in accuracy (MDA) value was
237 used to rank the top ten predictors from each model on a scale of ten to one – from most to least important – such
238 that the predictor with the highest MDA was assigned a score of 10, and predictors not in the top ten were
239 assigned a score of zero. These variable rankings were then summed across the 10 rebalancing iterations to give
240 a 10-model summary score of variable importance, ranging from 100-1. The ten highest scoring variables from
241 the ten-model summary score were then reassigned a final ranking of first to tenth. Next, we created a "lean"
242 ranking, using only those predictors ranked first through tenth within each collection year. To do this, we
243 performed an additional ten iterations of the aforementioned balancing and tuning process with the top ten
244 predictors, thereby reducing excess noise in the variable ranking assessment caused by including a large number
245 of candidate predictors. Because we hypothesized that the inclusion of human infection as a predictor of bovine
246 infection would strongly influence the predictive capacity of our RF models due to a presumed association
247 between bovine and human infection, we also conducted ten iterations of a sensitivity analysis for each collection
248 year that excluded the human infection variables from the algorithm. The ability of the full, lean and sensitivity RF
249 models to predict infection status was assessed using accuracy, sensitivity and the receiver operator curve (ROC)

250 area under the curve (AUC), respectively. In the case of disagreement or a tie when comparing the chosen
251 performance metrics, kappa and specificity were subsequently compared to select the top performing model for
252 each year.

253 Simple logistic regression analyses were performed to assess the direction of association between the
254 top predictors and bovine infection, dividing continuous variables into tertiles to assess for potential non-linear
255 relationships. The direction of association was recorded for the top predictors within each collection year, using a
256 p-value of <0.2 to indicate weak evidence of a between-group difference. In the case that no difference was
257 indicated between tertile groups at the p<0.2 level, the predictors were further divided into quartiles and re-
258 assessed. If still no evidence for a between group difference was identified using quartiles, this point was noted in
259 the results. Density plots by infection status were also examined for a subset of predictors that were found to have
260 a change in the direction of association across the collection years. Stata 15 and R Studio 4.0 were used for all
261 analyses (RStudio Team, 2020; StataCorp, 2015).

262 **3. Results**

263 *3.1 Bovine infection prevalence and population characteristics*

264 Villages with high bovine infection prevalence (>30%) were present in all collection years. While some
265 villages were found to have repeatedly high infection prevalence in bovines across all years despite treatment
266 programs and aggressive control measures, other villages emerged as high prevalence villages in 2016 after
267 years of low infection prevalence in bovines. Figure 1 shows a map of bovine infection distribution by village
268 across two counties.

269 The bovine study population was comprised of fewer water buffaloes than cattle (18.5% of all bovines
270 were water buffalo in 2007; 15.5% in 2010), was predominantly female (86.3% in 2007; 87.2% in 2010), and
271 ranged widely in age from less than a year to 26 years old.

272 Table 2 compares the distributions of each predictor variable by bovine infection status, for 2007, 2010
273 and 2016.

274 There were several notable shifts in the distribution of predictors over time in our study region. Access to
275 improved sanitation increased across our study period, rising from 22.8% of households reporting improved
276 sanitation in 2007 to 52.1% with improved sanitation by 2016. Notably, the percent of bovines infected was
277 roughly equal across the two sanitation groups in 2007, but by 2016, a greater proportion of bovines were infected

if they belonged to households with unimproved sanitation (23.5%) compared to the households with improved sanitation (16.2%).Household assets also increased over time, as approximately half of all villages had a mean asset score of less than two in 2007, whereas in 2010 and 2016, none of the included villages had a mean asset score that low. Across the study period, there was a steady increase in the percent of households reporting planting rice (69.1% in 2007; 71.5% in 2010; 81.7% in 2016), and other summer crops (77.2% in 2007; 98% in 2010; 100% in 2016), as well as the cultivated area of rice and dry summer crops. Night soil use on rice and winter crops decreased over time: the proportion of households reporting any night soil use on rice crops dropped from 35.6% to 11.7% between 2007 and 2016, and for winter crops it dropped from 58.3% to 12.6%. By contrast, the proportion of night soil users for dry summer crops remained relatively constant across the study period (52.4% in 2007; 53.5% in 2010; 50.7% in 2016). Notably, the average amount of night soil being applied to crops dropped across the study period for all crop types.

3.2 Predictors of bovine infection

The predictor rankings were relatively stable across the full models, lean models and sensitivity models within each year, though there was more variability across collection years (Figure 2). For example, animal reservoirs were frequently among the top ten predictors in 2016, but only a household's pig ownership (2010) and bovine density in the surrounding village (both 2007 and 2010) were among the top ten in earlier years.

The density of bovine in a village was one of the most consistent predictors: it was scored in the top ten in at least one RF analysis from each year, all of which indicate that an increase in bovine density in the surrounding village corresponds with an increase in bovine infection risk. Pig ownership was associated with a decrease in bovine infection risk in both years that it was assessed.

Agricultural practices was the most frequently ranked predictor category across all years. Specifically, the household area of winter crops planted, the mean area of rice planted in the surrounding village, and the mean amount of night soil applied to dry summer crops in the surrounding village were all among the top ten predictors in all models. Additionally, the total household area of summer crops planted and the village mean area of winter crops were also all among the top ten predictors in at least one of the three model types for 2007, 2010 and 2016. However, despite the inter-year agreement indicating the importance of agricultural predictors, the direction of association between the top agricultural predictors and bovine infection was not always consistent across years. In 2007, increases in all key agricultural predictors were associated with an increase in bovine infection risk, apart

306 from night soil use on winter crops. By contrast, in 2010 and 2016 our models indicate a mixture of positive and
307 negative associations across the key agricultural predictors, and in one instance (household winter crop area in
308 2010), no evidence of a relationship was found.

309 Figure 3 depicts changes in the distribution of different agricultural practices by bovine infection status
310 between 2007 and 2016. Most prominent among these is dry summer crop farming, which saw a notable increase
311 in the total and mean area of crop being planted by households and villages between 2007 and 2016 (Figure 3,
312 panel B). In 2007 and 2016 the area of dry summer crops cultivated at the household and village-level was
313 positively associated with bovine infection, but in 2010 it was not. There was a general decrease in the amount of
314 night soil being applied to rice, winter crops and, to a lesser extent, dry summer crops, between 2007 and 2016
315 (Figure 3, panels D-F). Greater night soil use on dry summer crops at the village level was associated with
316 increased bovine infection risk in 2007 and 2016, but not in 2010. Less night soil was used on rice, compared to
317 other crops, and night soil use on rice was not a key predictor of bovine infection in any year.

318 There are several other key predictors that stand out in one or more collection years. Human infection
319 prevalence in the surrounding village was among the top five predictors of bovine infection in 2007 and 2010, and
320 the number of infected humans within the household was among the top ten predictors in 2010. In all cases, an
321 increase in human infections was associated with an increase in bovine infections. When the human infection
322 predictors were removed for the sensitivity analysis, the rankings of the remaining predictors did not shift
323 substantially in any collection year. Bovine age was also among the top predictors in 2007, and the probability of
324 infection was found to increase with age. In all of the 2010 analyses, the number of hatch tests was an important
325 predictor of bovine infection, a feature not shared by the 2007 and 2016 analyses. This may be related to the
326 lower proportion of bovines with three hatch test results in 2010 (64.5%), as compared to 2007 (70.6%) and 2016
327 (73.2%).

328 Of the three different analyses performed (full, lean and sensitivity) for each collection year, the full
329 models (i.e. those that included the full list of predictors available in a given year) tended to perform the best
330 (Table 3, Figure 4). Overall, the models had high accuracy values, with the top performing full models producing a
331 maximum accuracy of 0.864 (95% CI: 0.79 – 0.92) in 2007, 0.816 (95% CI: 0.68 – 0.91) in 2010, and 1.0 (0.81 –
332 1.0) in 2016. Furthermore, despite some variation in model performance across the ten iterations of RF models
333 for each analysis year, ultimately the models were relatively stable. For example, across each of the ten iterations

of full model analyses, the AUC range was relatively narrow: 0.724 – 0.75 in 2007, 0.816 – 0.863 in 2010, and 0.982 – 1.0 in 2016. However, due to class imbalance in the reserved test datasets (see the no information rate (NIR) in Table 3), the Kappa value is also a useful performance metric for our models, as this takes class imbalance into account. According to the benchmarks laid out by Landis and Koch (1977), the Kappa statistics from our 2007 analyses suggest a “Fair” level of agreement (0.21 – 0.40) between our best RF models and the true known values in 2007. For 2010, the highest Kappa statistic came from the full predictor analysis, with a Kappa of 0.463, indicating a “Moderate” level of agreement (0.41 – 0.60) between the prediction model and the reserved test dataset (Landis and Koch, 1977). In 2016, both the full and sensitivity models achieved perfect prediction (Kappa = 1) for the test dataset in at least one of the ten model iterations, whereas the Kappa statistic for the top performing lean model was 0.853, or “Almost Perfect”, according to Landis & Koch (Landis and Koch, 1977).

Because the full list of predictors changed slightly across the collection years, a supplemental analysis was conducted in which only predictors that were available in all three collection years were included in the RF models. This analysis demonstrated that 1) intra-year rankings and extra-year patterns did not change substantially, and 2) agricultural variables remained the most prominent predictor category when comparing across the entire study period. See Supplementary Figure S1 for details of the supplemental analysis. **4. Discussion**

Agricultural risk factors, bovine density, human infection, pig ownership and bovine age were all found to be key predictors of bovine infection status in one or more of the years assessed in this analysis. Of the agricultural risk factors, night soil use on dry summer crops, the village-level area of rice crops, and both the household and village-level areas of dry summer and winter crops were each ranked among the top five predictors for one or more collection years in our RF models. Interestingly, for 2007, all of the ranked agricultural variables except night soil use on winter crops were associated with an increase in bovine infection risk in our logistic regression assessments, whereas in 2010 and 2016, these agricultural factors were found to be variably positively and negatively associated with infection. For example, the area of rice cultivated in a village was positively associated with bovine infection in 2007 and 2010, but negatively associated in 2016. This finding may be related to changing norms and interventions that have taken hold in recent years as a result of increasing awareness of the potential risks posed by both bovines as a reservoir of schistosomiasis, and specific agricultural

practices. For example, across our study period, we saw a steady increase in the percent of households planting rice (69.1% in 2007; 71.5% in 2010; 81.7% in 2016), and a simultaneous decrease in the percent of households applying any night soil to their rice crops (35.6% in 2007, 14.7% in 2010; 11.7% in 2016). These shifting norms in rice production and night soil use likely resulted in a decrease in the overall concentration of night soil on rice crops within our study villages, which in turn, may help to explain why the village-level rice crop area shifts from having a positive association with bovine infection in 2007 and 2010, to a negative association by 2016.

Assessments conducted in China early in the new millennium repeatedly highlighted bovines as a key source of environmental contamination and as the main animal reservoir of *S. japonicum* in the country (Gray et al., 2009a; Gray et al., 2007; Guo et al., 2006). Beginning in 2004, a new government-led approach to eliminating schistosomiasis transmission in China was adopted, which – in conjunction with infrastructure improvements in rural areas and several new schistosomiasis elimination interventions – featured replacing bovines with machinery in agricultural production (Liu et al., 2017). Thus, the negative associations that were found intermittently between bovine infection and some of our agricultural variables in 2010 and 2016 may be linked to the added precautions that were being adopted when bovines were being used for agriculture, or because bovines were being reallocated for other purposes (e.g. beef production) as machinery became the norm for large crop areas or those deemed high risk (e.g. wet rice crops).

Dry summer crops stand out as a key agricultural risk factor for bovine infection. The area of dry summer crops cultivated and the quantity of night soil applied to dry summer crops were key predictors of bovine infection in all years, and positively associated with bovine infection in 2007 and 2016. It is unclear why this is: it is possible that the cultivation of such crops provides snail habitat and/or the use of night soil on dry summer crops facilitates infection of snails living near the field (in our study region, more night soil is applied to dry summer crops than rice), leading to bovine infection. A downward trend in night soil use (total and mean number of buckets) on crops can be observed in Figure 3 (parts D-F), though notably, we do not see any substantial shift in the overall proportion of households that reported any night soil use on summer crops over the years (52.4% in 2007; 53.5% in 2010; 50.7% in 2016) (Table 2). Increasing recognition of the potential risks posed by night soil use during our study period (Carlton et al., 2011) may have also contributed to some decreases in night soil use and/or the more careful treatment of night soil prior to field applications. However, the continued practice of applying night soil to dry summer crops, paired with the steady increase in the total area of summer crops being planted by villagers

over the study period (see Figure 3, part B) may help to explain why night soil use on summer crops is positively association with bovine infection status in 2016.

Bovine density in the surrounding village was in the top ten predictors of our RF models, and was positively associated with bovine infection in our regression models in all collection years. These findings align well with the existing literature that points to bovines as the most important reservoir of *S. japonicum* infection in China (Gray et al., 2009a), and suggests that being in close proximity to higher densities of bovine hosts may correspond with increasing infection risk, as has been found for other bovine pathogens (Guo et al., 2006; Meadows et al., 2018; Spencer et al., 2015). However, it is worth noting that household-level bovine ownership was not among the top predictors in any of our RF models, highlighting that the larger-scale lens (i.e. village-level analysis scale) may be particularly important to future investigations and control strategies. Likewise, recent informal interviews with residents in our study sites have revealed that bovines are infrequently kept near the home, as allowing bovines to graze (and defecate) freely is an economical and efficient way of raising bovines, further illustrating that the household scale may not always be broad enough to capture larger scale trends. Instead, villagers opt to bring their bovines to the mountains to graze during the day, which subsequently presents more opportunities for contact between bovines from different households, and may ultimately result in more widespread environmental contamination (e.g. bovine feces washed into nearby irrigation ditches after precipitation).

In the developmental stages of this analysis, we hypothesized that human infection prevalence and the number of infected people in the household would be among the top predictors of bovine infection status, given the known link between human schistosomiasis and bovine reservoirs (e.g. Gray et al., 2009a). It was therefore somewhat surprising to find that household-level human infection was only highly ranked in 2007 within our RF models, while village human infection prevalence was ranked in the top ten in 2007 and 2010, and neither household-level human infection nor village-level human infection prevalence were important predictors in 2016. One potential explanation for the changing importance of human infection as a predictor of bovine infection could be related to the aforementioned bovine-removal phenomenon, in which bovines are increasingly being removed from the village area and brought to alternative mountain locations for grazing. As this act would likely result ~~in an~~in an overall drop in time humans spent in bovine grazing areas, one might expect to see a separation of the bovine and human infection cycles if snail habitats are common in bovine grazing sites. In fact, the drop in the

important rankings of human infection status in 2016 coincides with a jump in the variable importance rankings for village-level bovine ownership (6th – 8th in 2007 and 2010; 1st -2nd in 2016), providing further support of the theory that bovines may be becoming increasing important reservoirs of schistosomiasis infection. On the other hand, an altogether different explanation for the differences in the 2016 rankings compared to 2007 and 2010 is that the 2016 data collection simply didn't have a large enough sample size to allow for the detection of a true relationship between relatively rare events.

As such, one limitation of this assessment was the relatively small sample sizes, particularly in 2016 (N=71), though to a lesser extent, 2010 (N=197) and 2007 (N=473), given the correspondingly large number of predictors that were included in the full predictor models (N=29, N=31, N=26, in 2007, 2010 and 2016 respectively). While RF models are generally acknowledged as being able to handle assessments of high dimensional data even with relatively small sample sizes (Biau and Scornet, 2016), it remains that small samples sizes can still give rise to the aforementioned issue of non-detection of rare events. Another limitation to this assessment is that RF models tends to favor continuous predictors over categorical measures, as they allow for a wider range of potential split points for classifying observations. For this reason, it is not particularly surprising that age was the only predictor from the individual/physical characteristics predictor group that was ranked among the top ten predictors, as the remaining individual characteristics were binary measures. Another notable limitation of the variable importance rankings used in RF models is that they become less reliable when predictors are highly correlated with one another (Strobl et al., 2008). This may be particularly important to the rankings ascribed to the agricultural variables, as correlation between the area of the different crop types planted and the amount of night soil used on each crop tended to be high across all collection years, with the highest predictor correlations found in the 2016 collection year (See Supplementary Figures S2 – S4 for correlation matrices). This is notable, as a higher degree of instability in the variable importance rankings was also found for 2016 as compared to 2010 or 2007, suggesting predictor correlation may be responsible. We therefore recommend that the variable rankings presented in this analysis be interpreted more holistically (e.g. agricultural variables are strong predictors of bovine infection), and advise caution when comparing unique variable ranking values against one another (e.g. rice crop area is less important than winter crop area).

Our main interests in this assessment were to 1) identify the best physical and environmental predictors of bovine *S. japonicum* infection within rural farming communities in Sichuan China, and 2) to ascertain whether

446 there are broader trends in bovine infection distribution across individual, household or village-levels scales or
447 over time. Our RF assessments have highlighted some key patterns that were repeated across multiple collection
448 years and multiple iterations of three different models. Agricultural factors and animal reservoirs – specifically,
449 high bovine density in the surrounding village – were repeatedly found to be among the top predictors of bovine *S.*
450 *japonicum* infection across all assessment years. Furthermore, human infection, pig ownership and bovine age
451 were also found to be strong predictors of bovine infection in at least one year. Taken together, these findings
452 highlight the potential utility of presumptively treating bovines residing in villages and households that engage in
453 high-risk agricultural practices, those of older age, or those belonging to villages with particularly high human
454 infection or bovine density. Additionally, village-level predictors tended to be better predictors of bovine infection
455 than household-level predictors, suggesting that future investigations and interventions may need to apply a
456 broad ecological lens in order to successfully extricate and address environmental sources of ongoing
457 transmission. In so doing, not only can unnecessary bovine morbidity and the associated economic losses be
458 reduced, but such measures will also help to minimize the potential for bovines to serve as another pathway to
459 persistent human schistosomiasis transmission.

460

461 **Acknowledgments:** We are grateful for the support and efforts of the field research team members from the
462 Institute of Parasitic Diseases and the county anti-schistosomiasis control stations for their efforts in collecting the
463 data presented here.

464 **Author Contributions:** Conceptualization: EC, YL, SP, AB; Data Curation: DL, EG; Formal Analysis: EG, SP,
465 KK, AB, EC; Funding Acquisition: EC, YL, SP; Investigation: YL, DL, EC; Methodology: EG, SP, EC, AB, KK;
466 Project Administration: YL, DL; Resources: EC, YL, DL; Validation: EG, SP, EC, KK; Writing – Original Draft
467 Preparation: EG, SP, EC, KJ; Writing – Review & Editing: EG, SP, KK, YL, DL, EC, AB, KJ.

468 **Data Availability:** Due to the inclusion of potentially identifying information (e.g. socio-economic indicators and
469 human infection status), access to study data must be requested through the Carlton Lab group to ensure the
470 protection of research subjects and compliance with all ethical guidelines. Please contact Elizabeth Carlton at
471 elizabeth.carlton@cuanschutz.edu for more information.

472 **Declaration of interests:** The authors have declared that no competing interests exist.

473 **Funding:** This research was supported by grants from the National Institute of Allergy and Infectious Diseases:
474 R01AI134673 (EJC, PI), R21AI115288 (EJC, PI) and R01AI068854 (Robert Spear, PI). The content is solely the
475 responsibility of the authors and does not necessarily represent the official views of the National Institutes of
476 Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation
477 of the manuscript.

478 **References**
479

480 Al-Gaabary, M., Osman, S., El-Tonoby, A., 2009. STUDIES ON PARAMPHISTOMIASIS IN RUMINANTS.
481 Kafrelsheikh Veterinary Medical Journal 7, 116-136.

482 Alemneh, T., Engdaw, Alemseged, G., Abuhay, 2015. OVERVIEW ON: SCHISTOSOMA INFECTION IN
483 CATTLE. 1-11.

484 Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25, 197-227.

485 Breiman, L., Cutler, A., 2011. Manual–setting up, using, and understanding random forests V4. 0. 2003. URL
486 https://www. stat. berkeley. edu/~ breiman/Using_random_forests_v4. 0. pdf.

487 Breiman, L.C., A.; Liaw, A.; Wiener, M. , 2018. Breiman and Cutler's Random Forests for Classification and

488 Regression, https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

489 Carlton, E.J., Bates, M.N., Zhong, B., Seto, E.Y.W., Spear, R.C., 2011. Evaluation of Mammalian and
490 Intermediate Host Surveillance Methods for Detecting Schistosomiasis Reemergence in Southwest China. PLoS
491 Negl Trop Dis 5.

492 Carlton, E.J., Liu, Y., Zhong, B., Hubbard, A., Spear, R.C., 2015. Associations between Schistosomiasis and the
493 Use of Human Waste as an Agricultural Fertilizer in China. PLoS Negl Trop Dis 9.

494 Chanie, M., Dejen, B., Fentahun, T., 2012. Prevalence of cattle schistosomiasis and associated risk factors in
495 Fogera cattle, south Gondar zone, Amhara national regional state, Ethiopia. Journal of Advanced Veterinary
496 Research 2, 153-156.

497 Colley, D.G., Loker, E.S., 2018. New Tools for Old Questions: How Strictly Human Are "Human Schistosomes"-
498 And Does It Matter? The Journal of infectious diseases 218, 344-346.

499 Defersha, T., Belete, B., 2018. The Neglected Infectious Disease, Bovine Schistosomiasis: Prevalence and
500 Associated Risk Factors for its Occurrence among Cattle in the North Gulf of Lake Tana, Northwest Ethiopia. J
501 Vet Med Health 2, 112.

502 Department of Disease Control, 2000. Textbook for Schistosomiasis Control, Shanghai Shanghai Publishing
503 House for Science and Technology, pp. 72-76.

504 Environmental Systems Research Institute (ESRI), 2017 ArcGIS Desktop Release, 10.5.1 ed, Redlands, CA.

505 Gebremeskel, A.K., Simeneh, S.T., Mekuria, S.A., 2017. Prevalence and Associated Risk Factors of Bovine
506 Schistosomiasis in Northwestern Ethiopia. World 7, 01-04.

507 Gordon, C.A., Kurscheid, J., Williams, G.M., Clements, A.C.A., Li, Y., Zhou, X.N., Utzinger, J., McManus, D.P.,
508 Gray, D.J., 2019. Asian Schistosomiasis: Current Status and Prospects for Control Leading to Elimination. Trop
509 Med Infect Dis 4.

510 Gray, D.J., Williams, G.M., Li, Y., Chen, H., Forsyth, S.J., Li, R.S., Barnett, A.G., Guo, J., Ross, A.G., Feng, Z.,
511 2009a. A cluster-randomised intervention trial against Schistosoma japonicum in the Peoples' Republic of China:
512 bovine and human transmission. PloS one 4, e5900.

513 Gray, D.J., Williams, G.M., Li, Y., Chen, H., Li, R.S., Forsyth, S.J., Barnett, A.G., Guo, J., Feng, Z., McManus,
514 D.P., 2007. A cluster-randomized bovine intervention trial against Schistosoma japonicum in the People's
515 Republic of China: design and baseline results. The American journal of tropical medicine and hygiene 77, 866-
516 874.

517 Gray, D.J., Williams, G.M., Li, Y., McManus, D.P., 2009b. Transmission Dynamics of Schistosoma japonicum in
518 the Lakes and Marshlands of China. PloS one 3, e4058.

519 Grimes, J.E.T., Croll, D., Harrison, W.E., Utzinger, J., Freeman, M.C., Templeton, M.R., 2015. The roles of water,
520 sanitation and hygiene in reducing schistosomiasis: a review. Parasites & vectors 8, 156.

521 Guo, J., Li, Y., Gray, D., Ning, A., Hu, G., Chen, H., Davis, G.M., Sleigh, A.C., Feng, Z., McManus, D.P., Williams,
522 G.M., 2006. A drug-based intervention study on the importance of buffaloes for human Schistosoma japonicum
523 infection around Poyang Lake, People's Republic of China. The American journal of tropical medicine and
524 hygiene 74, 335-341.

525 He, Y., Xu, S., Shi, F., Shen, W., Hsü, S., Hsü, H., 1992. Comparative studies on the infection and maturation of
526 schistosoma japonicum in cattle and buffaloes. Current Zoology 38, 266-271.

527 He, Y.X., Salafsky, B., Ramaswamy, K., 2001. Host--parasite relationships of Schistosoma japonicum in
528 mammalian hosts. Trends Parasitol 17, 320-324.

529 Katz, N., Chaves, A., Pellegrino, J., 1972. A simple device for quantitative stool thick-smear technique in
530 Schistosomiasis mansoni. Rev Inst Med Trop Sao Paulo 14, 397-400.

531 Kebede, A., Dugassa, J., Haile, G., Wakjira, B.M., 2018. Prevalence of bovine of schistosomosis in and around
532 Nekemte, East Wollega zone, Western Ethiopia. Journal of Veterinary Medicine and Animal Health 10, 123-127.

533 Kittur, N., King, C.H., Campbell, C.H., Kinung'hi, S., Mwinzi, P.N.M., Karanja, D.M.S., N'Goran, E.K., Phillips,
534 A.E., Gazzinelli-Guimaraes, P.H., Olsen, A., Magnussen, P., Secor, W.E., Montgomery, S.P., Utzinger, J.,
535 Walker, J.W., Binder, S., Colley, D.G., 2019. Persistent Hotspots in Schistosomiasis Consortium for Operational
536 Research and Evaluation Studies for Gaining and Sustaining Control of Schistosomiasis after Four Years of Mass
537 Drug Administration of Praziquantel. The American journal of tropical medicine and hygiene 101, 617-627.

538 Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33,
539 159-174.

540 Li, H., Dong, G.D., Liu, J.M., Gao, J.X., Shi, Y.J., Zhang, Y.G., Jin, Y.M., Lu, K., Cheng, G.F., Lin, J.J., 2015.
541 Elimination of schistosomiasis japonica from formerly endemic areas in mountainous regions of southern China
542 using a praziquantel regimen. Veterinary parasitology 208, 254-258.

543 Li, Y.S., McManus, D.P., Lin, D.D., Williams, G.M., Harn, D.A., Ross, A.G., Feng, Z., Gray, D.J., 2014. The
544 Schistosoma japonicum self-cure phenomenon in water buffaloes: potential impact on the control and elimination
545 of schistosomiasis in China. International journal for parasitology 44, 167-171.

546 Liang, S., Yang, C., Zhong, B., Qiu, D., 2006. Re-emerging schistosomiasis in hilly and mountainous areas of
547 Sichuan, China.

548 Liu, Y., Zhong, B., Wu, Z.-S., Liang, S., Qiu, D.-C., Ma, X., 2017. Interruption of schistosomiasis transmission in
549 mountainous and hilly regions with an integrated strategy: a longitudinal case study in Sichuan, China. Infectious
550 Diseases of Poverty 6, 79.

551 Lulie, B., Guadu, T., 2014. Bovine schistosomiasis: A threat in public health perspective in Bahir Dar town,
552 northwest Ethiopia. Acta Parasitologica Globalis 5, 1-6.

553 McCauley, E.H., Majid, A.A., Tayeb, A., Bushara, H.O., 1983. Clinical diagnosis of schistosomiasis in sudanese
554 cattle. Tropical Animal Health and Production 15, 129-136.

555 Meadows, A.J., Mundt, C.C., Keeling, M.J., Tildesley, M.J., 2018. Disentangling the influence of livestock vs. farm
556 density on livestock disease epidemics. Ecosphere 9, e02294.

557 RColorBrewer, S., Liaw, M.A., Package 'randomForest'.

558 Ross, A.G., Sleigh, A.C., Li, Y., Davis, G.M., Williams, G.M., Jiang, Z., Feng, Z., McManus, D.P., 2001.
559 Schistosomiasis in the People's Republic of China: prospects and challenges for the 21st century. Clin. Microbiol.
560 Rev. 14, 270-295.

561 RStudio Team, 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, 2020.

562 Rudge, J.W., Webster, J.P., Lu, D.B., Wang, T.P., Fang, G.R., Basáñez, M.G., 2013. Identifying host species
563 driving transmission of schistosomiasis japonica, a multihost parasite system, in China. Proceedings of the
564 National Academy of Sciences of the United States of America 110, 11457-11462.

565 Song, L.G., Wu, X.Y., Sacko, M., Wu, Z.D., 2016. History of schistosomiasis epidemiology, current status, and
566 challenges in China: on the road to schistosomiasis elimination. Parasitology research 115, 4071-4081.

567 Southgate, V.R., 1997. Schistosomiasis in the Senegal River Basin: before and after the construction of the dams
568 at Diama, Senegal and Manantali, Mali and future prospects. J Helminthol 71, 125-132.

569 Spencer, S.E., Besser, T.E., Cobbold, R.N., French, N.P., 2015. 'Super' or just 'above average'? Supershedders
570 and the transmission of Escherichia coli O157:H7 among feedlot cattle. J R Soc Interface 12, 0446.

571 StataCorp, 2015. Stata Statistical Software: Release 15. StataCorp LP, College Station, TX.

572 Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random
573 forests. BMC Bioinformatics 9, 307.

574 Tan, T.K., Low, V.L., Lee, S.C., Panchadcharam, C., Kho, K.L., Koh, F.X., Sharma, R.S.K., Jaafar, T., Lim, Y.A.L.,
575 2015. Detection of Schistosoma spindale ova and associated risk factors among Malaysian cattle through
576 coprological survey. Japanese Journal of Veterinary Research 63, 63-71.

577 Tsega, M., Derso, S., 2015. Prevalence of bovine schistosomiasis and its associated risk factor in and around
578 Debre Tabor town, north west of Ethiopia. Europ J Biol Sci 7, 108-113.

579 Van Dorssen, C.F., Gordon, C.A., Li, Y., Williams, G.M., Wang, Y., Luo, Z., Gobert, G.N., You, H., McManus,
580 D.P., Gray, D.J., 2017. Rodents, goats and dogs – their potential roles in the transmission of schistosomiasis in
581 China. Parasitology 144, 1633-1642.

582 World Health Organization, 2020. Schistosomiasis: Key Facts, Fact sheets.

583 Xu, J., Steinman, P., Maybe, D., Zhou, X.N., Lv, S., Li, S.Z., Peeling, R., 2016. Evolution of the National
584 Schistosomiasis Control Programmes in The People's Republic of China. Adv. Parasitol. 92, 1-38.

585 Xu, S., Shi, F., Shen, W., Lin, J., Wang, Y., Lin, B., Qian, C., Ye, P., Fu, L., Shi, Y., 1993. Vaccination of bovines
586 against schistosomiasis japonica with cryopreserved-irradiated and freeze-thaw schistosomula. Veterinary
587 parasitology 47, 37-50.

588 Yihunie, A., Urga, B., Alebie, G., 2019. Prevalence and risk factors of bovine schistosomiasis in Northwestern
589 Ethiopia. BMC veterinary research 15, 12.

590 Zhang, L.J., Xu, Z.M., Yang, F., Dang, H., Li, Y.L., Lü, S., Cao, C.L., Xu, J., Li, S.Z., Zhou, X.N., 2021. [Endemic
591 status of schistosomiasis in People's Republic of China in 2020]. Zhongguo Xue Xi Chong Bing Fang Zhi Za Zhi
592 33, 225-233.

593 Zheng, J., Guo, J.G., Wang, X.F., Zhu, H.Q., 2000. Relationship of the livestock trade to schistosomiasis
594 transmission in mountainous area. Zhongguo ji sheng chong xue yu ji sheng chong bing za zhi = Chinese journal
595 of parasitology & parasitic diseases 18, 146-148.

596 Zhou, G., Shan, L., Shi-Zhu, L., Hui, D., Li-Juan, Z., Jing, X., 2018. Analysis on the situation of schistosome
597 infections in floating population in national schistosomiasis surveillance sites of China. Zhongguo xue xi chong
598 bing fang zhi za zhi= Chinese journal of schistosomiasis control 30, 124-130.

599 Zhou, Y.B., Liang, S., Jiang, Q.W., 2012. Factors impacting on progress towards elimination of transmission of
600 schistosomiasis japonica in China. Parasit Vectors 5, 275.

601
602
603
604
605
606
607
608
609
610
611

612
613 **Figure Legends**

614

615 **Figure 1. Village-level prevalence of schistosomiasis in bovines in 2007, 2010 and 2016.** The prevalence of
616 schistosomiasis infection in bovines is indicated for each village included in 2007 (left), 2010 (center) and 2016
617 (right). The darker the shade, the higher the prevalence. While villages in the lower left corner of each map were
618 found to have repeatedly high infection prevalence, other villages achieved 0% infection in bovines in later years
619 (upper left and center right of maps), while still others emerged as high prevalence villages in 2010 (center of
620 maps) and 2016 (upper left of maps) after years of low infection prevalence in bovines. Service Layer Credits:
621 World Terrain Base Sources: Esri, USGS, NOAA. OpenStreetMap Data Extracts for China, Asia: Data/Maps
622 Copyright 2018 Geofabrik GmbH and OpenStreetMap Contributors.

623 *Note: Figure 1 has a color version (online) and a grayscale version (print).*

624

625 **Figure 2.** V**ariable importance rankings and direction of association for candidate predictors of bovine *S.***
626 ***japonicum* infection in 2007, 2010 and 2016.** Variable importance rankings are based on a composite of mean
627 decrease in accuracy scores for 10 random forest (RF) models for each model type (full, lean and sensitivity) and
628 collection year. The direction of association was determined through logistic regression, using tertile categories
629 for continuous variables to assess evidence for non-linearity. A p-value of <0.2 was used to indicate evidence of a
630 between-group difference, and, when a between group difference was found, the direction of association is
631 indicated. See Supplementary Table S2 for detailed logistic regression results.

632 *Note: Figure 2 will be in color for both the online and print versions.*

633

634 **Figure 3. Changes in agricultural practices and the relationship between bovine infection and agricultural**
635 **predictors over time.** For each of the agricultural predictors included in this analysis, boxplots are used to
636 represent the distribution of uninfected (white) and infected bovines, for household-level (left) and village-level
637 variables (right) in 2007, 2010 and 2016.

638 *Note: Figure 3 has a color version (online) and a grayscale version (print).*

Figure 4. ROC for the best performing models. The ROC curves and corresponding AUC are shown for the top performing full, lean and sensitivity models for A) 2007 in the top panel, B) 2010 in the middle panel, and C) 2016 in the bottom panel.

Note: Figure 4 has a color version (online) and a grayscale version (print).

Click here to access/download;Table;BovinePredictorsPaper_Tables.docx

Table 1. Summary of predictor variables included in the analysis.

Predictor list	Predictor category [a]	Scale of analysis	Years available	Variable type
Bovine type (water buffalo vs. cow)	Physical/biological	Individual	2007, 2010	Binary
Bovine sex	Physical/biological	Individual	2007, 2010	Binary
Bovine age	Physical/biological	Individual	2007, 2010	Continuous
Number of hatch tests [b]	Physical/biological	Individual	All	Discrete
County of residence	Physical/biological	Household	All	Binary
Prevalence of human infection in village	Potential human sources	Village	All	Continuous
Number of infected human household members	Potential human sources	Household	All	Discrete
Percent of households with improved sanitation	Socio-economic indicators	Village	All	Continuous
Household has improved sanitation (y/n) [c]	Socio-economic indicators	Household	All	Binary
Village mean asset score (0-9)	Socio-economic indicators	Village	All	Continuous
Household asset score (0-9)	Socio-economic indicators	Household	All	Discrete
Percent of households that own cats	Animal reservoirs/sources	Village	All	Continuous
Household cat ownership	Animal reservoirs/sources	Household	All	Binary
Percent of households that own dogs	Animal reservoirs/sources	Village	All	Continuous
Household dog ownership	Animal reservoirs/sources	Household	2007, 2010	Binary
Mean number of pigs per household	Animal reservoirs/sources	Village	2010, 2016	Continuous
Household pig ownership	Animal reservoirs/sources	Household	2010	Discrete
Mean number of bovines per household	Animal reservoirs/sources	Village	All	Continuous
Household owns other bovines	Animal reservoirs/sources	Household	All	Discrete
Mean area of rice planted per household	Agricultural risk factors	Village	All	Continuous
Area of rice planted by household	Agricultural risk factors	Household	All	Continuous
Mean area of dry summer crop planted per household	Agricultural risk factors	Village	All	Continuous
Area of dry summer crop planted by household	Agricultural risk factors	Household	All	Continuous
Mean area of winter crop planted per household	Agricultural risk factors	Village	All	Continuous
Area of winter crop planted by household	Agricultural risk factors	Household	All	Continuous
Mean number of buckets of night soil used on rice crops	Agricultural risk factors	Village	All	Continuous
Number of buckets of night soil used on rice crops	Agricultural risk factors	Household	All	Discrete
Mean number of buckets of night soil used on dry summer crops	Agricultural risk factors	Village	All	Continuous
Number of buckets of night soil used on dry summer crops	Agricultural risk factors	Household	All	Discrete
Mean number of buckets of night soil used on winter crops	Agricultural risk factors	Village	All	Continuous
Number of buckets of night soil used on winter crops	Agricultural risk factors	Household	All	Discrete

[a] Predictors were grouped into five categories relevant to bovine *S. japonicum* infection risk probability: 1) physical/biological characteristics (e.g. old-age); 2) potential human sources of environmental schistosomes (e.g. human *S. japonicum* infection prevalence in the bovine's household); 3) socio-economic indicators (e.g. prevalence of improved sanitation systems in the surround village); 4) potential animal reservoirs/sources of infections (e.g. prevalence of dog ownership in the surrounding village); 5) agricultural risk factors (e.g. a household's total rice crop area).

[b] Because not all bovines produced three stool samples, and examination of a greater number of stool samples can increase the probability of detecting infection, the number of hatch tests used on a given bovine was also included as a predictor.

[c] Improved sanitation was defined as access to an improved toilet in the household, including a biogas digester or a three-compartment toilet.

Table 2. Tabulation of all predictors by bovine infection status.

	2007			2010			2016		
	N Positive	**Total tested**	**% Positive**	**N Positive**	**Total tested**	**% Positive**	**N Positive**	**Total tested**	**% Positive**
Total	**63**	**473**	**13.32%**	**34**	**197**	**17.26%**	**14**	**71**	**19.72%**
Bovine type									
Water buffalo	7	86	8.14%	5	29	17.24%	ND [a]	ND [a]	
Cattle	54	380	14.21%	28	158	17.72%	ND [a]	ND [a]	
Missing	2	7	28.57%	1	10	10.00%			
Bovine sex									
Female	50	320	15.63%	29	163	17.79%	ND [a]	ND [a]	
Male	9	51	17.65%	4	24	16.67%	ND [a]	ND [a]	
Missing	4	102	3.92%	1	10	10.00%			
Bovine age [b]									
<2	3	43	6.98%	6	35	17.14%	ND [a]	ND [a]	
2 to 4	26	220	11.82%	9	59	15.25%	ND [a]	ND [a]	
5+	28	137	20.44%	18	93	19.35%	ND [a]	ND [a]	
Missing	6	73	8.22%	1	10	10.00%			
Hatch tests (N)									
1	3	59	5.08%	0	29	0.00%	1	10	10.00%
2	8	80	10.00%	6	41	14.63%	2	9	22.22%
3	52	334	15.57%	28	127	22.05%	11	52	21.15%
County									
1	15	191	7.85%	17	71	23.94%	3	21	14.29%
2	48	282	17.02%	17	126	13.49%	11	50	22.00%
Prevalence of human infection in village									
0-2.53%	14	147	9.52%	0	76	0.00%	4	27	14.81%
2.54- 11.11%	13	158	8.23%	11	57	19.30%	8	29	27.59%
≥11.12%	36	168	21.43%	23	64	35.94%	2	15	13.33%
Number of infected human household members									
0	38	366	10.38%	19	154	12.34%	12	61	19.67%
1+	19	84	22.62%	13	29	44.83%	2	10	20.00%
Missing	6	23	26.09%	2	14	14.29%	0	0	--
Percent of households with improved sanitation									
<10%	14	180	7.78%	3	24	12.50%	0	0	0.00%
10-50%	38	215	17.67%	29	133	21.80%	6	24	25.00%
≥50%	11	78	14.10%	2	40	5.00%	8	47	17.02%
Household has improved sanitation									
No	47	365	12.88%	27	129	20.93%	8	34	23.53%
Yes	16	108	14.81%	7	68	10.29%	6	37	16.22%
Village mean asset score (0-9)									
<2	39	225	17.33%	0	0	0.00%	0	0	0.00%
2-3	24	248	9.68	26	143	18.18%	2	23	8.70%
≥4	0	0	0.00%	8	54	14.81%	12	48	25.00%
Household asset score (0-9)									
0 - 2	22	184	11.96%	4	24	16.67%	2	10	20.00%
2 - 3	31	211	14.69%	15	89	16.85%	3	14	21.43%
4+	10	78	12.82%	15	84	17.86%	9	47	19.15%
Percent of households that owns dogs									
<70%	17	98	17.35%	2	51	3.92%	7	50	14.00%
70-85%	34	217	15.67%	18	94	19.15%	7	21	33.33%
≥85%	12	158	7.59%	14	52	26.92%	0	0	0.00%
Household dog ownership [c]									
No	6	67	8.96%	3	29	10.34%	2	7	28.57%
Yes	57	404	14.11%	31	168	18.45%	12	64 [c]	18.75%
Missing	0	2	0.00%	0	0	--	0	0	--

Percent of households that owns cats									
<45%	12	109	11.01%	12	71	16.90%	4	36	11.11%
45-60%	21	124	16.94%	14	73	19.18%	5	13	38.46%
≥60%	30	240	12.50%	8	53	15.09%	5	22	22.73%
Household cat ownership									
No	17	155	10.97%	12	69	17.39%	5	23	21.74%
Yes	46	312	14.74%	22	128	17.19%	9	48	18.75%
Missing	0	6	0.00%	0	0	--	0	0	--
Mean number of pigs per household									
0	ND [a]	ND [a]		1	21	4.76%	11	20	55.00%
0.01 - 1	ND [a]	ND [a]		26	119	21.85%	1	45	2.22%
≥ 1	ND [a]	ND [a]		7	57	12.28%	2	6	33.33%
Household pig ownership [c]									
0	ND [a]	ND [a]		27	116	23.28%	14	64 [c]	21.88%
1+	ND [a]	ND [a]		7	81	8.64%	0	7	0.00%
Mean number of bovines per household									
< 0.5	5	116	4.31%	4	59	6.78%	10	57	17.54%
0.5 – 1	40	300	13.33%	19	99	19.19%	4	14	28.57%
≥ 1	18	57	31.58%	11	39	28.21%	0	0	0.00%
Household bovine ownership									
0	49	334	14.67%	20	126	15.87%	8	38	21.05%
1+	14	139	10.07%	14	71	19.72%	6	33	18.18%
Mean area of rice planted per household (mu)									
<0.75	14	143	9.79%	9	90	10.00%	7	18	38.89%
0.75-1.5	19	183	10.38%	12	48	25.00%	5	41	12.20%
≥1.5	30	147	20.41%	13	59	22.03%	2	12	16.67%
Area of rice planted by household (mu)									
0	16	146	10.96%	5	56	8.93%	5	13	38.46%
<2	24	195	12.31%	16	84	19.05%	4	31	12.90%
2+	23	132	17.42%	13	57	22.81%	5	27	18.52%
Mean area of dry summer crops planted per household (mu)									
<1	11	184	5.98%	1	14	7.14%	0	3	0.00%
1 -2.5	44	195	22.56%	23	110	20.91%	2	26	7.69%
≥2.5	8	94	8.51%	10	73	13.70%	12	42	28.57%
Area of dry summer crops planted by household (mu)									
0	8	108	7.41%	0	4	0.00%	0	0	--
<3	39	242	16.12%	18	99	18.18%	2	26	7.69%
3+	16	123	13.01%	16	94	17.02%	12	45	26.67%
Mean area of winter crops plant planted per household (mu)									
<2	21	149	14.09%	17	77	22.08%	12	53	22.64%
2-2.75	18	178	10.11%	11	66	16.67%	0	2	0.00%
≥2.75	24	146	16.44%	6	54	11.11%	2	16	12.50%
Area of winter crops planted by household (mu)									
<2	14	142	9.86%	9	52	17.31%	7	29	24.14%
2-3	33	223	14.80%	18	97	18.56%	7	29	24.14%
4+	16	108	14.81%	7	48	14.58%	0	13	0.00%
Mean number of buckets of night soil used on rice crops per household									
<1	3	58	5.17%	29	146	19.86%	8	41	19.51%
1-9.9	21	153	13.73%	4	38	10.53%	6	27	22.22%
≥10	39	262	14.89%	1	13	7.69%	0	3	0.00%
Number of buckets of night soil used on rice crops by household									
0	40	304	13.16%	29	168	17.26%	8	53	15.09%

1+	22	168	13.10%	5	29	17.24%	1	7	14.29%
Missing	1	1	100.0%	0	0	--	5	11	45.45%
Mean buckets of night soil used on dry summer crops per household									
<10	13	190	6.84%	24	99	24.24%	1	15	6.67%
10-30	18	109	16.51%	7	47	14.89%	13	47	27.66%
≥30	32	174	18.39%	3	51	5.88%	0	9	0.00%
Number of buckets of night soil used on dry summer crops by household									
0	31	248	12.50%	24	105	22.86%	6	36	16.67%
1-25	6	53	11.32%	5	32	15.63%	5	22	22.73%
>25	26	172	15.12%	5	60	8.33%	3	13	23.08%
Mean number of buckets of night soil used on winter crops per household									
<10	1	26	3.85%	24	121	19.83%	14	71	100.0%
10-30	19	218	8.72%	7	41	17.07%	0	0	--
≥30	43	229	18.78%	3	35	8.57%	0	0	--
Number of buckets of night soil used on winter crops by household									
0	27	197	13.71%	31	137	22.63%	9	62	14.52%
1-26	5	79	6.33%	0	16	0.00%	1	3	33.33%
>26	31	197	15.74%	3	44	6.82%	4	6	66.67%

[a] No data (ND). Data was not collected on bovine sex, age or type in 2016, or on pig ownership in 2007.

[b] Although predictors are broken into categories to facilitate a comparison of the distributions across the study period, most were assessed as continuous variable in our RF models. See Table 1 for details on variable formulation for the RF models.

[c] Variables with >90% of observations taking on a single value were excluded from the RF assessment. This exclusion criteria applied twice in 2016: >90% of the included households reported owning at least one dog, and >90% of households owned zero pigs

Table 3. Comparison of model performance metrics for the top performing model [a] from the full, lean and sensitivity analyses in 2007, 2010 and 2016.

Performance metrics	2007 Models			2010 Models			2016 Models		
	Full	**Lean**	**Sens.**	**Full**	**Lean**	**Sens.**	**Full**	**Lean**	**Sens.**
Accuracy	0.864	0.864	0.856	0.816	0.816	0.816	1	0.944	1
95% CI	0.79 - 0.92	0.79 - 0.92	0.78 - 0.91	0.68 - 0.91	0.68 - 0.91	0.68 - 0.91	0.81 - 1	0.73 - 1	0.81 - 1
NIR [b]	0.864	0.864	0.864	0.837	0.837	0.837	0.778	0.778	0.778
Kappa [c]	0.313	0.313	0.246	0.463	0.360	0.416	1	0.853	1
Sensitivity	0.313	0.313	0.25	0.75	0.5	0.505	1	1	1
Specificity	0.951	0.951	0.951	0.829	0.878	0.625	1	0.929	1

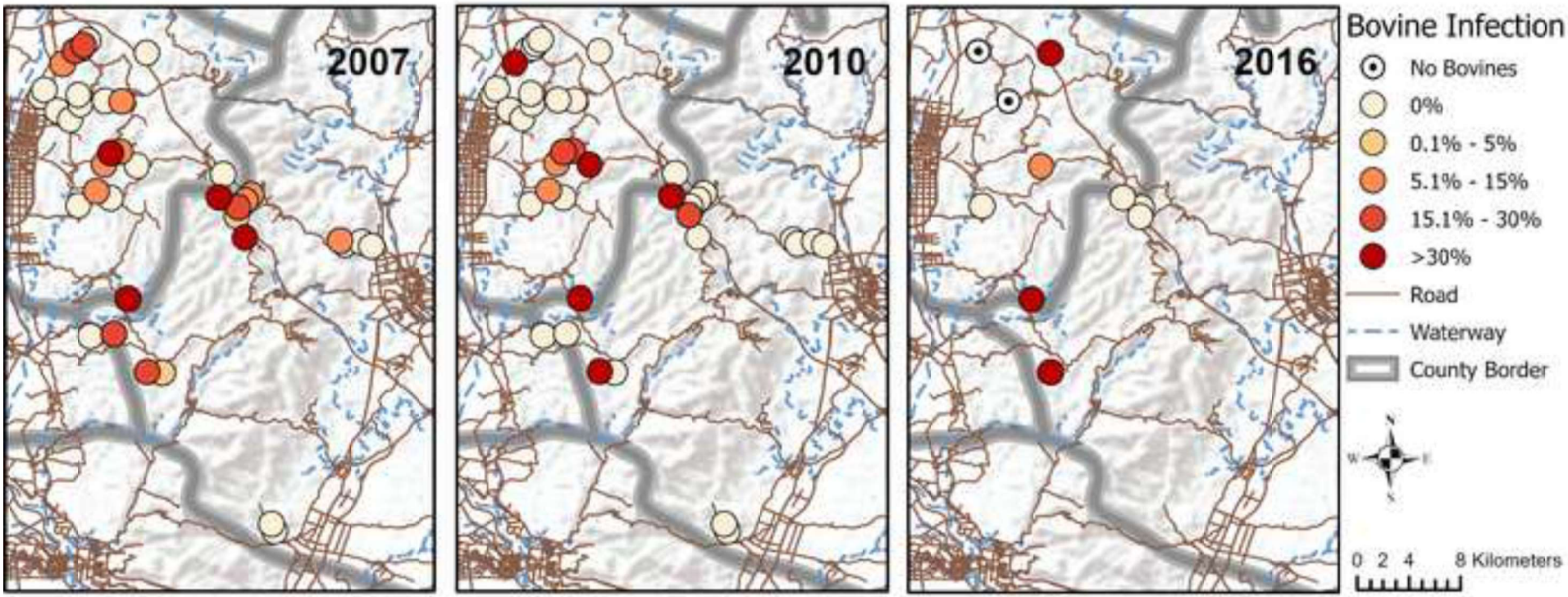
[a] The top performing model was defined as the one with the highest accuracy for each analysis type (full, lean & sensitivity) and collection year (2007, 2010, 2016). In the case of a tie for the highest accuracy value, the sensitivity, ROC AUC, kappa and specificity were subsequently compared to select the top performing model for each analysis type and year.

[b] No Information Rate

[c] Due to the high degree of imbalance between the outcome classes across the study period, the Kappa values is a useful metric for our models, as it helps to correct bias that results when rewarding the prediction of the majority class. The benchmark values outlined by Landis & Koch (1977) are useful here for determining the relative strength of the predictive models: <0.00 = Poor; 0.00 – 0.20= Slight; 0.21 – 0.40 = Fair; 0.41 – 0.60 = Moderate; 0.61 – 0.81 = Substantial; 0.81 – 1.0 = Almost Perfect.

Figure1_color

Click here to access/download;Figure;Figure1_Color.jpg



Click here to access/download;Figure;Figure1_Grayscale.jpg

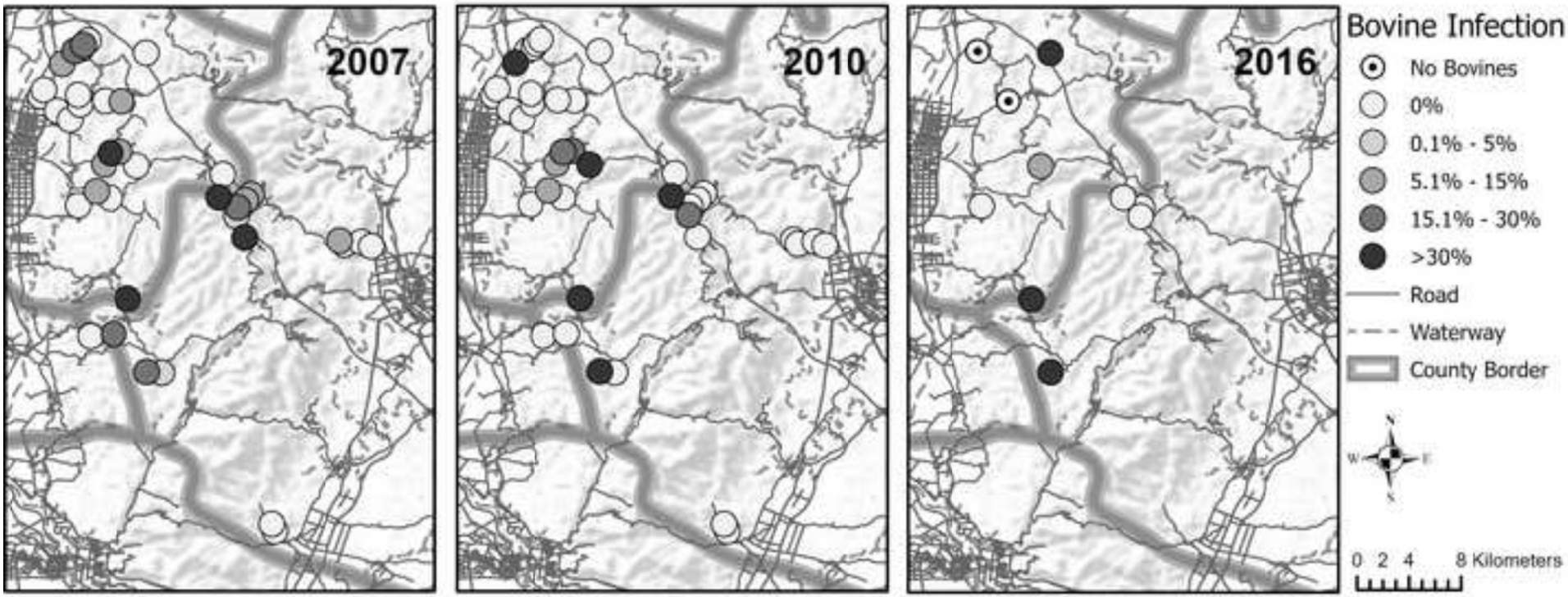


Figure2

Click here to access/download;Figure;Figure2_IN_COLOR_ONLY.pdf

Predictor categories	2007			2010			2016		
	Full	Lean	Sens.	Full	Lean	Sens.	Full	Lean	Sens.
Physical/biological characteristics									
Bovine type							ND	ND	ND
Bovine sex							ND	ND	ND
Bovine age	3 ↑	1 ↑	2 ↑				ND	ND	ND
Number of hatch tests				2 ↑	2 ↑	1 ↑			
County of residence									
Potential human sources of schistosomes									
Human infection prevalence in the village	5 ↑	3 ↑	Excl.	1 ↑	1↑	Excl.			Excl.
Number of infected human household members			Excl.	6 ↑	7 ↑	Excl.			Excl.
Socio-economic indicators									
Village prevalence of improved sanitation						9 ∩			
Household has improved sanitation (y/n)									
Village mean asset score (0-9)			10 ↓				8 ↑	5 ↑	8 ↑
Household asset score (0-9)						10 ∅			
Potential animal reservoirs/sources of infection									
Village prevalence of cat ownership							9* ↑	10↑	10 ↑
Household cat ownership									
Village prevalence of dog ownership							9* ↑	9 ↑	
Household dog ownership							Excl.	Excl.	Excl.
Village mean number of pigs owned	ND	ND	ND				3 ↓	7 ↓	4 ↓
Household pig ownership	ND	ND	ND	7 ↓	8 ↓	6 ↓	Excl.	Excl.	Excl.
Village mean number of bovines owned	8 ↑	8 ↑	6 ↑			8 ↑	2 ↑	1 ↑	1 ↑
Household bovine ownership									
Agricultural risk factors									
Village mean rice area	6 ↑	5 ↑	3 ↑	5 ↑	5 ↑	3* ↑	5 ↓	8 ↓	5 ↓
Household rice area				8 ↑	10↑	7 ↑			
Village mean dry summer crop area	2 ↑	7 ↑	4 ↑				6 ↑	6 ↑	7 ↑
Household dry summer crop area	4 ↑	6 ↑	5 ↑	4 ↓	4 ↓	3* ↓			9 ↑
Village mean winter crop area	1 ↑	2 ↑	1 ↑	10↓	6 ↓		1 ↓	2 ↓	2 ↓
Household winter crop area	7 ↑	4 ↑	7 ↑	9 ∅	9 ∅	5 ∅	7 ↓	4 ↓	6 ↓
Village night soil rice: mean # buckets									
Household night soil rice: # buckets									
Village night soil summer dry crop: mean # buckets	9 ↑	10↑	8 ↑	3 ↓	3 ↓	2 ↓	4 ↑	3 ↑	3 ↑
HH night soil summer dry crop: # buckets									
Village night soil winter crop: mean # buckets	10* U	9 U	9 U						
Household night soil winter crop: # buckets	10* U								

Symbol key:

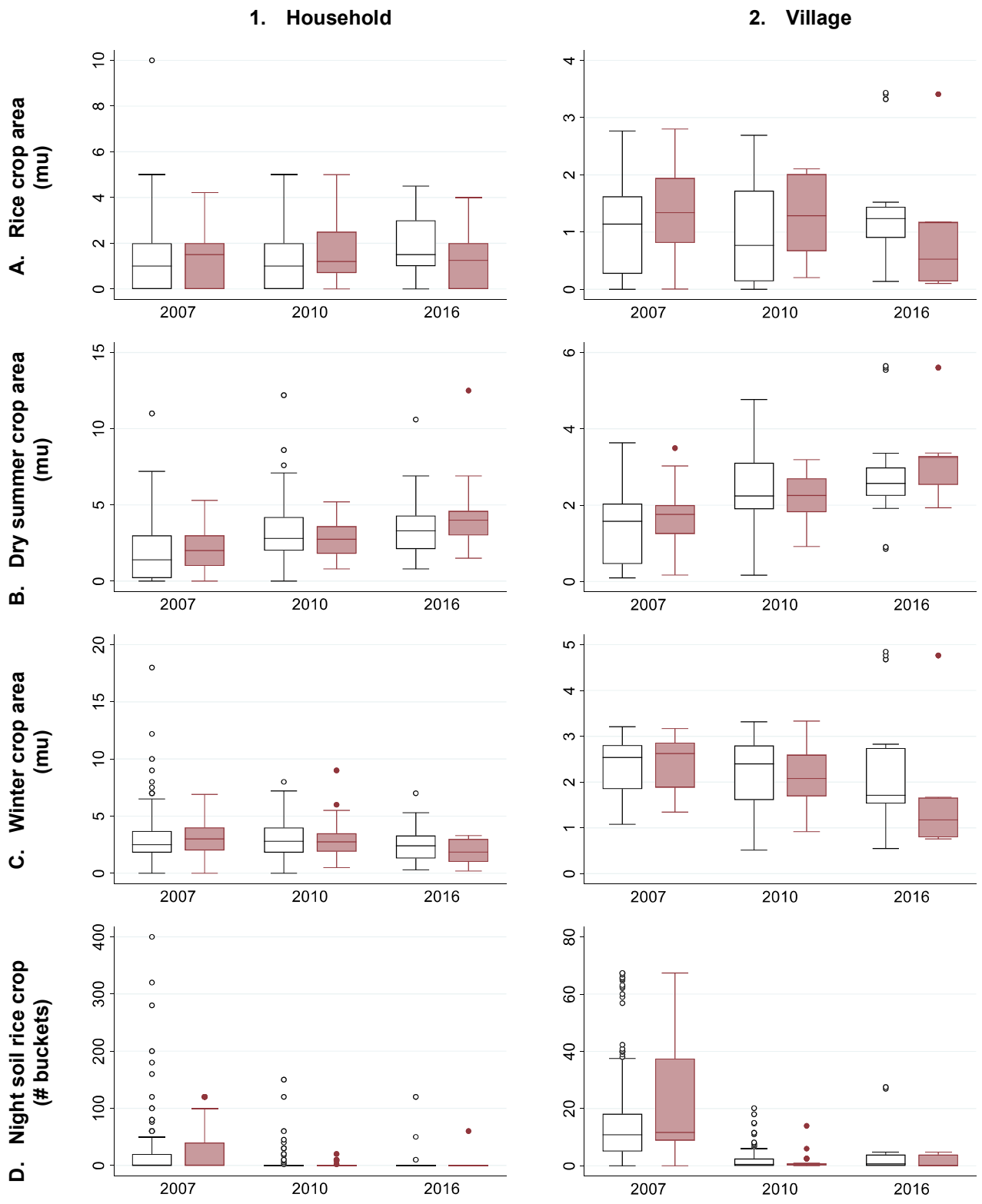
- ↑ Positive association
- ↓ Negative association
- ∩ Non-linear association (Rise-Fall)
- U Non-linear association (Fall-Rise)
- ND Data not collected
- Excl. Variable excluded from model
- * Tied for importance rank
- ∅ No evidence for a significant between - group difference found

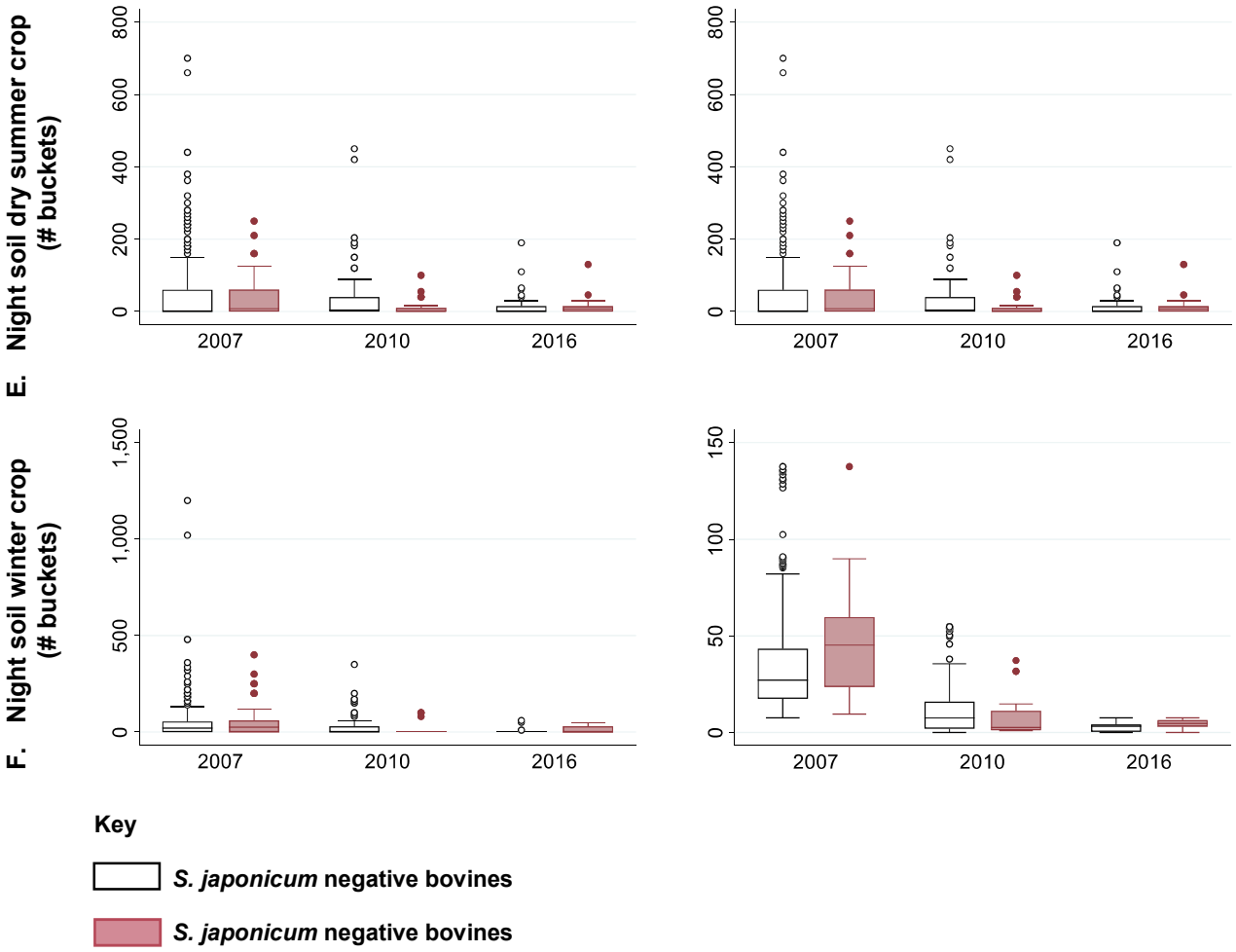
Color key: Variable importance rankings (1st-10th) by scale

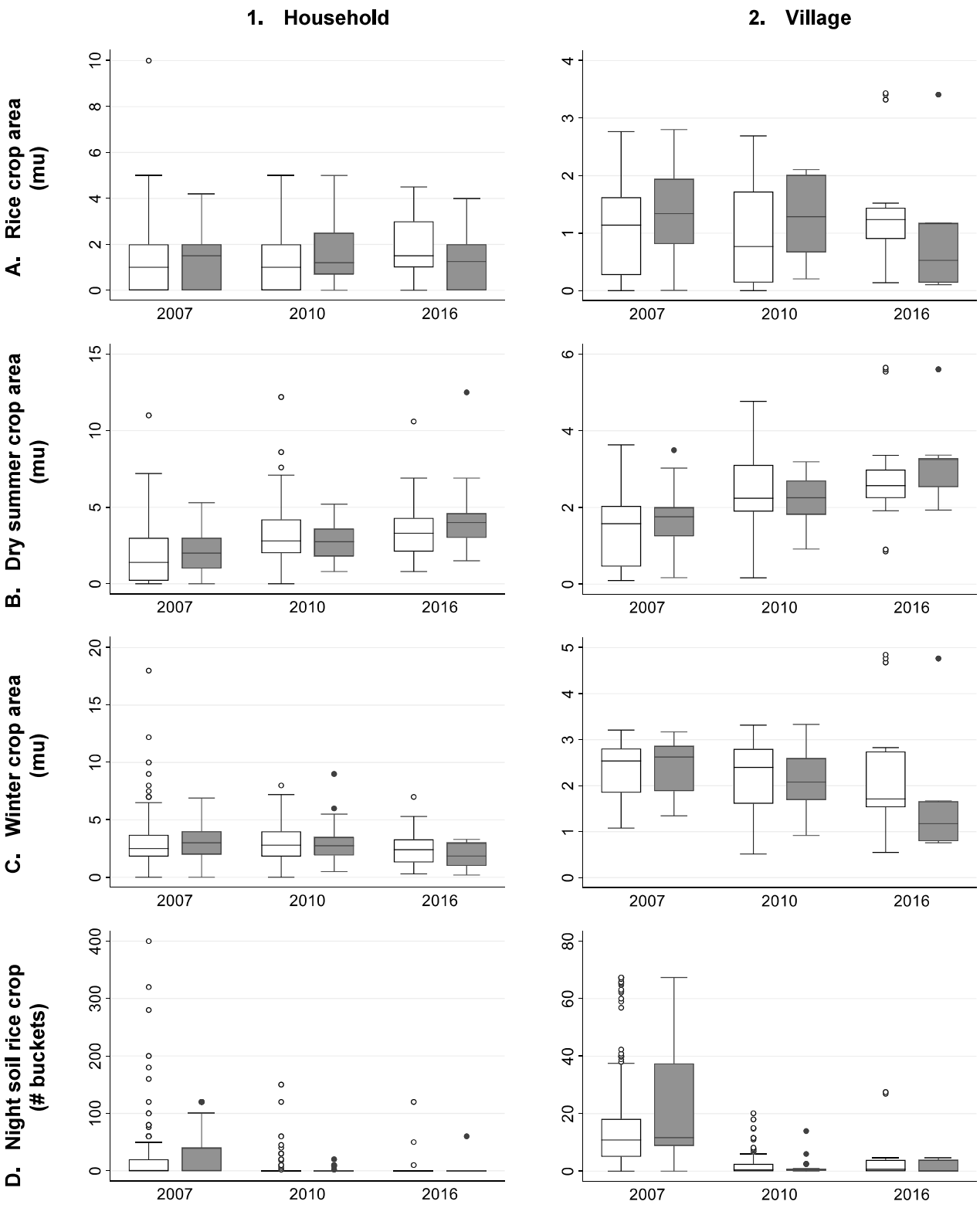
Individual	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th
Household	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th
Village	1st – 2nd	3rd – 4th	5th – 6th	7th – 8th	9th – 10th

= Not ranked 1-10 for the given model/year

Click here to access/download;Figure;Figure3_Color.pdf




E.
Night soil dry summer crop (# buckets)
0
200
400
600
800
2007
2010
2016
F.
Night soil winter crop (# buckets)
0
500
1,000
1,500
0
50
100
150
2007
2010
2016
Key
S. japonicum negative bovines
S. japonicum negative bovines


1. Household
2. Village
A. Rice crop area (mu)
B. Dry summer crop area (mu)
C. Winter crop area (mu)
D. Night soil rice crop (# buckets)
2007
2010
2016

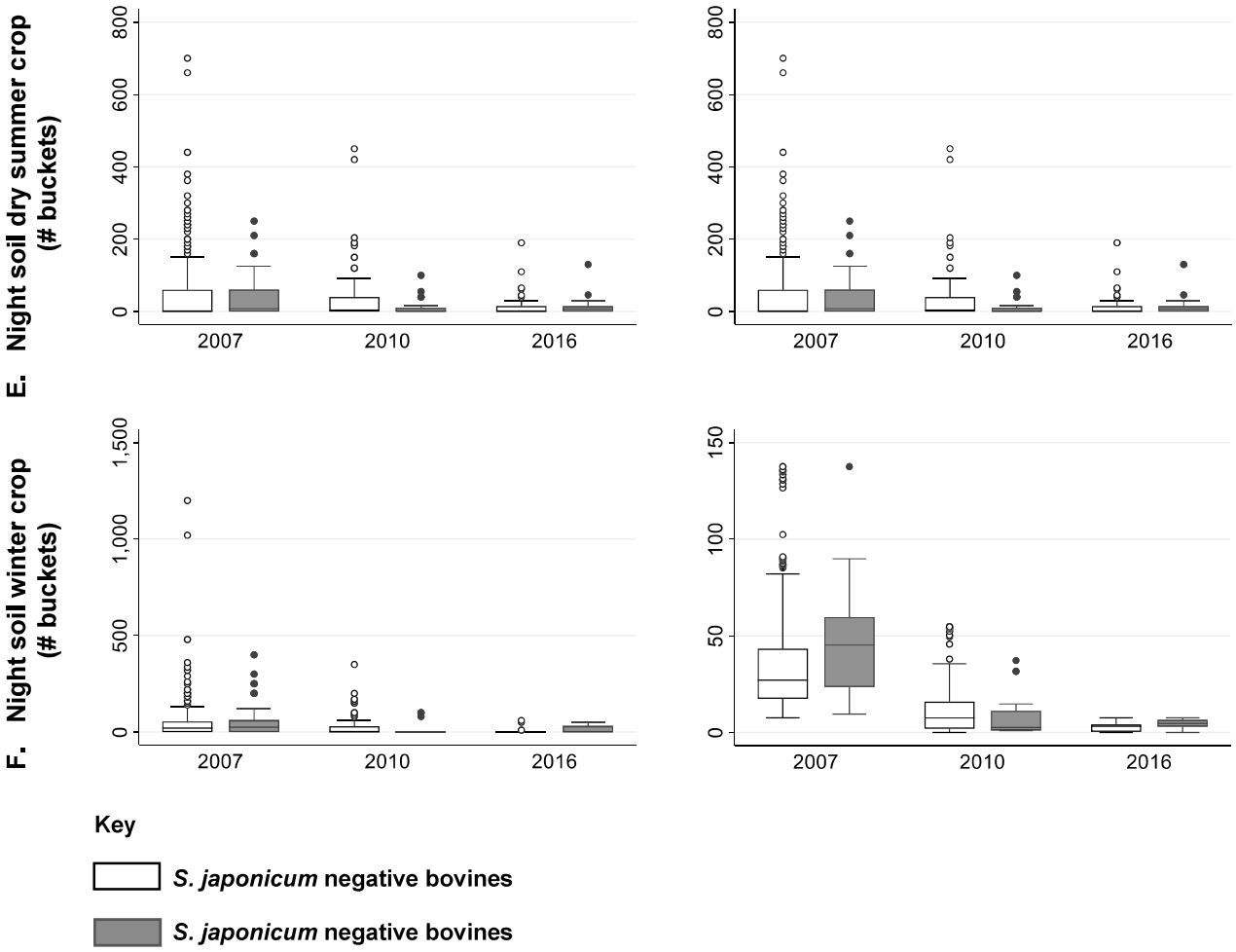
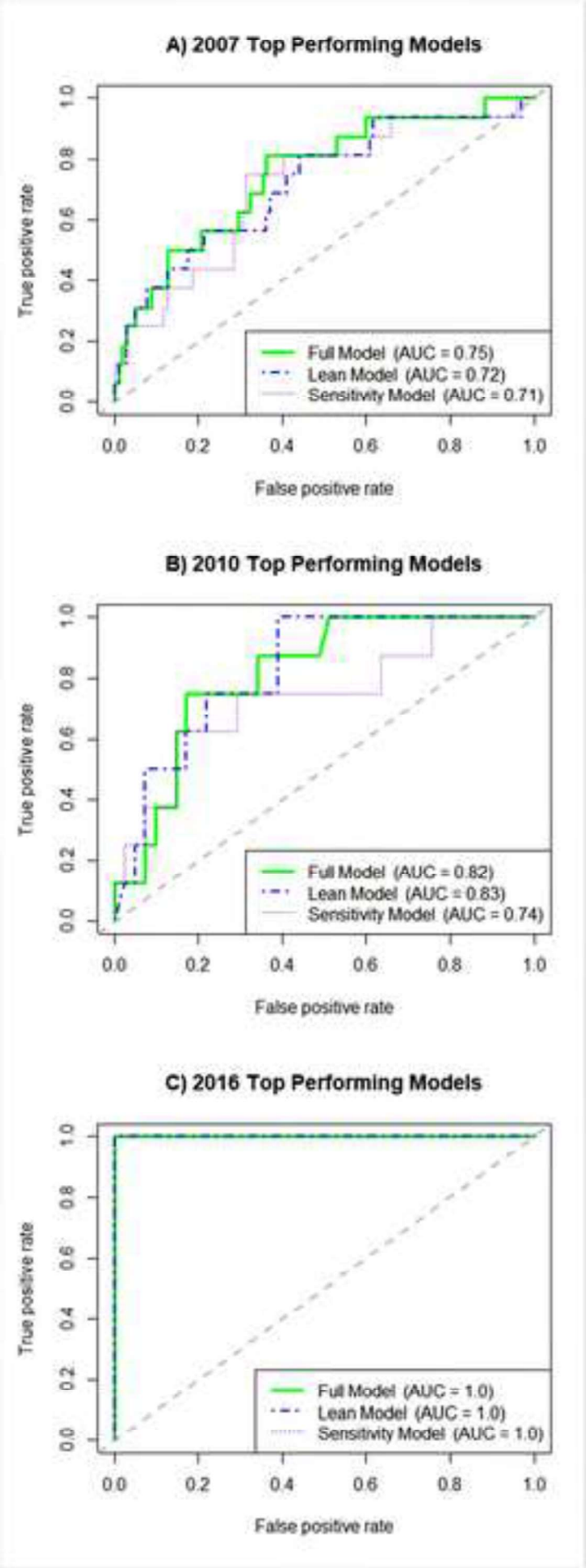
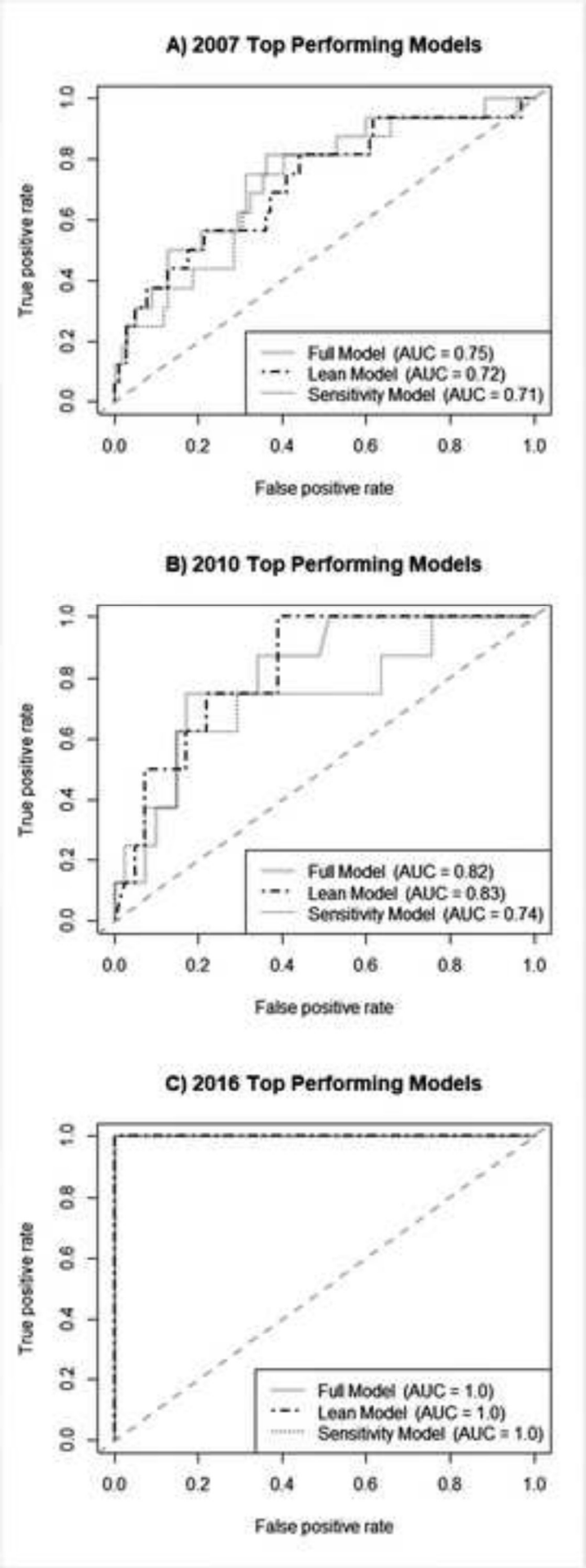

E. Night soil dry summer crop (# buckets)
F. Night soil winter crop (# buckets)
0
200
400
600
800
500
1,000
1,500
50
100
150
2007
2010
2016
Key
S. japonicum negative bovines
S. japonicum negative bovines

Figure4_color

Click here to access/download;Figure;Figure4_color.jpg




A) 2007 Top Performing Models
True positive rate
False positive rate
Full Model (AUC = 0.75)
Lean Model (AUC = 0.72)
Sensitivity Model (AUC = 0.71)
B) 2010 Top Performing Models
True positive rate
False positive rate
Full Model (AUC = 0.82)
Lean Model (AUC = 0.83)
Sensitivity Model (AUC = 0.74)
C) 2016 Top Performing Models
True positive rate
False positive rate
Full Model (AUC = 1.0)
Lean Model (AUC = 1.0)
Sensitivity Model (AUC = 1.0)