UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science (BCA)

Siim Karel Koger

Dark diversity estimation based on a single matrix of binary observations

Bachelor's Thesis (9 ECTS)

Supervisors: Ardi Tampuu
Supervisor: Raul Vicente Zafra
Supervisor: Carlos Pérez Carmona

Tartu 2020

Dark diversity estimation based on a single matrix of binary observations

Abstract:

Ecological theory and nature conservation have traditionally relied solely on observed local diversity. However, the biodiversity of a site also includes the absent species that are present in the surrounding region and can potentially inhabit the site's particular ecological conditions. These unobserved species constitute the "dark diversity" of the site. Dark diversity is by definition unobservable and can only be estimated - in binary fashion or as a degree of certainty about species membership.

This thesis compares the effectiveness of several implementations of non-negative matrix factorization **(**NMF) and basic autoencoders (a type of artificial neural network) to generate probabilistic values about a species' membership in a specific site based on a single matrix of binary observations.

We find that it is possible to generate a suitability matrix that is highly correlated with the underlying suitability values with both methods and that using autoencoders specifically for dark diversity predictions has a lot of potential for even more future improvements.

Keywords:

Dark diversity, non-negative matrix factorization, scikit-learn, autoencoder, Keras

CERCS:

P170 Computer science

Tumeda elurikkuse hindamine üheainsa binaarsete väärtustega vaatlus maatriksi põhjal

Lühikokkuvõte:

Ökoloogiliste süsteemide teooria ja looduskaitse teadus on traditsiooniliselt tuginenud lokaalsele mitmekesisusele. Bioloogilise mitmekesisuse alla kuuluvad aga ka liigid, kes potentsiaalselt sobiksid konkreetsesse kooslusesse, aga miskipärast siiski seal ei esine - neid puuduvaid liike nimetame liigifondi "tumedaks elurikkuseks". Tume elurikkus ei ole definitsiooni poolest

mõõdetav, kuid seda on võimalik määrata läbi liigifondi määratlemise; seda on võimalik hinnata nii binaarselt kui ka tõenäosuslikult.

Käesolev bakalaureusetöö võrdleb mitut implementatsiooni mitte-negatiivsest maatriksi faktoriseerimisest (eng. *NMF*) ja kolme erineva arhitektuuriga autokooderit tumeda elurikkuse tõenäosuslikkuse määratlemiseks põhinedes ühele vaatlus maatriksile (mis ütleb kas liik eksisteerib teatud piirkonnas või mitte).

Töö käigus leiame, et mõlema uuritud meetodiga on võimalik genereerida maatriks, mis on tugevas korrelatsioonis tõeste sobivus väärtustega ning just autokooderis leidub suur potentsiaal edasistes uuringutes veelgi paremate tulemuste saavutamiseks.

Võtmesõnad:

Tume elurikkus, mitte-negatiivne maatriksi faktorisatsioon, scikit-learn, autokooder, Keras

CERS:

P170 Arvutiteadus

Contents

1. Introduction

The biodiversity of a site consists of species present, but also includes the absent species that are present in the surrounding region and can potentially inhabit those particular ecological conditions. These unobserved species were branded as “dark diversity” (Pärtel et al., 2011). Dark diversity can be used, for example, to counteract biodiversity loss and to estimate the restoration potential of ecosystems, and relating local and dark diversities enables comparisons between regions, ecosystems, and taxonomic groups, and to evaluate the roles of local and regional processes in ecological communities. (Bennett and Pärtel, 2017; Carmona et al., 2019).

Unlike present species, dark diversity is by definition unobservable and can only be estimated. Such estimations can be done either in a binary fashion, where species are either ascribed (1) or not (0) to dark diversity or in a probabilistic way in which the degree of certainty about species membership to dark diversity is assessed.

Many of the methods designed to predict dark diversity do it in a binary fashion, however, they require a threshold to filter which species are included in dark diversity. The selection of thresholds remains arbitrary, can affect the results, and is often difficult to justify (Karger et al., 2016; Lewis et al., 2016).

Dark diversity is better defined as a degree of certainty about a species potential membership in a site, but even though arguments for using probabilistic instead of binary approach has long been recognized, methods using this approach are only recently being developed (Carmona et al., 2019; Pärtel et al., 1996).

1.1. Contribution

This study aims to advance the development of probabilistic methods to estimate dark diversity by using two well-known methods from the field of statistics and machine learning, non-negative matrix factorization and autoencoders, to generate a probabilistic dark diversity matrix from a single species presence-absence matrix. Methods’ results are compared to each other.

1.2. Thesis Outline

Section 2 of the thesis gives an overview of dark diversity and habitat suitability and distribution modeling. It also gives an overview of other similar problems that were used as an inspiration to

solve the problem covered in this thesis. Section 3 covers the methods used in previous works on dark diversity and explains the data, techniques used, and why these were chosen. Section 4 shows the results of the methods used and compares them. The conclusion and future prospects are covered in Section 5.

2. Related work

This section covers the past works related to the subject of this problem: **dark diversity**. It also discusses how the problem introduced in the thesis may be approached as a kind of a recommendation system and introduces some methods used for recommendation systems that give ideas how to approach the particular challenge tackled in this thesis.

2.1. Dark diversity

Relating observed and unobserved species enables comparisons between different sites, ecosystems, and taxonomic groups, and to evaluate the roles of local and regional processes in ecological communities (Pärtel et al., 2011).

The term "**dark diversity**" was coined in the 2011 paper "Dark Diversity: Shedding Light on Absent Species". It describes the species in an ecosystem that are absent during an observation but could potentially inhabit those particular ecological conditions (Pärtel et al., 2011). For example, if a bird community in a specific region has been sampled, dark diversity includes all bird species from adjacent regions that are currently absent in the study site but can potentially disperse to and colonize it. If the focal site is in a forest, many of the regional species that are typical from forest habitats will probably be part of the local dark diversity. Conversely, many of the species typical from crops are unlikely to be part of the dark diversity of our forest site.

Many of the methods used for predicting dark diversity do it in a binary fashion - either a species is present (1) in the dark diversity of a study site or not (0). Binary classification requires thresholds to be chosen to decide which species to include in dark diversity and despite the efforts to make this procedure trustworthy, the selection remains arbitrary and is often difficult to justify (Carmona et al., 2019; Karger et al., 2016; Lewis et al., 2016).

Since dark diversity is unobservable, a probabilistic estimation about our degree of belief that a species belongs to it makes more sense than a binary one. One of the methods (and the one that works the best) for this approach is using **species co-occurrence patterns**, leaning on the knowledge that species that share similar ecological requirements are bound to co-occur a site. Some of the implementations that use this method are Beals index, a favorability transformation (further development based on Beals), and hypergeometric distribution (Carmona et al., 2019).

2.1.1. Beals index

This method assigns to each study site and species the probability of the species being present, which is computed by combining information of the species actually found in the community and their patterns of co-occurrence with the species in interest (Carmona et al., 2019).

The Beals probability index that a species ***j*** is present in a community ***i*** can be estimated as follows (Lewis et al., 2016; Münzbergová and Herben, 2004):

$$P_{ij} = \frac{1}{S_i - I_{ij}} \sum_{k \neq j} \frac{N_{jk} I_{ik}}{N_k}$$

where

- S_i is the number of species at community ***i***.
- I_{ij} is the incidence (0, 1) of species ***j*** at community ***i***.
- N_{jk} is the number of joint occurrences of species ***j*** and ***k***.
- I_{ik} is the incidence (0, 1) of species ***k*** at site ***i***.
- N_k is the number of occurrences of species ***k***.

The probabilities predicted by the Beals index are correlated with the frequency of the species in the considered dataset. This is problematic because the fact that a certain species is rarely observed in a site is not an indicator that the species is not part of the dark diversity. The probability that a species will be observed in a site where it is currently absent depends on the suitability of the local conditions and factors related to dispersal (regional frequency, dispersal ability) (Carmona et al., 2019).

Accordingly, the Beals index has been used in studies aiming to predict species appearances in the near future without distinguishing habitat suitability (Karger et al., 2016). When studying dark diversity, however, we are only interested in species suitability. This has led some authors to recommend setting a species-specific probability threshold, which effectively creates a binary index and lacks the preferred notion of dark diversity (Beals, 1984; Carmona et al., 2019).

2.1.2. Favorability index

An alternative to the Beals index that avoids thresholding and makes the probabilities independent of species frequency is the favorability index proposed by (Real et al., 2006). It is a way to correct the raw Beals index so that predictions for each species become independent from the prevalence of the species in the dataset. As mentioned, Beals predicts higher probabilities for species with higher prevalence, but this should not have an effect on predicting probabilistic suitability for dark diversity as we want to know to what degree a site is "suitable" for a species, regardless of how frequent the species has been in observations.

2.1.3. Hypergeometric distribution

For each pair of species, their realized number of co-occurrences with random expectations can be compared. Let us consider two species ***i*** and ***j***; the probability ***M*** that they co-occur in a number of sites is given by the mass function of the hypergeometric distribution (Griffith et al., 2016):

$$P_{ij=M} = \frac{\binom{n_i}{M}\binom{N-n_i}{n_j-M}}{\binom{N}{n_j}}$$

where ***ni*** and ***nj*** are the total number of occurrences of species ***i*** and ***j***, respectively, and ***N*** is the total number of sites sampled. The mean of this distribution denotes the expected number of co-occurrences between species ***i*** and ***j***:

$$\overline{M_{ij}} = \frac{n_i n_j}{N}$$

If the number of actual co-occurrences is greater than expected by chance, the two species are positively associated, and vice versa. We can estimate this departure from expected simply by subtracting mean from actual co-occurrences:

$$ES_{ij} = M_{ij} - \overline{M_{ij}}$$

ES, however, does not convey information on the strength of the association between two species. For this, we can estimate standardized effect sizes (SES) by dividing the effect size by the square root of the variance of the hypergeometric distribution (the standard deviation):

$$Var_{ij} = \left(\frac{n_i n_j}{N}\right)\left(\frac{N - n_i}{N}\right)\left(\frac{N - n_j}{N - 1}\right)$$

$$SES_{ij} = \frac{Effect\ size}{\sqrt{Var_{ij}}}$$

SES indicates how many standard deviations the observed number of co-occurrences is from the expected value. They can then be expressed as probabilities (*Pij*) by confronting the SES value with the cumulative normal distribution function with mean=0 and standard deviation=1. Probabilities close to 1 indicate that the two species are positively associated, whereas probabilities close to 0 indicate that the two species are negatively associated; intermediate values denote a random association. This procedure can be applied to all pairs of species to build a symmetric indication matrix reflecting the strength of the association between all species pairs. The indication matrix can then be used to predict the probabilistic dark diversity of a given site (*k*) for which we know the observed diversity. This probability can be estimated for each of the absent species in the site (i.e. all species in the dataset that were not present in the site) simply by averaging the indication values of the species actually present in the community:

$$P_{ki} = \frac{1}{S_k}\sum_{j \neq i}^{S} P_{ij} I_{kj},$$

where *Sk* is the total number of species found in site *k*, *Ikj* reflects the incidence (0, 1) of the indicator species *j* in site *k*, and *S* is the total number of species in the region. Hence, the probability of an absent species belonging to the dark diversity of a site is high if it tends to have positive associations with those species that are present, and negative associations result in a low probability of membership. (Carmona et al., 2019)

2.2. Recommendation systems

The problem covered by this thesis is similar to what one might have while developing a recommendation system. A recommendation system is a subclass of an information filtering system that tries to predict the preference of a subject towards an item - the suggestions relate to various decision-making processes, such as what items to buy, which ads to show to customers, which people could know each other on social media. (Hanani et al., 2001; Resnick and Varian, 1997)

These systems take data of users (their app-usage history, their likes-dislikes, their browser history, etc.) and a product (in case of YouTube - video titles, description, duration, popularity, etc.) and predict what a user might like the most. In the same way, this thesis uses a matrix of species and study sites to predict how likely a species is to be a part of dark diversity (how much would a species that has not yet been observed “likes” the study site).

2.2.1. Collaborative filtering

One approach to the design of recommender systems that has wide use is collaborative filtering. The system generates recommendations using information about rating profiles for different users or items. A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items without requiring an "understanding" of the item itself (Breese et al., 2013, 1998).

2.2.2. Matrix factorization

Matrix factorization is a decomposition of a matrix as the product of two or more matrices with certain properties. In particular, it can be used as a collaborative filtering method used in recommender systems. Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices (Koren et al., 2009). A multitude of matrix factorization approaches have been proposed for recommender systems but the idea behind them stays the same - to represent users and items in a lower-dimensional latent space. In this thesis non-negative matrix factorization (**NMF**) is used.

2.2.3. Non-negative matrix factorization

An implementation of matrix factorization algorithms where a matrix **V** is factored into two (or more) matrices **W** and **H**, with the property that all three matrices have no negative elements. When multiplying matrices, the dimensions of the factor matrices may be significantly lower than those of the product matrix and it is this property that forms the basis of NMF as it can remove the noise and keep the signal (Dhillon and Sra, 2005).

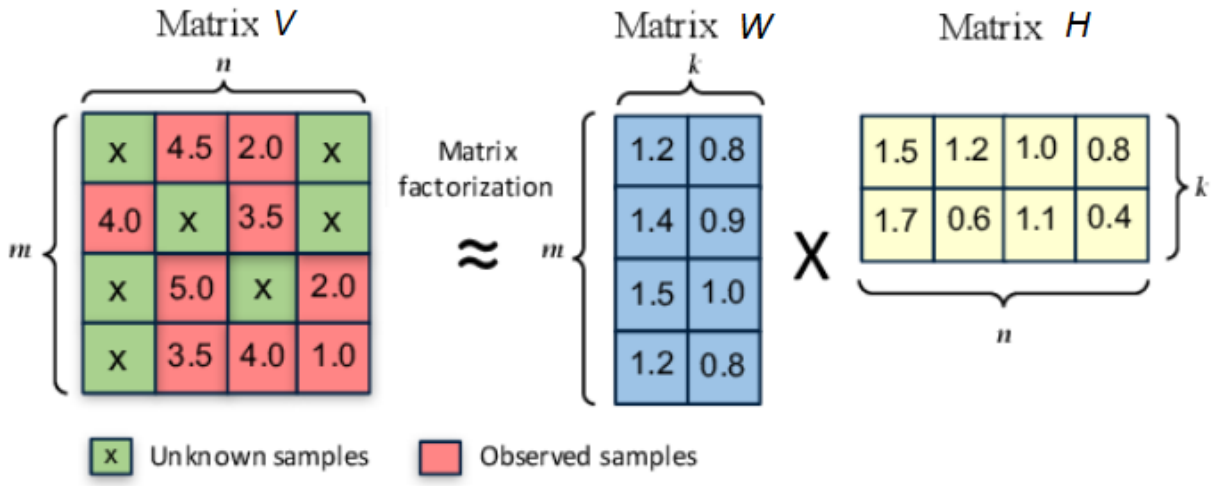

Figure 1. Illustration of approximate non-negative matrix factorization: the matrix **V** is represented by the two smaller matrices **W** and **H**, which, when multiplied, approximately reconstruct **V**.

2.2.4. Autoencoders

Another method used successfully in collaborative filtering is autoencoders. An autoencoder aims to learn, in an unsupervised manner, a representation for a set of data by training the network to ignore signal noise. It has an internal (hidden) layer that describes a code used to represent the input, and it is constituted by two main parts: an encoder that maps the input into the code, and a decoder that maps the code to a reconstruction of the original input. (Charte et al., 2018; Kramer, 1991)

An autoencoder is a neural network that learns to copy its input to its output but performing the copying task perfectly would simply duplicate the signal, and this is why autoencoders usually are

restricted in ways that force them to reconstruct the input approximately, preserving only the most relevant aspects of the data in the copy. (Kramer, 1991; Schmidhuber, 2015)

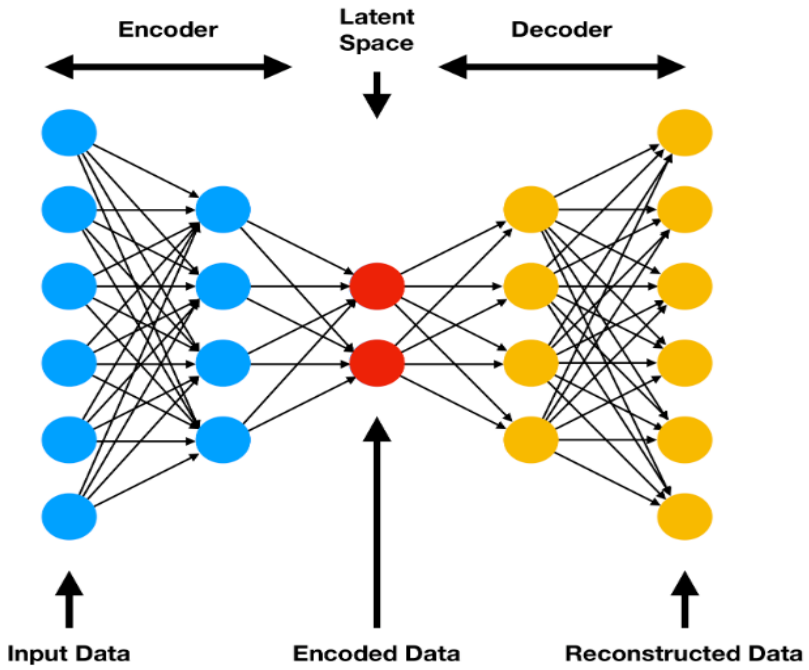

Figure 2. Illustration of the overall structure of an autoencoder. Autoencoder consists of an input layer and output layer which have the same number of nodes. It also has a bottleneck "Encoded Data" layer, which should help with noise reduction and dimensionality reduction. Additional encoder and decoder layers may be added and they are usually in a mirror image from each other.

3. Methodology

This section gives a detailed overview of all the techniques used and the experiments done. It starts by explaining the dataset - how it looks, how it was generated, how it is used in the context of this particular problem, and what is the baseline of the work. Followed by describing the chosen implementations of methods described in the "Related work" section (NMF and autoencoder).

3.1. Proposed methods

The goal of this thesis is to generate a matrix of probabilistic values about a species' likelihood of being a part of the communities. In a real-world scenario, one would only have the observations matrix so when developing a model with any method then only the information about the observation matrix can be used (an example of what an observation matrix would look like is depicted by Table 1). The underlying suitabilities can only be used during testing the models but the information it contains can not be applied to make the model better (Table 2 is a depiction of a possible "true suitability" or "ground truth" matrix, which contains the actual values of every species for every study site).

Table 1. An example of an observation matrix (species presence-absence matrix). The values are either 1 (Sp was observed >= 1 times in Site) or 0 (Sp has not been observed at all in Site)

Observations matrix example (with M species and N sites, where M and N are positive integers)					
	Sp 1	Sp 2	Sp 3	...	Sp M
Site 1	0	0	0	...	1
Site 2	1	0	0	...	0
Site 3	0	0	1	...	0
...	...	...	...	...	...

Site N	1	1	0	...	1

Table 2. An example of an underlying true suitabilities. The values are in range $[0, 1]$.

True suitability (ground truth) example (with M species and N sites, where M and N are positive integers)					
	Sp 1	Sp 2	Sp 3	...	Sp M
Site 1	0.112	0.344	0.054	...	0.966
Site 2	0.671	0.111	0.001	...	0.011
Site 3	0.001	0.555	0.363	...	0.555
...	...	...	...	…	…
Site N	0.423	0.899	0.011	...	0.873

Table 3. An example of a generated suitability matrix. The values are in range $[0, 1]$.

Generated suitabilities example (with M species and N sites, where M and N are positive integers)					
	Sp 1	Sp 2	Sp 3	...	Sp M
Site 1	0.002	0.744	0.014	...	0.999
Site 2	0.571	0.021	0	...	0.011
Site 3	0.011	0.453	0.211	...	0.444
...	...	...	...	...	...

Site N	0.323	0.677	0.877	...	1

This thesis proposes two methods to reproduce the underlying true suitability matrix. First is non-negative matrix factorization - a version of the matrix factorization algorithm with only positive generated values for each cell of the matrix. The second method is using autoencoders which are widely used in recommender systems (Zhang et al., 2020). Autoencoders are in a way a generalization over NMF, which is forced to essentially use one matrix multiplication to go from input to lower-dimension representation, and one multiplication to go back; autoencoders can get complex quickly but in this thesis a version of it is presented which should be the easiest architecture to find non-linear patterns.

Both methods are most often used to regenerate the original matrix to fill in the missing values. However, in this thesis' case the matrix should not be regenerated very accurately as this is not the goal here; we want to generate a matrix (Table 3 depicts a possible generated matrix) that is as close as possible to underlying true suitabilities rather than the binary observation matrix itself. We restrict the perfect input regeneration in ways that the reconstruction of input is only an approximation, preserving only the most relevant aspects of the data in the copy.

The predictability is tested by measuring Spearman correlation (to see how well does the model find features between species and habitation zones) and mean absolute error (to see how much the model under- or overestimates the values) between true suitability and generated suitabilities.

3.2. Dataset

Probabilistic suitability can by definition only be predicted; the underlying suitability (as depicted by **Table 2**) is in reality present only on idea level and can not be measured. It is useful to train models on simulated data, because in that case the generated suitability matrix can be compared against the true underlying suitability, which was used for the simulations.

The data used in this thesis came from using a virtual landscape created by (Carmona et al., 2019) containing different habitats and a set of species with different suitability for these habitats and which allows communities to develop by following some simple rules for a period of time. This simulation is in turn based on (Jõks and Pärtel, 2019). The difference between the data used in

their paper and this thesis is just the amount of data generated (more cells, smaller plots, and more species).

For this thesis, a **500 x 500 grid** was created and divided into **10000 plots, each encompassing 5 x 5 cells.** Cells can either contain an individual or be empty. Individuals acted according to simple rules that corresponded to some of the basic processes that determine diversity. Among these processes, selection depended on the suitability of each species to each plot. For this, the same value was assigned for the environment to all the cells in the same plot, which was drawn from a normal distribution with $\mu=0$ and $\sigma=5$. **A set of 1000 species** was created, with each species having an optimal value in the environment drawn from a uniform distribution from -10 to 10; all individuals of a species had the same value (i.e. there was no intraspecific variability). Once these values were assigned, the distance between each community's environment and each species optimum was estimated, considering the environment as a circular variable. Suitability indicates how close an environment is to the optimum of a given species; suitability was 1 when the environment value in the plot was equal to the species optimum and decreased towards 0 as distance increased (following a normal distribution). Simulations started with an empty grid (no individuals present) and were run for 5250 sequential cycles. In each cycle, the same processes took place as described in (Carmona et al., 2019).

The problem that this thesis is trying to tackle begins with a simulated species presence-absence (observation) matrix and an underlying suitability matrix using which the presence-absence matrix was generated. The values in the presence-absence matrix are binary - 0 if the species has not been observed in the particular site and 1 if it has been observed one or more times (as depicted in **Table 1**., if **M = 1000** and **N = 10000**). The values in the underlying true suitability matrix are real number values in range $[0, 1]$ (as shown in **Table 2**., if **M = 1000** and **N = 10000**).

3.3. Baseline

The first baseline for this thesis are the previous results from the study by (Carmona et al., 2019), where the same type of simulation was used for data generation. In that work a hypergeometric distribution method was used for dark diversity prediction. However, the simulated worlds are of different size and also other parameters differ between that work and the work presented here. As such, we can only qualitatively compare with this baseline, verifying the results are at least of

similar quality. Direct comparison of performance metrics does not make sense, as the underlying data is different.

The second baseline is not a model, but just a sanity check. We compare the unedited observation matrix against the underlying true suitabilities. If this would give a better result than the matrices generated by models proposed by this thesis, we'd certainly know the models are not effective.

3.4. NMF

For this thesis non-negative matrix factorization instead of normal MF was chosen as it doesn't make sense for our data to be represented in negative values (generated probabilities must stay in range **[0, 1]**).

3.4.1. Scikit learn

Python library Scikit-learn includes an implementation of NMF (sklearn.decomposition.NMF). During hyperparameter search different ranks and max-iteration values were tested; tolerance of stopping condition stayed at default 10e-6.

Using Scikit-learn's NMF.

```
factorizer = NMF(init='nndsvd', n_components=rank, max_iter=max_iter_num, solver="cd", tol=10e-6)
W = factorizer.fit_transform(observation_matrix)
H = factorizer.components_
estimated_matrix = np.dot(W, H)
```

3.4.2. Nimfa

Nimfa is a Python library for NMF. It includes implementations of several factorization methods, initialization approaches, and quality scoring. Both dense and sparse matrix representations are supported. In this thesis three different implementations of NMF from Nimfa were tested.

NMF itself is not really designed with a binary input matrix in mind so Nimfa was chosen as an alternative to the basic implementation of Scikit-learn. From the library ICM, BMF and Standard NMF were tested.

3.5. Autoencoders

Python deep learning library Keras was used to build autoencoders for this thesis. Three different versions of autoencoders were built and tested (described below on **Figure 3.-5.**); let's call them **A**, **B** and **C** respectively. All three use **sigmoid activation function** on the output layer, **linear activation function** on the middle layer, **binary cross entropy** as the loss function, **adadelta** as gradient descent optimization algorithm, and for all of them, bias is not used. During training the input vectors were presented in mini-batches and the distribution of vectors in the batches was random.

An input vector represents a single habitation site with each node representing an observation about specific species (if the species was observed or not); the output layer represents the predicted suitability for every species in this same site. Input and output layers are the same size - 1000 nodes.

During hyperparameter search we played around with the number of nodes in the bottleneck (middle) layer (2 to 4 nodes to match the ranks of NMF-s that gave the best results; more in "Results" section), the number of epochs trained and the batch size; all of which were chosen empirically.

3.5.1. Autoencoder A (shallow linear)

The first autoencoder is as basic as it gets. **A shallow linear network** with just the input, output and middle "bottleneck" layers. As the application of autoencoders for dark diversity problem is among newer approaches, there were no expectations for their performance so as simple architecture was chosen as possible while still trying to replicate what NMF might do.

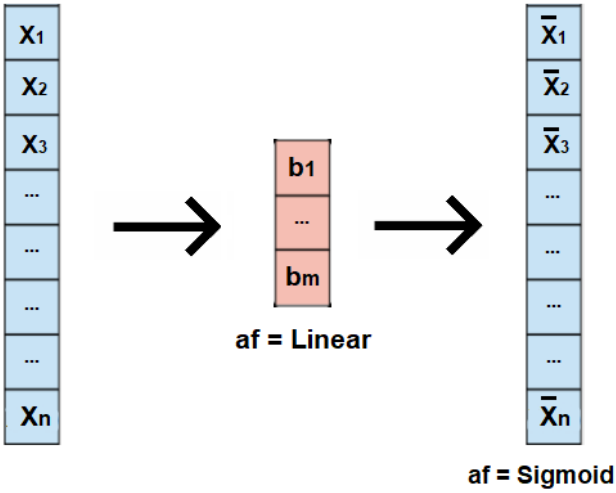

Figure 3. af=”activation function”; X is an array of species’ suitability values $xi \in X$, and $\bar{X}$ is an array of generated suitability values $\bar{xi} \in \bar{X}$ where $i \in [0, n]$ and $n = 1000$. bj is a value of j’th node in the middle layer where $j \in [0, m]$; experiments were done with values $m \in [2, 4]$.

Using Keras’ Sequential class.

```
autoencoder = Sequential()
autoencoder.add(Dense(encoding_dim, activation="linear", input_shape=(input_dim,), use_bias = False))
autoencoder.add(Dense(input_dim, activation="sigmoid", use_bias = False))
autoencoder.compile(metrics=['accuracy'], loss='binary_crossentropy', optimizer='adadelta')
```

3.5.2. Autoencoder B (deep linear)

Compared to **autoencoder A,** this autoencoder uses two additional hidden layers with an empirical choice of 32 nodes each. Linear activation function was used for the first and third hidden layers.

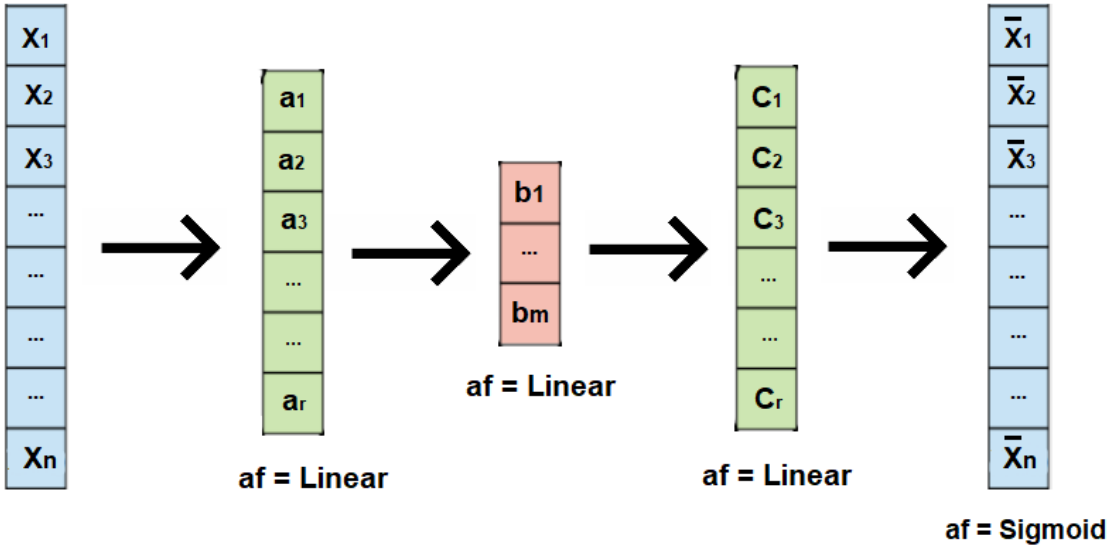

Figure 4. af=”activation function”; X is an array of species’ suitability values $xi \in X$, and $\bar{X}$ is an array of generated suitability values $\bar{xi} \in \bar{X}$ where $i \in [0, n]$ and $n = 1000$. bj is a value of j’th node in the middle layer where $j \in [0, m]$; experiments were done with values $m \in [2, 4]$. ak is a value of k’th node in the first middle layer and ck is a k’th value in the third middle layer where $k \in [0, r]$ and $r = 32$.

Using Keras’ Sequential class.

```
autoencoder = Sequential()
autoencoder.add(Dense(32, activation="linear", input_shape=(input_dim,), use_bias = False))
autoencoder.add(Dense(encoding_dim, activation="linear", use_bias = False))
autoencoder.add(Dense(32, activation="linear", use_bias = False))
autoencoder.add(Dense(input_dim, activation="sigmoid", use_bias = False))
autoencoder.compile(metrics=['accuracy'], loss='binary_crossentropy', optimizer='adadelta')
```

3.5.3. Autoencoder C (deep non-linear)

This autoencoder uses the same hidden layers as the **deep linear autoencoder** with the difference that the first and third hidden layers use **RELU** (rectified linear unit) instead of linear activation function.

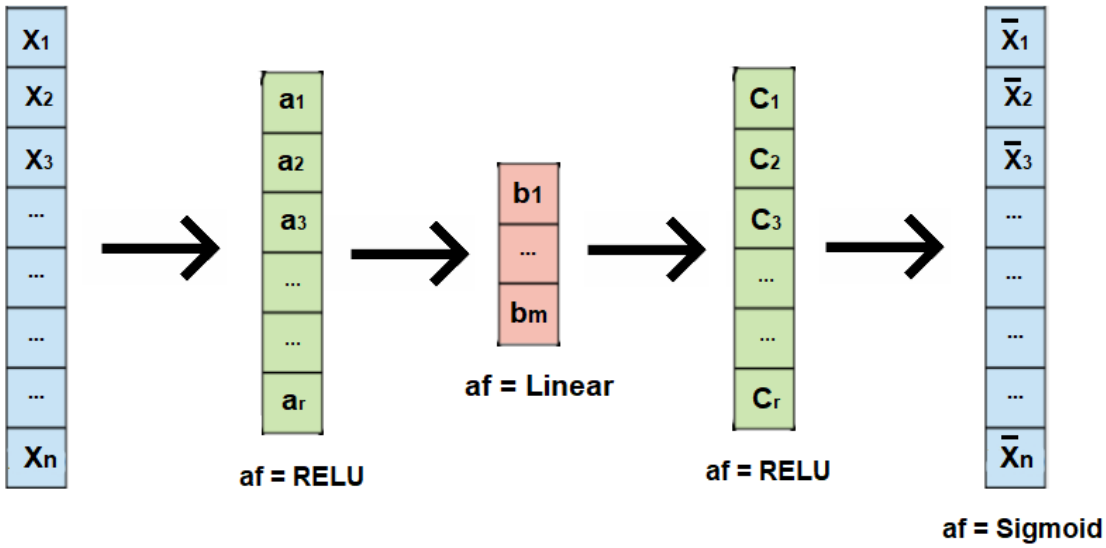

Figure 5. af=”activation function”; X is an array of species’ suitability values $xi \in X$, and $\bar{X}$ is an array of generated suitability values $\bar{xi} \in \bar{X}$ where $i \in [0, n]$ and $n = 1000$. bj is a value of j’th node in the middle layer where $j \in [0, m]$; experiments were done with values $m \in [2, 4]$. ak is a value of k’th node in the first middle layer and ck is a k’th value in the third middle layer where $k \in [0, r]$ and $r = 32$.

Using Keras’ Sequential class.

```
autoencoder = Sequential()
autoencoder.add(Dense(32, activation="relu", input_shape=(input_dim,), use_bias = False))
autoencoder.add(Dense(encoding_dim, activation="linear", use_bias = False))
autoencoder.add(Dense(32, activation="relu", use_bias = False))
autoencoder.add(Dense(input_dim, activation="sigmoid", use_bias = False))
autoencoder.compile(metrics=['accuracy'], loss='binary_crossentropy', optimizer='adadelta')
```

4. Results

The following section goes over the results obtained from applying the methods described in the "Methodology" section. How the results were measured and how a recalibration was used to better the results.

4.1. Testing

The goal of this thesis was to generate a matrix which describes a probability of species being part of the dark diversity of a certain site. The results (the generated suitability matrix; like depicted Table 3) can be compared against the true suitability (as depicted by Table 2) matrix using MAE (mean absolute error) and Spearman correlation. MAE will show how much the models under- or overestimate the predicted values. Spearman is chosen because we have no presumption of the error distribution and the linearity of the relationship between predicted and underlying values.

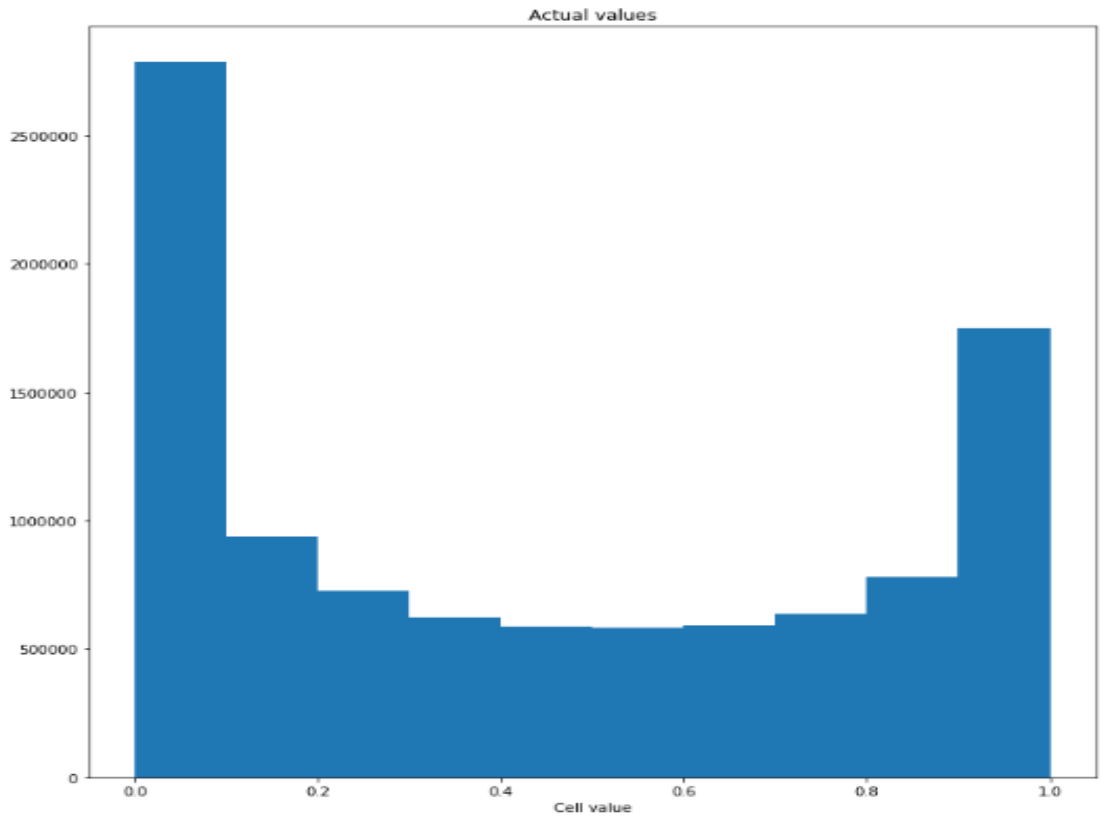

Figure 6. True suitabilities value distribution. The x-axis is divided into 10 equal length sections between numbers 0 and 1. The y-axis shows the number of values that fall into a certain section.

4.2. NMF

The following are the results of Scilearn's NMF models built with parameters as described in the "Methodology" section. Scilearn's NMF always produced better and faster results (correlation and MAE wise) than Nimfa implementations tested (ICM, MBF, Standard NMF), so in this section only Scilearn's results are covered, as they were chosen for deeper analysis and robustness testing. It was determined empirically that NMF models using rank 2 yielded the best results (highest Spearman correlation and lowest MAE). When the rank is increased then the correlation rapidly drops

Table 4. MAE, Spearman correlation, and Pearson correlation of **the most successful NMF model** using different numbers of max iterations. Using rank=2.

Iterations	Mean absolute error	Spearman correlation
1	0.342	0.708
5	0.337	0.756
10	0.330	0.760
25	**0.335**	**0.761**
50	**0.335**	**0.761**
100	**0.335**	**0.761**
1000	**0.335**	**0.761**

Table 5. MAE, Spearman correlation, and Pearson correlation of NMF model with rank=5.

Iterations	Mean absolute error	Spearman correlation
1	0.433	0.701
5	0.432	0.750

10	0.432	0.752
25	0.432	0.750
50	0.432	0.755
100	0.432	0.754
1000	0.432	0.759

NMF with rank=2 is clearly better in MAE than NMF with rank=5 and higher ranks tend towards the basic baseline of just comparing with the observation matrix, because higher rank reconstructs the matrix more accurately. However, this is not necessarily what we want; the reconstruction should be lossy in hopes that some noise is lost and the signal is preserved.

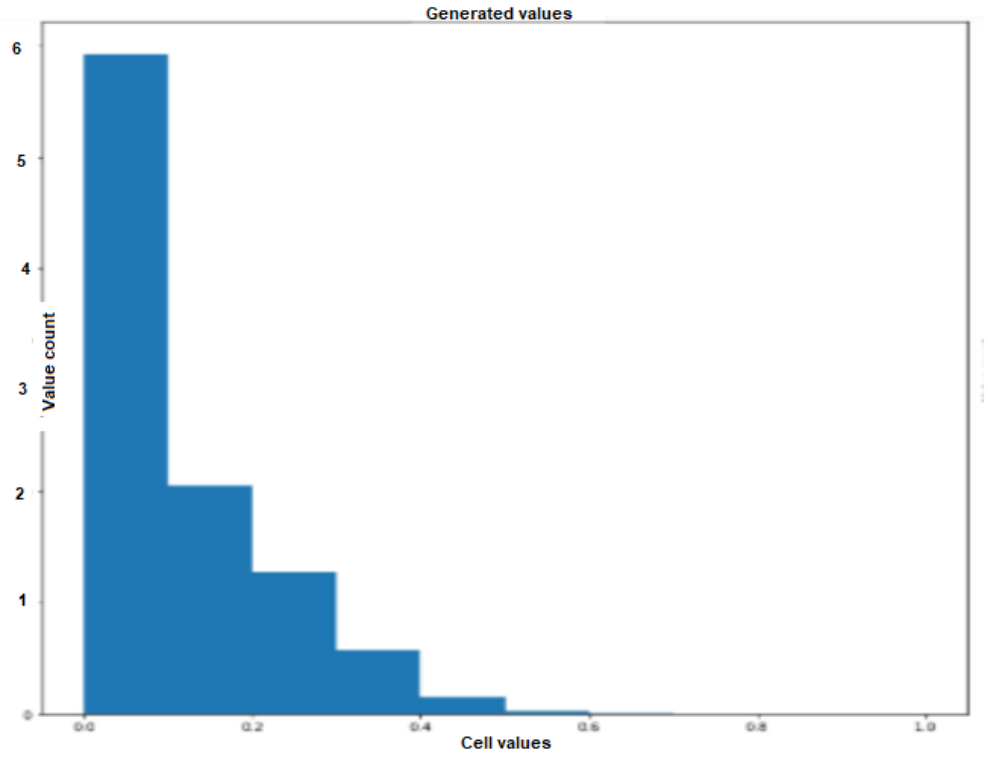

Figure 7. Value distribution of the NMF with best Spearman correlation. Rank=2, max iterations = 1000. Y-axis represents value count in millions.

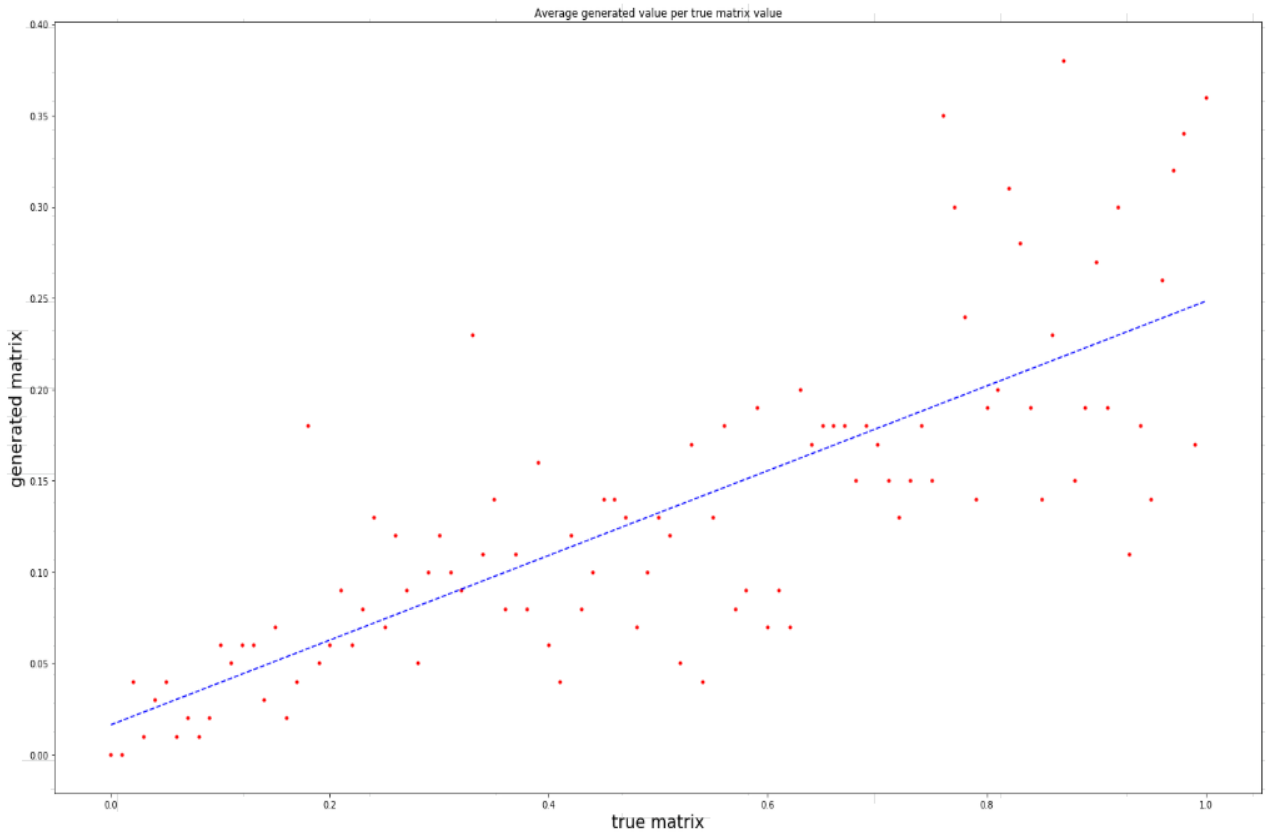

Figure 8. Best case NMF generated values compared against true suitability values. X-axis depicts the true suitability values in the underlying matrix and is divided into 100 equal length parts (0.01 each). Y-axis is the generated values; there are 10mil of them so to make it visually readable only the average values per bin are presented denoted with a red dot. Trendline is drawn in blue

4.3. Autoencoders

Following are the results of the autoencoders A, B, C described in the "Methodology" section. For all the results below, batch size 64 was chosen empirically, as this seemed to give the best result for most runs; also 32, 50, 100, 500, 1000 were tested - the first three gave very similar results to 64; the bigger the batch size the more epochs were needed to reach the similar level of Spearman correlation as the best combination of batch size 64 and epochs 3000.

Similarly, for all the results below the middle layer has 2 nodes, as it always gave better results than with 3 or 4 nodes.

Table 6. MAE (mean absolute error) and Spearman correlation of autoencoder A outputs with the true suitabilities.

Epochs	Mean absolute error	Spearman correlation
50	0.430	0.117
100	0.426	0.170
500	0.422	0.244
1000	0.419	0.552
3000	0.417	0.649

Table 7. MAE (mean absolute error) and Spearman correlation of autoencoder B outputs with the true suitabilities.

Epochs	Mean absolute error	Spearman correlation
50	0.422	0.207
100	0.425	0.190
500	0.420	0.553
1000	0.420	0.563
3000	0.425	0.550

Table 8. MAE (mean absolute error) and Spearman correlation of autoencoder C outputs with the true suitabilities.

Epochs	Mean absolute error	Spearman correlation
50	0.421	0.354
100	0.421	0.405

500	0.421	0.431
1000	0.418	0.843
3000	0.415	0.868

The most complex autoencoder, type C, gave better results for both MAE and Spearman correlation compared to architectures A and B when comparing models which use the same number of epochs and batch size for training. For further analysis only the best autoencoder architecture (model C) was chosen.

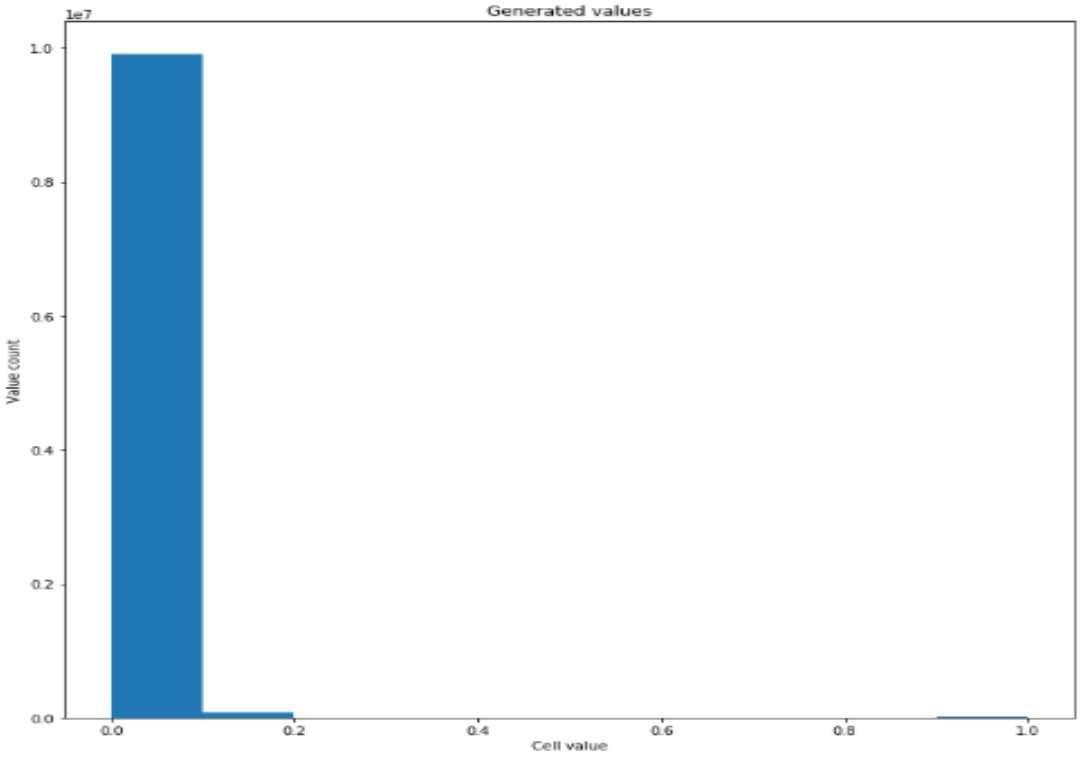

Figure 9. Autoencoder C value distribution. 3000 epochs, batch size 64. This distribution is very skewed towards the low values, as compared to the true suitability distribution (cf. Figure 1)

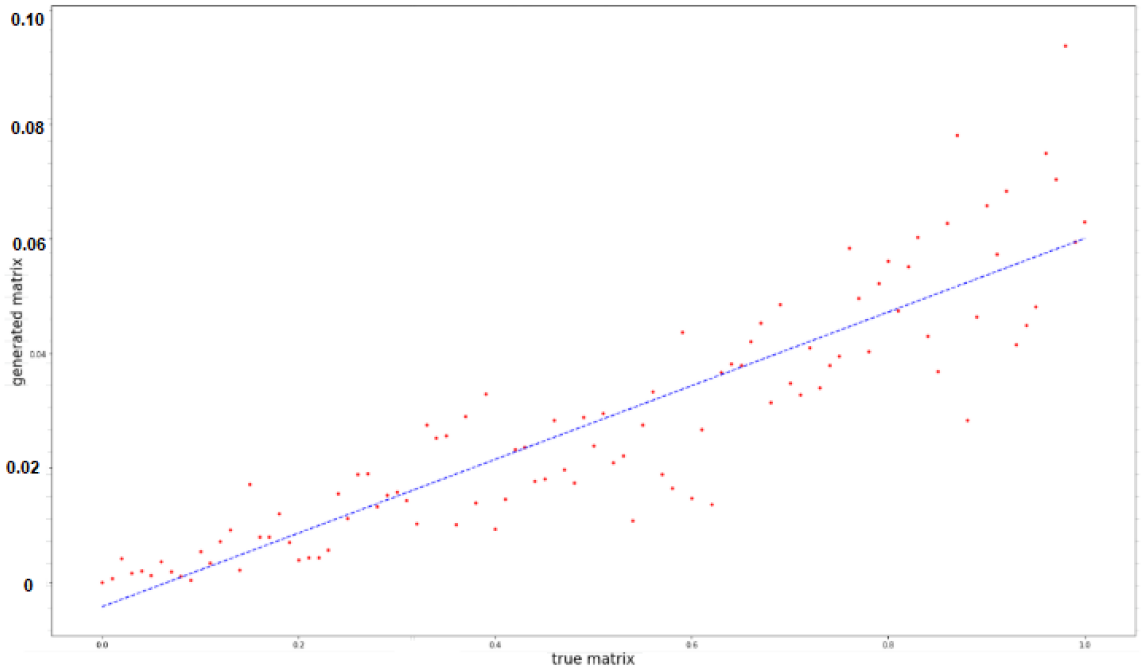

Figure 10. Autoencoder C generated values compared against true suitability values. X-axis depicts the true suitability values in the underlying matrix and is divided into 100 equal length parts (0.01 each). Y-axis is the generated values; there are 10mil of them so to make it visually readable only the average values per bin are presented denoted with a red dot. Trendline is drawn in blue

4.4. Rescaling predictions

The best results of autoencoders were considerably better than best cases of NMF when it came to **Spearman correlation** (**0.868 and 0.761 respectively**). However, according to **MAE** the best case of NMF outperformed the best autoencoder (**0.335 and 0.419 respectively**). In reality this means that autoencoders allow better ordering of which species is more suited to an environment, but NMF probably predicts the numeric values more accurately. MAE=0.335 of NMF means that the suitabilities predicted are on average 33.5% off. Furthermore, the autoencoder's MAE=0.419 is just a little better than if we were to measure MAE=0.435 of the untouched observation matrix against underlying true suitabilities (aka if we just used the observation matrix).

As seen from the histograms above, the models systematically underestimate the values. The higher the underlying suitability, the larger the underestimation magnitude gets. As seen on **Figure 7**, this relationship is rather linear and it is reasonable to believe that scaling predictions linearly

would solve the problem. However, the underlying true suitability values can not be used for determining the **scaling factor**, as in real life these are not known. For the same reason, we cannot use traditional calibration methods that rely on comparing predictions with the underlying values. The thesis proposes to upscale the generated predictions by dividing the values with the average prediction for the species present and clipping the values to 1.

Following are the results of the same (best) NMF and autoencoder models described above but with results scaled as determined by the proposed method. We can now observe that the MAE has fallen considerably for both NMF and autoencoder, and the value distribution graph looks much closer to underlying suitability distribution.

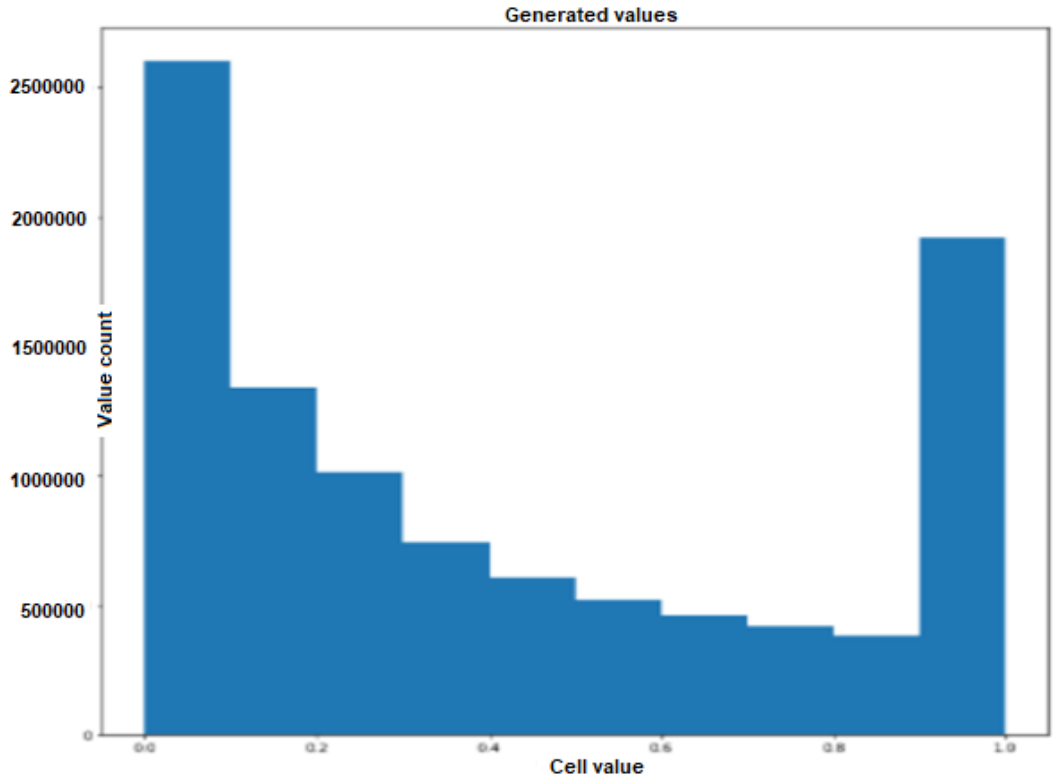

Figure 11. Value distribution of NMF model with the best Spearman correlation **with rescale method applied**. MAE=0.185. Spearman=0.761.

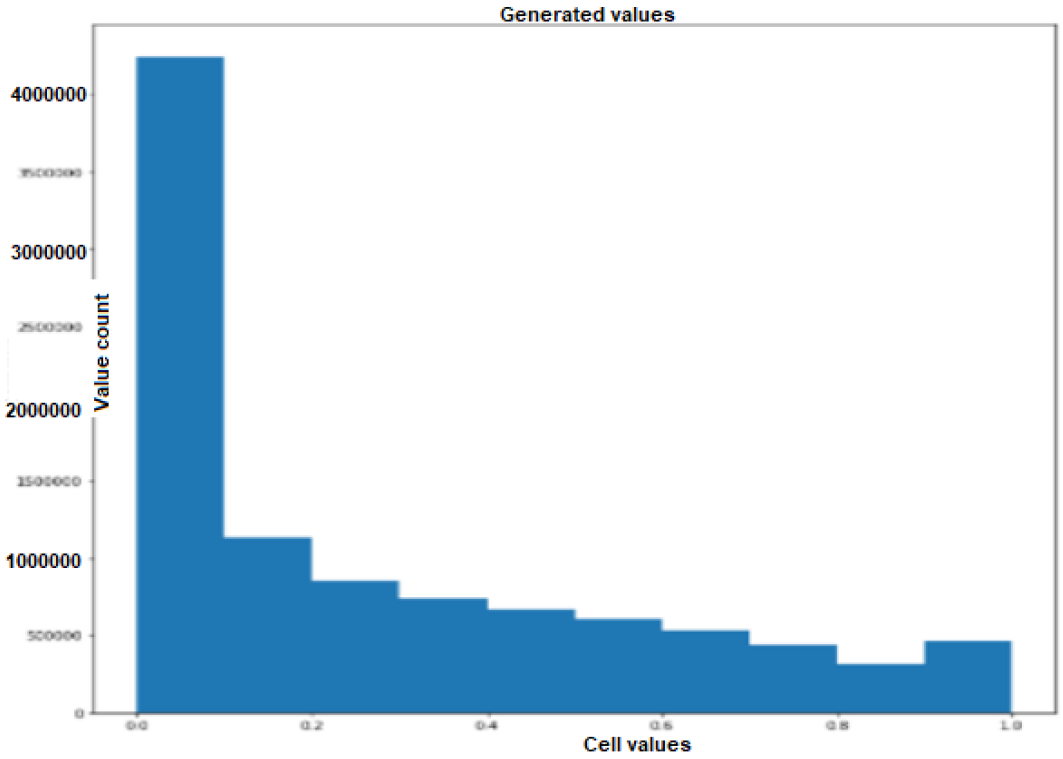

Figure 12. Autoencoder C value distribution **with rescale method applied**. 3000 epochs, batch size 64. MAE=0.192. Spearman=0.867.

After it was confirmed that the rescale method works well, it was discovered by accident that the deep non-linear autoencoder run for **5000 epochs**, after rescaled, gives even better **MAE=0.154** value than the best case rescaled NMF (**MAE=0.185**) and rescaled deep non-linear autoencoder run for 3000 epochs (**MAE=0.192**); **however the correlation drops** a bit from 3000 epochs version **Spearman=0.867 to Spearman=0.856**. As this drop in MAE was discovered so late into the research, no further tweaking took place, but it shows that there is room for even better results.

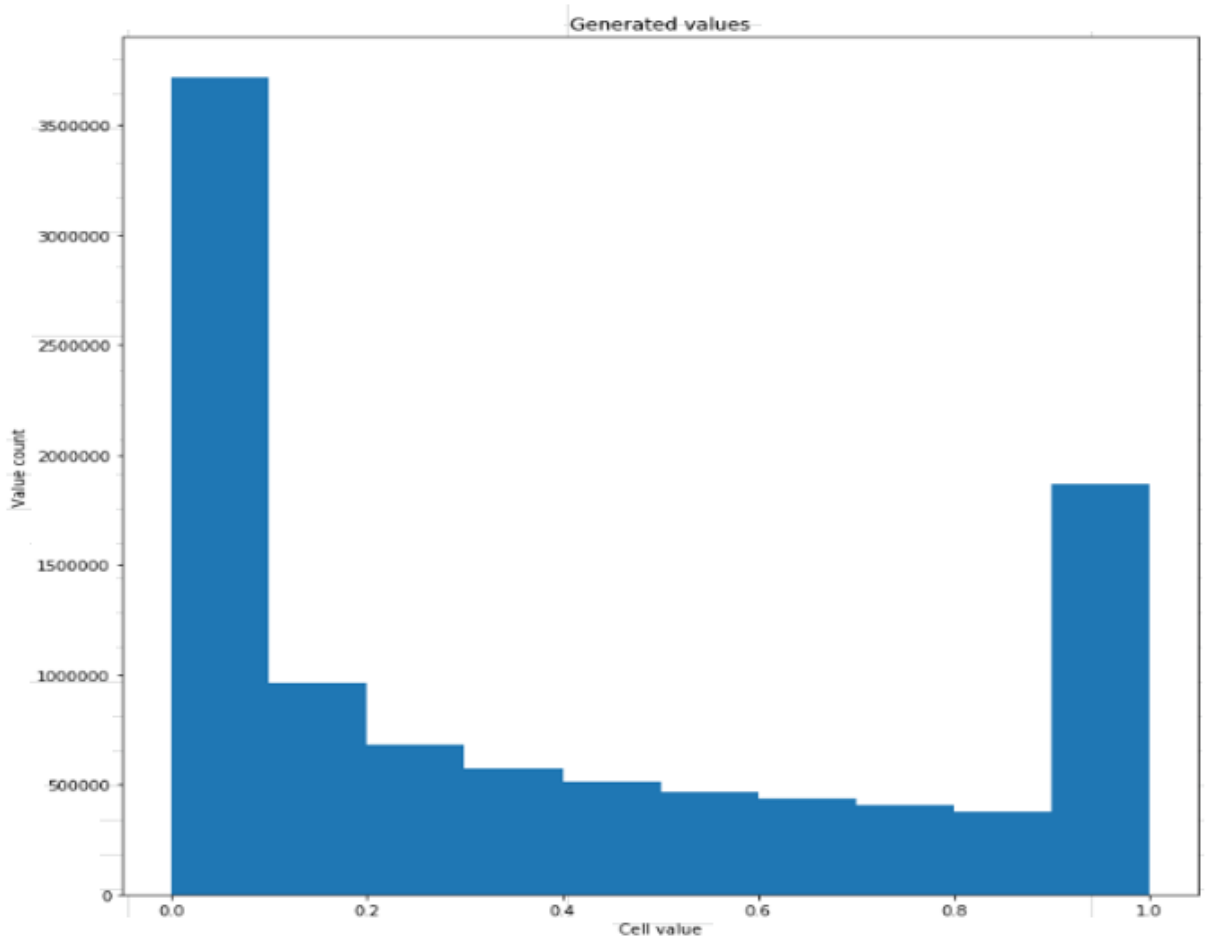

Figure 13. Autoencoder C value distribution **with rescale method applied**. 5000 epochs, batch size 64. MAE=0.154. Spearman=0.856.

4.5. Model robustness and reliability

To check models' generalization 50 different observation matrices were generated (via running a new simulation) from the same underlying true suitability matrix to test the robustness of our approaches (model training and scaling). Repeating the experiments allows us to see if the models are robust enough to discover the underlying suitabilities with similar accuracy regardless of the randomness of which species get established (and are observed).

The results for NMF stayed relatively the same with the **average Spearman of 0.762 and standard deviation of 0.008**. The results for autoencoder stayed relatively the same with the average Spearman of **0.862** and standard deviation of **0.004**. From these figures the autoencoder

results seem more predictable; must be mentioned though, that for NMF all the generated matrices were tested, for autoencoder only 20 of them due to much longer training time.

4.6. Discussion

This thesis was aiming to advance the development of probabilistic methods to estimate dark diversity (the absent part of the site-specific species pool) using machine learning techniques NMF (non-negative matrix factorization) and autoencoders (a type of neural network) which are widely used in recommender systems, especially collaborative filtering. These methods were chosen because we approached the problem as a kind of recommendation system. Predicting dark diversity should not take into account the number of species occurrences but just the species suitability for some habitation zone – using just those methods and the observation matrix we have no bias towards the species that turn up more often during observations.

The thesis showed that it is possible to get highly correlating values with both NMF and autoencoders, but the most complex (deep non-linear) autoencoder did the job considerably better, when it came to correlation with true suitabilities. Without the proposed rescaling method, the best NMF gave considerably better MAE value than the best autoencoder - rescaling works very well for both.

NMF is quite limited in what types of tendencies it can discover in the data; as mentioned above, however, only a few implementations of NMF were tested from Nimfa library, so it's possible that some other implementations (or better chosen hyper parameters) give better results.

The extent to which the three types of autoencoders' results differed, and how well the non-linear deep autoencoder performs, gives a reason to look into this approach more and try out more complex versions of autoencoders (or other collaborative filtering methods) as the architectures tried in this thesis are very simple. More complex architectures might yield even better results in this problem and others. Also, as autoencoders are capable of learning very complex functions then in future works it could be tested if their performance is maintained if the underlying simulation were more complex.

5. Conclusion

This thesis developed machine learning models using non-negative matrix factorization (NMF) and autoencoders (using three different architectures) that predict species probabilistic dark diversity values in a study site from a single binary-valued observation matrix. As expected, the autoencoder with the most complex architecture (cf. Figure 5) gave the highest Spearman correlation value; rescaled (**spearman = 0.856**) or not (**spearman = 0.868**). Without the proposed rescale method (as described in section 4.4), NMF gets lower (better) results for Mean Average Error (**MAE = 0.335**) than autoencoders but when rescale was done then there was one tested autoencoder (cf. Figure 13) which got better MAE (**MAE = 0.154**) than any NMF test **(MAE = 0.185**).

The result show that it is possible to tackle dark diversity problem with well known machine learning algorithms, and it seems there is a huge potential for further development of autoencoders in this field as the ones covered by this thesis had very simple architecture and were still able to get the great results (especially with the rescale) in both measures.

References

Beals, E.W., 1984. Bray-Curtis Ordination: An Effective Strategy for Analysis of Multivariate Ecological Data, in: MacFadyen, A., Ford, E.D. (Eds.), Advances in Ecological Research. Academic Press, pp. 1–55. https://doi.org/10.1016/S0065-2504(08)60168-3

Bennett, J.A., Pärtel, M., 2017. Predicting species establishment using absent species and functional neighborhoods. Ecol. Evol. 7, 2223–2237. https://doi.org/10.1002/ece3.2804

Breese, J.S., Heckerman, D., Kadie, C., 2013. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. ArXiv13017363 Cs.

Breese, J.S., Heckerman, D., Kadie, C., 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering.

Carmona, C.P., Szava-Kovats, R., Pärtel, M., 2019. Estimating probabilistic dark diversity based on the hypergeometric distribution. bioRxiv 636753. https://doi.org/10.1101/636753

Charte, D., Charte, F., García, S., del Jesus, M.J., Herrera, F., 2018. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. Inf. Fusion 44, 78–96. https://doi.org/10.1016/j.inffus.2017.12.007

Dhillon, I.S., Sra, S., 2005. Generalized nonnegative matrix approximations with Bregman divergences, in: Proceedings of the 18th International Conference on Neural Information Processing Systems, NIPS'05. MIT Press, Vancouver, British Columbia, Canada, pp. 283–290.

Griffith, D.M., Veech, J.A., Marsh, C.J., 2016. cooccur: Probabilistic Species Co-Occurrence Analysis in R. J. Stat. Softw. 69, 1–17. https://doi.org/10.18637/jss.v069.c02

Hanani, U., Shapira, B., Shoval, P., 2001. Information Filtering: Overview of Issues, Research and Systems. User Model User-Adapt Interact 11, 203–259. https://doi.org/10.1023/A:1011196000674

Jõks, M., Pärtel, M., 2019. Plant diversity in Oceanic archipelagos: realistic patterns emulated by an agent-based computer simulation. Ecography 42, 740–754. https://doi.org/10.1111/ecog.03985

Karger, D.N., Cord, A.F., Kessler, M., Kreft, H., Kühn, I., Pompe, S., Sandel, B., Cabral, J.S., Smith, A.B., Svenning, J.-C., Tuomisto, H., Weigelt, P., Wesche, K., 2016. Delineating probabilistic species pools in ecology and biogeography. Glob. Ecol. Biogeogr. 25, 489–501. https://doi.org/10.1111/geb.12422

Koren, Y., Bell, R., Volinsky, C., 2009. Matrix Factorization Techniques for Recommender Systems. Computer 42, 30–37. https://doi.org/10.1109/MC.2009.263

Kramer, M.A., 1991. Nonlinear principal component analysis using autoassociative neural networks. AIChE J. 37, 233–243. https://doi.org/10.1002/aic.690370209

Lewis, R.J., Szava-Kovats, R., Pärtel, M., 2016. Estimating dark diversity and species pools: an empirical assessment of two methods. Methods Ecol. Evol. 7, 104–113. https://doi.org/10.1111/2041-210X.12443

Münzbergová, Z., Herben, T., 2004. Identification of suitable unoccupied habitats in metapopulation studies using co-occurrence of species. Oikos 105, 408–414. https://doi.org/10.1111/j.0030-1299.2004.13017.x

Netflix Update: Try This at Home [WWW Document], n.d. URL https://sifter.org/~simon/journal/20061211.html (accessed 5.6.20).

Pärtel, M., Szava-Kovats, R., Zobel, M., 2011. Dark diversity: shedding light on absent species. Trends Ecol. Evol. 26, 124–128. https://doi.org/10.1016/j.tree.2010.12.004

Pärtel, M., Zobel, M., Zobel, K., van der Maarel, E., 1996. The Species Pool and Its Relation to Species Richness: Evidence from Estonian Plant Communities. Oikos 75, 111–117. https://doi.org/10.2307/3546327

Real, R., Barbosa, A.M., Vargas, J.M., 2006. Obtaining Environmental Favourability Functions from Logistic Regression. Environ. Ecol. Stat. 13, 237–245. https://doi.org/10.1007/s10651-005-0003-3

Resnick, P., Varian, H.R., 1997. Recommender systems. Commun. ACM 40, 56–58. https://doi.org/10.1145/245108.245121

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Netw. 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Zhang, G., Liu, Y., Jin, X., 2020. A survey of autoencoder-based recommender systems. Front. Comput. Sci. 14, 430–450. https://doi.org/10.1007/s11704-018-8052-6

Appendix

Link to the repository that holds jupyter notebook that include the code used to generate the models used in this thesis (please remember, the code style was not thought about as it was just used by one person): https://github.com/siimkkoger/dark_diversity_bsc

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, Siim Karel Koger,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,
   Dark diversity estimation based on a single matrix of binary observations,
   supervised by Ardi Tampuu, Raul Vicente Zafra and Carlos Pérez Carmona.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Siim Karel Koger
08/05/2020