Instructions to run the dispeRsal function to predict seed dispersal distances

1. Pre-requirements

The statistical software R is needed for using the dispeRsal function. R is a commando-line based statistical software package that is freely available. For downloading and installing it, and to find introductory manuals go to:

http://www.r-project.org/

dispeRsal relies on mixed effect models as implemented in the available nlme package (Pinheiro et al. 2011). If you have not installed this library yet, you can do so by writing in your R console:
`install.packages(nlme)`

dispeRsal also makes use of functions from the Taxonstand library (Cayuela et al. 2012), but because these needed to be slightly modified, the Taxonstand library does not have to be installed but the modified code is automatically read in with the dispeRsal function and dataset (see section 3).

2. Data preparation

For the dispeRsal function to work in the desired way, your data has to follow a specific structure and format.

Your dataset should have species as rows and further data organized in columns. The order of the columns is not important but the names of the columns and the attributes in them have to be exactly as described below, otherwise the R function can not match your dataset to the predictive model. Species can have multiple entries with different dispersal syndromes.

Your dataset has to include taxonomical and trait information. At least data on dispersal syndrome needs to be provided, but consider that predicting dispersal distances from just one trait will largely affect the accuracy of the predictions.

Taxonomical information

Only the species names (without authorship) need to be supplied in a single column. The dispeRsal function then uses the modified Taxonstand functions to look up the provided species names at www.theplantlist.org to resolve the synonymy for the species and assign them to families. Then, species are assigned to orders, following APG III for angiosperms (The Angiosperm Phylogeny Group 2009), and Christenhuzs and others (2011) for gymnosperms.

Trait information

The trait data can be a combination of the following traits.

- dispersal syndrom (DS)
  Categorical trait with the following possible field values:
  - animal - all vertebrate dispersed seeds, without distinction between attached (epichorous) and ingested (endochorous) dispersal of seeds

    - ant - seeds ant-dispersed, often with the help of an elaiosome, i.e. a nutrient-rich seed appendage
    - ballistic - seeds are dispersed by some kind of 'explosive' mechansism of the mother plant
    - wind.none - seeds are dispersed usually by wind, but have no special adaptation for this dispersal vector
    - wind.special - seeds are dispersed by wind with the help of seed appendages such as wings and pappi

- growth form (GF)
  Categorical trait with the following possible field values:
    - tree
    - shrub
    - herb

- releasing height (RH)
  log10 transformed continuous data (in m)

- seed mass (SM)
  log10 transformed continuous data (in mg)

- terminal velocity (TV)
  log10 transformed continuous data (in m/s)

Once your data has been arranged in the demonstrated way (see Table 1 for an example) save your file in comma separated file format (.csv) to the working directory of your R workspace. This can be done for example by using Excel or similar software to prepare your dataset and then store the data in this file format.

Table 1. The head of a table, including the header row, of what a possible dataset would look like.

Species	GF	DS	SM	TV	RH
Abies alba	tree	wind.special	1.836	NA	1.7
Banksia spp.	shrub	wind.special	NA	NA	NA
Abutilon theophrasti	herb	wind.none	0.954	NA	NA
Acacia cyclops	shrub	ant	1.481	NA	NA
…	…	…	…		…

Read in your dataset to your R workspace from your working directory.

```
data.predict <- read.table("thenameofyourdatafile.csv", header = TRUE, sep = ";", dec = ".")
```

Note that depending on where you are on the world there are different .csv formats used. In the US, the separator between columns is defined as a comma, whereas in Europe a semicolon is used. Therefore here `sep = ";"`. The `dec` argument specifies if you used a period or a comma as the decimal sign in your continuous data.

3. Reading in data and using the dispeRsal function

When you have successfully read in your data it is now time to read in the necessary functions and data to calculate predictions of maximum dispersal distances of your species set. These are stored in the dispeRsal.rda file which you simply load into your R workspace with

```
load("dispeRsal.rda")
```

This will load the following objects into the current R environment:

`dispeRsal` function to predict dispersal distances for data provided by the user (see below)

`model.data` dataframe that is used for the predictive models

`own.data` example dataframe for demonstrating the functionality of the dispeRsal function (see section 5 in this document)

`OrderFamilies` dataframe containing the assignment of taxonomic families to orders

`TPLckMod, TPLMod` modified versions of the TPLck and TPL functions from the Taxonstand package for handling synonymies in data provided by the user

The `dispeRsal` function to predict dispersal distances for your dataset has 6 arguments that need to be passed to it:

```
dispeRsal(data.predict, model = NULL, CI = FALSE, random = TRUE,
tax = "family", write.result = FALSE)
```

`data.predict` the dataset with taxonomic and trait information that you prepared as described above and that you loaded into your R workspace in the previous step.

`model` either an integer value of (1,2,3,4,5) that specifies which predictive model is to be used or a vector of trait abbreviations to specify your own model (e.g. `c("DS", "SM")`. We added the latter possibility for convenience but want to stress that depending on what model you specify the accuracy of predicting dispersal distances might be greatly decreased, especially if you use just one explanatory variable in your model.

`1` = model 1: dispersal distance = DS + GF + TV
`2` = model 2: dispersal distance = DS + GF + SM + RH
`3` = model 3: dispersal distance = DS + GF + RH
`4` = model 4: dispersal distance = DS + GF + SM
`5` = model 5: dispersal distance = DS + GF

`CI` if TRUE, a lower and upper confidence limit are given for the predicted dispersal distances. Currently, this is only implemented for the predicted values from the fixed effects.

`random` should the taxonomy of the species be modeled as a random variable in a linear mixed model (`TRUE`) or should only a simple linear model be fitted (`FALSE`). Defaults to `TRUE`.

`tax` when `random = TRUE`, defines if only the order (`"order"`) or order and family (`"family"`) are considered for the random component of the linear mixed model. Defaults to `"family"`.

`write.result` when TRUE, writes the output data frame that contains the predicted dispersal distances to a file called predictedDD.txt in your working directory. If a mixed effects model is chosen, a second file called unmatched.txt is written to the working directory. This file contains the species names of the species that could not be matched to the taxonomy on www.theplantlist.org and have therefore been dismissed from the prediction. Defaults to FALSE, i.e. the data frame is just screened on the R commando window and no output file is written.

A possible call to the function dispeRsal would be for instance:

```
dispeRsal(data.predict, model = 4, CI = false, random = TRUE, tax
= "family", write.result = TRUE)
```

This would cause R to calculate dispersal distances for the dataset that you read in as described above using the model where dispersal distances are predicted from the traits growth form, dispersal syndrome and seed mass. A linear mixed model would be used that accounts for the random structure of order and family in your data. The output data frame would be written to a file called predictedDD.txt.

4. The output from calling dispeRsal

Initially, information about the current version of dispeRsal is displayed in the commando window including a notice whether a newer version is available for download.

The main output file (predictedDD.txt) is a data frame with the following columns:

`Species` – the name of the species for which dispersal distance was predicted. These are the species from the user-provided dataset, minus the species for which the predictive model could not be applied because they lacked information on a certain trait. Species can have multiple entries when they were provided with multiple dispersal syndromes in the user's dataset.

`Order` – order of the species

`Family` – family of the species

`DS` – the dispersal syndrome for which the dispersal distance has been predicted.

`log10MDD_Family` – predicted dispersal distances calculated from a linear mixed model with family and order as nested random variables. If some species can not be assigned to a family in the predictive model dataset, the function automatically tries to match the species at the order level and predicts the dispersal distance according to the random effect of the order. If also the order of the species is not in the predictive model dataset, dispersal distance is predicted using a simple linear model for prediction.

`Log10MDD_Family _lwrCL` – as `Log10MDD_lwrCL` but calculated on the basis of `log10MDD_Family`. Only given if argument `CI` has been set to `TRUE`.

`Log10MDD_Family _uppCL` – as `Log10MDD_uppCL` but calculated on the basis of `log10MDD_Family`. Only given if argument `CI` has been set to `TRUE`.

`log10MDD_Order` - predicted dispersal distances calculated from a linear mixed model with order as random variable (i.e. `tax = "order"`) or dispersal distance predicted according to the order effect because no match at the family level was found (see `log10MDD_Family`).

`Log10MDD_Order_lwrCL` – as `Log10MDD_lwrCL` but calculated on the basis of `log10MDD_Order`. Only given if argument `CI` has been set to `TRUE`.

`Log10MDD_Order_uppCL` – as `Log10MDD_uppCL` but calculated on the basis of `log10MDD_Order`. Only given if argument `CI` has been set to `TRUE`.

`log10MDD` – predicted dispersal distances when a simple linear model has been used for prediction (i.e. `random = FALSE`) or when predictions can not be made from a mixed effect model because family and/or orders in the dataset do not match the names in the predictive model data.

`Log10MDD_lwrCL` – the lower confidence limit for the predicted dispersal distance from the fixed effect only model (i.e. log10MDD). Only given if argument `CI` has been set to `TRUE`.

`Log10MDD_uppCL` – the upper confidence limit for the predicted dispersal distance from the fixed effect only model (i.e. log10MDD). Only given if argument `CI` has been set to `TRUE`.

`log10MDD_measured` – if there was an overlap of species from your dataset and with the species used for the predictive models, the empirical dispersal distance from the predictive model dataset is given here, but only if the dispersal syndrome for the species was the same as in the predictive model dataset.

Additionally to the predicted dispersal distances, a list of species names is given that could not be matched to the taxonomy at www.theplantlist.org. This list is NULL if the predictive model was a linear model without random effects. It is empty if all species could be matched.

5. An example session with dispeRsal

To make you familiar with dispeRsal and its functionality we give here a short walk through with an example dataset for which maximum dispersal distances will be predicted. This dataset is provided with the dispersal.rda file that you already read in. The name of the dataset is `own.data`.

Have a look at it by typing its name into the R console:

```
own.data
```

As you can see, taxonomic data is provided in the form of species names in a single column. There is also data on dispersal syndrome, growth form, seed mass, releasing height and terminal velocity. Hence, all four different predictive models can be run with the example data. However, because for some species not all data is available, dispersal distances can not be predicted with all the models for all species. Note also that for some species there are double entries that differ only in their dispersal syndrome. It is possible to predict different dispersal distances for different dispersal syndromes within the same species.

As the simplest model, we could run a prediction without taxonomy as a random variable, and only dispersal syndrome and growth form as predictors.

```
dispeRsal(own.data, model = 5, random = FALSE)
```

The result data frame gives you the species names, family names, order names, dispersal syndrome, one single column with the predicted dispersal distances and a column with real maximum dispersal distances measured for the particular species.

Let us next look at the 'best model', i.e. predictors are dispersal syndrome, growth form and terminal velocity, and taxonomy is included as a random effect at the family level. Also we will use CI argument to get confidence intervals for predictions.

```
dispeRsal(own.data, model = 1, random = TRUE, CI = TRUE, tax =
"family")
```

You will recognize that there are less species now in the output data frame, because terminal velocity, which is needed as a predictor in predictive model 1, is not available for many species in `own.data`. This time, there are three different columns with predicted dispersal distances. Log10MDD_Family gives predicted maximum dispersal distances for the species, whose taxonomy could be matched to the predictive models at the family level. The function also matches the species at the next higher level, i.e. at the order, because for the *Viola* species no match could be made at the Family level. The predicted distances from this step are given in log10MDD_Order. When neither of these matches at the taxonomical level could be achieved,

dispersal distances would be predicted from the fixed effects (log10MDD). The column log10MDD_measured gives the dispersal distance for species that overlap with species in the predictive model dataset. Additionally, there are four columns showing the lower and upper confidence limits for predictions at the Family level (log10MDD_Family_lwrCL, log10MDD_Family_uppCL) and Order level (log10MDD_Order_lwrCL, log10MDD_Order_uppCL).

6. References

The Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* **161**:105-121.

Christenhusz, M. J., J. L. Reveal, A. Farjon, M. F. Gardner, R. R. Mill, and M. W. Chase. 2011. A new classification and linear sequence of extant gymnosperms. *Phytotaxa* **19**:55-70.

Cayuela, L., I. Granzow-de la Cerda, F. S. Albuquerque, and D. J. Golicher. 2012. TAXONSTAND: An R package for species names standardisation in vegetation databases. Methods in Ecology and Evolution **3**:1078-1083.

Pinheiro, J., D. Bates, D, S. DebRoy, D. Sarkar, and R Development Core Team (2011) nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-102. http://cran.r-project.org/web/packages/nlme/index.html.