MR. ALISTAIR GRAHAM AUFFRET (Orcid ID : 0000-0002-4190-4423)
MR. SIMON JAKOBSSON (Orcid ID : 0000-0003-1703-0145)
DR. JAMES M BULLOCK (Orcid ID : 0000-0003-0529-4020)

Article type : Application
Editor : Sarah Goslee

APPLICATIONS

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

Alistair G. Auffret[1,2]*, Adam Kimberley[1], Jan Plue[1], Helle Skånes[1], Simon Jakobsson[1], Emelie Waldén[1] Marika Wennbom[1], Heather Wood[1], James M. Bullock[3], Sara A. O. Cousins[1], Mira Gartz[1], Danny A.P. Hooftman[3,4], Louise Tränk[1]

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

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

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

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

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

Abstract

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

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

3. *HistMapR* contains a number of functions that can be used to semi-automatically digitize historical land-use according to a map's colours, as defined by the RGB bands of the raster image. We test the method on different historical land-use map series and compare results to manual digitizations.

4. Digitization is fast, and agreement with manually-digitized maps of around 80-90% meets common targets for image classification. We hope that the ability to quickly classify large areas of historical land-use will promote the inclusion of land-use change into analyses of biodiversity, species distributions and ecosystem services.

Keywords

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

Introduction

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

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

The *HistMapR* package contains a set of functions (**Table 1**) that allow a fast and accurate digitization of historical land-use maps in R (R Development Core Team 2015). Calling functions from the *raster* package (Hijmans 2016), our method uses the RGB values (Red, Green and Blue; 0-255) of historical map images to classify land use according to map colours.

HistMapR workflow

Single maps - District Economic map of Sweden

The District Economic Map series (AKA The Hundred map; Swedish: *Häradsekonomiska kartan;* 1859-1934) describes major land use, settlements and infrastructure (**Figure 1a-b**). We digitized 11 × 105 km^2 maps from Södermanland county (scale 1:20 000). Each map was first smoothed using *smooth_map*, with a *darkValue* of 150 and a *window.size* of 33 × 4 m^2 pixels. We then defined the colours for the land-use categories forest, arable land and meadow/dwelling (water was digitized using a modern GIS layer, see *Additional steps* below). Clicking 10 times within each category from across the image ensured that the full range of colour tones in each category was sampled. The colour table was then inspected to make sure that there were no misplaced clicks. For land-use classification, we first tested the effects of changing the *errors* and *exceptions* arguments. The colour table could also be rearranged to determine which categories should take precedence over others in the case of overlap, as categories are assigned from the first row of the colour table and down, meaning that if a pixel contains RGB values falling within the range of several categories, it is to the category in the final row of the table that the pixel is assigned. When the best combination of arguments was found, as determined through plotting in the R environment, classifications were written as GeoTiffs for inspection in a GIS program. Computation time was approximately 15 minutes per map on a standard computer, excluding time spent assigning colours with *click_sample* and testing.

Batch processing - The Economic map of Sweden

This map series (*Ekonomiska kartan*) was published 1935-1978 covering the whole of Sweden. In southern Sweden, each sheet covers 25 km^2 at the 1:10 000 scale. The maps consist of a monochrome aerial orthomosaic, with arable land, gardens and pasture on former

arable fields coloured yellow, and additional information such as roads, larger buildings and boundaries in black (**Figure 1c-d**). We classified 7069 maps from the 15 southernmost counties in Sweden, corresponding to an area of 176 725 km^2, at a 1 m^2 resolution. Maps were split according to county, and then visually inspected and split into a number of groups using a file manager or GIS program according to the relative colour tones present in the map. For most counties, this resulted in 5-20 groups containing anything from a few up to 329 maps. Within each group, a representative map was digitized as above into arable land etc. (yellow), forest (darker shades - trees present in the map image) and other open land (lighter shades – no trees). For smoothing, *darkValue*s were typically 80-120 and a *window.size* was 25. Classification settings of the selected map were tested on another map within the same group before running the method in a for-loop or computer cluster to digitize all maps in the group unsupervised. Computation time was 5-10 minutes per sheet. These batched classifications were inspected in a GIS program and groups or individual maps re-run with different settings as needed. Water was added using a modern vector layer as described below.

Additional steps

As different maps from the same series may require different colour table arrangements to achieve optimal results, maps must be reclassified so that raster categories match across all maps in the series. Additionally, in both Swedish map series, surface water was not denoted in a way which meant that they could be adequately classified as a separate land-use category using their RGB values. We used the function *gdal_rasterize* in the package *gdalUtils* (Greenberg & Mattiuzzi 2015) to burn a modern water vector layer onto the digitized raster. If doing this, users should note that any major natural or anthropogenic changes in surface water since the historical map could cause inaccuracies in the final output.

The *HistMapR* package and documentation are hosted at https://github.com/AGAuffret/HistMapR/. Detailed example scripts and input maps are available on Figshare (Auffret *et al.* 2017).

Verification of *HistMapR* outputs

Methods

To verify the *HistMapR* method, we compared digitized maps with manually-digitized versions of the same maps that had been reclassified to match our output categories. For the District Economic Map this was the corresponding area from Cousins *et al.* (2015). For the Economic Map we used 34 manually-digitized maps from across the study region, 0.79-139 km^2 in area, which were either digitizations of the Economic maps (Gartz 2015; J. Plue unpublished data), or stereographic interpretations of the underlying aerial photographs (Skånes & Bunce 1997; Cousins & Eriksson 2008; Cousins 2009). Manual digitizations were rasterized using *gdal_rasterize*, before pixels in both digitizations were aggregated by a factor of five to try to reduce the effect of differences in georeferencing, then 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 maps and the fraction of pixels that were categorized as each category in the *HistMapR* digitizations. Finally, the total fraction of each map assigned to each category in each digitization was calculated, and the root-mean-square deviation (RMSD) of cover between digitizations was calculated for each map category as well as all categories combined.

The Land Utilisation Survey of Great Britain

For additional verification of the method, we digitized one sheet of the UK Land Utilisation Survey (Stamp 1931) and compared outputs to Hooftman & Bullock (2012). As only one sheet was digitized, direct comparison with the other map series is difficult, so the workflow and results are published in **Supporting Information Appendix 1**.

Results

Agreement for both map series was generally around 80-90%, with the majority of pixels in each land-use category being classified to the same category in both *HistMapR* and manual digitizations (**Figure 2**). Overall share of land-use categories was very similar across methods. For the District Economic Map, deviation (RMSD) was 4.6% for all categories combined, with values of 2%, 6%, and 6% for arable, meadow and forest categories respectively (**Figure 3a**). Deviation in the Economic map was 9% for all categories, 4% for arable, 12% for open land and 12% for forest (**Figure 3b**). Agreement across methods for the UK Land Utilisation survey map sheet was high both at the pixel and whole map level (**Appendix 1**).

Discussion

HistMapR facilitates the rapid and accurate semi-automated digitization of historical land-use maps. Pixel-level agreement between *HistMapR* and manual digitizations was high **(Fig 2a-b)**, meeting commonly-set targets for land-cover classification accuracy (Foody 2002). Deviation of fractional cover of land-use categories between digitization methods was generally low, both at the overall and category level (**Fig 3a-b**), while time savings were significant. The almost 1700 km^2 study area of the District Economic map took around two

months to manually digitize for Cousins *et al.* (2015) compared to 1-2 days' work using *HistMapR*, while time savings can be multiplied many times over with the batch-processing used for the Economic map. The extra case study digitizing the UK map demonstrates that the method can be applied in different types of landscapes with a higher number of land-use categories (**Appendix 1**).

Despite good results, there were sources of error. In the District Economic map series, disagreement arose due to map age and poor scan quality, resulting in colour variation within and between land-use categories in each map sheet (**Figure 2a, 3a**). The disagreement relating to forest and open land in the Economic maps (**Figure 2b, 3b**) was largely due to the fact that 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 the active determination of 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 the underlying 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. We therefore point out that in many cases, disagreement between *HistMapR* and manual digitizations does not equate to our maps being incorrect, and that actual agreement with the input maps themselves may often have been even higher than reported here. With *HistMapR*, users can tailor classification to suit their specific research questions and the potential sources of error arising from the historical map of interest.

We believe that *HistMapR* is highly useful for a range of applications in ecology. Other semi-automated classification systems exist, but HistMapR benefits from being open source, while using the R environment facilitates the repetition of digitization using for loops and cluster computing, as well as the resulting classifications being immediately available for further analysis. Although manual land-use classification results in a more accurate and detailed digital representation of historical maps, *HistMapR* offers the possibility to efficiently classify broad land-use categories over large areas. This could lead to a greater understanding one of the major drivers of changes in species diversity and distributions, enabling better predictions of future responses to change.

Acknowledgements

We are grateful to R. Hijmans for creating the *raster* package upon which our method heavily relies. Swedish maps ©Lantmäteriet made available to Stockholm University on licence I2014/00691. Many thanks go to the Swedish OpenStreetMap community for georeferencing the Economic Map. A. Smith and P. Platts gave useful help and advice. Thanks also to the reviewers and editors for helpful advice regarding both the manuscript and package development. This work is funded by the Swedish research council Formas (2015-1065).

Author contributions

AGA conceived the project. AGA and AK developed the method. AGA, AK, SJ, JP, HS, EW, MW, HW tested the method and digitized maps. JMB, SAOC, DAPH, MG, JP, HS, LT manually digitized maps used for verification. AGA analyzed the data and led the writing with help from 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 are available from Figshare, along with the manually-digitized maps used for verification (Auffret *et al.* 2017). 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).

Supporting Information

Appendix 1. Testing *HistMapR* on the Land Utilisation Survey of Great Britain

References

Auguie, B. (2016) gridExtra: Miscellaneous Functions for "Grid" Graphics. *R package version 2.2.1*, url: http://CRAN.R-project.org/package=gridExtra.

Cousins, S.A.O. (2009) Landscape history and soil properties affect grassland decline and

plant species richness in rural landscapes. *Biological Conservation*, **142**, 2752–2758.

Cousins, S.A.O., Auffret, A.G., Lindgren, J. & Tränk, L. (2015) Regional-scale land-cover change during the 20th century and its consequences for biodiversity. *AMBIO*, **44**, 17–27.

Cousins, S.A.O. & Eriksson, O. (2008) After the hotspots are gone: Land use history and grassland plant species diversity in a strongly transformed agricultural landscape. *Applied Vegetation Science*, **11**, 365–374.

Foody, G.M. (2002) Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**, 185–201.

Gartz, M. (2015) Plantdiversitet på svenska slåtterängar : En GIS-analys med kulturella perspektiv. *Bachelor thesis in Physical Geography at Stockholm University*.

Greenberg, J.A. & Mattiuzzi, M. (2015) gdalUtils: Wrappers for the Geospatial Data Abstraction Library (GDAL) Utilities. *R package version 2.0.1.7*, url: http://CRAN.R-project.org/package=gdalUtils.

Hijmans, R.J. (2016) raster: Geographic Data Analysis and Modeling. *R package version 2.5-8*, url: http://CRAN.R-project.org/package=raster.

Hooftman, D.A.P. & Bullock, J.M. (2012) Mapping to inform conservation: A case study of changes in semi-natural habitats and their connectivity over 70 years. *Biological Conservation*, **145**, 30–38.

Jiang, M., Bullock, J.M. & Hooftman, D.A.P. (2013) Mapping ecosystem service and biodiversity changes over 70 years in a rural English county. *Journal of Applied Ecology*, **50**, 841–850.

Newbold, T., Hudson, L.N., Arnell, A.P., Contu, S., Palma, A.D., Ferrier, S., Hill, S.L.L., Hoskins, A.J., Lysenko, I., Phillips, H.R.P., Burton, V.J., Chng, C.W.T., Emerson, S., Gao, D., Pask-Hale, G., Hutton, J., Jung, M., Sanchez-Ortiz, K., Simmons, B.I., Whitmee, S., Zhang, H., Scharlemann, J.P.W. & Purvis, A. (2016) Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science*, **353**, 288–291.

R Development Core Team. (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Saar, L., Takkis, K., Pärtel, M. & Helm, A. (2012) Which plant traits predict species loss in calcareous grasslands with extinction debt? *Diversity and Distributions*, **18**, 808–817.

Skånes, H.M. & Bunce, R.G.H. (1997) Directions of landscape change (1741–1993) in Virestad, Sweden — characterised by multivariate analysis. *Landscape and Urban Planning*, **38**, 61–75.

Stamp, D.L. (1931) The Land Utilisation Survey of Britain. *The Geographical Journal*, **78**, 40–47.

Swetnam, R.D. (2007) Rural land use in England and Wales between 1930 and 1998:

Mapping trajectories of change with a high resolution spatio-temporal dataset. *Landscape and Urban Planning*, **81**, 91–103.

Wickham, H. (2009) *ggplot2 - Elegant Graphics for Data Analysis*. Springer, New York.

Willcock, S., Phillips, O.L., Platts, P.J., Swetnam, R.D., Balmford, A., Burgess, N.D., Ahrends, A., Bayliss, J., Doggart, N., Doody, K., Fanning, E., Green, J.M.H., Hall, J., Howell, K.L., Lovett, J.C., Marchant, R., Marshall, A.R., Mbilinyi, B., Munishi, P.K.T., Owen, N., Topp-Jorgensen, E.J. & Lewis, S.L. (2016) Land cover change and carbon emissions over 100 years in an African biodiversity hotspot. *Global Change Biology*, **22**, 2787–2800.

Figure 1.

Examples of input (©Lantmäteriet) and output maps from (a-b) the District Economic map and (c-d) the Economic map.

Figure 2.

Total fraction of pixels assigned to the same land-use category from manual and HistMapR digitizations, and the fraction of pixels in each manual digitization that are assigned to each map category in the HistMapR digitization of (a) the Swedish District Economic map series (11 maps) and (b) the Swedish Economic map series (34 manual digitizations). Boxes represent upper and lower quartiles, thick lines show the median, and whiskers the dataset range without outliers (observations falling outside the quartiles +/- 1.5 × the interquartile range)**.** Colours match original map shadings.

Figure 3.

Fraction of land cover assigned to each land-use category in (a) the Swedish District Economic map series (11 maps) and (b) the Swedish Economic map series (34 manual digitizations). Colours match original map shadings.

Table 1. Brief descriptions of the *HistMapR* functions. More details in the package documentation (https://github.com/AGAuffret/HistMapR/).

Function	**Description**
smooth_map	Applies Gaussian smoothing to an input raster map calling the *focal* function from the *raster* package, assigning each pixel in each RGB channel the mean value from a user-defined window of *n* pixels surrounding the target pixel (*window.size* argument). Pixels with RGB values below a user-defined threshold can also be removed (*dark.rm* and *darkValue* arguments), 'smoothing over' small patches of dark colour such as place names and property boundaries. Returns a smoothed raster image.
click_sample	Defines the colours for user-specified land-use categories from the smoothed map, calling *raster*'s *click* function. Returns a colour table containing maximum, median and minimum RGB values for each category.
plot_colour_table	Visualizes colour tables for user inspection, calling functions from *gridExtra* (Auguie 2016) and *ggplot2* (Wickham 2009).
class_map	Assigns each pixel in the smoothed input raster to a land-use category according to the colour table produced by *click_sample*. The range of RGB values in each category can be expanded by *n* standard errors due to the likelihood that the most extreme RGB values for each category were not clicked (*errors* argument). Pixels that do not fall within any category can be assigned to an existing category or left unclassified (*exceptions* argument). Classifications can both be plotted within the R environment

	and written to disk as a GeoTiff.