Predicting Aggregate Properties of Soil Communities vs. Community Structure in an Agricultural Setting

Damjan Demšar¹, Sašo Džeroski¹, Marko Debeljak¹, Paul Henning Krogh²

Abstract

Increasing amounts of environmental data are being collected. With environmental data, we often encounter the situation of having to predict several target variables of similar type, such as biomasses of different species. This situation is usually handled by computing an aggregate target variable (like total biomass or a biodiversity measure) and then predicting the aggregate variable. Another possible (but rarely taken) approach is to model all target variables and then calculate the aggregate variable from the model outputs. In this paper, we try to answer the question whether the simpler approach of producing one model for the aggregate target variable is worse than the more complex approach of producing multiple models and then calculating the aggregate variable from the model outputs. We do this by taking a dataset describing the agricultural events and soil biological parameters as independent variables and a set of microarthropod species biomasses as dependent variables. We calculated several aggregate target variables such as total biomass, Shannon biodiversity and species richness from the original data. We build models to predict these directly, and also build separate predictive models for the biomass of the microarthropod species and calculate the aggregate target variables from the outputs of these models. We compared the aggregate variables calculated from the measured data, the aggregate variables predicted directly and the aggregate variables calculated from the outputs of the models for individual species using the Pearson correlation coefficient and two additional error measures. Our results show, that in most cases first calculating the aggregate variables, and then learning models to predict these directly yields better results than modeling individual species and then calculating the aggregate variables from the predictions of these models.

1. Introduction

With increasing amounts of (environmental) data being collected, more and more data sets have more than one dependent variable. In such cases, we can construct a separate model for each of the dependent variables, which can be useful for simulating the behavior of the system, but is too complex for other uses. Another possibility is to use one of the modeling systems that are capable of predicting several variables with one model (e.g. a system for building predictive clustering trees CLUS [1]). A far more common approach is to aggregate all dependent variables into one aggregate, compound variable, and then build a model for that aggregate variable. Examples of aggregate variables describing an ecosystem can be total biomass, species richness, biodiversity for a set of species data, energy input, etc. Such aggregate type models (e.g. either a complex set of models or a simple model of aggregate variable) can be used as sub-models in decision support model for assessing environmental, economical and social effects of application of genetically modified corn in agricultural practice. [2]. Description of the effects on soil microarthropods in such a decision support model would based on a set of small models describing the soil microarthropod community by aggregate variables like total biomass of microarthropods or their Shannon biodiversity. Since there is also the possibility of using several single species models, we were interested in differences between single species models and aggregate type of models.

In order to detect the possible loss of accuracy due to different modelling approaches, we performed a series of experiments, making single species models and aggregate type of models as well. We compared

¹ Department of Knowledge Systems, Jožef Stefan Institute, Ljubljana, Slovenia
² Department of Terrestrial Ecology, National Environmental Research Institute, Roskilde, Denmark
the results of the single species models predicting aggregate variables (e.g. biomass) and the results of the aggregated models with the aggregated variables calculated from the original data. In order to evaluate the performance of both modeling approach, the Pearson correlation coefficient, relative absolute error and root relative square error were calculated for each type of modeling approach, and the comparisons of performance’s measures were performed.

2. Data

Dataset describes four experimental farming systems with different crops (Tab. 2), located at the Foulum experimental station, Denmark (15 fields in total) over the period 1989 to 1993. Two systems are conventional systems with pesticide use, the other two are organic ones with no pesticide use. Dataset consists from 534 microarthropod samples [3].

The data available for the study (Table 1) include different agricultural measures (e.g. example, packing, tillage; fertilizer and pesticide use, crops planted and cattle grazing). Cropping and grazing history is available for the last three years of sampling period. Dataset contains environmental variables describing properties of the sites from where the samples of soil microarthropods have been collected. The variables used to make models of microarthropods are listed in Table 1.

In order to eliminate the effect of non linear decay of tillage by microarthropods, the following transformation of tillage was applied:

\[
\text{tillage} = 10 \times \text{time since tillage (in months)}
\]

where power \( i \) depends on the type of tillage (deep to subshallow tillage: \( i = 2 \), shallow tillage: \( i = 4 \)).

The attributes describing current and previous crops are binary attributes, where only one attribute in a set takes the value 1, and all others are set to 0. There are 4 sets: current crops and 3 sets of previous crops (abbreviated crop names are listed in Table 2).

The same sampling method was used for each field. The distance between each sample was 5 m and all samples were collected within a 20x20 m area. The distance to hedges and ditches was at least 10 m. Samples were taken in the upper 5.5 cm soil layer. The sampling containers measured 6 cm in diameter. Sampling was done using a split soil corer and extraction was performed using a MacFadyen high gradient heat extractor. The dataset includes data about the measured species (Table 3), all species belong to the collembola group (springtails, 40 species). The 40 species were primary dependent variables, and were used to calculate the Shannon biodiversity index:

\[
H = -\sum p_i \log p_i
\]

where \( p_i \) represents the proportion of species \( i \) in the sample and \( S \) represents total number of species in sample (number of species with biomass over zero).

Addition to this, partial sums for the following Collembolans functional groups were calculated:

- Euedaphic (EU) - true soil living animals,
- Hemiedaphic (HE) - lives in the litter layer or in the upper few cm of the soil,
- Epiedaphic (EP) - lives on the soil and in habitats on top of the soil,
- Epi to Hemi (EP-HE) - species that cannot be clearly grouped into EP or HE groups,
- Hemi to Eu (HE-EU) - species that cannot be clearly grouped into HE or EU groups.
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>actsit_mon</td>
<td>age of actual situation (time in months since current crop was sown or last crop was harvested)</td>
</tr>
<tr>
<td>samp_time</td>
<td>1 = March - April, 2 = May - June, 3 = July - August, 4 = September - November</td>
</tr>
<tr>
<td>Current_crop=X</td>
<td>binary attributes describing current crop (see Table 2 for possible values of X)</td>
</tr>
<tr>
<td>crop1=X</td>
<td>binary attributes describing last years crop (see Table 2 for possible values of X)</td>
</tr>
<tr>
<td>crop2=X</td>
<td>binary attributes describing crop two years ago (see Table 2 for possible values of X)</td>
</tr>
<tr>
<td>AC, MC, CAC, WIC, PC_X</td>
<td>current (AC) / previous crop (AC_1, AC_2 and AC_3) is annual crop</td>
</tr>
<tr>
<td>pesticide</td>
<td>pesticide (1 = fields in a rotation where pesticides are used; 0 = no pesticide use)</td>
</tr>
<tr>
<td>tr_packing</td>
<td>packing (months since) transformed using $\frac{10^{(x-10\text{ months})}}{10}$</td>
</tr>
<tr>
<td>tr_shal_till</td>
<td>shallow tillage (weed harrowing etc) 0-5 cm layer (months since) transformed using $\frac{10^{(x-10\text{ months})}}{10}$</td>
</tr>
<tr>
<td>tr_subshal_till</td>
<td>shallow tillage 5-10 cm layer (months since) transformed using $\frac{10^{(x-10\text{ months})}}{10}$</td>
</tr>
<tr>
<td>tr_deep_till</td>
<td>deep tillage (plowing, rotation) &gt;10 cm layer (months since) transformed using $\frac{10^{(x-10\text{ months})}}{10}$</td>
</tr>
<tr>
<td>fert_lev</td>
<td>low=0, normal=1, high=2.</td>
</tr>
<tr>
<td>fert_type</td>
<td>n=0, solid=1, liquid=2.</td>
</tr>
<tr>
<td>fert_time</td>
<td>fertilization time (months since) transformed using $\frac{10^{(x-10\text{ months})}}{10}$</td>
</tr>
<tr>
<td>cu</td>
<td>cattle (cattle grazing on the field)</td>
</tr>
<tr>
<td>sh</td>
<td>sheep (sheep grazing on the field)</td>
</tr>
<tr>
<td>grazing</td>
<td>animals grazing on the field</td>
</tr>
<tr>
<td>si</td>
<td>silage/hay (the crops was intended for sillage or hay)</td>
</tr>
<tr>
<td>sf</td>
<td>stubble field</td>
</tr>
<tr>
<td>se</td>
<td>seed bed – bare field, seeds planted less than one month ago</td>
</tr>
<tr>
<td>seha</td>
<td>bare field harrowed</td>
</tr>
<tr>
<td>sepl</td>
<td>bare field plowed</td>
</tr>
<tr>
<td>ca</td>
<td>cattle (cattle grazing in last year)</td>
</tr>
<tr>
<td>sotr_1</td>
<td>any soil treatment (tillage and similar) one year ago (0 - none, 1 - in spring or autumn)</td>
</tr>
<tr>
<td>sotr_2</td>
<td>any soil treatment (tillage and similar) two years ago (0 - none, 1 - in spring or autumn)</td>
</tr>
<tr>
<td>sotr_3</td>
<td>any soil treatment (tillage and similar) three years ago (0 - none, 1 - in spring or autumn)</td>
</tr>
</tbody>
</table>

Table 6: The available attributes
Table 2: Possible crops (some combinations are possible for example ba-clgr – winter barley with clover grass undersown).

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Crop</th>
</tr>
</thead>
<tbody>
<tr>
<td>ba</td>
<td>winter barley</td>
</tr>
<tr>
<td>be</td>
<td>beets/carrots</td>
</tr>
<tr>
<td>cc</td>
<td>catch crop</td>
</tr>
<tr>
<td>ch</td>
<td>clover</td>
</tr>
<tr>
<td>chgr</td>
<td>chicory+grass</td>
</tr>
<tr>
<td>clgr</td>
<td>clover+grass</td>
</tr>
<tr>
<td>fa</td>
<td>fallow</td>
</tr>
<tr>
<td>gr</td>
<td>grass</td>
</tr>
<tr>
<td>le</td>
<td>leeks</td>
</tr>
<tr>
<td>lu</td>
<td>lupins</td>
</tr>
<tr>
<td>pe</td>
<td>peas</td>
</tr>
<tr>
<td>po</td>
<td>potatoes</td>
</tr>
<tr>
<td>ra</td>
<td>rape</td>
</tr>
<tr>
<td>rd</td>
<td>radish</td>
</tr>
<tr>
<td>rh</td>
<td>rye</td>
</tr>
<tr>
<td>sba</td>
<td>spring barley</td>
</tr>
<tr>
<td>swh</td>
<td>spring wheat</td>
</tr>
<tr>
<td>tc</td>
<td>triticale</td>
</tr>
<tr>
<td>wc</td>
<td>whole crop</td>
</tr>
<tr>
<td>sa</td>
<td>oats</td>
</tr>
</tbody>
</table>

Table 3: The modeled species (springtails).

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Species Group</th>
<th>Species</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itigr</td>
<td>EP-HE</td>
<td>Isotoma tigrina</td>
<td>EU</td>
</tr>
<tr>
<td>Iang</td>
<td>EP-HE</td>
<td>Isotoma anglicana</td>
<td>EU</td>
</tr>
<tr>
<td>Hprr</td>
<td>EP-HE</td>
<td>Ceratophysella denticulata</td>
<td>EU</td>
</tr>
<tr>
<td>Hzra</td>
<td>EP-HE</td>
<td>Cryptopygus thermophilus</td>
<td>HE</td>
</tr>
<tr>
<td>Iang</td>
<td>EP-HE</td>
<td>Isotomurus palustris</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus lanunginosus</td>
<td>HE</td>
</tr>
<tr>
<td>Itang</td>
<td>EP-HE</td>
<td>Tomocerus flavescens</td>
<td>EP</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Itang</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
<tr>
<td>Xarma</td>
<td>EP-HE</td>
<td>Lepidocyrtus cyaneus</td>
<td>EU</td>
</tr>
<tr>
<td>Hsuc</td>
<td>EP-HE</td>
<td>Ceratophysella succinea</td>
<td>HE</td>
</tr>
</tbody>
</table>

3. Experiment setup
In order to estimate the effects of using aggregated target variables models (aggregated models) instead of using a complex set of original target variable models (single-species models) the whole dataset was split according to the total collembola variable into 10 stratified subsets of data. Each case was present in only one data subset. From 10 subsets of data we created a set of 10 learning and 10 testing datasets (each fold was included in 9 learning and one testing dataset – for example fold 3 was a part of learning datasets 0, 1, 2, 4, 5, 6, 7, 8, 9 and was the testing dataset 3). From those global learning and testing datasets we created...
one set of data for each of the (original and aggregate) target variables. Using Weka [6] (in particular M5’ [5] – a reimplementation of C4.5 [4] and linear regression) and CLUS [1] modeling tools we make regression trees models with constants and linear equations in leaves (using M5’), linear regression models using greedy and M5 attribute selection methods and predictive clustering trees for all original target variable in one tree and for each target variable separately (using CLUS). In this way we created models for each classifier, each target variable and each of the 10 different datasets (and for the trees 10 different settings), ending up with over 10000 models. Since every sample was in one of the ten test datasets, we get a prediction for each target variable for each sample and for each classifier. Using the predictions for the original target variables (species) we calculated predictions for aggregated variables. Comparing both the predicted (straight from the models) aggregate variables and the calculated (from original variables models) aggregate variables with the original values of aggregate variables (calculated from the measured values of the original values) Pearson correlation coefficient, relative absolute error and root relative square error were calculated. Based on these models’ performances, the comparison between models and modeling approaches was done.

4. Results

Looking at the Pearson correlation coefficient of the aggregate variables calculated from the models of single species (CalcVsReal in figures) and the Pearson correlation coefficient of the aggregate variables predicted from the models (SingleVsReal in figures) (Figures 1-4), we can see that in the most cases the maximum correlation achieved by the different classifiers on different target aggregate variables, does not vary much between calculated and predicted aggregate variables. Usually there is difference in favor of aggregate variables predicted from one model over the ones calculated from an outcome of a set of models.

![Fig. 1: The maximum correlation of aggregate variables (calculated and) predicted by model trees.](image-url)

The performance of M5’ model trees (that is regression trees with linear models in leaves) (Fig.1) shows that using one model gives better results then using a set of models for predicting all target aggregate variables. According to the results presented in Fig. 1, the aggregate variable EP has the lowest correlation coefficient. Looking at the different performance measures (relative average error (Figure 5), root relative square error and also for other classifiers), the relative errors are more or less the same for aggregated variables which are the sum of species measurements/predictions. However the relative errors of two
aggregate variables that are calculated in another way (the species richness and biodiversity) are much bigger. The results show that summing the outcomes of models can reduce the errors of the models, while other more complex equations do not have that effect. The biggest difference occurs when modeling species richness, which is the count of occurring (non-zero) species. This is understandable, since only slight errors in the species models around zero point can heavily influence the outcome of the calculated species richness.

Fig. 2: The maximum correlation of aggregate variables (calculated and) predicted by regression trees.

Fig. 3: The maximum correlation of aggregate variables (calculated and) predicted by linear regression with greedy attribute selection.

Using regression trees (with constants in the leaves) gives greater differences (Figure 2) than model trees. The differences increase when complex equations are used and when we try to predict more demanding variables (EP). On the other hand the results of linear regression with greedy attribute selection (using M5 attribute selection does not change results) gives similar results to using model trees. But Figure 3 shows for the first time that using a set of models for predicting the group of euedaphic collembolans gives slightly better performance than a single-species model.
The maximum correlation of aggregate variables (calculated and) predicted by CLUS single prediction trees.

The relative average error of aggregate variables (calculated and) predicted by model trees.

The evaluation of the results achieved with CLUS show that the models for prediction of all species and all aggregate variables with one tree gives are deformed trees with only one leaf, therefore we used CLUS to predict one variable at one time. The results of CLUS (Figure 4) show that the biggest difference occurs when variables aggregated by using more complex equations like biodiversity and species richness are predicted. This pattern is the same for regression trees. Like with the linear regression models, some of the results show slight benefits of using a set of models instead of one model. However those differences are small, and much smaller than the differences that can benefit one model over a set of models.

5. Conclusions

This research was conducted to estimate the lose of performance when single species models for prediction aggregate variables (e.g. biomass) and the aggregated models with the aggregated variables calculated from the original data were applied on the same dataset. The evaluation of models' performance
based on real dataset with 40 original target variables, from which we calculated 8 aggregate target variables. Original dataset was split into 10 folds, and we run 10-fold cross-validation for every target variable (original and aggregate) model. From the predictions of the original target variable models the aggregate variables were calculated. They were compared with the original aggregated data. Based on these comparisons we may conclude that using the aggregated data is often better than using a set of single species models.

If data are aggregated first then the measurement errors are reduces more than modelling (and measurement) errors are reduced by the aggregating the results of single species models. However, on some occasions using the set of models produces slightly better results, but their performance measures can not justify this modeling approach compared to the application of models created on aggregated data. Therefore in most cases first calculating the aggregate variables and then learning models to predict these directly, yields better results than modeling individual species and then calculating the aggregate variables from the predictions of these models.

Acknowledgements
This work is supported by the ECOGEN project funded by the Fifth European Community Framework Programme: Quality of Life and management of living resources contract no QLK5-CT-2002-01666 and DARCOF, Nature quality in organic farming.

Bibliography