

1 **HistMapR: Rapid digitization of historical land-use maps in R**

2

3 Alistair G. Auffret[1,2]*, Adam Kimberley[1], Jan Plue[1], Helle Skånes[1], Simon Jakobsson[1],

4 Emelie Waldén[1] Marika Wennbom[1], Heather Wood[1], James M. Bullock[3], Sara A. O.

5 Cousins[1], Mira Gartz[1], Danny A.P. Hooftman[3,4], Louise Tränk[1]

6

7 [1] Biogeography and Geomatics, Department of Physical Geography, Stockholm University, 10691

8 Stockholm, Sweden

9

10 [2] Department of Biology, University of York, York, YO10 5DD, UK

11

12 [3] NERC Centre for Ecology & Hydrology, Benson Lane, Wallingford, Oxfordshire OX10 8BB,

13 UK.

14

15 [4] Lactuca: Environmental Data Analyses and Modelling, 1112 NC, Diemen. The Netherlands.

16

17 *Corresponding author. alistair.auffret@natgeo.su.se

18

19

20

21

22

23

24

25

26 **Abstract**

27 **1.** Habitat destruction and degradation represent serious threats to biodiversity, and quantification of
28 land-use change over time is important for understanding the consequences of these changes to
29 organisms and ecosystem service provision.

30 **2.** Comparing land use between maps from different time periods allows estimation of the
31 magnitude of habitat change in an area. However, digitizing historical maps manually is time-
32 consuming and analyses of change are usually carried out at small spatial extents or at low
33 resolutions.

34 **3.** We developed a method to semi-automatically digitize historical land-use maps using the R
35 environment. We created a number of functions that use the existing *raster* package to classify land
36 use according to a map's colours, as defined by the RGB channels of the raster image. The method
37 was tested on three different types of historical land-use map and results were compared to manual
38 digitisations.

39 **4.** Our method is fast, and agreement with manually-digitised maps of around 80-92% meets
40 common targets for image classification. We hope that the ability to quickly classify large areas of
41 historical land-use will promote the inclusion of land-use change into analyses of biodiversity,
42 species distributions and ecosystem services.

43

44 **Keywords**

45 Biodiversity, Habitat destruction, Fragmentation, Historical Ecology, Landscape Ecology, Land-use
46 change, Mapping, Species Distribution Modelling

47

48

49

50

51 **Introduction**

52 Historical land-use maps represent valuable sources of information in ecology. In addition to the
53 estimation of land-use change over time (Skånes & Bunce 1997; Swetnam 2007), historical map
54 data are commonly coupled with species observations to relate land-use change to changes in
55 biodiversity (Saar *et al.* 2012; Cousins *et al.* 2015; Hooftman, Edwards & Bullock 2016) and
56 ecosystem services over time (Jiang, Bullock & Hooftman 2013; Willcock *et al.* 2016).

57
58 At present, most studies involving the analysis of historical land-use are carried out at landscape
59 scales (Swetnam 2007; Cousins 2009; Saar *et al.* 2012), while analyses at larger spatial scales are
60 uncommon (Hooftman & Bullock 2012; Cousins *et al.* 2015; Willcock *et al.* 2016). This is because
61 digitization of historical land-use maps most commonly involves the time-consuming manual
62 delineation of different land-cover types on scanned, georeferenced historical maps using a desktop
63 GIS or illustration program. As a result, historical land-use (change) rarely features in analyses of
64 biodiversity and species distributions following environmental change at large spatial scales (Hill *et*
65 *al.* 2002; Powney *et al.* 2014), despite the acknowledgment of land-use change as the principal
66 determinant of biodiversity loss worldwide (Newbold *et al.* 2016)

67
68 The *HistMapR* package contains a set of functions that allow a fast and accurate digitization of
69 historical land-use maps in R (R Development Core Team 2015). Map colours are defined by the
70 combination of values (0-255) of the RGB (Red, Green and Blue) channels of a raster image.
71 Calling functions from the *raster* package (Hijmans 2016), our method uses RGB values to classify
72 land use according to user-defined colours. We describe the method, before demonstrating it using
73 three historical map series and comparing outputs to manual digitizations.

74

75

76 **Materials and methods**

77 *Classification method*

78 Step 1. Image smoothing (function: *smooth_map*)

79 Scanned historical paper maps contain inconsistencies in colour due to variations in map
80 production, age and the quality of scanning. The first function applies a Gaussian smoothing to the
81 input raster, calling the *focal* function from the *raster* package. Each pixel in each RGB channel is
82 assigned the mean value from a user-defined window of n pixels surrounding the target pixel. In
83 addition, RGB values below a user-defined threshold can be removed. This allows small patches of
84 dark colour, for example denoting place names and property boundaries to be ‘smoothed over’ so
85 that they do not interfere with the land-use classification. Finally, the smoothed raster is cut to the
86 dimensions of the input raster to remove the halo effect, which occurs where the smoothing process
87 spreads pixel values into any non-image areas of the raster.

89 Step 2. Assign user-defined colours (function: *click_sample*)

90 This function requires the user to define the colours for each land-use category from the smoothed
91 map, calling the *raster* package's *click* function. Clicking a number of times within each category
92 from across the image ensures that the full range of colour tones is sampled. A colour table
93 containing maximum, median and minimum RGB values for each category is produced, and the
94 associated colours can be plotted for inspection with the extra function *plot_colour_table*, calling
95 functions from the packages *gridExtra* (Auguie 2016) and *ggplot2* (Wickham 2009).

97 Step 3. Test classification and write to raster file (function: *class_map*)

98 Each pixel in the smoothed raster is then assigned to a land-use category according to the colour
99 table produced in the previous step (so-called parallelepiped classification). Categories are assigned
100 from the first row of the colour table and down, meaning that if a pixel contains RGB values falling

101 within the range of several categories, it is to the nethermost category in the table that the pixel is
 102 assigned in the final classification. Rearranging the colour table allows the user to choose which
 103 categories should take precedence over others in the case of overlap. Additionally, the range of
 104 RGB values in each category can be expanded by a chosen number of standard errors to account for
 105 the likelihood that the most extreme RGB values for each category were not clicked in the previous
 106 step. The effects of various standard errors can be visually inspected using *plot_colour_table*.
 107 Finally, the RGB values of some pixels are likely to fall outside all categories in the colour table.
 108 These exceptions can be assigned to an existing category or left unclassified. The effects of
 109 rearranging the colour table, assigning error values and exception categories can be assessed by
 110 plotting within the R environment and by writing to a raster file for examination in a GIS program.

111

112 Additional steps

113 Different maps from the same series may require different colour table arrangements to achieve
 114 optimal results. In such cases maps must be reclassified so that raster categories match among maps
 115 prior to analysis and joining maps together to cover larger areas. Additionally, in two of the three
 116 historical map series below, surface water was not denoted in a way which meant that they could be
 117 adequately classified as a separate land-use category using their RGB values. In these cases we used
 118 the function *gdal_rasterize* in the package *gdalUtils* (Greenberg & Mattiuzzi 2015) to burn a
 119 modern water vector layer onto the digitized raster. The *HistMapR* package and documentation are
 120 hosted at <https://github.com/AGAuffret/HistMapR/>. Detailed example scripts and input maps are
 121 available on Figshare (Auffret *et al.* 2017).

122

123 *Case study examples*

124 Dorset, UK - The Land Utilisation Survey of Great Britain (1930s)

125 The UK Land Utilisation Survey was led by Stamp (1931). Sheet 140 over Weymouth and

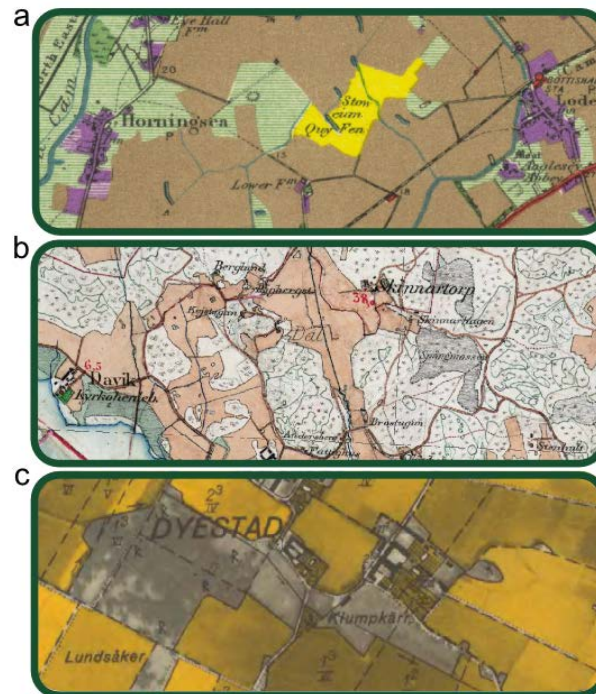
Dorchester covers was 800 km² and was mapped at the 1:63 360 scale, depicting the following land-use types: [1] forest and woodland, [2] arable land, [3] meadow and permanent grass, [4] heath and moorland, [5] gardens, orchards and allotments, [6] urban and industrial areas, [7] inland water (**Figure 1a**). We used *HistMapR* to classify land use on this map into a raster file containing these seven categories at a resolution of 8 m². Total computing time was around 30 minutes on a standard computer, excluding time spent assigning colours with *click_sample* and testing. We then compared our digitized map to Hooftman & Bullock (2012), who manually-digitized maps over the county of Dorset. To be able to compare outputs we first reclassified land-use categories, merging [5] & [6] to match the manual digitization. The manual digitization was rasterized using *gdal_rasterize*, and both digitizations were aggregated by a factor of five to try to reduce the effect of differences in georeferencing, before being masked by each other using *raster's mask* function to ensure that they had the same extent. Total agreement between the two digitizations was calculated by identifying the fraction of corresponding pixels that were classified into the same category. We also calculated the fraction of pixels assigned to each map category in the manually-digitized map the fraction of pixels that were categorized as each category in the HistMapR digitization. Finally, the total fraction assigned to each category in each digitization, and the root-mean-square deviation (RMSD) of cover between digitizations was calculated.

143

144 Södermanland, Sweden - District Economic map of Sweden (1859-1934)

145 This map series (AKA The Hundred map; Swedish: *Häradsekonomska kartan*) describes major
146 land use, settlements and infrastructure (**Figure 1b**). We digitized 11 maps in the county of
147 Södermanland (scale 1:20 000) that were manually digitized by Cousins *et al.* (2015). Each map
148 covers approximately 105 km², and the manual digitization classified land use into nine categories.
149 We classified land use into the general categories of forest, arable land, meadow/dwelling and water
150 (using a modern layer, see above) at a 4 m² resolution. Computing time was approximately 15

151 minutes per map. Comparison of the *HistMapR* and the manual digitization was performed as above
152 for each map sheet, with RMSD also calculated for each category individually.



153
154 **Figure 1.**
155 *Illustrative examples from (a) the Land Utilisation Survey of Great Britain, (b) the Swedish District*
156 *Economic map and (c) the Swedish Economic map. Map images copyright (a) Audrey N. Clark, (b-*
157 *c) the Swedish Agency Lantmäteriet.*

159 Southern Sweden - The Economic map of Sweden (1935-1978)

160 The Economic map series (*Ekonomiska kartan*) was successor to the District Economic Map,
161 published 1935-1978 and covering the whole of Sweden. In southern Sweden, each sheet covers 25
162 km² at the 1:10 000 scale. The maps consist of a monochrome aerial orthomosaic, with arable land,
163 gardens and pasture on former arable fields coloured yellow, and additional information such as
164 roads, larger buildings and boundaries in black (**Figure 1c**). We classified 7069 maps from the 15
165 southernmost counties in Sweden, corresponding to an area of 176 725 km², at a 1 m² resolution.

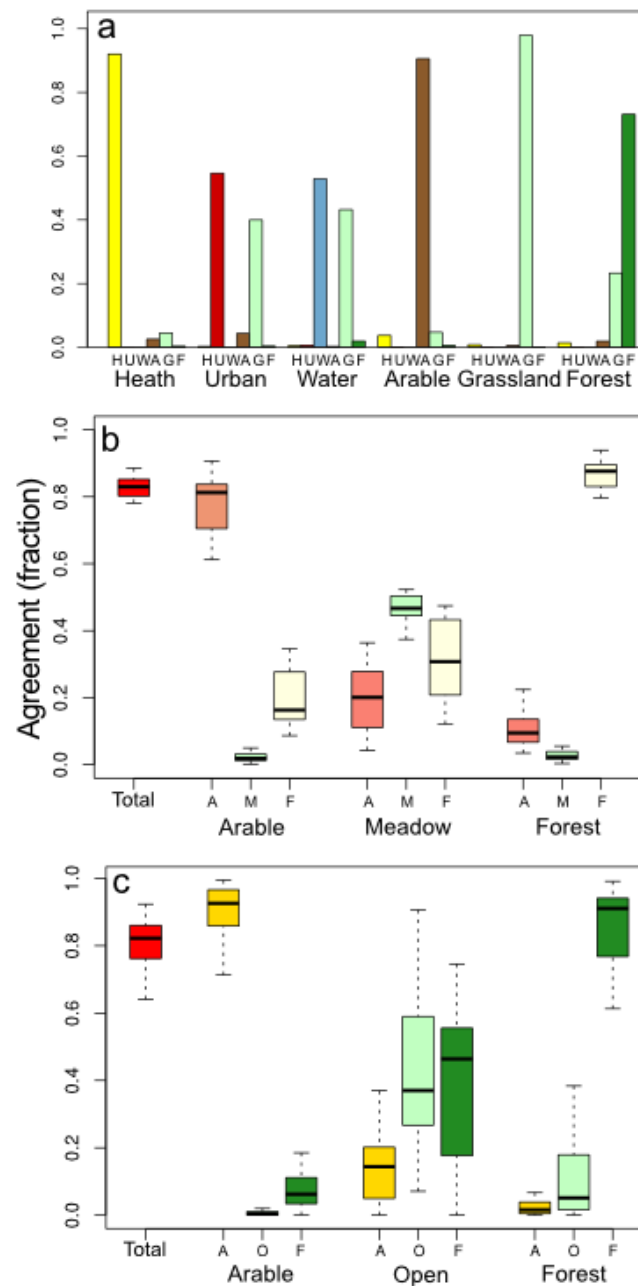
166 Maps were split according to county, and then visually inspected and split into a number of groups
 167 using a file manager or GIS program according to the relative colour tones present in the map. For
 168 most counties, this resulted in 5-20 groups containing anything from a few up to 329 maps. Within
 169 each group, a representative map was digitized using *HistMapR* into arable land etc. (yellow), forest
 170 (darker shades - trees present in the map image) and other open land (lighter shades – no trees).
 171 Classification settings of the selected map were tested on another map within the same group before
 172 running the method in a for-loop or computer cluster to digitize all maps in the group unsupervised.
 173 Computation time was 5-10 minutes per sheet. These batched classifications were inspected in a
 174 GIS program and groups or individual maps re-run with different settings as needed. Water was
 175 added using a modern vector layer as described above. For verification, we took 34 manually-
 176 digitized maps from across the study region, 0.79-139 km² in area. These were either digitizations
 177 of the Economic maps themselves (Gartz 2015; J. Plue unpublished data), or stereographic
 178 interpretations of contemporary aerial photographs (Skånes & Bunce 1997; Cousins & Eriksson
 179 2008; Cousins 2009). Land-use categories were changed to match our classification and
 180 comparisons were carried out as described above.

181

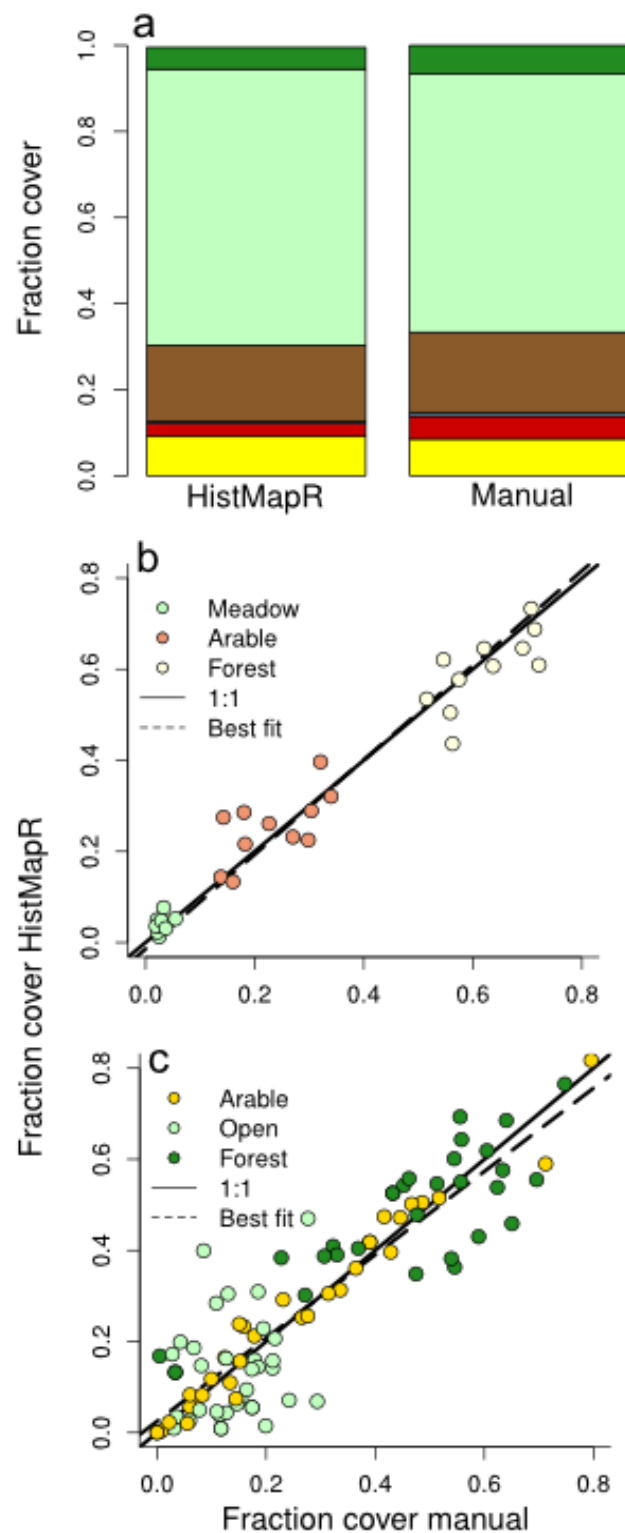
182

183 **Results**

184 We found *HistMapR* to be both fast and straightforward. Comparing outputs with manual
 185 digitizations showed good agreement for all three map series. The map over Dorset showed a 92%
 186 overall agreement at the pixel level, with the majority of pixels in each land-use category being
 187 classified to the same category across digitizations (**Figure 2a**). Agreement in the Swedish map
 188 series was generally around 80-90% (**Fig 2b-c**), with pixels again mostly classified into the same
 189 categories using both methods. In the Economic map, there was less agreement for open areas
 190 compared to arable and forested land.



191 **Figure 2.**
192 *Fraction of pixels assigned to the same land-use category from manual and HistMapR digitizations,*
193 *and the fraction of pixels in each manual digitisation that are assigned to each map category in the*
194 *HistMapR digitisation of (a) the Land Utilisation Survey of Great Britain (1 map), (b) the Swedish*
195 *District Economic map series (11 maps) and (c) the Swedish Economic map series (34 manual*
196 *digitizations). Boxes represent upper and lower quartiles, thick lines show the median, and whiskers*
197 *the dataset range without outliers (observations falling outside the quartiles $\pm 1.5 \times$ the*
198 *interquartile range). Colours match original map shadings.*



199

200 **Figure 3.**

201 *Fraction of land cover assigned to each land-use category in (a) the Land Utilisation Survey of*

202 *Great Britain (1 map), (b) the Swedish District Economic map series (11 maps) and (c) the Swedish*

203 *Economic map series (34 manual digitizations). Colours match original map shadings.*

204
205 Overall share of land-use categories was very similar between the *HistMapR* and manual
206 digitizations. In Dorset, deviation (RMSD) across categories for the whole map was 2%, with
207 grassland over-represented at the expense of small proportions of most other categories (**Figure 3a**).
208 For the District Economic Map, deviation was 4.6% for all categories combined, with values of 2%,
209 6%, and 6% for arable, meadow and forest categories respectively (**Figure 3b**). Deviation in the
210 Economic map was 9% for all categories, 4% for arable, 12% for open land and 12% for forest
211 (**Figure 3c**).

212

213

214 Discussion

215 We have developed a method for a rapid and accurate semi-automated classification of historical
216 land-use using open-source software. Pixel-level agreement between *HistMapR* and manual
217 digitizations was high for all map series (**Fig 2a-c**), meeting commonly-set targets for land-cover
218 classification accuracy (Foody 2002). Deviation of fractional cover of land-use categories between
219 digitizations was usually within a few percent, both at the overall and category level (**Fig 3a-c**),
220 while time savings were significant. We estimate that the manual digitisation of our classified area
221 of the Dorset map (Hooftman & Bullock 2012) took approximately 3-4 weeks to complete,
222 including familiarization with the area and partial method development, compared to the 30-minute
223 *HistMapR* digitization. The almost 1700 km² study area of the District Economic Map took around
224 two months to manually digitize for Cousins *et al.* (2015) compared to 1-2 days' work using our
225 method.

226

227 Despite good results, there were sources of error, which differed between map series. On the
228 smaller-scale UK map, land-use categories that largely consisted of small and linear elements, such

as urban areas (including roads) and surface water were more affected by the smoothing function. This meant that the boundary areas of these land-use types differed in colour from core areas and therefore were often classed according to the exceptions argument, in this case grassland (**Figure 2a**). In the District Economic map series, disagreement at the pixel level arose due to map age and poor scan quality, resulting in variation within and between land-use categories in each map sheet (**Figure 2b**). The disagreement relating to forest and open land across digitization for the Economic maps (**Figure 2c, 3c**) was largely due to the character of the map series. Only arable land, gardens and pasture on former arable fields were formally mapped, with other land-use types only visible as part of the underlying image. This means that all manual digitizations involved users to actively determine the level of tree-cover needed to discriminate parcels of wooded from open land. On the other hand, discrimination between relatively-darker and lighter colours (forest and open land, respectively) could only take place at the whole map level when using the *HistMapR* method, and pixels were then classified as such regardless of patch size. Furthermore, over one-third of 34 the manual digitizations used for comparison were based on corresponding aerial photographs rather than the Economic maps themselves, meaning that in several cases arable fields in the Economic maps were classified as open grassland in the manual digitization and vice versa, thus introducing an additional source of disagreement.

Our results show that despite some pixel-level disagreement, the resulting effect on relative cover of land-use types is generally low (**Figure 3**). It is also important to point out that disagreement between *HistMapR* and manual digitizations does not equate to our maps being incorrect. The issue of delineating land-use categories is a problem for any land-cover classification. With *HistMapR*, users can tailor classification to suit their specific research questions and minimise other potential sources of error according to the historical map in question. Moreover, imperfect georeferencing between digitizations meaning that layers do not always perfectly overlap leads us to believe that

254 actual agreement may even be higher. Our digitizations also represent raw outputs from the R
255 environment, and the potential for improving small-scale accuracy with other GIS tools remains,
256 while retaining significant time savings compared to manual digitization.

257
258 Although manual land-use classification results in a more accurate and detailed digital
259 representation of historical maps, our method is highly useful for a range of applications in ecology.
260 To efficiently classify broad land-use categories over large areas is extremely valuable for
261 quantifying the magnitude of habitat loss over time. This could lead to a greater understanding of
262 the anthropogenic drivers of changes in species diversity and distributions, enabling better
263 predictions of future responses to change at multiple spatial scales.

264

265

266 **Acknowledgements**

267 We are grateful to R. Hijmans for creating the *raster* package upon which our method heavily relies.
268 For the Dorset map, scanned images were provided through <http://www.VisionofBritain.org.uk>,
269 showing material from The Land Utilisation Survey of Great Britain, 1933-49, ©Audrey N. Clark.
270 The manually-digitised map was taken from georeferenced scans of Dorset created by JMB and
271 DAPH under DEFRA licence 10001880. Swedish maps ©Lantmäteriet made available to
272 Stockholm University on licence I2014/00691. Many thanks go to the Swedish OpenStreetMap
273 community for georeferencing the Economic Map. A. Smith and P. Platts gave useful help and
274 advice. This work is funded by the Swedish research council Formas (2015-1065).

275

276

277 **Author contributions**

278 AGA conceived the project. AGA and AK developed the method and created the functions. AGA,

AK, SJ, JP, HS, EW, MW, HW tested the method and digitised maps. JMB, SAOC, DAPH, MG, JP, HS, LT manually digitised maps used for verification. AGA analysed the data and led the writing in close consultation with AK. All co-authors assisted with edits and approve publication.

Data accessibility

Code and example scripts

The *HistMapR* package and documentation are hosted at <https://github.com/AGAuffret/HistMapR/>. Detailed example scripts and input maps are available from Figshare <http://dx.doi.org/10.17045/sthlmuni.4649854> (Auffret *et al.* 2017).

Maps

All Swedish District Economic and Economic maps that we digitized using our method are also available from Figshare for download and use, along with the manually-digitized maps used for verification (Auffret *et al.* 2017). The Dorset maps are under 3rd party copyright.

Scanned Swedish historical maps can be found at <http://historiskakartor.lantmateriet.se/en> (Accessed: 2 February 2017). We used Lantmäteriet's open-access terrain map for contemporary water layers, available from <https://www.lantmateriet.se/sv/Kartor-och-geografisk-information/Kartor/oppna-data/hamta-oppna-geodata/> (In Swedish; Accessed: 2 February 2017).

304 References

- 305 Auffret, A.G., Kimberley, A., Plue, J., Skånes, H., Jakobsson, S., Waldén, E., Wennbom, M., Wood,
306 H., Bullock, J.M., Cousins, S.A.O., Hooftman, D.A.P., Gartz, M. & Tränk, L. (2017) Data
307 from: HistMapR: Rapid digitization of historical land-use maps in R. *figshare data*
308 *repository* doi: 10.17045/sthlmuni.4649854.
- 309 Auguie, B. (2016) gridExtra: Miscellaneous Functions for “Grid” Graphics. *R package version*
310 *2.2.1*, url: <http://CRAN.R-project.org/package=gridExtra>.
- 311 Cousins, S.A.O. (2009) Landscape history and soil properties affect grassland decline and plant
312 species richness in rural landscapes. *Biological Conservation*, **142**, 2752–2758.
- 313 Cousins, S.A.O., Auffret, A.G., Lindgren, J. & Tränk, L. (2015) Regional-scale land-cover change
314 during the 20th century and its consequences for biodiversity. *AMBIO*, **44**, 17–27.
- 315 Cousins, S.A.O. & Eriksson, O. (2008) After the hotspots are gone: Land use history and grassland
316 plant species diversity in a strongly transformed agricultural landscape. *Applied Vegetation*
317 *Science*, **11**, 365–374.
- 318 Foody, G.M. (2002) Status of land cover classification accuracy assessment. *Remote Sensing of*
319 *Environment*, **80**, 185–201.
- 320 Gartz, M. (2015) Plantdiversitet på svenska slätterängar : En GIS-analys med kulturella perspektiv.
321 *Bachelor thesis in Physical Geography at Stockholm University*.
- 322 Greenberg, J.A. & Mattiuzzi, M. (2015) gdalUtils: Wrappers for the Geospatial Data Abstraction
323 Library (GDAL) Utilities. *R package version 2.0.1.7*, url: [http://CRAN.R-](http://CRAN.R-project.org/package=gdalUtils)
324 [project.org/package=gdalUtils](http://CRAN.R-project.org/package=gdalUtils).
- 325 Hijmans, R.J. (2016) raster: Geographic Data Analysis and Modeling. *R package version 2.5-8*, url:
326 <http://CRAN.R-project.org/package=raster>.
- 327 Hooftman, D.A.P. & Bullock, J.M. (2012) Mapping to inform conservation: A case study of changes
328 in semi-natural habitats and their connectivity over 70 years. *Biological Conservation*, **145**,
329 30–38.
- 330 Hooftman, D.A.P., Edwards, B. & Bullock, J.M. (2016) Reductions in connectivity and habitat
331 quality drive local extinctions in a plant diversity hotspot. *Ecography*, **39**, 583–592.
- 332 Jiang, M., Bullock, J.M. & Hooftman, D.A.P. (2013) Mapping ecosystem service and biodiversity
333 changes over 70 years in a rural English county. *Journal of Applied Ecology*, **50**, 841–850.
- 334 Newbold, T., Hudson, L.N., Arnell, A.P., Contu, S., Palma, A.D., Ferrier, S., Hill, S.L.L., Hoskins,
335 A.J., Lysenko, I., Phillips, H.R.P., Burton, V.J., Chng, C.W.T., Emerson, S., Gao, D., Pask-
336 Hale, G., Hutton, J., Jung, M., Sanchez-Ortiz, K., Simmons, B.I., Whitmee, S., Zhang, H.,
337 Scharlemann, J.P.W. & Purvis, A. (2016) Has land use pushed terrestrial biodiversity beyond
338 the planetary boundary? A global assessment. *Science*, **353**, 288–291.
- 339 R Development Core Team. (2015) *R: A Language and Environment for Statistical Computing*. R
340 Foundation for Statistical Computing, Vienna.

- 341 Saar, L., Takkis, K., Pärtel, M. & Helm, A. (2012) Which plant traits predict species loss in
342 calcareous grasslands with extinction debt? *Diversity and Distributions*, **18**, 808–817.
- 343 Skånes, H.M. & Bunce, R.G.H. (1997) Directions of landscape change (1741–1993) in Virestad,
344 Sweden — characterised by multivariate analysis. *Landscape and Urban Planning*, **38**, 61–
345 75.
- 346 Stamp, D.L. (1931) The Land Utilisation Survey of Britain. *The Geographical Journal*, **78**, 40–47.
- 347 Swetnam, R.D. (2007) Rural land use in England and Wales between 1930 and 1998: Mapping
348 trajectories of change with a high resolution spatio-temporal dataset. *Landscape and Urban*
349 *Planning*, **81**, 91–103.
- 350 Wickham, H. (2009) *ggplot2 - Elegant Graphics for Data Analysis*. Springer, New York.
- 351 Willcock, S., Phillips, O.L., Platts, P.J., Swetnam, R.D., Balmford, A., Burgess, N.D., Ahrends, A.,
352 Bayliss, J., Doggart, N., Doody, K., Fanning, E., Green, J.M.H., Hall, J., Howell, K.L.,
353 Lovett, J.C., Marchant, R., Marshall, A.R., Mbilinyi, B., Munishi, P.K.T., Owen, N., Topp-
354 Jorgensen, E.J. & Lewis, S.L. (2016) Land cover change and carbon emissions over
355 100 years in an African biodiversity hotspot. *Global Change Biology*, **22**, 2787–2800.