

Statistical matching for conservation science

Schleicher, Judith; Eklund, Johanna; Barnes, Megan ; Goldman, Jonas;
Oldekop, Johan A.; Jones, Julia P.G.

Conservation Biology

DOI:

[10.1111/cobi.13448](https://doi.org/10.1111/cobi.13448)

Published: 01/06/2020

Peer reviewed version

[Cyswllt i'r cyhoeddiad / Link to publication](#)

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):

Schleicher, J., Eklund, J., Barnes, M., Goldman, J., Oldekop, J. A., & Jones, J. P. G. (2020).
Statistical matching for conservation science. *Conservation Biology*, 34(3), 538-549.
<https://doi.org/10.1111/cobi.13448>

Hawliau Cyffredinol / General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1 **Title: A good match? The appropriate use of statistical matching in conservation**
2 **impact evaluation**

3 **Impact statement:** We provide a step-by-step guide for using matching in conservation
4 impact evaluation; highlighting pitfalls and potential improvements

5 **Running head:** A good match?

6 **Keywords:** causal inference, conservation effectiveness, spill-over, spatial autocorrelation,
7 counterfactual

8 **Word count:** 6,183 words.

9 **Authors:** Judith Schleicher^{a,*}, Johanna Eklund^b, Megan Barnes^c, Jonas Geldmann^d, Johan A.
10 Oldekop^e and Julia P. G. Jones^f

11 ^a Department of Geography, University of Cambridge, UK; email: judith.schleicher@geog.cam.ac.uk.

12 ^b Department of Geosciences and Geography, Helsinki Institute of Sustainability Science, Faculty of
13 Science, PO Box 64 (Gustaf Hällströmin katu 2A), FI-00014 University of Helsinki, Finland; email:
14 johanna.f.eklund@helsinki.fi

15 ^c School of Biology, The University of Queensland, St Lucia, QLD, 4067, Australia; email:
16 meganbarnes84@gmail.com

17 ^d Conservation Science Group, Department of Zoology, University of Cambridge, Downing St.,
18 Cambridge CB2 3EJ, UK; email: jg794@cam.ac.uk

19 ^e School of Environment, Education and Development, University of Manchester, Oxford Road,
20 Manchester, M13 9PL, UK; email: johan.oldekop@manchester.ac.uk

21 ^f College of Engineering and Environmental Sciences, Thoday Road, Deniol Road, Bangor University,
22 LL57 2UW, UK; email: julia.jones@bangor.ac.uk

23 * Corresponding author: judith.schleicher@geog.cam.ac.uk; +44 (0)7570086101

24 **Acknowledgements:** This paper resulted from a symposium JS and JE organised at the
25 European Congress of Conservation Biology 2018, for which we received funding from the
26 FinCEAL programme. We thank Diogo Veríssimo, Rachel Carmenta, Janet Lessmann,
27 Alejandro Guizar and Stephanie Hernandez for helpful comments on earlier versions of this
28 paper, and governmental and non-governmental organizations in Peru for data that
29 provided the basis for Figure 2. JS was supported by the Economic and Social Research

30 Council (grant number ES/I019650/1). JPGJ thanks Fitzwilliam College and the Leverhulme
31 Trust (grant RPG-2014-056). JE thanks the Kone foundation for funding. JG was supported by
32 European Union's Horizon 2020 Marie Skłodowska-Curie programme (No 676108) and
33 Villum Fonden (VKR023371).

34 **ABSTRACT**

35 The awareness of the need for robust impact evaluations in conservation is growing, and
36 statistical matching techniques are increasingly being use to assess the impacts of
37 conservation interventions. Used appropriately, matching approaches are powerful tools, but
38 they also pose potential pitfalls. We present important considerations and best practice when
39 using matching in conservation science. We identify three steps in a matching analysis. The
40 first step requires a clear theory of change to inform selection of treatment and controls,
41 accounting for real world complexities and potential spill-over effects. The second step
42 involves selecting the appropriate covariates and matching approach. The third step is
43 assessing the quality of the matching by carrying out a series of checks. The second and third
44 steps can be repeated and should be finalized before outcomes are explored. Future
45 conservation impact evaluations could be improved by increased planning of evaluations
46 alongside the intervention, better integration of qualitative methods, considering spill-over
47 effects at larger spatial scales, and more publication of pre-analysis plans. This will require
48 more serious engagement of conservation scientists, practitioners and funders to mainstream
49 robust impact evaluations into conservation. We hope that this paper will improve the quality
50 of evaluations, and help direct future research to continue to improve the approaches on offer.

51 INTRODUCTION

52 There have been numerous calls for conservation science to provide a stronger evidence base
53 for policy and practice (Pullin & Knight 2001; Sutherland et al. 2004; Baylis et al. 2016). Rigorous
54 impact assessments of conservation interventions is vital to prevent wasting conservation
55 resources (Ferraro & Pattanayak 2006), and tackling rapid biodiversity loss. While the
56 importance of establishing counterfactuals (what would have happened in the absence of an
57 intervention) to generate more precise, and less biased, estimates of conservation impacts is
58 increasingly recognized (Baylis et al. 2016), robust impact evaluations remain limited in number
59 and scope (Schleicher 2018).

60

61 It is seldom feasible, or even desirable, to randomly implement conservation interventions for
62 ethical, logistical and political reasons. Experimental evaluations are therefore likely to remain
63 rare (Baylis et al. 2016; Pynegar et al. 2018; Wiik et al. 2019). However, methodological
64 advances to improve causal inference from non-experimental data have helped to better
65 attribute conservation impacts (Ferraro & Hanauer 2014a). These methods emulate
66 experiments by identifying treatment and control groups with similar observed and
67 unobserved characteristics (Rosenbaum & Rubin 1983; Stuart 2010). Among the range of non-
68 experimental approaches available for impact evaluations, each with their strengths and
69 weaknesses (see Table 1), 'matching' approaches are playing an increasingly important role in
70 conservation science (e.g. Andam et al. 2008; Nelson & Chomitz 2011; Naidoo et al. 2019).

71

72 Matching comprises a suite of statistical techniques aiming to improve causal inference of
73 subsequent analyses. They do so by identifying 'control' units that are closely 'matched' to
74 'treatment' units according to pre-defined measurable characteristics (covariates), and a

75 measure of similarity (Gelman & Hill 2007; Stuart 2010). Selecting comparable units of analysis
76 (e.g. sites, individuals, households or communities) is important when conservation
77 interventions are not assigned randomly. This is because units exposed to the intervention
78 (treatment units), and those not exposed (control units) can differ in characteristics that
79 influence the allocation of the treatment (i.e. where an intervention occurs, or who receives it)
80 and the outcome of interest (e.g. species population trends, deforestation rates, changes in
81 poverty levels). These characteristics are commonly referred to as confounding factors. For
82 example, habitat conditions *before* an intervention can influence both the likelihood of the
83 intervention being carried out in a specific location, and habitat condition *after* the
84 intervention's implementation.

85 Matching has two main applications in impact evaluation. First, where researchers seek to
86 evaluate the impact of an intervention *post hoc*, matching can reduce differences between
87 treatment and control units, and help isolate intervention effects. For example, when
88 examining protected area (PA) effects on deforestation, distance from population centers
89 (remoteness) is a likely confounder: remote sites tend to be more likely designated as
90 protected, and less prone to deforestation because they are harder to reach (Joppa & Pfaff
91 2009). Second, matching can be used to inform study design and data collection prior to the
92 implementation of an intervention. For example, to evaluate how a planned conservation
93 intervention affects local communities, matching can be used to identify appropriate control
94 and treatment communities to monitor effects before and after the intervention's
95 implementation (Clements et al. 2014).

96 Matching is a powerful statistical tool, but not a magic wand. The strengths and weaknesses
97 of matching relative to alternative methods should be considered carefully, and its use

98 optimized to maximize the benefits. Given the rapid rise in the use of matching approaches in
99 conservation science, there is an urgent need for reviewing best practices and bringing
100 together the diverse technical literature, mostly from economics and statistical journals
101 (Imbens & Wooldridge 2009; Abadie & Cattaneo 2018), for a conservation science audience.
102 The few existing related papers targeted at a conservation audience have focused on the
103 conceptual underpinnings of impact evaluations (Ferraro & Hanauer 2014a; Baylis et al. 2016),
104 without providing specific methodological insights. We address this gap by providing an
105 overview of matching and key methodological considerations for the conservation science
106 community. We do so by drawing on the wider literature and our own collective experience
107 using matching in conservation impact evaluations. We focus on important considerations
108 when using matching, outline best practices, and highlight key methodological issues that
109 deserve further attention and development.

110

111 **IMPORTANT CONSIDERATIONS WHEN USING MATCHING IN CONSERVATION IMPACT** 112 **EVALUATION**

113 **Three key steps when using matching for impact evaluations**

114 As with any statistical analysis, matching studies require careful design (Stuart 2010; Ferraro &
115 Hanauer 2014a). We identify three main steps for a matching analysis (Figure 1). *The first step*
116 *involves identifying units exposed to the treatment and those not. The second step* consists of
117 *selecting appropriate covariates and the specific matching approach. The third step* involves
118 *running the matching analysis and assessing the quality of the match (Table 2). Steps 2 and 3*
119 *should be repeated iteratively until the matching has been optimized. Only then should the*
120 *matched data be used for further analysis. Doing so is important in post hoc analyses to avoid*

121 selecting a matching approach that produces a desired result (Rubin 2007). We elaborate a
122 number of key considerations involved at each of these steps (see Figure 1) below.

123

124 **Defining treatment and control units (Step 1)**

125 *A 'theory of change' is needed to make impact evaluation possible*

126 The strength of the causal inference in observational studies relies on a clear understanding
127 of the mechanism through which interventions influence outcomes of interest. Rival
128 explanations should be carefully considered and, if possible, eliminated. Therefore, although
129 impact evaluation is an empirical exercise, it requires a strong theory-based explanation and
130 model of the causal pathways linking the intervention to the outcomes of interest (Ferraro &
131 Hanauer 2014b). This theoretical model is often referred to as a 'theory of change' (also called
132 'causal chain' or 'logic model'). It comprises a theoretical understanding of how a treatment
133 interacts with the social-ecological system it is embedded in (Qiu et al. 2018). This
134 understanding is required to successfully argue that a causal pathway runs from the
135 intervention to the outcome of interest (and not *vice versa*). For example, the expansion of a
136 PA network might lead to the development of tourism infrastructure, which might also result
137 in poverty reduction (Ferraro & Hanauer 2014b; den Braber et al. 2018). However, causality
138 could run in the opposite direction: the development of tourism infrastructure close to a PA
139 might be the outcome of reduced poverty as local communities invest revenue.

140 *Real world complexity cannot be ignored*

141 Conservation interventions are seldom implemented in simple settings where the impacts of
142 one intervention can be easily separated from others. A thorough understanding of the study
143 area and context is essential for identifying appropriate treatment and control units. Typically,

144 conservation interventions are implemented in a landscape where potential treatment and
145 control units have been exposed to a range of different interventions. The availability of
146 spatially-explicit datasets identifying where interventions have been implemented, is
147 inconsistent: spatial information for some interventions are much more readily available than
148 for others (Oldekop et al. 2019). Teasing apart the effects of specific interventions can therefore
149 be challenging. In the Peruvian Amazon for example, there are few land areas with no formal
150 or informal land use restrictions, and these often overlap (Figure 2). This hinders the isolation
151 of one particular treatment-type (e.g. PA) and identifying appropriate control units (e.g. non-
152 protected land without land use restrictions). Indeed, the few matching studies that have
153 accounted for differences between land use restrictions have found that the degree to which
154 conservation interventions can be considered effective is influenced by how control areas are
155 defined and selected (Gaveau et al. 2012; Schleicher et al. 2017). Conservation impact
156 assessments could be improved by being more explicit about what the alternative land uses
157 to the conservation interventions are, and why specific controls were selected.

158 *'Spill-over' should be considered in the selection of controls*

159 A central assumption in matching studies is that the outcome in one unit is not affected by
160 the treatment in other units (Rubin 1980). However, this assumption does not always hold.
161 There are many situations where outcomes in treatment units may 'spill-over' and affect
162 outcomes in control units, either positively or negatively (Ewers & Rodrigues 2008; Baylis et al.
163 2016). For example, increased fish population in no-take zones might spill-over into adjacent
164 non-protected habitats, a case of positive spill-over that is part of the design of no-take marine
165 PAs. This would mask the positive impact of the intervention by reducing the difference
166 between treatment and potential control units. In addition, fishing effort may be displaced

167 from a no-take zone into potential control areas (negative spill-over). One might thus wrongly
168 conclude that the intervention was successful, despite there being no overall reduction in
169 fishing effort. In studies evaluating the impact of PAs on deforestation, negative spill-overs
170 (also called 'leakage') have usually been accounted for by excluding buffer zones around
171 treatment areas, so that they cannot be included as controls (Andam et al. 2008). However,
172 leakage effects can vary across landscapes (Robalino et al. 2017), and take place over larger
173 geographical scales, which have so far not been accounted for in matching studies.

174 **Selecting covariates and matching approach (Step 2)**

175 *The selection of matching covariates should be informed by the theory of change*

176 A key assumption in non-experimental studies is that selection to the treatment should be
177 independent of potential outcomes (known as the 'conditional ignorability assumption';
178 Rosenbaum & Rubin, 1983). If factors affecting treatment assignment can be ignored, all
179 confounding factors should have been controlled for, and the study should not suffer from
180 hidden bias (i.e. not be very sensitive to potential missing variables). Therefore, matching
181 analyses should ideally include all covariates likely to impact both the selection to the
182 treatment and the outcome of interest (e.g. remoteness, as how remote a piece of land is will
183 affect the likelihood of it being designated as PA and also deforested). Researchers should
184 thus carefully consider which covariates are likely related to the outcome. It is better to err on
185 the side of caution by including a covariate if the researcher is unsure of its likely role as a
186 confounder. However, it is important that no variables likely to have been influenced by the
187 outcome of interest are used as part of the matching process (Stuart 2010), so matching should
188 only include variables pre-dating the intervention or time-invariant variables. Creating a table
189 of all possible confounding factors and how they relate to the selection and outcome variables,

190 can help organize this process (e.g. Schleicher et al. 2017). Running regression analyses prior
191 to matching or plotting the results of a Principal Component Analysis (PCA) can also inform
192 covariate selection. PCA can help visualize how treatment and outcome relate to the selected
193 covariates by showing which combination of covariates explain the outcomes observed in
194 different units of analysis, and whether treatment and outcome show similar patterns (Eklund
195 et al. 2016).

196 *Selection of the matching approach and how it is implement should be carefully considered*

197 There are various matching approaches, all with strengths and weaknesses. It is difficult to
198 assess *a priori* which method is the most appropriate for a given study. Thus, testing a suite of
199 different matching methods to evaluate which produces the best balance (see Step 3 Figure
200 1), instead of relying on any one method, can be useful (e.g. Oldekop et al. 2018). Matching
201 approaches include Mahalanobis, Propensity Score, Genetic and Full Matching (Stuart 2010;
202 lacus et al. 2012; Diamond & Sekhon 2013). Mahalanobis and Propensity Score matching are
203 particularly commonly used in conservation science, and there is growing interest in the use
204 of Genetic matching. Mahalanobis matching calculates how many standard deviations a unit
205 is from the mean of other units (e.g. Rasolofoson et al. 2015). In contrast, Propensity Score
206 matching combines all covariates into a single distance measure that estimates the probability
207 of units receiving the treatment (e.g. Carranza et al. 2013). Genetic matching automates the
208 iteration process (Diamond & Sekhon 2012) by optimizing balance diagnostics, rather than
209 mean standardized distance (e.g. Hanauer & Canavire-Bacarreza 2015). Full matching uses a
210 Propensity Score to match multiple control units to treatment unit and *vice versa*, and is
211 particularly well suited when analyzing balanced datasets with similar number of treatment
212 and control units (e.g. Oldekop et al. 2019). The development and testing of matching

213 approaches remains an active research area with some strongly arguing for one method over
214 another (King & Nielsen 2019).

215

216 Each of these methods can be configured in multiple ways, requiring a series of additional
217 decisions about: (1) *Treatment-Control ratio*: the ratio of treatment to control units used during
218 matching (i.e. whether to use a one-to-one match or to match one treatment unit to several
219 control units), (2) *Replacement*: whether control units can be used multiple times or not (i.e.
220 match with or without replacement), (3) *Weighting*: the relative importance placed on retaining
221 as many treatment units or control units in the analysis as possible (with some approaches
222 applying sampling weights to give more importance to certain units and adjust for unbalanced
223 datasets), (4) *Calipers*: whether to set bounds (called 'calipers') on the degree of difference
224 between treatment and control units, (5) *Order*: the order in which matches are selected (e.g.,
225 at random or in a particular order) (Lunt 2014), and (6) *Exact matching*: whether or not to only
226 retain units with the exact same covariate value. Exact matching using continuous covariates
227 typically results in many treatment units being excluded because no control units with identical
228 values are found. This can increase bias because data is being systematically discarded. It is
229 thus better suited for categorical variables.

230 *Inference can only be made for the region of 'common support'*

231 In some cases, treatments may be so closely interlinked with potential confounders that no
232 good matches exist. For example, if intact habitat remains only on mountain tops and all
233 mountain tops are protected, it would be impossible to separate the contribution of location
234 from that of the intervention itself, as there are no controls with similar habitat available that
235 are not protected (Green et al. 2013). Matching therefore depends on a substantial overlap in

236 relevant covariates between units exposed to the intervention and potential controls. This
237 overlap is known as the region of 'common support'. An assessment of common support early
238 on in the matching process can be a good filter to determine whether matching will be useful.
239 When using the Propensity Score, it is simple to discard potential control units with scores
240 outside the range of the treatment group. Visual diagnostics, including the Propensity Score
241 distribution, are a simple and robust way of diagnosing any challenges with common support
242 (Caliendo & Kopeinig 2008; Lechner 2000; see Figure 1 and Table 2). Where many potential
243 control units need to be discarded, it can be helpful to define the discard rule based on one
244 or two covariates rather than the Propensity Score (Stuart 2010). If many treatment units must
245 be discarded because no appropriate control units can be found, the research question being
246 answered by the analysis is likely to be different from the one that was being asked to begin
247 with. This needs to be acknowledged. In some cases, it will simply not be possible to use
248 matching to evaluate the impact of an intervention on an outcome of interest, requiring the
249 use of alternative quantitative or qualitative methods (e.g. Green et al. 2013).

250

251 **Assessing the quality of the matching (Step 3)**

252 *The quality of the match achieved must be explored and reported*

253 Matching provides no guarantee that biases have been sufficiently addressed. It is therefore
254 important to assess the quality of the match and to report relevant statistics (see Figure 1 and
255 Table 2). In fact, an advantage of using matching rather than standard regression, is that it
256 highlights areas of the covariate distribution where there is not sufficient common support
257 between treatment and control groups to allow effective inference without substantial
258 extrapolation (Gelman and Hill 2007). When assessing the performance and appropriateness

259 of a match, three key features should be assessed and reported: (1) how similar are the
260 treatments and controls after matching (covariate balance), (2) how similar is the pre-match
261 treatment to the post-match treatment (large dissimilarities can potentially increase bias), and
262 (3) the number of treatment units that were matched and discarded during matching. In
263 addition, when matching is done with replacement, it is prudent to check the selection rate of
264 matched controls, to ensure that there is no oversampling of specific controls. The best
265 matching method will be the one that keeps the post-matched treatment as similar to the pre-
266 matched treatment as possible, while ensuring maximum similarity between post-match
267 treatment and control units, and removing the least number of observations in the process.
268 The proportion of covariates that have met a user-specified threshold for balance and the
269 covariate with the highest degree of imbalance, have been shown to be effective indicators in
270 diagnosing imbalance and potential bias (Stuart et al. 2013). Standard tests and visualizations
271 that explore match quality have been widely published in the statistical, economics, health and
272 political science literatures (e.g. Harris & Horst 2016; Rubin 2001). It is useful to combine both
273 numeric and visual diagnostics (see Table 2 for examples) (Caliendo & Kopeinig 2008; Stuart
274 2010; Harris & Horst 2016).

275

276 A central assumption underlying the use of matching approaches is that any difference
277 between treatment and control populations remaining after matching are due to treatment
278 effects alone. Validating this assumption rests on a robust theory of change, and a careful
279 selection of covariates. However, even if all known sources of potential bias have been
280 controlled for, unknown mechanisms might still confound either treatment or outcomes.
281 Checks to assess whether post-matching results are sensitive to potential unmeasured

282 confounders (e.g. Rosenbaum bounds; Rosenbaum 2007), allow one to evaluate the amount
283 of variation that an unmeasured confounder would have to explain to invalidate the results.

284 *The robustness of matching results to spatial autocorrelation should be considered*

285 Conservation interventions, and most data used to assess their impacts, have a spatial
286 component. A key assumption of many statistical tests is that units of observation are
287 independent from each other (e.g. Dormann et al. 2007; Haining 2003). Yet, this assumption is
288 easily violated when using spatial data: units of observation that are closer together in space
289 are often more similar to each other than units of observation that are further apart. Such
290 spatial dependency, referred to as spatial autocorrelation (SAC), is often not discussed or
291 explicitly tested for in conservation matching studies, despite being a well-recognized
292 phenomenon (Legendre 1993; Dormann et al. 2007). While it is unclear how matching affects
293 SAC, SAC can clearly affect impact estimations. For example, studies modeling deforestation
294 have shown that the spatial coordinates of a data point are among the top predictors of
295 deforestation (Green et al. 2013; Schleicher et al. 2017). Some matching studies in the
296 conservation literature have acknowledged the potential resulting bias, and have attempted
297 to account or test for any potential effects linked to the spatial sampling framework (e.g.
298 Carranza et al. 2013; Schleicher et al. 2017; Oldekop et al. 2019). We call for increased attention
299 to SAC when evaluating place-based interventions. Steps to test for SAC could include Moran's
300 I tests, semi-variograms, correlograms, and spatial plots of model residuals (Schleicher et al.
301 2017; Oldekop et al. 2019). These could be used to test for SAC of post-matching analyses and
302 treatment assignment (e.g. by testing SAC of Propensity Score models). SAC could also be
303 tested separately in the treatment and control groups before and after matching. If significant
304 SAC remains after matching, it would be a strong indication that it needs to be accounted for

305 in any post-matching regression, something that could be confirmed through inspection of
306 spatial patterns of model residuals (Dormann et al. 2007; Zuur et al. 2009; Oldekop et al. 2019).

307 **Post-matching analyses**

308 Matching is often used as a data pre-processing step (Ho et al. 2007). If matching perfectly
309 reduces the difference between treatment and control units to zero, or the residual variation
310 is close to random and uncorrelated with treatment allocation and the outcome of interest,
311 then the average treatment effect can be measured as the difference in the outcome between
312 treatment and control units. However, in most instances matching reduces - but does not
313 eliminate - differences between treatment and control units. It is often followed by regression
314 analyses to control for any remaining differences between treatment and control units (Imbens
315 & Wooldridge 2009). Where longitudinal panel data is available, matching can be combined
316 with a difference-in-difference research design (e.g. Jones & Lewis 2015; Table 1). Combining
317 matching with other statistical methods in this way tends to generate treatment effect
318 estimates that are more accurate and robust than when using any one statistical approach
319 alone (Blackman 2013).

320

321 **MOVING FORWARD**

322 The increasing use matching approaches in conservation science has great potential to
323 rigorously inform what works in conservation. However, while matching approaches are a
324 powerful tool that can improve causal inference, they are not a silver bullet. We caution against
325 using matching approaches without a clear understanding of their strengths and weaknesses.
326 Looking to the future, we highlight clear avenues for improving the use of matching in

327 conservation studies. This includes developing robust theories of change, incorporating real
328 world complexities, careful selection of matching variables and approaches, assessing the
329 quality of matches achieved, and accounting for SAC. Conservation impact evaluation would
330 benefit by increased evaluation planning alongside conservation interventions, better
331 integration of qualitative approaches with quantitative matching-based methods, further
332 consideration of how spill-over effects should be accounted for, and more publications of pre-
333 analysis plans. We explore each of these in turn.

334 *Post hoc* evaluations are often necessary in conservation as there is a pressing policy need to
335 explore the impacts of past interventions. However, there are limits to what statistical analyses
336 can do *post hoc* to overcome problems in the underlying study design of an impact evaluation
337 (Ferraro & Hanauer 2014a). More integration of impact evaluations within intervention
338 implementations is needed to address and account for biases in where interventions are
339 located. Occasionally, this may provide the opportunity for experimental evaluation (Pynegar
340 et al. 2018; Wiik et al. 2019). More commonly, where this is not possible or desirable, good
341 practice should be to explore and consider potential controls using matching from as early as
342 possible. Innovative funding is needed to allow researchers to work alongside conservation
343 practitioners throughout their intervention to incorporate rigorous impact evaluation from the
344 start (Craigie et al. 2015).

345 Matching does not provide certainty about causal links, and on its own does not likely provide
346 insights into the mechanism by which an intervention had an impact. This highlights the
347 importance of making use of the diverse set of evaluation approaches and data sources
348 available. This includes the important, but often overlooked, contribution that qualitative data
349 can make to impact evaluation and counterfactual thinking. For example, incorporating

350 qualitative data can provide depth in understanding, identify hypotheses, and help clarify
351 potential reasons why an effect of an intervention was or was not found. Process tracing, realist
352 evaluation, assessment of exceptional responders and contribution analyses are all suited for
353 exploring the mechanisms by which an intervention led to an outcome (Collier 2011; Lemire
354 et al. 2012; Westthorp 2014; Meyfroidt 2016; Post & Geldmann 2018). Qualitative Comparative
355 Analysis can also be useful for exploring what factors needed to be present to achieve
356 successful outcomes, or how impacts vary among different groups and circumstances
357 (Korhonen-Kurki et al. 2014).

358 There are remarkably few explicit assessments of the importance of spill-over effects beyond
359 intervention boundaries at different spatial scales (Pfaff & Robalino 2017). While impact
360 evaluations on deforestation rates commonly avoid selecting control pixels from a pre-defined
361 buffer area around an intervention, the size of the buffer are seldom based on a clear
362 justification. We know of no matching studies that explicitly account for spill-over effects over
363 larger spatial scales. This is despite the need to account for spill-overs to assess whether a net
364 reduction in conservation pressure has taken place, instead of simply displacing it elsewhere
365 (Pfaff & Robalino 2012). For example, stronger implementation of logging rules in one region
366 of Brazil shifted pressures to other regions (Dou et al. 2018) and China's national logging bans
367 mean that timber demand is being met through imports from Indonesia (Lambin & Meyfroidt
368 2011). Accounting for these effects is inherently complex as many factors complicate the ability
369 to account for effects over large spatial scales, including demand and supply dynamics,
370 feedback cycles, and behavioral adaptation (Ferraro et al. 2019) – and will require further
371 collective, interdisciplinary thinking and methodological developments.

372 Increasingly, there is a push for researchers in a number of fields to publish pre-analyses plans
373 (e.g. Nosek et al. 2018), which lay out hypotheses identified *a priori*, and proposed analyses
374 before the effects are assessed (Bauhoff & Busch 2018). The aim of pre-analyses plans is to
375 reduce the risk of HARKing (Hypothesising After Results are Known; Kerr 1998). As there are
376 many potential acceptable ways to select appropriate matches, there are benefits in publishing
377 the matching and planned analysis before carrying it out.

378 Given continuous loss of biodiversity despite considerable conservation efforts, there is an
379 urgent need to take impact evaluations more seriously, learn from other disciplines, and
380 improve our practices as a conservation science community. The increasing interest in the use
381 of counterfactual approaches for evaluating conservation impacts is therefore a very positive
382 development. There is an important role for conservation practitioners, funders and academics
383 to encourage this development and to mainstream rigorous impact evaluations into
384 conservation practice. Furthermore, there is certainly a need to increase the capacity of
385 conservation scientists and practitioners in both the conceptual and technical challenges of
386 impact evaluation, including by incorporating impact evaluation and counterfactual thinking
387 in postgraduate training of future conservationists. We hope that this paper will help both
388 improve the general quality of evaluations being undertaken, and direct future research to
389 continue to improve the approaches currently on offer.

390 **LITERATURE CITED**

- 391 Abadie A, Cattaneo MD. 2018. Econometric methods for program evaluation. *Annual Review*
392 *of Economics* **10**:465–503.
- 393 Alix-Garcia JM, Sims KRE, Orozco-olvera VH, Costica LE. 2018. Payments for environmental
394 services supported social capital while increasing land management. *Proceedings of the*
395 *National Academy of Sciences of the United States of America* **115**:7016–7021.
- 396 Andam KS, Ferraro PJ, Pfaff A, Sanchez-Azofeifa GA, Robalino JA. 2008. Measuring the
397 effectiveness of protected area networks in reducing deforestation. *Proceedings of the*
398 *National Academy of Sciences of the United States of America* **105**:16089–94. Available
399 from
400 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567237&tool=pmcentrez&](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567237&tool=pmcentrez&rendertype=abstract)
401 [rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567237&tool=pmcentrez&rendertype=abstract).
- 402 Bauhoff S, Busch J. 2018. Does Deforestation Increase Malaria Prevalence? Evidence from
403 Satellite Data and Health Surveys. 480, Center for Global Development Working Paper.
- 404 Baylis K, Honey-Rosés J, Börner J, Corbera E, Ezzine-de-Blas D, Ferraro PJ, Lapeyre R, Persson
405 UM, Pfaff A, Wunder S. 2016. Mainstreaming Impact Evaluation in Nature Conservation.
406 *Conservation Letters* **9**:58–64.
- 407 Blackman A. 2013. Evaluating Forest Conservation Policies in Developing Countries Using
408 Remote Sensing Data: An Introduction and Practical Guide. *Forest Policy and Economics*
409 **34**:1–16.
- 410 Caliendo M, Kopeinig S. 2008. Some practical guidance for the implementation of propensity
411 score matching. *Journal of Economic Surveys* **22**:31–72.
- 412 Carranza T, Balmford A, Kapos V, Manica A. 2013. Protected Area Effectiveness in Reducing
413 Conversion in a Rapidly Vanishing Ecosystem: The Brazilian Cerrado. *Conservation Letters*
414 **7**:216–223. Available from <http://doi.wiley.com/10.1111/conl.12049>.
- 415 Clements T, Suon S, Wilkie DS, Milner-Gulland EJ. 2014. Impacts of Protected Areas on Local
416 Livelihoods in Cambodia. *World Development* **64**:S12–S134. Elsevier Ltd. Available from
417 <http://dx.doi.org/10.1016/j.worlddev.2014.03.008>.
- 418 Collier. 2011. Understanding Process Tracing. *Political Science and Politics* **44**:823–30.
- 419 Craigie ID, Barnes MD, Geldmann J, Woodley S. 2015. International funding agencies: potential
420 leaders of impact evaluation in protected areas? *Philosophical Transactions of the Royal*
421 *Society B: Biological Sciences* **370**:20140283.
- 422 den Braber B, Evans KL, Oldekop JA. 2018. Impact of protected areas on poverty , extreme
423 poverty, and inequality in Nepal. *Conservation Letters*:e12576.
- 424 Diamond A, Sekhon JS. 2012. Genetic Matching for Estimating Causal Effects: A General
425 Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of*
426 *Economics and Statistics*.

- 427 Diamond A, Sekhon JS. 2013. Genetic Matching for Estimating Causal Effects: A General
428 Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of*
429 *Economics and Statistics* **95**:932–945.
- 430 Dormann CF et al. 2007. Methods to account for spatial autocorrelation in the analysis of
431 species distributional data: a review. *Ecography* **30**:609–628.
- 432 Dou Y, da Silva RFB, Yang H, Jianguo L. 2018. Spillover effect offsets the conservation effort in
433 the Amazon. *Journal of Geographical Science* **28**:1715–1732.
- 434 Eklund J, Blanchet FG, Nyman J, Rocha R, Virtanen T, Cabeza M. 2016. Contrasting spatial and
435 temporal trends of protected area effectiveness in mitigating deforestation in
436 Madagascar. *Biological Conservation* **203**:290–297.
- 437 Ewers RM, Rodrigues ASL. 2008. Estimates of reserve effectiveness are confounded by leakage.
438 *Trends in Ecology and Evolution* **23**:113–116.
- 439 Ferraro PJ, Hanauer MM. 2014a. Advances in Measuring the Environmental and Social Impacts
440 of Environmental Programs. *Annual Review of Environment and Resources* **39**:495–517.
441 Available from [http://www.annualreviews.org/doi/abs/10.1146/annurev-environ-](http://www.annualreviews.org/doi/abs/10.1146/annurev-environ-101813-013230)
442 [101813-013230](http://www.annualreviews.org/doi/abs/10.1146/annurev-environ-101813-013230).
- 443 Ferraro PJ, Hanauer MM. 2014b. Quantifying causal mechanisms to determine how protected
444 areas affect poverty through changes in ecosystem services and infrastructure.
445 *Proceedings of the National Academy of Sciences* **111**:4332–4337.
- 446 Ferraro PJ, Pattanayak SK. 2006. Money for nothing? A call for empirical evaluation of
447 biodiversity conservation investments. *PLoS biology* **4**:e105. Available from
448 <http://www.ncbi.nlm.nih.gov/pubmed/16602825>.
- 449 Ferraro PJ, Sanchirico JN, Smith MD. 2019. Causal inference in coupled human and natural
450 systems **116**:5311–5318.
- 451 Gaveau DLA, Curran LM, Paoli GD, Carlson KM, Wells P, Besse-Rimba A, Ratnasari D, Leader-
452 Williams N. 2012. Examining protected area effectiveness in Sumatra: importance of
453 regulations governing unprotected lands. *Conservation Letters* **5**:142–148.
- 454 Gelman A, Hill J. 2007. *Data analysis using regression and multilevel/hierarchical models*.
455 Cambridge University Press, Cambridge, UK.
- 456 Green JMH, Larrosa C, Burgess ND, Balmford A, Johnston A, Mbilinyi BP, Platts PJ, Coad L. 2013.
457 Deforestation in an African biodiversity hotspot: Extent, variation and the effectiveness of
458 protected areas. *Biological Conservation* **164**:62–72. Available from
459 <http://dx.doi.org/10.1016/j.biocon.2013.04.016>.
- 460 Haining RP. 2003. *Spatial data analysis: Theory and Practice*. Cambridge University Press,
461 Cambridge, UK.
- 462 Hanauer MM, Canavire-Bacarreza G. 2015. Implications of heterogeneous impacts of protected
463 areas on deforestation and poverty. *Philosophical Transactions of the Royal Society B:*
464 *Biological Sciences* **370**:20140272.

- 465 Harris, Horst. 2016. A Brief Guide to Decisions at Each Step of the Propensity Score Matching
466 Process. *Practical Assessment, Research & Evaluation* **21**. Available from
467 <https://pareonline.net/getvn.asp?v=21&n=4>.
- 468 Ho DE, Imai K, King G, Stuart EA. 2007. Matching as nonparametric preprocessing for reducing
469 model dependence in parametric causal inference. *Political Analysis* **15**:199–236.
- 470 Iacus S, King G, Porro G. 2012. Causal Inference without Balance Checking: Coarsened Exact
471 Matching. *Political Analysis* **20**:1–24.
- 472 Imbens GW, Wooldridge JM. 2009. Recent Developments in the Econometrics of Program
473 Evaluation. *Journal of Economic Literature* **47**:5–86.
- 474 Jones KW, Lewis DJ. 2015. Estimating the counterfactual impact of conservation programs on
475 land cover outcomes: The role of matching and panel regression techniques. *PLoS ONE*
476 **10**:e0141380.
- 477 Joppa LN, Pfaff A. 2009. High and far: biases in the location of protected areas. *PLoS ONE*
478 **4**:e8273.
- 479 Kerr NL. 1998. HARKing: Hypothesizing After the Results are Known. *Personality and Social*
480 *Psychology Review* **2**:196–217.
- 481 King G, Nielsen R. 2019. Why Propensity Scores Should Not Be Used for Matching. *Political*
482 *Analysis*:1–20. Available from <http://j.mp/2ovYGsW>.
- 483 Korhonen-Kurki K, Sehring J, Brockhaus M, Di M, Sehring J, Brockhaus M, Di M. 2014. Enabling
484 factors for establishing REDD+ in a context of weak governance weak governance.
485 *Climate Policy* **14**:1–20.
- 486 Lambin EF, Meyfroidt P. 2011. Global land use change , economic globalization , and the
487 looming land scarcity. *Proceedings of the National Academy of Sciences of the United*
488 *States of America* **108**:3465–3472.
- 489 Lechner M. 2000. A Note on the Common Support Problem in Applied Evaluation Studies.
490 2001–01, Univ. of St. Gallen Economics Discussion Paper.
- 491 Legendre P. 1993. Spatial Autocorrelation: Trouble or New Paradigm? *Ecology* **74**:1659–1673.
- 492 Lemire ST, Nielsen SB, Dybdal L. 2012. Making contribution analysis work: A practical
493 framework for handling influencing factors and alternative explanations. *Evaluation*
494 **18**:294 –309.
- 495 Liscow ZD. 2013. Do property rights promote investment but cause deforestation? Quasi-
496 experimental evidence from Nicaragua. *Journal of Environmental Economics and*
497 *Management* **65**:241–261.
- 498 Lunt M. 2014. Selecting an appropriate caliper can be essential for achieving good balance
499 with propensity score matching. *American Journal of Epidemiology* **179**:226–235.
- 500 Meyfroidt P. 2016. Approaches and terminology for causal analysis in land systems science.
501 *Journal of Land Use Science* **11**:501–522.

- 502 Naidoo R et al. 2019. Evaluating the impacts of protected areas on human well-being across
503 the developing world. *Science Advances* **5**:eaav3006.
- 504 Nelson A, Chomitz KM. 2011. Effectiveness of strict vs. multiple use protected areas in reducing
505 tropical forest fires: a global analysis using matching methods. *PLoS ONE* **6**:e22722.
- 506 Nosek BA, Ebersole CR, Dehaven AC, Mellor DT. 2018. The preregistration revolution.
507 *Proceedings of the National Academy of Sciences of the United States of America*
508 **2017**:2600–2606.
- 509 Oldekop J, Sims K, Karna B, Whittingham M, Agrawal A. 2019. Reductions in deforestation and
510 poverty from decentralized forest management in Nepal. *Nature Sustainability*:in press.
- 511 Oldekop JA, Sims KRE, Karna B, Whittingham MJ, Agrawal A. 2018. An upside to globalization:
512 International migration drives reforestation in Nepal. *Global Environmental Change*
513 **52**:66–74.
- 514 Pfaff A, Robalino J. 2012. Protecting forests, biodiversity, and the climate: predicting policy
515 impact to improve policy choice. *Oxford Review of Economic Policy* **28**:164–179.
- 516 Pfaff A, Robalino J. 2017. Spillovers from Conservation Programs. *Annual Review of Resource*
517 *Economics* **9**:299–315.
- 518 Post G, Geldmann J. 2018. Exceptional responders in conservation. *Conservation Biology*
519 **32**:576–583.
- 520 Pullin AS, Knight TM. 2001. Effectiveness in Conservation Practice: Pointers from Medicine and
521 Public Health. *Conservation Biology* **15**:50–54. Available from
522 <http://dx.doi.org/10.1111/j.1523-1739.2001.99499.x>.
- 523 Pynegar EL, Jones JPG, Gibbons JM, Asquith NM. 2018. The effectiveness of Payments for
524 Ecosystem Services at delivering improvements in water quality: lessons for experiments
525 at the landscape scale. *PeerJ* **6**:e5753.
- 526 Qiu J et al. 2018. Evidence-based causal chains for linking health, development and
527 conservation actions. *BioScience* **68**:182–193.
- 528 Rasolofoson R a., Ferraro PJ, Jenkins CN, Jones JPG. 2015. Effectiveness of Community Forest
529 Management at reducing deforestation in Madagascar. *Biological Conservation* **184**:271–
530 277.
- 531 Robalino J, Pfaff A, Villalobos L. 2017. Heterogeneous Local Spillovers from Protected Areas in
532 Costa Rica. *Journal of the Association of Environmental and Resource Economists* **4**:795–
533 820.
- 534 Rosenbaum PR. 2007. Sensitivity Analysis for m-Estimates, Tests, and Confidence Intervals in
535 Matched Observational Studies. *Biometrics* **63**:456–464.
- 536 Rosenbaum PR, Rubin DB. 1983. The Central Role of the Propensity Score in Observational
537 Studies for Causal Effects. *Biometrika* **70**:41–55.
- 538 Rosenbaum PR, Silber JH. 2009. Amplification of Sensitivity Analysis in Matched Observational

- 539 Studies. *American Statistical Analysis* **104**:1398–1405.
- 540 Rubin A. 2007. Improving the teaching of evidence-based practice: Introduction to the Special
541 Issue. *Research on Social Work Practice* **17**:541–547.
- 542 Rubin D. 1980. Bias reduction using Mahalanobis metric matching. *Biometrics* **36**:293–298.
- 543 Rubin DB. 2001. Using propensity scores to help design observational studies: Application to
544 the tobacco litigation. *Health Services & Outcomes Research Methodology* **2**:169–188.
- 545 Schleicher J. 2018. The environmental and social impacts of protected areas and conservation
546 concessions in South America. *Current Opinion in Environmental Sustainability* **32**:1–8.
- 547 Schleicher J, Peres CA, Amano T, Llactayo W, Leader-Williams N. 2017. Conservation
548 performance of different conservation governance regimes in the Peruvian Amazon.
549 *Scientific Reports* **7**:11318.
- 550 Sills EO et al. 2015. Estimating the Impacts of Local Policy Innovation: The Synthetic Control
551 Method Applied to Tropical Deforestation. *PLoS ONEoS ONE* **10**:e0132590.
- 552 Stuart EA. 2010. Matching methods for causal inference: A review and a look forward. *Statistical*
553 *Science* **25**:1–21.
- 554 Stuart EA, Lee BK, Leacy FP. 2013. Prognostic score-based balance measures can be a useful
555 diagnostic for propensity score methods in comparative effectiveness research. *Journal*
556 *of Clinical Epidemiology* **66**:S84.
- 557 Sutherland WJ, Pullin AS, Dolman PM, Knight TM. 2004. The need for evidence-based
558 conservation. *Trends in ecology & evolution* **19**:305–8.
- 559 Westhorp. 2014. *Realist Impact Evaluation: an introduction*. London, UK.
- 560 Wiik E, D'Annunzio R, Pynegar E, Crespo D, Asquith N, Jones JPG. 2019. Experimental evaluation
561 of the impact of a payment for environmental services program on deforestation.
562 *Conservation Science and Practice*.
- 563 Zuur A, Saveliev AA, Ieno EN, Smith GM, Walker N. 2009. *Mixed Effects Models and Extensions*
564 *in Ecology with R*. Springer Verlag, New York.
- 565

TABLES AND FIGURES:

Table 1. Commonly used non-experimental, quantitative impact evaluation approaches with the pros and cons of their use in environmental management or conservation.

Method	When can it be used?	Pros	Cons
Matching*	When baseline information on confounding factors (those affecting both selection to the treatment and outcomes) are available for both treatment and control units (e.g. Andam et al. 2008).	Relatively low data requirements and lends itself to integration with other approaches when used as a data pre-processing step.	Assumes balance in observable covariates reflects balance in unobserved covariates, i.e. that there are no unobserved confounders.
Before-After-Control-Impact (Difference-in-Difference)	When data before and after treatment implementation can be collected from replicated treatment and 'control' units (e.g. Pynegar et al. 2018).	Controls for time invariant variables and for variables that change over time but affect both treatment and control groups equally.	Assumes a parallel trend in outcome between treatment and controls (confounding factors in this case are those affecting treatment assignment and changes in outcome over time).
Regression discontinuity	When selection to the intervention follows a sharp assignment rule (e.g., participants above a certain threshold are selected into the treatment; Alix-Garcia et al. 2018).	Strong causal inference.	Outcomes can only be calculated for units close to the cut-off (i.e. data from only a small sub-group of units are used).
Instrumental Variables	When treatment assignment is correlated with the error term (endogeneity), a third variable (the instrument) that is correlated with treatment but uncorrelated with the error term can be used instead of the treatment (e.g. Liscow 2013).	Helps to overcome endogeneity.	Suitable instruments can be hard to find.
Synthetic Control	When the intervention has only occurred in a single unit of observation information from a potential pool of controls can be synthesised to generate a single artificial counterfactual (e.g. Sills et al. 2015).	Can be conducted when large numbers of treatment units are not available.	Credibility relies on a good pre-implementation fit for the outcome of interest between treated unit and synthetic control.

* Matching can be used to identify control units for comparison with treatment units as a method for impact evaluation, but is often used to improve the rigor of other approaches. For example, matching can be used to select 'control' units for difference-in-differences analysis.

Table 2. Example diagnostics for the checks (suggested in Figure 1) part of a matching analysis to assess the quality of the matching and robustness of the post-matching analysis.

Check	Example diagnostic	Explanation and purpose	Example visualizations
Check 1: Balance	Mean values and standardized mean differences before and after matching	Test whether differences among treatment and control populations are meaningful. Compare covariate means and deviations for treatment and control units (before and after matching) to assess whether a matching has improved balance (similarity between treatment and control units). After matching mean covariate values should be similar and the standardized mean difference should ideally be close to zero. Standardized mean values of <0.25 are often deemed acceptable, but thresholds of 0.1 are more effective at reducing bias (Stuart 2010; Stuart et al. 2013).	Love plots and propensity score distributions before and after matching (e.g. Figure 1, Oldekop et al. 2019)
Check 2: Spatial autocorrelation	Moran's I and spatial distribution of post-matching analysis residuals	Moran's I values of the post-matching analysis should not be significantly different from zero to demonstrate low levels of spatial autocorrelation. Plotting the spatial distribution of post-matching analysis residuals can help visualize whether there is a spatial pattern to the error term.	Correlograms, semi-variograms and bubble plots (Figure 1, Oldekop et al. 2019)
Check 3: Hidden Bias	Rosenbaum bounds	Assess sensitivity of post-matching estimate to presence of an unobserved confounder. Rosenbaum bounds help to determine how much an unobserved covariate would have to affect selection into the treatment to invalidate the post-matching result (Rosenbaum 2007).	Amplification Plots (Rosenbaum & Silber 2009)

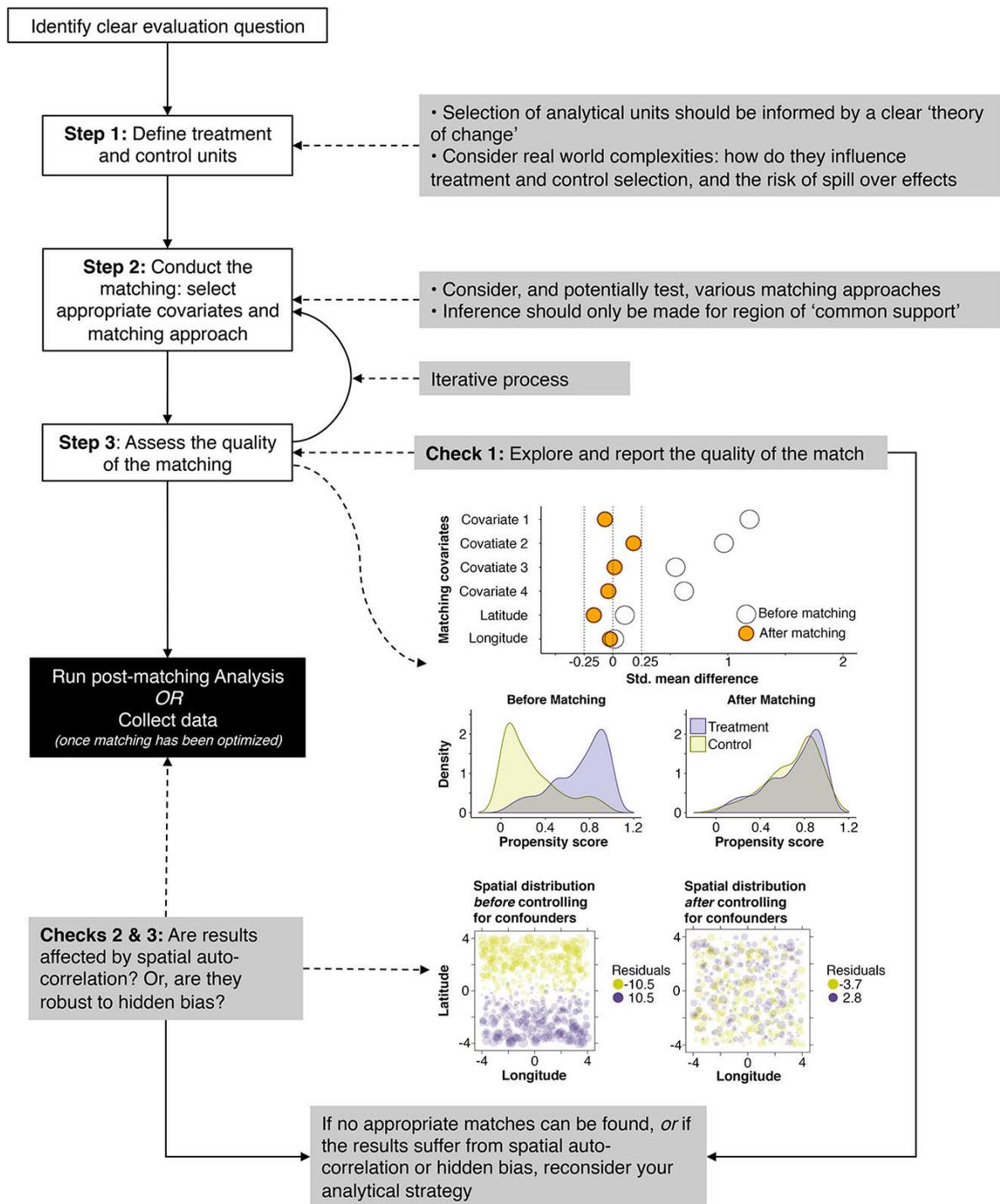


Figure 1. Visual representation of the suggested workflow, including key steps of a matching analysis, potential checks (see Table 2) and visual diagnostics of the matching process.

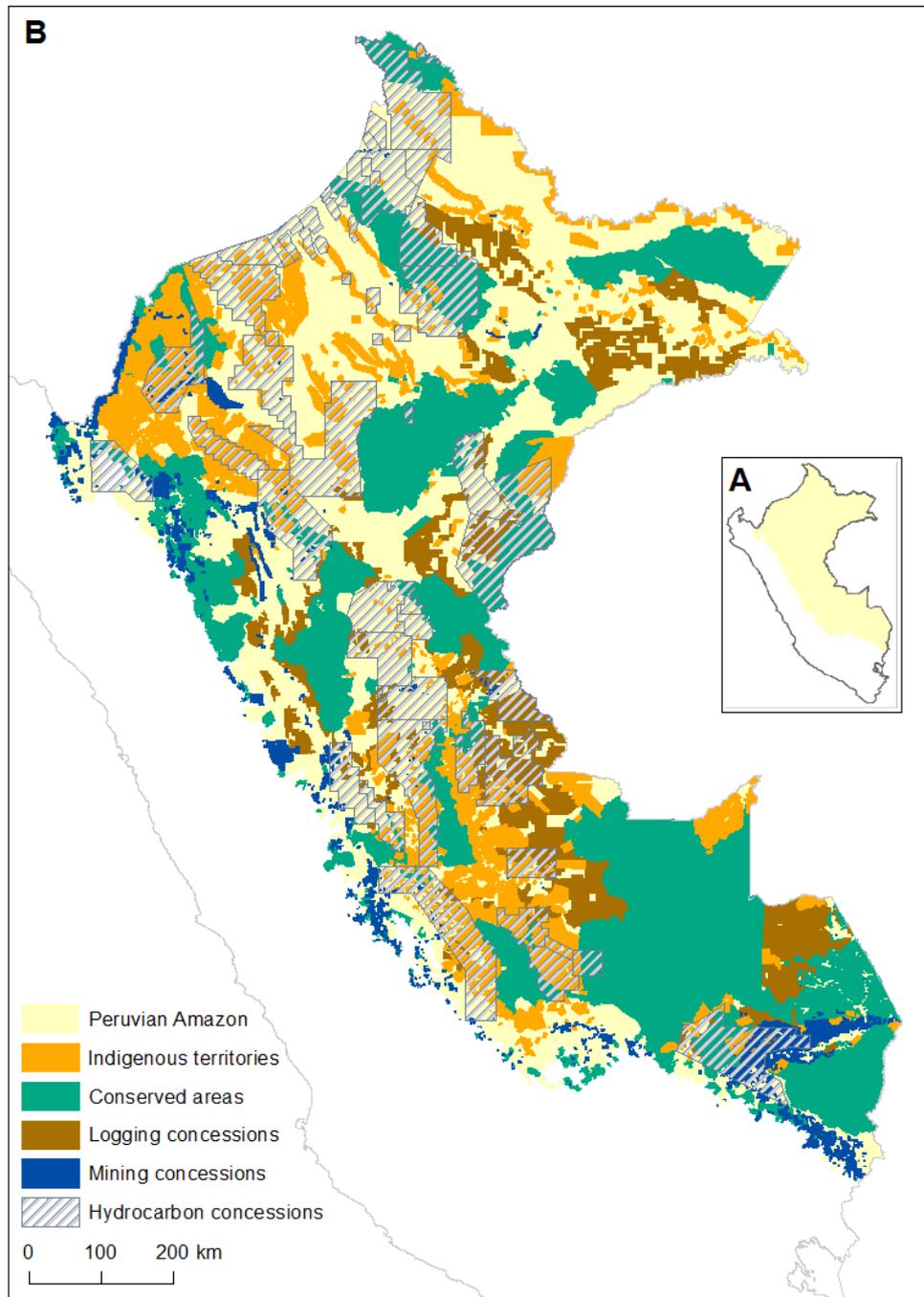


Figure 2. Map of (A) Peru and (B) the Peruvian Amazon with the main land use designations in 2011 to 2013. Conserved areas include government protected areas (PAs), conservation concessions, ecotourism concessions, concessions of non-timber forest products and territorial reserves. In an analysis of the impacts of PAs, Indigenous Territories and conservation concessions on deforestation rates, the decision of what to consider as appropriate control areas from which to select control pixels is far from straight forward given the multiple, and in part overlapping, land use designations (Schleicher et al. 2017).

Figure Legends:

Figure 1. Visual representation of the suggested workflow, including key steps of a matching analysis, potential checks (see Table 2) and visual diagnostics of the matching process.

Figure 2. Map of (A) Peru and (B) the Peruvian Amazon with the main land use designations in 2011 to 2013. Conserved areas include government protected areas (PAs), conservation concessions, ecotourism concessions, concessions of non-timber forest products and territorial reserves. In an analysis of the impacts of PAs, Indigenous Territories and conservation concessions on deforestation rates, the decision of what to consider as appropriate control areas from which to select control pixels is far from straight forward given the multiple, and in part overlapping, land use designations (Schleicher et al. 2017).