Machine learning for ecosystem services

Willcock, Simon; Martinez-Lopez, Javier; Hooftman, Danny; Bagstad, Kenneth; Balbi, Stefano; Marzo, Alessia; Prato, Carlo; Sciandrello, Saverio; Signorello, Giovanni; Voigt, Brian; Villa, Ferdinando; Bullock, James; Athanasiadis, Ioannis

Ecosystem Services

DOI: 10.1016/j.ecoser.2018.04.004

Published: 01/10/2018

Version created as part of publication process; publisher's layout; not normally made publicly available

Cyswllt i'r cyhoeddiad / Link to publication

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):
Willcock, S., Martinez-Lopez, J., Hooftman, D., Bagstad, K., Balbi, S., Marzo, A., Prato, C., Sciandrello, S., Signorello, G., Voigt, B., Villa, F., Bullock, J., & Athanasiadis, I. (2018). Machine learning for ecosystem services. Ecosystem Services, 33 pt B, 165-174. https://doi.org/10.1016/j.ecoser.2018.04.004

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.
Machine learning for ecosystem services

Simon Willcock a,b,⁎, Javier Martínez-López c, Danny A.P. Hoofman d,e, Kenneth J. Bagstad f, Stefano Balbi c, Alessia Marzo g, Carlo Prato g, Saverio Scandrello e, Giovanni Signorello g, Brian Voigt h, Ferdinando Villa c,i, James M. Bullock e, Ioannis N. Athanasiadis j

⁎ Corresponding author at: School of Environment, Natural Resources and Geography, Bangor University, United Kingdom.
E-mail addresses: s.willcock@bangor.ac.uk (S. Willcock), javier.martinez@bc3research.org (J. Martínez-López), danny.hoofman@lactuca.nl (D.A.P. Hoofman), kjbagstad@pusgs.gov (K.J. Bagstad), stefano.balbi@bc3research.org (S. Balbi), alessia.marzo@unic.it (A. Marzo), s.scandrello@unic.it (S. Scandrello), g.signorello@unic.it (G. Signorello), brian.voigt@uvm.edu (B. Voigt), Ferdinando.Villa@bc3research.org (F. Villa), jmbul@ceh.ac.uk (J.M. Bullock), ioannis.athanasiadis.info (J.N. Athanasiadis).

https://doi.org/10.1016/j.ecoser.2018.04.004
2212-0416/© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Many scientific disciplines are taking an increasingly integrative approach to planetary problems such as global climate change, food security and human migration (Baziliana et al., 2011; Bullock et al., 2017). To address such challenges, methods and practices are becoming more reliant on large, interdisciplinary data repositories often collected in cutting-edge ways, for example via citizen scientists or automated data collection (Isaac et al., 2014). Recent developments in information technology have expanded modelling capabilities, allowing researchers to maximise the utility of such ‘big data’ (Lokers et al., 2016). Here, we focus on one of these developments: data-driven modelling (DDM). DDM is a type of empirical modelling by which the data about a system are used to create models, which use observed systems’ states as inputs for estimating some other system state(s), i.e., outputs (Jordan and Mitchell, 2015; Witten et al., 2016). Thus, DDM is the process of identifying useful patterns in data, a process sometimes previously referred to as knowledge discovery in databases (Fayyad et al., 1996). This process consists of five key steps: 1) understanding the research goal, 2) selecting appropriate data, 3) data cleaning, pre-processing and transformation, 4) data mining...
(creating a data driven model), and 5) interpretation/evaluation
(Fayyad et al., 1996) (Fig. 1). A variety of methods for data mining
and analysis are available, some of which utilise machine learning
algorithms (Witten et al., 2016; Wu et al., 2014) (Fig. 1). A machine
learning algorithm is a process that is used to fit a model to a data-
set, through training or learning. The learned model is subse-
quently used against an independent dataset, in order to
determine how well the learned model can generalise against the
unseen data, a process called testing (Ghahramani, 2015; Witten
et al., 2016). This training–testing process is analogous to the cal-
ibration–validation process associated with many process-based
models.

In general, machine learning algorithms can be divided into two
main groups (supervised- and unsupervised-learning; Fig. 1), sep-
aratly by the use of explicit feedback in the learning process (Blum
and Langley, 1997; Russell and Norvig, 2003; Tarca et al., 2007).
Supervised-learning algorithms use predefined input-output pairs
and learn how to derive outputs from inputs. The user specifies
which variables (i.e., outputs) are considered dependent on others
(i.e., inputs), which sometimes indicates causality (Hastie et al.,
2009). The machine learning toolbox includes several linear and
non-linear supervised learners, predicting either numeric outputs
(regressors) or nominal outputs (classifiers) (Table 1). An example
of supervised machine learning that is familiar to many ecosystem
service (ES) scientists is using a general linear model, whereby the
user provides a selection of input variables hypothesised to predict
values of an output variable and the general linear model learns to
reproduce this relationship. The learning process needs to be fine-
tuned through a process, as for example in the case of stepwise
selection where an algorithm selects the most parsimonious
best-fit model (Yamashita et al., 2007). However, note that

| Category               | Task                               | Common algorithms                                      |
|------------------------|------------------------------------|--------------------------------------------------------|
| Unsupervised learning  | (learning without feedback from a trainer) | k-means, Apriori, PCA                                   |
| Supervised learning    | (learning past actions/ decisions with a trainer) | Bayesian Networks, Decision Trees, Neural Networks, Linear Regression, Perceptron |

Table 1
A simplified summary of machine learning algorithms (categorised as supervised and unsupervised).

...
between variables (Ghahramani, 2015). However, for some socio-ecological systems, but instead capture relationships that models do not represent the underlying processes within. Users of DDM should be aware that uncertainty is useful, whilst estimates of model uncertainty (i.e., the cross-validation statistic). Whilst the learning process, often accompanied with an estimation of the variable drift (Gama et al., 2004). Model validation/testing, which has yet to become standard practice within the ES modelling community (Baveye, 2017; Hamel and Bryant, 2017), is an integral part of the machine learning process within DDM. This is vital as DDM can result in overfitting, which occurs when the model learns the training data well (i.e., a close fit to the training data), but performs poorly on independent test data (Clark, 2003).

To assess the quality of the learning process, machine learning algorithms use various methods (summarised in Witten et al. (2016)) to ensure that the results are generalizable and avoid overfitting. For example, k-fold cross validation allows for fine-tuning of model performance (Varma and Simon, 2006; Wiens et al., 2008). This approach maximises the data availability for model training by dividing the data into k subsets and using k-1 subsets to train the model whilst retaining a subset for independent validation. This process is repeated k times so that all available data have been used for validation exactly once. The results of the k-folds are then combined to produce metrics of quality for the machine learning process, often accompanied with an estimation of the model uncertainty (i.e., the cross-validation statistic). Whilst the goodness-of-fit parameter used varies within DDM (e.g., root mean square error is used extensively within regression models, but the standard error is more commonly used in Bayesian machine learning (Cheung and Rensvold, 2002; Uusitalo, 2007)), it provides the user with a transparent estimate of model uncertainty. Whilst estimates of uncertainty are useful, users of DDM should be aware that such models do not represent the underlying processes within socio-ecological systems, but instead capture relationships between variables (Ghahramani, 2015). However, for some datasets and model applications (see Section 4 for further details), DDM can produce more accurate models than process-based models, as the latter may suffer from an incomplete representation of the socio-ecological processes (Jordan and Mitchell, 2015; Tarca et al., 2007). Finally, as with any modelling, DDM depends on the quality of the training and testing datasets used; whilst some extreme cases or outliers might get ignored during DDM, the quality of the information supplied to the machine learning algorithms should be verified beforehand (Galelli et al., 2014).

The aim of this paper is to demonstrate the utility of DDM to the ES community. We present two examples of DDM using Bayesian networks (a supervised learning technique), as implemented in the Waikato Environment for Knowledge Analysis machine learning software (Weka; http://www.cs.waikato.ac.nz/ml/weka/; Frank et al., 2016; Hall et al. (2009)), used both standalone and as part of the Artificial Intelligence for Ecosystem Services (ARIES; http://aries.integratedmodelling.org; Villa et al. (2014)) modelling platform. We chose Bayesian network methods as uncertainty metrics describing both the model fit and the grid-cell uncertainty can be calculated (Aguilera et al., 2011; Landuyt et al., 2013; Uusitalo, 2007). Our Weka example focuses on firewood use in South Africa, and is comparable to conventional ES models recently published by Willcock et al. (submitted for publication). Using ARIES, we model biodiversity value within Sicily, and demonstrate how DDM can make use of volunteered geographical information by incorporating data from Open Street Maps into the machine learning process. In both examples, we highlight how model structure and uncertainty computed in the machine learning process supplement and enhance the value of the results reported to the user.

2. Methods

For the first example, we used Weka, an open-source library of machine learning algorithms (Frank et al., 2016; Hall et al., 2009), to create a model capable of identifying the upper quartile of sites for firewood use in South Africa. We chose this example as: 1) firewood use is of high policy relevance in sub-Saharan Africa (Willcock et al., 2016); 2) robust spatial data on firewood use are available within South Africa and may, for some municipalities, provide a comparable context to other parts of sub-Saharan Africa, which are often more vulnerable but data deficient (Hamann et al., 2015); 3) models ranking the relative importance of different sites were rated as useful to support ES decision-making by nearly 90% of experts in sub-Saharan Africa (Willcock et al., 2016); and 4) multiple conventional models have recently been run for this ES covering this spatial extent (see Willcock et al. (submitted for publication) for full details).

The firewood use data are freely available (Hamann et al., 2015) and are based on the South African 2011 population census, which provides proportions of households per local municipality using a specific ES (similar data are available for a set of other ES; see www.statssa.gov.za for all 2011 census output). For this paper we used the proportion of households that use collected firewood as a resource for cooking (Hamann et al., 2015). To derive a measure of total resource use, we multiplied the proportion of use by the 2011 official census municipal population size (from www.statssa.gov.za) as: [% households using a service] x [municipal population size]. We then divided this value by the area of each local municipality to provide an estimate of firewood use density, ensuring that model inputs are independent of the land area of the local municipality.

To utilise Bayesian networks, the decision variable (firewood use density) had to be converted into a categorical (nominal) attribute; note, the categories created during this process are unordered. The goal of this task was to predict the areas in the upper quartile, reflecting demand from decision-makers for identification of the most important sites for ES production and, once identified, enabling these areas to be prioritised for sustainable management (Willcock et al., 2016). Thus, the firewood use density data were categorised within the highest 25% (Q4) and the lowest 75% (Q1-Q3) quartiles using Weka’s Discretize filter to create ranges of equal frequencies (four in our case). Out of the generated quartiles, the three lower ones were merged with the MergeTwoValues filter. To ensure like-for-like comparisons between our DDM and conventional models, we provided the machine learning algorithms with the same user supplied input data used to model firewood within Willcock et al. (submitted for publication) (Table 2). Since most Bayesian network inference algorithms can use only categorical data as inputs, the input data were discretised by grouping their values in five bins of equal frequencies. Selecting the number of bins is a design choice and may impact model output (Friedman and Goldszmidt, 1996; Nojavan et al., 2017). As such, the sensitivity of the modelled output to variable bin numbers warrants future investigation, but is beyond the scope of this first-order introduction to machine learning for ES.
The municipal-scale inputs into the Weka machine learning algorithms to estimate firewood use in South Africa. Overfitting is avoided by first training the algorithm on subsets of these data and then testing against the remaining data. 

| Attribute         | Description                                                                 |
|-------------------|------------------------------------------------------------------------------|
| LCAgriculture     | The proportion of agricultural land area, derived from GeoTerraImage (2015) |
| LCForest          | The proportion of forested land area, derived from GeoTerraImage (2015)     |
| LCGrassland       | The proportion of grassland land area, derived from GeoTerraImage (2015)    |
| LCUrban           | The proportion of urban areas, derived from GeoTerraImage (2015)             |
| LCWater           | The proportion of water bodies area, derived from GeoTerraImage (2015)       |
| COFirewood        | The proportion of area on which firewood can be produced (Forest, Woodland, Savanna), derived from GeoTerraImage (2015) |
| OPotential        | The proportion of protected natural areas, derived from the World Database on Protected Areas (www.protectedplanet.net) |
| MOCarbon          | Mean amount of carbon stored per hectare, as calculated in Willcock et al. (submitted for publication) |
| OGrowthDay        | Average number of growing days in the area as driven by the relationship between rainfall and evapotranspiration, as calculated in Willcock et al. (submitted for publication) |
| ZScholesA         | A metric of the nutrient-supplying capacity of the soil (Scholes, 1998)    |
| ZScholesB         | A metric of the nutrient-supplying capacity of the soil (Scholes, 1998)    |
| ZScholesD         | Scholes (1998) land use correction, as calculated in Willcock et al. (submitted for publication) |
| ZSlope            | This is the mean slope in the area, based on the global 90-m digital elevation model downloaded from CGIAR-CSI (srtm.csi.cgiar.org/SELECTION/inputCoord.asp) |
| Population_density| The municipal population based on the South African 2011 census (www.statssa.gov.za) |
| Firewood_density  | Observed firewood use for cooking from the South African 2011 census (Hammann et al., 2015) |

3. Results

In the first example, the results for all configurations of the DDM created for firewood use in South Africa had a classification accuracy above 80% (see Appendix 1). The model predictions are statistically significant with a confidence level of 0.05 (two tailed) when compared to the ZeroR classifier (a baseline classifier that always predicts the majority class). Using ArcGIS v 10.5.1, we spatially mapped the outputs of the most accurate Bayesian network DDM (Figs. 3, 4; Appendix 3). The confusion matrix for this model shows that 186 out of the 226 local municipalities were correctly classified (an overall classification accuracy of 82%), and, out of 56 municipalities classified in the upper quartile (Q4), 36 were correct predictions (64% recall [i.e. the percentage of the most important sites for firewood ES correctly identified], comparable with conventional modelling methods evaluated against independent data [Table 3; Willcock et al. (submitted for publication); Appendix 3]. The DDM also produces probabilistic outputs for the respective inputs (Appendix 4).

For biodiversity value in Sicily, 43% of the testing subsample was correctly classified into 1 of 10 biodiversity value categories, with a majority of the incorrectly classified results falling into immediately close numeric ranges (Appendix 5). During a workshop in June 2017, the same Sicilian experts that provided the training set (a team of five including an academic conservationist, an academic ornithologist, an academic botanist and an expert on agricultural biodiversity) qualitatively evaluated the output in non-sampled but well-known regions and deemed it a distinct improvement on previously computed biodiversity value assessments, built through conventional GIS overlapping and interpolation techniques; an assessment that was embraced by other participants from both local governmental and conservation institutions (Fig. 2). As the map reflects the human assessment of biodiversity value rather than objective measurements, the consensus of experts and practitioners was deemed equivalent to a satisfactory validation. The confusion matrix (Appendix 5) shows how the majority of misclassifications are between similar value categories. For example, 73% of test data were predicted within one class above or below their actual class, and 84% of test data were correctly classified within two classes above and below their actual class. A Spearman Rho test...
Fig. 2. The relative value of terrestrial biodiversity in Sicily estimated by a) inverse distance weighted interpolation of observed values and b) Bayesian networks using data-driven modelling. Both original (white) biodiversity value observations and the additional sites of low biodiversity value (black) are shown as points.

Fig. 3. Diagrammatic representation of the machine-learned Bayesian network model of firewood use in South Africa (see Table 2 for category codes). The structure of the model was informed by the machine learning algorithm with no predetermined restrictions.
highlights the significant correlation between the ranked model and validation data categories (Rho: 0.58; p-value <0.001). The root-mean-squared error of the model prediction was also computed and resulted in a value of 0.26 (Hyndman and Koehler, 2006).

4. Discussion

Lack of credibility, salience and legitimacy are the major reasons for the 'implementation gap' between ES research and its incorporation into policy- and decision-making (Clark et al., ...
Comparing recall of DDM outputs with conventional models when producing estimates of firewood use in South Africa. Outputs from conventional models of varying complexity were validated using independent data (see Willcock et al. (submitted for publication) for full model descriptions and model complexity analysis). DDM outputs were validated using k-fold cross validation (see Section 2). Comparing recall of DDM outputs with conventional models when producing estimates of firewood use in South Africa. Outputs from conventional models of varying complexity were validated using independent data (see Willcock et al. (submitted for publication) for full model descriptions and model complexity analysis). DDM outputs were validated using k-fold cross validation (see Section 2).

Table 3

| Model                      | Model criteria                          | Recall for the upper quartile of firewood use (%) |
|----------------------------|----------------------------------------|--------------------------------------------------|
| Bayesian network within Weka (Frank et al., 2016; Hall et al., 2009) | Assignment threshold = 50%                | 64.3                                              |
| Conventional model A (Complexity score: 2; Willcock et al. (submitted for publication)) | Assignment threshold = 75%                | 90.9                                              |
| Conventional model B (Complexity score: 4; Willcock et al. (submitted for publication)) | Gridcell size = 1 km                      | 75.0                                              |
| Conventional model C (Complexity score: 4; Willcock et al. (submitted for publication)) | Gridcell size = 10 km                     | 73.2                                              |
| Conventional model D (Complexity score: 36; Willcock et al. (submitted for publication)) | Gridcell size = 1 km                      | 75.0                                              |
| Conventional model E (Complexity score: 31; Willcock et al. (submitted for publication)) | Gridcell size = 10 km                     | 76.8                                              |
| Conventional model F (Complexity score: 35; Willcock et al. (submitted for publication)) | Gridcell size = 55.6 km                   | 76.8                                              |
| Conventional model G (Complexity score: 3; Willcock et al. (submitted for publication)) | Gridcell size = 5 km                       | 53.6                                              |

* Models have been anonymised as identification of the best specific model for a particular use is likely to be location specific and may shift as new models are developed (Willcock et al., submitted for publication).

2016; Olander et al., 2017; Wong et al., 2014). A lack of uncertainty information and the inability to run models in data-poor environments and/or under conditions where underlying processes are poorly understood may contribute to the implementation gap. However, DDM can help to address these current shortcomings in ES modelling. Here, we have demonstrated that DDM is feasible within ES science and is capable of providing estimates of uncertainty. For our South African case study, the machine learning algorithms were able to produce a modelled output of comparable accuracy to conventional modelling methods when using the same input variables, despite our DDM using data at a much coarser (local municipality) scale (Table 3). Using the spatially attributed uncertainty (i.e., the probability of each local municipality being in Q4), decision-makers would be able to set their own level of acceptable uncertainty. In our example, since we have two categorical bins (i.e., Q1-3 and Q4), any local municipality with a modelled Q4 probability over 0.5 is assigned to the Q4 category. This assignment threshold can be varied; e.g., it is possible to state that municipalities where modelled Q4 probability is less than 0.25 or greater than 0.75 are likely to be grouped within Q1-3 and Q4 respectively, and to admit that we are less certain for the remaining municipalities. In our example, this would result in a 96% (135 out of 140) categorisation accuracy for Q1-3 and a 91% (30 out of 33) categorisation accuracy for Q4, with 53 local municipalities left uncategorised due to uncertainty.

Thus, using Bayesian networks and machine learning, we are able to convey to decision-makers not only which sites show the highest ES use or value, but also how confident we are in our estimate at each site (Aguilera et al., 2011; Chen and Pollino, 2012; Landuyt et al., 2013). This information allows decision-makers to 1) apply an assignment threshold of their choosing to the modelled output before making a policy- or management-decision, and 2) use their own judgement for potentially contentious decisions, where uncertainty is higher (Olander et al., 2017). For example, whilst it is perhaps obvious that sites where we are highly certain that there is high ES value should be appropriately managed, it is unclear which sites should be the next highest management priority. Given a limited budget, it is a medium-ES value site with high certainty more or less worthy of management than a potentially high-value site with medium or low certainty? Decision-makers show both capacity and willingness to engage with the uncertainty information should these data be made available (McKenzie et al., 2014; Scholes et al., 2013; Willcock et al., 2016), even when results may indicate high levels of uncertainty. This is illustrated by a Sicilian case study, in which decision-makers, when advised of the relatively low overall classification accuracy (43%), accepted it as predictions were close to their actual value (i.e. 73% of test data were predicted within one class above or below their actual class) and were viewed as an improvement on previous estimates (Fig. 2). Thus, providing estimates of uncertainty should become standard practice within the ES community (Hamel and Bryant, 2017).

There are both advantages and disadvantages to using machine learning algorithms for the ‘data mining’ step of DDM (Fayyad et al., 1996). As highlighted above, machine learning algorithms provide indications of uncertainty that could usefully support decision-making. However, similar uncertainty metrics can also be obtained using conventional modelling (i.e., via the confidence intervals surrounding regressions (Willcock et al., 2014) or Bayesian belief networks (Balbi et al., 2016)). Similar to conventional modelling, the performance of model algorithms substantially depends on the parameters, model structure and algorithm settings applied (Zhang and Wallace, 2015). For example, many machine learning algorithms require categorical data and so potentially an additional step of data processing whereby continuous data are discretised. In our South African case study, we divided firewood use data into five bins but acknowledge that the number of bins may affect model performance and the impact of this warrants further investigation (Friedman and Goldszmidt, 1996; Nojavan et al., 2017; Pradhan et al., 2017). However, a variety of machine learning algorithms are available (Table 1) and not all of them required discretised data (Jordan and Mitchell, 2015; Witten et al., 2016). Furthermore, for our firewood models, we used machine learning to create the model structure. Structural learning can yield better performing models (i.e., all our South African model configurations had a classification accuracy above 80%; Appendix 1) and may highlight relationships that have not yet been theorised (or have previously been discarded) (Gibert et al., 2008; Suominen and Toivanen, 2016). However, the obtained structures (Fig. 3) may not be causal and could confuse end-users (Schmidhuber, 2015). Thus, predefined network structures may be preferred for applications where causality is particularly important. Further generalisations useful for ES modellers considering machine learning algorithms include the following: 1) Multi-classification problems may have lower accuracy – as highlighted by comparing our South African (2 category output, 82% accuracy) and Sicilian (10 category output, 43% accuracy) examples – the more categories in the modelled output, the lower the apparent accuracy. Thus, the number of categories in the output should be considered when interpreting the model accuracy metric. For example, a random model with a two category output and a four category output will be accurate 50% and 25% of the time respectively. Thus, a machine-learned model with an accuracy of 40% is...
poor if the output had two categories, but learned more (and so might be of more use) if a four category output was being considered; 2) Supervised learning can be used when drivers are known – for example, with no a priori assumptions, unsupervised learning could cluster beneficiaries into groups, but these may not match known beneficiary groups (i.e., livelihoods) and so might be difficult to interpret (Schmidthuber, 2015). Supervised learning can be used to align the outputs from machine learning algorithms with decision-maker specified beneficiary groups; 3) machine learning algorithms are best applied to the past and present, but not the future – Although machine learning algorithms can detect strong relationships, accurately describing past events and providing useful predictions where process-based understanding is lacking (Jean et al., 2016), the relationships identified may not be causally linked and so may not hold when extrapolating across space or time (Mullainathan and Spiess, 2017). Thus, where the process is well understood, DDM is unlikely to be more appropriate than conventional process-based models (Jordan and Mitchell, 2015). Understanding the caveats and limitations of machine learning algorithms is important before the algorithms are used for DDM.

A further critique of DDM is that it can appear as a ‘black box’ in which the machine learning processes are not clear to the user and so they could widen the implementation gap (Clark et al., 2016; Olander et al., 2017; Wong et al., 2014). However, we have demonstrated that utilisation of machine learning algorithms can be transparent and replicable. For example, Bayesian networks allow the links between data to be visualised (Fig. 3) (Aguilera et al., 2011; Chen and Pollino, 2012; Landuyt et al., 2013). The stand-alone Weka software is user friendly and requires minimal expertise, and ease of use has been further simplified within the ARIES software as DDM can be run merely by selecting a spatiotemporal modelling context and then using the ‘drag-drop’ function to start the machine learning process (Villa et al., 2014). Machine learning and machine reasoning (Bottou, 2014) are facilitated within the ARIES system through semantic data annotation, which makes data and models machine readable and allows for automated data selection and acquisition from cloud-hosted resources, as well as automated model building (Villa et al., 2017). To ensure that this complex process remains transparent, the Bayesian network is described using a provenance diagram (Fig. S2), characterising the DDM process, i.e., which data and models were selected by ARIES (Fig. 1). Furthermore, work has begun to enable the ARIES software to produce automated reports that describe the DDM process and modelling outputs in a readily understandable language (see Appendix 2 for a preliminary automated report for the ARIES example used in this study). Advances such as this may enable decision-makers to run and interpret ES models with minimal support from scientists, potentially increasing ownership in the modelled results and closing the implementation gap (Olander et al., 2017).

The DDM process encourages scientists to use as much data as possible to generate the highest quality knowledge. Machine learning algorithms provide a tool by which ‘big data’ can be incorporated into ES assessments (Hampton et al., 2013; Lokers et al., 2016; Richards and Tunçer, 2017). For example, using the ARIES software, we demonstrated how Open Street Map data can be included in the machine learning process (Haklay and Weber, 2008). Whilst future research is needed to determine how much data is actually needed, it is clear that ES scientists must contribute to and make use of large datasets to participate in the information age (Hampton et al., 2013), particularly where data are standardised and made machine-readable (Villa et al., 2017). Using machine learning algorithms to interpret big data may help provide a wide range of ES information across the variety of temporal and spatial scales required by decision-makers (McKenzie et al., 2014; Scholes et al., 2013; Willcock et al., 2016). There has been a recent call-to-arms within the ES modelling community to shift focus from models of biophysical supply towards understanding the beneficiaries of ES and quantifying their demand, access and utilisation of services, as well as the consequences for well-being (Bagstad et al., 2014; Poppy et al., 2014). Combining social science theory and data to explain the social-ecological processes of ES co-production, use and well-being consequences will likely result in substantial improvements to ES models (Bagstad et al., 2014; Díaz et al., 2015; Pascual et al., 2017; Suich et al., 2015; Willcock et al., submitted for publication). Such social science data are sometimes available at large scales (e.g., via national censuses) but, with some notable exceptions (e.g., Hamann et al. (2016, 2015)), are rarely used within ES models (Egoh et al., 2012; Martínez-Harms and Balvanera, 2012; Wong et al., 2014). The process of DDM guides researchers in how to incorporate big data into ES models, scaling up results from sites to continents (Hampton et al., 2013; Lokers et al., 2016). DDM allows an interdisciplinary approach across a large scale and so may help guide global policy-making, e.g., within the Intergovernmental Science-Policy Platform for Biodiversity and Ecosystem Services (IPBES: www.ipbes.net).

In conclusion, DDM could be a useful tool to scale up ES models for greater policy- and decision-making relevance. DDM allows for the incorporation of big data, producing interdisciplinary models and holistic solutions to complex socio-ecological issues. It is crucial that the approach and results of machine learning algorithms are conveyed to the user to enhance transparency, including the uncertainty associated with the modelled results. In fact, we hope that the validation of ES models becomes standard practice with the ES community for both process-based and DDM. In the future, automation of the modelling processes may enable users to run ES models with minimal support from scientists, increasing ownership in the final output. Such automation should be accompanied by transparent provenance information and procedures for a computerised system to select context-appropriate data and models. Taken together, the advances described here could help to ensure ES research contributes to and inform ongoing policy processes, such as IPBES, as well as national-, subnational-, and local-scale decision making.

Contributions

SW, INA, JML, KJB, SB, BV & FV conceiving the paper. SW, DAPH, JMB & INA carried out the South African case study. JML, SB, AM, CP, SS, CS & FV carried out the Sicilian case study. SW & INA wrote the manuscript, with comments and revisions from all other authors.

Acknowledgements

This work took place under the ‘WISER: Which Ecosystem Service Models Best Capture the Needs of the Rural Poor?’ project (NE/L001322/1), funded by the UK Ecosystem Services for Poverty Alleviation program (ESPA; www.espa.ac.uk). ESPA receives its funding from the UK Department for International Development, the Economic and Social Research Council and the Natural Environment Research Council. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. We would like to thank the two anonymous reviewers whose comments improved the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ecoser.2018.04.004.
