Bioversity International

Predictive characterization of crop wild relatives and landraces

Technical guidelines version 1

I. Thormann, M. Parra-Quijano, D.T.F. Endresen, M.L. Rubio-Teso, J.M. Iriondo and N. Maxted

PGR SECURE

Bioversity International is a global research-for-development organization. We have a vision – that agricultural biodiversity nourishes people and sustains the planet.

We deliver scientific evidence, management practices and policy options to use and safeguard agricultural biodiversity to attain sustainable global food and nutrition security. We work with partners in low-income countries in different regions where agricultural biodiversity can contribute to improved nutrition, resilience, productivity and climate change adaptation.

Bioversity International is a member of the CGIAR Consortium – a global research partnership for a food secure future.

www.bioversityinternational.org

Citation: Thormann, I., Parra-Quijano, M., Endresen, D.T.F., Rubio-Teso, M.L., Iriondo, M.J. and Maxted, N. 2014. Predictive characterization of crop wild relatives and landraces. Technical guidelines version 1. Bioversity International, Rome, Italy.

ISBN 978-92-9255-004-2

Cover photo: Collecting crop wild relatives in a field of wheat landraces in Morocco. N. Maxted

Bioversity Headquarters
Via dei Tre Denari 472/a
00057 Maccarese (Fiumicino)
Rome, Italy
Tel. (39-06) 61181
Fax. (39-06) 61979661
bioversity@cgiar.org

Predictive characterization of crop wild relatives and landraces

Technical guidelines version 1

I. Thormann, M. Parra-Quijano, D.T.F. Endresen, M.L. Rubio-Teso, J.M. Iriondo and N. Maxted

IT: Bioversity International, Rome, Italy

MPQ: International Treaty on Plant Genetic Resources for Food and Agriculture – FAO, Rome, Italy

DTFE: GBIF-Norway, Natural History Museum, University of Oslo, Norway

MLRT and JMI: Universidad Rey Juan Carlos, Mostoles, Spain

NM: University of Birmingham, UK

Acknowledgements

The ’Predictive characterization of crop wild relatives and landraces. Technical guidelines version 1‘ were developed within the framework of the project PGR Secure ‘Novel characterization of crop wild relative and landrace resources as a basis for improved crop breeding’, which was funded through the European Union’s Seventh Framework Programme for research, technological development and demonstration under Grant agreement no. 266394.

Bioversity International is grateful to those who collaborated in developing and testing datasets, R-scripts and environmental profiles within the PGR Secure project:

Dias, Sónia – Bioversity International, Italy

Fernandez, Sara – Universidad Politécnica de Madrid, Spain

Mancini, Chiara – Bioversity International, Italy

van Etten, Jacob – Bioversity International, Costa Rica

Kell, Shelagh – University of Birmingham, UK

Final language editing on behalf of Bioversity International was by Thorgeir Lawrence.

Layout on behalf of Bioversity International was by Ana Laura Cerutti.

Contents

Abbreviations used in the text ii

1. Introduction 1
- 1.1 Predictive characterization 1
- 1.2 Focused Identification of Germplasm Strategy 2
- 1.3 Ecogeographical filtering method 3
- 1.4 Calibration method 5
- 1.5 Structure and application of guidelines 5
- 1.6 Software requirements 6

2. Data preparation 7
- 2.1 Data compilation 7
- 2.2 Data cleaning 9
- 2.3 Geographical location data assessment 10
- 2.4 Georeferencing occurrences with location data 12
- 2.5 Quality evaluation of coordinates 12

3. Application of the ecogeographical filtering method 14
- 3.1 Ecogeographical land characterization map 14
- 3.2 Environmental profile 17
- 3.3 Selection of occurrence set 21
- 3.4 Use with biotic resistance traits as variables 25

4. Application of the calibration method 26
- 4.1 Installation of required R packages 26
- 4.2 Addition of climate data 27
- 4.3 Trait data compilation 28
- 4.4 Preparation of training and test set 30
- 4.5 Model calibration 31
- 4.6 Model testing 33
- 4.7 Selection of occurrence set 36

5. Final remarks 37

References 38

Abbreviations used in the text

C&E	characterization and evaluation
CWR	crop wild relative
DEM	Digital Elevation Model
ECPGR	European Cooperative Programme for Plant Genetic Resources
ELC	ecogeographical land characterization [map]
FIGS	Focused Identification of Germplasm Strategy
GBIF	Global Biodiversity Information Facility
GIS	Geographical Information System
GRIN	Germplasm Resources Information Network
Iar	De Martonne aridity index
ICARDA	International Centre for Agricultural Research in the Dry Areas
IPK	Leibniz Institute for Plant Genetics and Crop Plant Research
LR	landrace
LR+	positive predictive likelihood ratio
MCPD	Multi-crop Passport Descriptors
PPV	positive predictive value

1. Introduction

The number of germplasm accessions conserved in *ex situ* genebanks has grown to a total of about 7.5 million. However, the lack of characterization and evaluation (C&E) data continues to be reported as one major limitation to the use of these plant genetic resources (FAO, 2010). Traditional C&E methods cannot catch up with this growing number of accessions. Even less data are available for crop wild relative (CWR) populations conserved *in situ* and for on-farm managed landraces (LRs). *In situ* C&E is not a routine activity of protected area managers or an integral part of on-farm conservation, as their implementation is complex and resource intensive (Guarino *et al.*, 2002). Biodiversity conservationists and managers of protected areas generally see the conservation as the end goal and will not focus on collecting the C&E data which is needed for systematic utilization of this material in breeding programmes. Farmer-based selection (on farm), in contrast, is generally less formal and both less dependent on availability of C&E data, and less likely to generate such data. Additional methods for characterization of populations, accessions, collections and conservation sites are required to enhance the utilization of plant genetic resources *in situ* and *ex situ*. Predictive characterization approaches, which build on geographical location and agro-ecological data, can optimize the search for populations and accessions with adaptive traits and characteristics.

1.1 Predictive characterization

C&E of CWR and LRs—essential for enhancing their conservation and use—has nearly always involved an element of prediction. In practice, breeders rarely choose accessions for field C&E randomly. Where possible, they select accessions they believe are likely to contain the desired traits. Advances in molecular and Geographical Information System (GIS) analysis techniques mean that predictions of which accessions are likely to contain desired traits are now significantly more objective (evidence-based) than previously. Collectively, the approaches that involve GIS analysis are referred to as predictive characterization and they present a more cost effective approach than traditional phenotypic C&E of the complete germplasm collection. They build on the hypothesis that the different environments exert divergent selective pressures on plant populations, and thus spatial genetic differentiation. CWR populations growing in a specific environment, or LRs that developed within a given environment, will possess a suite of adaptive traits shaped by selection pressures unique to these environments.

Predictive characterization methods are predictive in the sense that they assign the potential of trait presence to uncharacterized germplasm (either *ex situ* or *in situ*) using (i) matching of biotic and abiotic characteristics associated with a collecting site; (ii) ecogeographical information associated with a collecting site; and (iii) previously recorded C&E data of trait occurrence associated with a set of locations different from those where the germplasm being examined has been collected. In each case a predictor is used to build a hypothesis that germplasm from a particular location will be genetically differentiated. The methods are represented in Schema 1.1 and are compared with the traditional characterization methods. Predictive characterization does not replace actual field trials, but considerably reduces the size of the field trials by reducing the set of candidate accessions which the breeder needs to screen before finding novel alleles for a target trait. This is achieved through the less expensive pre-screening against an environmental profile. The predictive methods therefore help to more efficiently utilize the limited and costly land area and human working time for the field screening.

1.2 Focused Identification of Germplasm Strategy

One of the first systematic applications of finding a predictive link between a resistance trait and a set of environmental parameters, named the Focused Identification of Germplasm Strategy (FIGS) (Mackay and Street, 2004; Street *et al.*, 2008), used biotic and abiotic matching techniques. FIGS was developed at the International Centre for Agricultural Research in the Dry Areas (ICARDA) based on early work by Michael Mackay in the 1980s and 1990s (Mackay 1986, 1990, 1995). The first FIGS studies were based on using scientific expert knowledge for matching environmental profiles that were known to be suitable for adaptations leading to the target trait properties in LRs growing in such locations. Areas that have high levels of aphids are likely to have higher levels of aphid resistance. Environmental profiles supportive of high aphid populations were used to identify resistance in the wheat crop to this pest (El Bouhssini *et al.,* 2011). The biotic matching method applies a series of filters based on expert knowledge to identify germplasm material with pest resistances by filtering out locations with environmental profiles suitable for the respective pest. The method can be illustrated using the work on Sunn pest resistance in wheat by El Bouhssini and co-workers (2009). Starting with 16,000 wheat LRs from different genebank collections, germplasm material collected in China, Pakistan and India was excluded based on expert knowledge and the reasoning that Sunn pest has only recently been reported here, providing too little time for adaptive resistance in the germplasm to evolve. This reduced the candidate set down to 6,328 LRs. The next filter excluded environments that are too dry for the Sunn pest insects to thrive (less than 280 mm precipitation per year), and environments with too low winter temperatures (below 10°C) reducing the candidate set further down to 1,502 LRs. In the final derived set of 534 LRs, 9 new and formerly uncharacterized resistant LRs were identified by ICARDA in field experiments during 2007 and 2008. Previous extensive series of field experiments conducted at ICARDA from 2000 to 2006 including more than 2,000 wheat LRs had not been able to identify wheat LRs with resistances. Khazaei and co-workers (2013) recently validated the FIGS prediction against actual and independent evaluation trial data for drought resistance in faba bean. Further studies using FIGS are summarized in Box 1.1.

FIGS methods have mainly been applied to major crops, in particular wheat and barley (Box 1.1). Building upon the foundation of the FIGS approach, studies that use ecogeographical information or previously recorded C&E data have also been developed and were tested for their applicability to CWR and LRs of minor crops within the context of the PGR Secure project ‘Novel characterization of CWR and LR resources as a basis for improved crop breeding‘ (http://www.pgrsecure.org/) (Thormann, 2012). These studies have explored the ecogeographical filtering method and the calibration method, which are the focus of these guidelines.

Box 1.1. Examples of predictive association studies and identification of pest and pathogen resistant and drought tolerant material through the use of FIGS

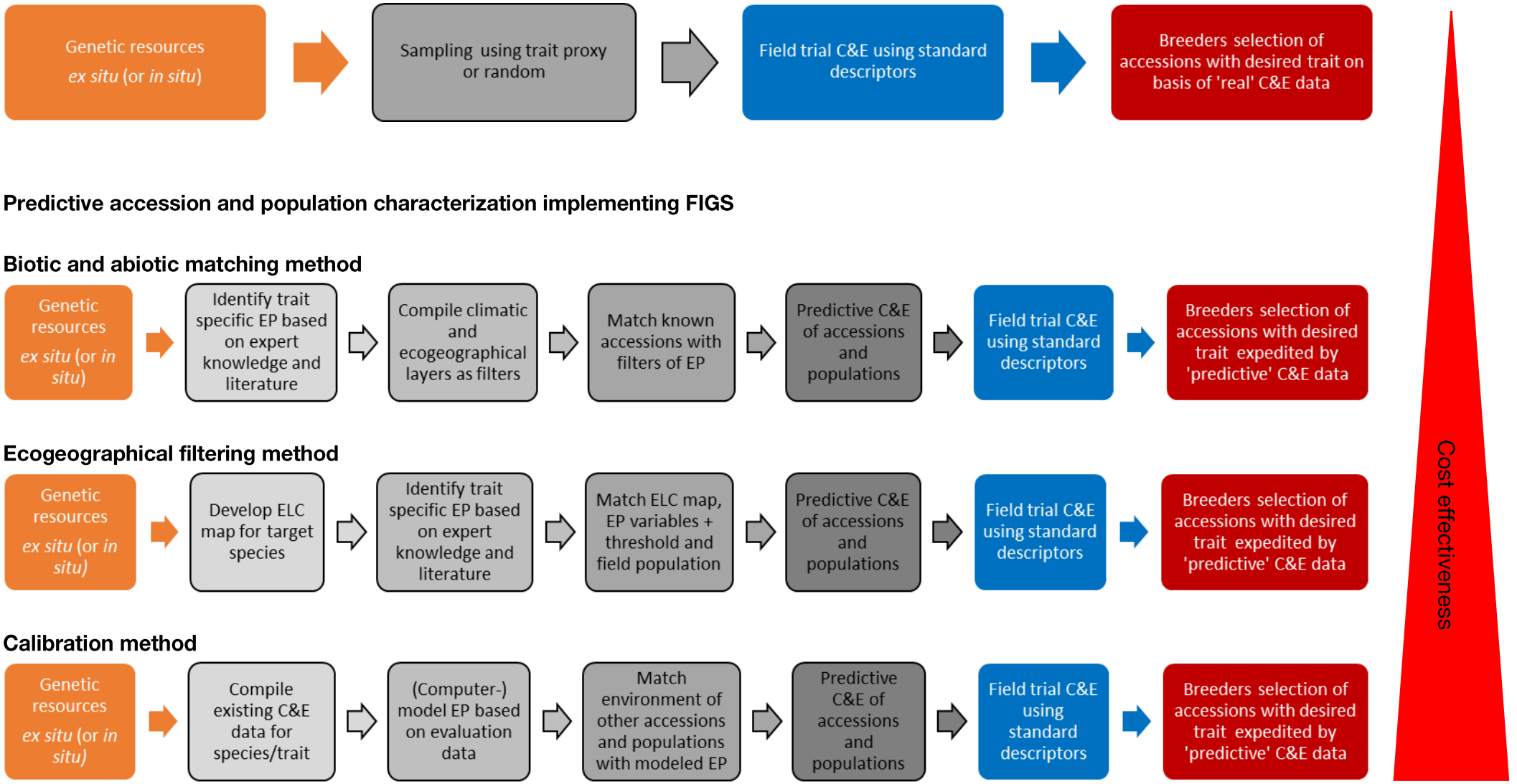
- Powdery mildew resistance in wheat (Bhullar *et al*., 2009; Kaur *et al*., 2008)
- Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* (El Bouhssini *et al*., 2009)
- Predictive association between trait data and ecogeographical data for Nordic barley landraces (Endresen, 2010)
- Sources of resistance in bread wheat to Russian wheat aphid, *Diuraphis noxia* (El Bouhssini *et al*., 2011)
- Predictive association between biotic stress traits and ecogeographical data for wheat and barley (Endresen *et al*. 2011)
- Wheat stem rust resistance linked to environmental variables (Bari *et al*., 2012)
- Resistance to stem rust (Ug99) in bread wheat and durum wheat (Endresen *et al*., 2012)
- Traits identified related to drought adaptation in *Vicia faba* genetic resources (Khazaei *et al*., 2013)
- Wheat yellow stripe rust resistance linked to environmental variables (Bari *et al*., 2014)

1.3 Ecogeographical filtering method

The ecogeographical filtering method combines the spatial distribution of the target taxon on an ecogeographical land characterization map (ELC) (Parra-Quijano, Iriondo and Torres, 2012; Parra-Quijano *et al*., 2012) with the ecogeographical characterization of those environments that are likely to impose selection pressure for the adaptive trait investigated, to filter occurrence records. In the predictive characterization context it uses a taxon-specific ELC map that is developed based on the variables most relevant for adaptation and for determining the species' distribution. This map aims at representing the adaptive scenarios that are present over the territory studied.

As a first step in this method, the ecogeographical categories from the ELC map are assigned to each occurrence record according to its coordinates and the records are then grouped according to their ELC map category. After all georeferenced occurrences have been ecogeographically characterized, the second step is to select occurrences from each group that comply with specific environmental requirements related to the traits of interest: the specific ecogeographical variables (geophysical, edaphic or bioclimatic) that best describe and delimit the environmental profile likely to impose selection pressure for the adaptive trait of interest. These are then used for further filtering to obtain a final subset of occurrences.

The filtering method only requires coordinates from collecting sites and can be applied to *in situ* CWR and LR occurrences as well as to *ex situ* accessions. Ideally it is applied at taxon level, but can also be used for a group of related taxa.



Key: C&E = Characterization and Evaluation; EP = Environmental Profile; ELC = Ecogeographical Land Characterization; FIGS = Focused Identification of Germplasm Strategy

1.4 Calibration method

The calibration method uses existing C&E data for the trait of interest, together with ecogeographical data specific to the environment at collecting sites from where these accessions were collected, to identify existing relationships between the trait and the environment. Based on these relationships, it calibrates a prediction model.

This prediction model is then applied to other non-evaluated accessions to identify those that, according to this model, are likely to have a higher probability of genetic adaptation for a target trait property. The model therefore aims to identify a subset that is more likely to show the target trait property than a subset merely selected randomly. The calibration method can be used when availability of evaluation data is not a limiting factor. The calibration method is used here only to provide a model that relates the target trait and the environment data, but does not provide biological explanations for any trait occurrence and expression. Cases exist where for example an identified pest resistance cannot be related to specific environments. In such cases, the calibration method might not be suitable for identification of any useful predictive relationship between the trait and any set of environmental variables. One example is the Rhizomania resistance in wild beet from Denmark. Resistance was found in accessions from an area in Denmark where the pathogen does not occur (Lothar Frese pers. comm.) and no disease incidence has been reported. In this case resistance to the pest cannot be the result of any adaptation to the pest. Another unrelated morphological or physiological trait might confer resistance to this pathogen. The use of the calibration method has been described in recent studies on morphological and agricultural traits in barley (Endresen 2010; Endresen *et al.*, 2011), wheat stem rust (Endresen *et al.*, 2011; Bari *et al.*, 2012; Endresen *et al.*, 2012).

1.5 Structure and application of guidelines

These guidelines follow the logical sequence of steps that need to be carried out to implement the ecogeographical filtering and the calibration method. They first address the compilation, quality assessment and improvement of the baseline data that are required for the application of both methods. The implementation of the respective methods is then dealt with in two subsequent chapters.

The choice of which method to use depends mainly on the trait of interest and the availability of data related to that trait. The lack of evaluation data frequently represents a bottleneck, and, depending on the taxon and target trait, is often not available at all or not available in sufficient quantity. This might be particularly the case for CWRs. The choice of the method will therefore default to ecogeographical filtering if no or insufficient evaluation data are available.

The taxonomic level at which the two methods are usually applied is the species or related-species group level, because the pattern of genetic diversity for adaptive traits may vary between species found across a range of ecogeographical constraints. The methods are based on the assumption that there are relationships between traits and environmental conditions determined by abiotic and biotic factors that determine population adaptation, and different species even within the same genus might react differently to the same conditions.

In particular for the calibration method, which is modelling an environment based on evaluation and climate data, it might be useful to study each specific taxon by separately calibrated prediction models, in order to exclude additional multi-taxon noise in the model. However, a potential use at genus level is provided by the ecogeographical filtering method when, for instance, occurrences are sought for a particular environmental condition that is supposed to generate potential resistances, like growing on soils with low pH and low organic matter content, where aluminum

concentration might be high and species thriving there might have specific traits that tolerate high aluminium concentration. The stronger the biotic or abiotic selection pressure in any locality, the more likely that populations present in that locality will have evolved mechanisms to mitigate the pressure and underpin population success in the face of that pressure.

The ecogeographical filtering method requires developing and using an ELC map. The ecogeographical region under study will most often be a coherent region, either a country, a part of a (very large and diverse) country or several contiguous countries. ELC maps can be prepared at the genus level to be applicable to a range of congeneric species. The calibration method looks at a specific trait that has already been identified in a specific species and seeks to model the relationship with the environment. Evaluation data can be available from different geographical regions and can be used for the modelling exercise. It is to be noted that these methods do not investigate any biological pathways or explanations for any identified relationship.

An example data set for each of the methods is available for download from the web sites indicated in the document. Both methods are implemented in R, an open source language and environment for statistical computing and graphics. Example R scripts are provided in the guidelines. The user can copy the R scripts into their working environment and test the functionalities of the script with the example data sets provided.

The CAPFITOGEN Program (http://www.planttreaty.org/capfitogen) for the Strengthening of the Capacities of the National Programmes on Plant Genetic Resources in Latin America, funded by the Spanish Government and implemented by the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), has developed a series of tools that improve, adapt and facilitate the use of methodologies for ecogeography and GIS (Parra-Quijano *et al.*, 2014; http://www.capfitogen.net/en/). These guidelines make use of the tools 'GEOQUAL' for the quality assessment of georeferenced passport data, and 'ELCmapas' for the generation of ELC maps.

1.6 Software requirements

- Spreadsheet software (e.g. Excel™)
- A web browser
- DIVA-GIS – a program for mapping and geographical data analysis (can be downloaded free of charge from http://www.DIVA-GIS.org/download)
- R – a language and environment for statistical computing and graphics. You can either work directly in the R-console or use the graphical user interface provided by RStudio
    - Download R from http://www.r-project.org/
    - Download RStudio from http://www.rstudio.com/
- CAPFITOGEN tools – series of tools that improve, adapt and facilitate the use of methodologies for ecogeography and GIS. Instructions for download and installation are provided at http://www.capfitogen.net/en/access/download-from-internet/.

2. Data preparation

The application of the ecogeographical filtering and the calibration method requires the thorough compilation of taxon passport data. As predictive characterization involves GIS analysis the passport data of accessions or populations must contain georeferences or a georeferencable location description, so that environmental data can be associated with the specific locations where the taxon occurs.

The data preparation section first addresses the compilation and cleaning of the taxon occurrence records. Subsequently, the geographical location data contained in the taxon occurrence records need to be assessed, and if necessary, occurrences need to be georeferenced, and the quality of the coordinates need to be evaluated. The quality and comprehensiveness of the occurrence and location data are essential for the subsequent analyses. It is anticipated that the compilation of the occurrence data and the validation and quality improvement of the location data may require a substantial number of days, depending on the quantity and quality of the source data. Other types of data such as environmental data or evaluation data are compiled depending on the method to be applied and are described accordingly in the corresponding chapters addressing the method. As the calibration method can only be used if evaluation data are available, the availability of evaluation data should be checked at an early stage, to decide if the use of this method can be considered.

2.1. Data compilation

Passport data about genebank accessions, herbarium specimens and field observations provide information about taxon occurrences. Table 1 provides examples of national and international online information systems from where occurrence data can be downloaded. The table includes the Germplasm Resources Information Network (GRIN) of the United States Department of Agriculture http://www.ars-grin.gov/npgs/acc/acc_queries.html as one example of a national online genebank database, and the online genebank database of the German Leibniz Institute for Plant Genetics and Crop Plant Research (IPK) as an example of an institutional online system. Several other national and institutional online databases exist and their number is growing. A list of online databases is provided, for example, by the European Cooperative Programme for Plant Genetic Resources (ECPGR) at http://www.ecpgr.cgiar.org/germplasm_databases/national_multicrop_databases.html. Also crop-specific online databases are available and a list of European crop databases is provided on the Web site of the ECPGR at http://www.ecpgr.cgiar.org/germplasm_databases.html. The resources listed here are examples. It is not within the scope of these guidelines to provide a comprehensive list of online databases, and such a list would quickly be outdated.

Table 2.1. Examples of online resources for taxon occurrence data

Resource	Type of data provided	URL (verified 12 Sep 2014)
Genesys – Gateway to genetic resources	Genebank accession passport data provided by the CGIAR genebanks, GRIN and EURISCO.	https://www.genesys-pgr.org/welcome
EURISCO – Catalogue of European *ex situ* plant collections	Genebank accession data of European genebanks from 43 countries.	http://eurisco.ecpgr.org/
GBIF – Global Biodiversity Information Facility	Herbaria, natural history museum collections, botanical garden collections and genebank accession passport data – worldwide global coverage.	http://www.gbif.org/
GRIN – Germplasm Resources Information Network of the United States Department of Agriculture	Genebank accession data of US *ex situ* collections.	http://www.ars-grin.gov/npgs/acc/acc_queries.html
GBIS-IPK – Genebank Information System of the IPK Gatersleben, Germany	Genebank accession data of the German genebank at the IPK	http://gbis.ipk-gatersleben.de/GBIS_I/

Passport data for the taxon of interest are compiled from online resources such as those presented in Table 2.1, or from other regional or national online databases. or both, that cover the taxon and geographical area of interest. Usually these online databases have a download function, through which the data can be downloaded in tab- or comma-delimited files (.txt or .csv files) or directly as an MS Excel™ spreadsheet. If no online data are available, the online databases do not cover adequately the taxon or geographical area under study, or do not have a data download function, you need to get in direct contact with database curators of those databases from which you need further data, or with institutes that can be expected to have taxon occurrence data. When downloading data or requesting data not available online, you should make sure that all available data fields describing the geographical location are included, as well as all available fields describing the origin and identity of the occurrence, and the time and details about the collecting or observation event. Fields describing the geographical location include the country of origin, latitude, longitude, altitude, administrative units at various levels and the site description. They are important for georeferencing (Section 2.4) and quality assessment of coordinates (Section 2.5). Fields describing the identity and origin of the occurrence are particularly important when you need to screen your data for duplicate records (Section 2.2).

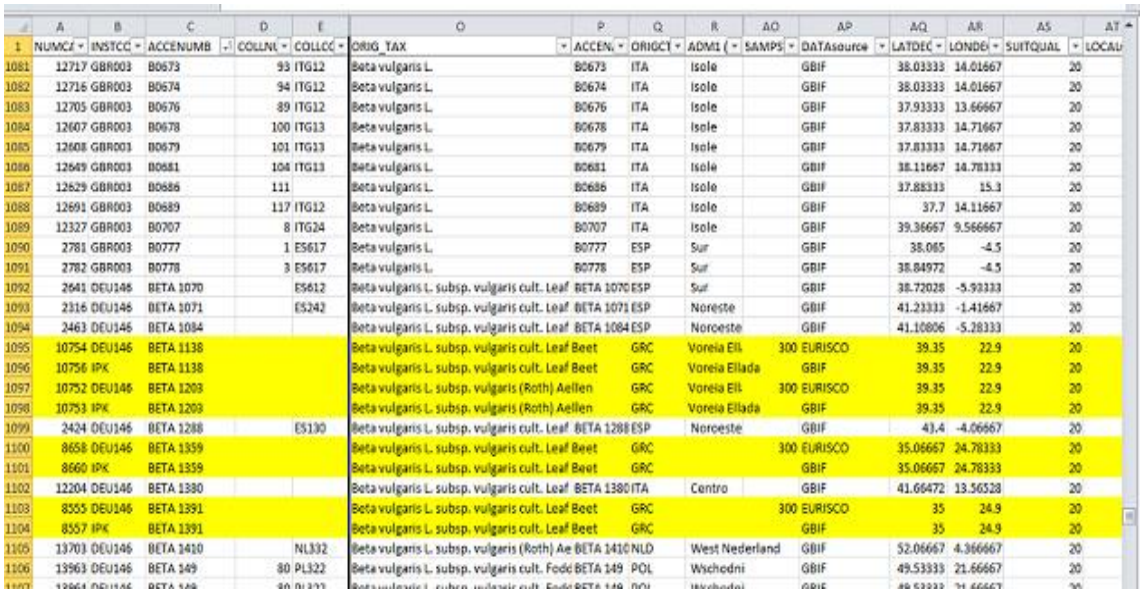
When data are obtained from more than one resource, the various datasets need to be merged. Data sources often provide varying amounts of data fields and these might be in a different column order. This needs to be adjusted for when merging datasets.

Records that provide no geographical information, i.e. lack coordinates as well as location description, are deleted from the dataset. A numeric ID is then assigned to each record to uniquely identify it for further reference.

The result of this first step is a unique table with available occurrence data that includes latitude+longitude or location description, or both, for each record. Each record is uniquely identified by an ID for later reference and identification of records.

2.2 Data cleaning

When data are compiled from various sources, the resulting data set is likely to contain duplicate records, which need to be removed. Duplicate records can be of two types. The first type of duplicates that should be removed are database records downloaded or obtained from different databases, e.g. from GBIF and Genesys (Figure 2.1), but which refer to the same genebank accession, herbarium specimen or *in situ* occurrence record. These can be identified through duplicate accession numbers. If the duplicate records do not contain exactly the same data, because not exactly the same data fields are available from each data source, or one record is more up-to-date than the other, it needs to be carefully judged which record to maintain. The second type of duplicates that should be identified are spatial duplicates, i.e. two rows with equal coordinates indicating the same presence or collecting site. Spatial duplicates are not necessarily errors in a database, as, for example, more than one LR might have been collected from the same site, but for the predictive characterization approaches it is required that a single accession or occurrence represent the collecting site and its spatial duplicates. It is therefore necessary to remove spatial duplicates.



	A	B	C	D	E	O	P	Q	R	AO	AP	AQ	AR	AS	AT
1	NUMCZ	INSTCC	ACCENUMB	COLLNU	COLLCC	ORIG_TAX	ACCEN	ORIGCT	ADM1 (	SAMPS	DATAsource	LATDEC	LONDE	SUITQUAL	LOCAL
1081	12717	GBR003	B0673	93	ITG12	Beta vulgaris L.	B0673	ITA	Isole		GBIF	38.03333	14.01667	20	
1082	12716	GBR003	B0674	94	ITG12	Beta vulgaris L.	B0674	ITA	Isole		GBIF	38.03333	14.01667	20	
1083	12705	GBR003	B0676	89	ITG12	Beta vulgaris L.	B0676	ITA	Isole		GBIF	37.93333	13.66667	20	
1084	12607	GBR003	B0678	100	ITG13	Beta vulgaris L.	B0678	ITA	Isole		GBIF	37.83333	14.71667	20	
1085	12608	GBR003	B0679	101	ITG13	Beta vulgaris L.	B0679	ITA	Isole		GBIF	37.83333	14.71667	20	
1086	12649	GBR003	B0681	104	ITG13	Beta vulgaris L.	B0681	ITA	Isole		GBIF	38.11667	14.78333	20	
1087	12629	GBR003	B0686	111		Beta vulgaris L.	B0686	ITA	Isole		GBIF	37.88333	15.3	20	
1088	12691	GBR003	B0689	117	ITG12	Beta vulgaris L.	B0689	ITA	Isole		GBIF	37.7	14.11667	20	
1089	12327	GBR003	B0707	8	ITG24	Beta vulgaris L.	B0707	ITA	Isole		GBIF	39.36667	9.566667	20	
1090	2781	GBR003	B0777	1	ES617	Beta vulgaris L.	B0777	ESP	Sur		GBIF	38.065	-4.5	20	
1091	2782	GBR003	B0778	3	ES617	Beta vulgaris L.	B0778	ESP	Sur		GBIF	38.84972	-4.5	20	
1092	2641	DEU146	BETA 1070		ES612	Beta vulgaris L. subsp. vulgaris cult. Leaf	BETA 1070	ESP	Sur		GBIF	38.72028	-5.93333	20	
1093	2316	DEU146	BETA 1071		ES242	Beta vulgaris L. subsp. vulgaris cult. Leaf	BETA 1071	ESP	Noreste		GBIF	41.23333	-1.41667	20	
1094	2463	DEU146	BETA 1084			Beta vulgaris L. subsp. vulgaris cult. Leaf	BETA 1084	ESP	Noroeste		GBIF	41.10806	-5.28333	20	
1095	10754	DEU146	BETA 1138			Beta vulgaris L. subsp. vulgaris cult. Leaf Beet		GRC	Voreia Ell	300	EURISCO	39.35	22.9	20	
1096	10756	IPK	BETA 1138			Beta vulgaris L. subsp. vulgaris cult. Leaf Beet		GRC	Voreia Ellada		GBIF	39.35	22.9	20	
1097	10752	DEU146	BETA 1203			Beta vulgaris L. subsp. vulgaris (Roth) Aellen		GRC	Voreia Ell	300	EURISCO	39.35	22.9	20	
1098	10753	IPK	BETA 1203			Beta vulgaris L. subsp. vulgaris (Roth) Aellen		GRC	Voreia Ellada		GBIF	39.35	22.9	20	
1099	2424	DEU146	BETA 1288		ES130	Beta vulgaris L. subsp. vulgaris cult. Leaf	BETA 1288	ESP	Noroeste		GBIF	43.4	-4.06667	20	
1100	8658	DEU146	BETA 1359			Beta vulgaris L. subsp. vulgaris cult. Leaf Beet		GRC		300	EURISCO	35.06667	24.78333	20	
1101	8660	IPK	BETA 1359			Beta vulgaris L. subsp. vulgaris cult. Leaf Beet		GRC			GBIF	35.06667	24.78333	20	
1102	12204	DEU146	BETA 1380			Beta vulgaris L. subsp. vulgaris cult. Leaf	BETA 1380	ITA	Centro		GBIF	41.66472	13.56528	20	
1103	8555	DEU146	BETA 1391			Beta vulgaris L. subsp. vulgaris cult. Leaf Beet		GRC		300	EURISCO	35	24.9	20	
1104	8557	IPK	BETA 1391			Beta vulgaris L. subsp. vulgaris cult. Leaf Beet		GRC			GBIF	35	24.9	20	
1105	13703	DEU146	BETA 1410		NL332	Beta vulgaris L. subsp. vulgaris (Roth) Ae	BETA 1410	NLD	West Nederland		GBIF	52.06667	4.366667	20	
1106	13963	DEU146	BETA 149	80	PL322	Beta vulgaris L. subsp. vulgaris cult. Fodd	BETA 149	POL	Wschodni		GBIF	49.53333	21.66667	20	

Figure 2.1. Example of duplicate records in merged downloads from GBIF and EURISCO.

If data are downloaded at genus level or for a taxon that can occur both as wild and under cultivation, it is necessary to identify LR or CWR records, or both, depending on the purpose of the study. The following approach can be used to filter for the desired material, as many source databases now follow, at least partially, agreed standards such as the Multi-crop Passport Descriptors v.2 (MCPD) (Alercia, Diulgheroff and Mackay, 2012). The MCPD descriptor for improvement status called SAMPSTAT is used to distinguish between LR and CWR.

- **Landraces** If working at genus level, select those species in the genus that are known to be cultivated. In the subspecies field, all subspecies, cultivars or varieties associated with cultivation are selected. If working at a species level, start directly with the check on the subspecies field. From these first selections, in both cases, only LRs (identified through the MCPD data field SAMPSTAT with value = 300) and those records with a blank in the SAMPSTAT field are considered. Breeding material and advanced cultivars (SAMPSTAT = 400 to 500) are excluded.

- **Wild populations** If working at genus level, select the wild populations identified by SAMPSTAT = 100, 200 or 999 of those species known to occur in the wild as well as cultivated, together with all records of species that only occur in the wild. It can help to narrow the data set initially to exclude all known cultivated material, as well as all records with SAMPSTAT values of 300, 400 or 500. In the subspecies field, all species, cultivars, varieties and hybrids associated with cultivation are discarded.
- If working at species level with a species for which cultivated and wild occurrences exist, discard all subspecies, varieties and cultivars known to exist only in cultivation. From the remaining records for taxon levels for which both wild and cultivated records could be available, select those identified by SAMPSTAT = 100, 200 or 999. For those species known to occur only in the wild, all records are included.
- If you decide to apply a conservative approach, records where the species name is not provided and the field contains “sp.”, the name of the genus or “?” are discarded.

The resulting dataset will contain records of LRs or CWR of the species or group of species of interest, which include the required minimum of location data, and with a minimum number of duplicates.

2.3 Geographical location data assessment

The quality and quantity of coordinates and the need for georeferencing needs to be assessed. Records should first be sorted into subsets based on the type of coordinates they contain:

- Subset a: Coordinates in sexagesimal format (e.g. 30°15′55″N);
- Subset b: Coordinates in decimal format (e.g. 30.265°); or
- Subset c: Records without coordinates or with low precision coordinates (degree level for sexagesimal format or without fraction for decimal format).

The following steps should then be carried out, depending on the data formats found in the data set:

(1) Coordinates in sexagesimal format (subset a) should be converted into coordinates in decimal format to read them with GIS software. The following formula can be used:

$$DC = h * (d + m/60 + s/3600)$$

where DC is the coordinate in decimal format; d is the degrees (°), m the minutes (′), and s the seconds (″) of the sexagesimal (base 60) system; h = 1 for the northern and eastern hemispheres and h = -1 for the southern and western hemispheres. For example, 30°30′0″S = -30.500. These calculations can be carried out in a spreadsheet program. Degrees, minutes, seconds and the hemisphere need to be separated into single columns without the respective symbols °, ’ and ”. Also, DIVA-GIS can be used, using the Geo-calculator under its Tools menu, but it requires checking coordinates one by one.

(2) Location description in all subsets should be distributed in separate hierarchical fields (ORIGCTY, ADM1, ADM2, ADM3, ADM4 and COLLSITE) (GADM, 2012). Not all databases contain all these single location fields, and location information is sometimes lumped into the COLLSITE field containing the location description. It is therefore often necessary to extract administrative information (ADM1, ADM2, ADM3, ADM4) from COLLSITE when this field contains all location data. The division of the location information into the appropriate fields is required to carry out the quality check of the coordinates described in Section 2.5 below. The passport data template

used in the GEOQUAL tool is available from http://www.capfitogen.net/en/MCDpassportFormat_FAO_Bioversity_2012_modified.xls. It is adapted from the MCPD and provides for each country and each administrative level (ADM1 to ADM4) the corresponding ADM name according to the GADM database of Global Administrative Areas (GADM, 2012).

(3) The decimal coordinates (i.e. converted subsets a and b) can be checked for errors with DIVA-GIS. DIVA-GIS has a 'check coordinates' facility under its data menu (Figure 2.2) that can help spot coordinates that contain errors. While some errors are easily spotted when plotting occurrences on a map, such as an occurrence falling in the middle of the ocean, other mistakes are not as easily spotted just by looking at the map, such as points not lying within the country of origin or within the administrative units provided in the data. The check coordinate function uses a method developed by Hijmans and co-workers (1999) to identify those potential errors. For further details refer to Section 3.9 in the DIVA-GIS manual, available from http://www.DIVA-GIS.org/docs/DIVA-GIS5_manual.pdf (accessed 1 August 2014).

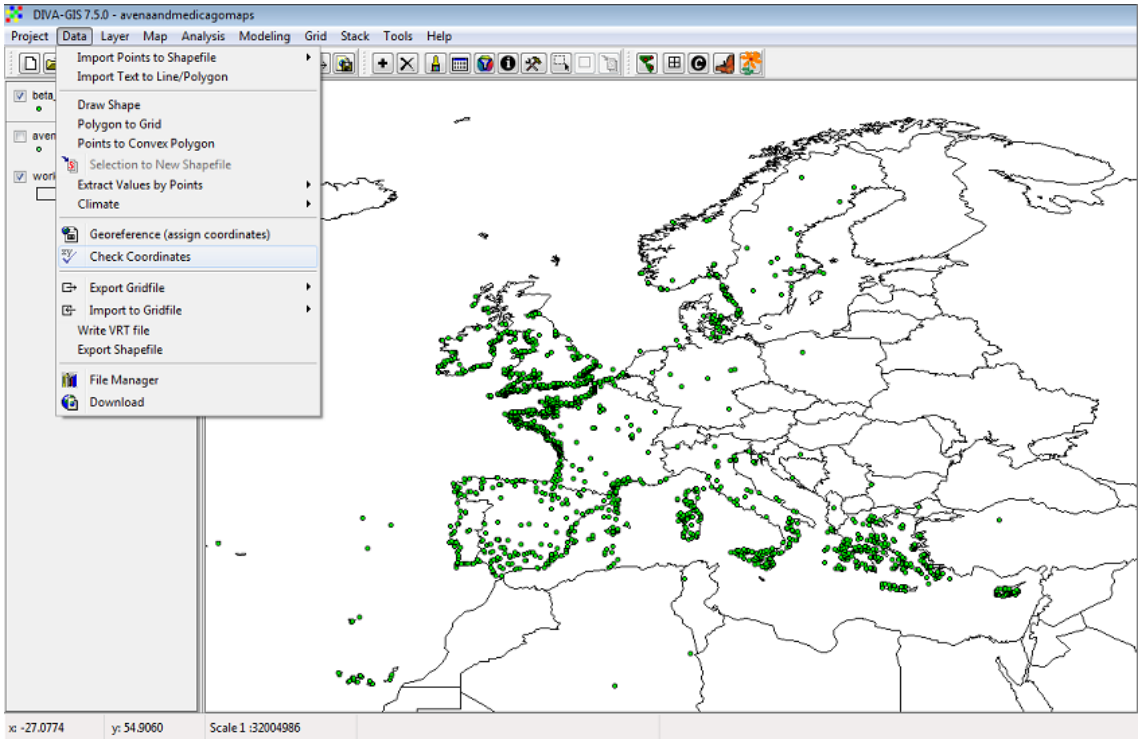


Figure 2.2. DIVA-GIS *check coordinate* option under its *data* menu. The map shows the distribution of wild *Beta* accessions in Europe with unchecked coordinates.

2.4 Georeferencing occurrences with location data

The use of predictive characterization methods requires that all occurrences included in the analysis are properly georeferenced and latitude and longitude are provided in decimal degrees. Those occurrences, for which only the location description but no geographical coordinates are provided, and those that have very low quality coordinates (i.e. subset c in Section 2.3 above) need to be georeferenced.

Georeferencing (Chapman and Wieczorek, 2006) can be done using GEOLocate (http://www.museum.tulane.edu/geolocate/). It analyses the collecting site descriptions to provide coordinates, together with an estimate of coordinate precision. Long lists of location descriptions can be georeferenced at once with the GEOLocate software, but it is recommended to check assigned coordinates as explained in the user's manual (http://www.museum.tulane.edu/geolocate/standalone/manual_ver2_0.pdf). Gazetteers and online mapping tools such as Google Maps (Google Inc., 2014) are also useful for georeferencing. Gazetteers can be downloaded from the DIVA-GIS web site at http://www.DIVA-GIS.org/gdata.

2.5 Quality evaluation of coordinates

All existing coordinates should be checked for quality and be assigned a quality rank through applying the geo-referencing quality evaluation procedure developed by Parra-Quijano and co-workers (2014). As predictive characterization methods build on the link between plant adaptation and environmental conditions of the collecting or occurrence site, the quality of coordinates becomes a key factor, since the higher the quality of coordinates, the higher the precision of the extracted ecogeographical information.

The georeferencing quality evaluation procedure is implemented in the CAPTIFOGEN tool GEOQUAL (see Sections 1.5 and 1.6) and examines and classifies coordinates and location description data according to three parameters, COORQUAL, SUITQUAL and LOCALQUAL, and then summarizes those in a final parameter TOTALQUAL.

COORQUAL assesses precision of coordinates and locality descriptions based on latitude and longitude, other fields that contain any coordinate quality information, and the collecting or observation date (if available). This parameter determines the intrinsic quality of the coordinates contained in the passport data, using a 0 to 20 scale, where:

- 0 = no coordinates
- 10 = coordinates at minute-level precision, with locality description at municipality level and collected between 1995 and 2000 (medium probability of GPS use)
- 20 = coordinates at second-level precision, with locality description below municipality level and collected after 2000 (maximum probability of GPS use)

SUITQUAL assesses occurrence at sites suitable for plants, assigning a quality value to coordinates according to how appropriate the collection site is for plant growth, based on land use maps. It differentiates the nature of the accession (wild or cultivated according to the SAMPSTAT descriptor). The range of values assigned is from 0 to 20:

- 0 = no coordinates
- 5 = points on inland or marine water
- 10 = points on wetlands
- 15 = points on bare or scarce vegetation or urban areas
- 20 = points on suitable areas

LOCALQUAL estimates concordance between locations described by passport data and locations derived from coordinates. Locality descriptions (at Country, State/Region, Province and Municipality levels) of occurrence data from the compiled database are compared with those extracted from coordinates by GIS. This process is based on a sequence of letter-comparing processes that are run in R.

- 0 = no coordinates
- 10 = coincidence only at Country level, without coincidence at State/Region, Province or Municipality levels
- 15 = coincidence at Country, State/Region and Province levels without coincidence at Municipality level
- 20 = fully coincidence (Country, State/Region, Province and Municipality).

TOTALQUAL is the final quality parameter, representing the sum of the results of COORDQUAL, SUITQUAL and LOCALQUAL evaluations. According to the range of possible values of these parameters, TOTALQUAL values range from 0 to 60. In order to facilitate the GEOQUAL results interpretation, the 0 to 60 scale is converted into a 0 to 100 scale and the new values are included in the TOTALQUAL100 parameter. Therefore in the case of TOTALQUAL100 parameter, values close to 0 are indicating poor georeferencing quality and values about 100 the best possible georeferencing.

GEOQUAL allows the user to upload a file with occurrence data, the quality of the coordinates is assessed and the file is returned to the user including additional columns reporting the values for COORQUAL, SUITQUAL, LOCALQUAL and TOTALQUAL. The occurrence data need to be formatted according to the MCPD. The columns in the file that is uploaded need to follow a certain sequence. A template is provided with the GEOQUAL tool, which can be used to prepare the data and which also indicates which data are obligatory and which data are recommended to be included in addition to the obligatory fields. The manual provided with the CAPFITOGEN tools contains additional useful details and instructions about the application. The template is also available from http://www.capfitogen.net/en/MCDpassportFormat_FAO_Bioversity_2012_modified.xls.

Assigning a coordinate quality rank to each record in your dataset will allow use of TOTALQUAL values as thresholds to exclude from further analyses records with poor geographical indications. As excluding poor quality coordinates can result in a very low number of records for the analysis, the threshold for including or excluding coordinates may need to be balanced with the number of resulting records.

The result of this exercise is a passport data set of known occurrences of the target taxon, with a minimum of duplicate records, and all with verified geographical coordinates.

3. Application of the ecogeographical filtering method

The ecogeographical filtering method has been developed and tested by working with abiotic traits, e.g. drought resistance or aluminum toxicity. Possible uses with biotic traits such as pest resistance are outlined in Section 3.4, together with considerations of the relevant data requirements.

In addition to the compilation of the occurrence data sets, the ecogeographical filtering method requires the identification of appropriate variables that characterize the environmental profile of sites in which the adaptive trait of interest is likely to develop in the target taxon and an ecogeographical land characterization (ELC) map, the development of which is discussed in Section 3.1 below.

As a first step in this method, the ecogeographical categories from the ELC map are assigned to each occurrence record according to its coordinates and the records are then grouped according to their ELC map category. After all georeferenced occurrences have been ecogeographically characterized, the second step is to select from each group occurrences that comply with specific environmental requirements related to the trait of interest. The specific ecogeographical variables that best describe and delimit the environmental profile likely to impose selection pressure for the adaptive trait of interest (see Section 3.2) are used for this filtering, which generates an ecogeographical core set. To obtain the final subset the ecogeographical variable of interest is used to rank the ecogeographical core set, based on an identified threshold value, and to select the records with the highest or lowest values of the variable of interest.

3.1 Ecogeographical land characterization map

ELC maps offer an objective and reproducible strategy for defining useful ecogeographical categories to identify potential variants in plant adaptation. The groups of variables used to construct ELC maps, i.e. climatic, edaphic and geophysical variables, are factors that might generate local adaptation. An ELC map delineates areas with similar environmental characteristics and aims at representing the adaptive scenarios that are present over the territory studied (Parra-Quijano, Iriondo and Torres, 2012). It is prepared for a specific country or region and can be developed also for a specific species or group of related species. Parra-Quijano, Iriondo and Torres (2012), for example, developed an ELC map for Spain and the Balearic Islands. The PGR Secure project (http://www.pgrsecure.bham.ac.uk/) developed genus-specific ELC maps at European level for *Avena*, *Brassica* and *Medicago*. An ELC map for *Beta* was developed in the AEGRO project (http://aegro.jki.bund.de/aegro/). It has been shown that ELC maps can be applied both to CWR and LR germplasm. They are particularly useful when developed for single-species or related-species groups (e.g. related grass or legume species (Parra-Quijano *et al.*, 2012).

Generic ELC maps developed for a specific territory try to structure the diversity of environmental conditions that is found in the territory that might be relevant to plants in general. ELC maps that are built specifically for particular species (crop)—like the ELC maps generated within the PGR Secure project—take into consideration the requirements, limitations and vulnerabilities of the taxon and represent the environmental adaptive scenarios occurring across the territory.

It should be noted here that the study could use pre-existing ecogeographical maps that may have been generated for other purposes, provided that they have an appropriate scale (the ecogeographical units could be too coarse or too small) and that they respond to the main factors that condition plant life. There are several types of maps that try to describe ecosystems, ecoregions or life zones. Most of them were built using diverse criteria (including anthropogenic factors) and specific crop or CWR

adaptive conditions probably were not considered. For this reason, the use of purpose-made ELC maps is recommended for the ecogeographical filtering method.

The steps that are generally required for the production of an ELC map are summarized below and more details are provided by Parra-Quijano, Iriondo and Torres (2012) and Parra-Quijano *et al.* (2014). They are implemented in the CAPFITOGEN tool ELCmapas, which is used here to generate ELC maps at country level.

- Environmental variables that are considered to be meaningful in terms of plant adaptation are compiled – and if necessary converted – in layers using DIVA-GIS, ArcGIS, QGIS or similar GIS tools.
- Variables selection process starts, trying to avoid highly correlated and/or collinear variables within each group (edaphic, geophysical and bioclimatic variables).
- For taxon-specific ELC maps, those variables that mainly influence adaptation of the taxon and therefore shape their distribution are identified.
- The selected variables are subjected to a cluster analysis. This clustering procedure will determine ecogeographical similarity among all the cells considered in the work frame. Then an objective process will establish the number of final clusters to maximize intergroup and minimize intragroup variation. There are two different methods available to determine the optimal number of groups in ELCmapas tool (CAPFITOGEN program), the elbow (Ketchen and Shook, 1996) and medoides (Kaufman and Rousseeuw, 1987; Rousseeuw, 1987). Both approaches can be implemented in R (Parra-Quijano *et al.*, 2014). Another alternative for clustering and optimal number of groups determining is the use of the Two Step Clustering (TSC) associated with the Bayesian Information Criterion in SPSS software (Parra-Quijano, Iriondo and Torres, 2012).
- The clusters generated for each group are combined to generate the ecogeographical categories.
- The map of ecogeographical units is generated as an ASCII file (.asc). The final ELC map shows each ecogeographical category in a different color. Each category from any ELC map can be defined in terms of means (for quantitative) or mode (for categorical) of the original ecogeographical variables (i.e. precipitation, temperature, soil type, slope, etc.).

As a first step, the environmental variables most relevant for adaptation and for determining the species' distribution need to be carefully selected. These are required to generate the ELC map and need to be determined prior to using the ELCmapas tool.

The identification of the variables that influence adaptation of the taxon and therefore shape the distribution, is a critical step in the development of the ELC map in order to distinguish and represent correctly the adaptive scenarios on the map. They are commonly identified based on a literature review and consultation with experts.

A literature search identifies scientific and technical publications that report environmental factors that influence, determine or limit the distribution of the taxon.

Experts working on the taxon should be identified and contacted to obtain their knowledge about factors that shape distribution of the species. Although it introduces some element of subjectivity, it supports the selection of the most important variables for the ELC map creation.

The factors identified often need to be matched to an available ecogeographical variable in order to support the development of the map.

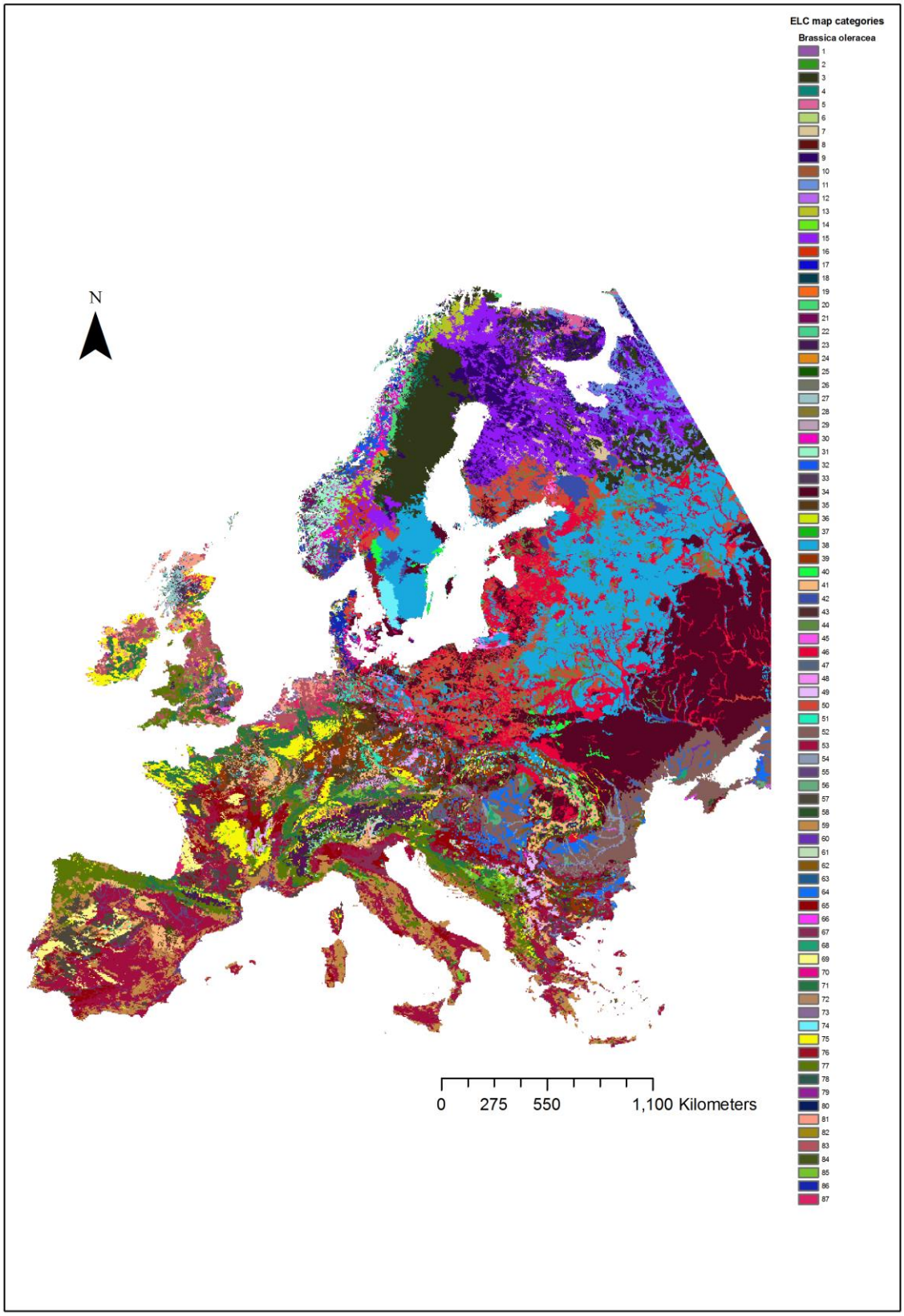


Figure 3.1. ELC map generated in the PGR Secure project for *Brassica* species

Once the list of potential variables has been developed, it is necessary to assess for each group of variables (edaphic, geophysical and bioclimatic) whether there are variables that could contribute redundant information. Those should be checked for co-linearity and correlations, and if significant correlation or co-linearity values are found, only one of the variables should be included. If all possible variables of one group are quantitative, a principal component analysis could help to understand the relationship among variables and help to select the most significant one.

Not more than five variables per edaphic, geophysical and bioclimatic variables group should be selected. ELCmapas includes a total of 105 variables among the three groups.

Figure 3.1 shows the ELC map for *Brassica* species developed during the PGR Secure project. The environmental variables used to generate the map are shown in Table 3.1.

ELCmapas generates maps that can be opened and read with DIVA-GIS. You need to convert the map into ASCII format with the function Export Gridfile in the DIVA-GIS Data menu in order to make the ELC map available for the R script used in Section 3.3.

The result of this step is a taxon-group-specific ELC map available in DIVA-GIS readable format and in ASCII format.

Table 3.1. Environmental variables used to generate the ELC map for *Brassica* species during the PGR Secure project

Bioclimatic variables	Edaphic variables	Geophysical variables
BIOCLIM 1	Topsoil texture (T_Texture)	Latitude
BIOCLIM 2	Topsoil pH (T_PH_H2O)	Longitude
BIOCLIM 6	Total exchangeable bases in topsoil (T_TEB)	Global irradiation on an optimal inclination (SOLARDOP)
BIOCLIM 11	Topsoil salinity (T_ECE)	Northness
BIOCLIM 12	Topsoil organic carbon (T_OC)	Eastness
PRECIP5		Slope degree (SLOPEDG)

3.2 Environmental profile

For the selection of LR or CWR occurrence subsets with potential interest for breeders and researchers working on specific abiotic resistances or tolerances, it is necessary to describe the specific environmental profiles in which the selected traits may have evolved in the taxon of interest. This requires identifying the most appropriate variables that characterize those environments. In addition, threshold values need to be identified for each variable, above or below which a specific site should be taken into consideration. Table 3.2 provides examples for traits, related variables and threshold values.

Table 3.2. Examples from the PGR Secure project for traits and variables for the project's target genera *Avena*, *Beta*, *Brassica* and *Medicago*

Genus	Identified abiotic trait	Identified variable(s)	Threshold value
Avena	Aluminium toxicity	Soil pH; Soil organic carbon content T_OC	< pH 5.5 < 1.2% T_OC
Beta	Drought	De Martonne aridity index (De Martonne, 1926), calculated based on temperature and precipitation of the three driest months (July, August and September in the Northern Hemisphere).	< 10
Brassica	Drought	De Martonne aridity index	< 10
	Salinity	Topsoil salinity (TSS) measured as electrical conductivity in dS/m (deciSiemens/metre)	> 4 dS/m Highest values in records with TSS > 4
		Mean temperature values for the driest months	
Medicago	Frost	BIOCLIM 11	colder than -2°C

The relevant environmental variables and their threshold values that can characterize environments where the trait of interest can evolve are mainly determined through literature searches and consultation with experts working on the taxon and/or trait of interest. Scientific and technical literature about the trait of interest within the taxon and geographical area of study can provide indications of the environment where the trait was found. Breeders and researchers working on the taxon and the trait should also be consulted, to obtain expert views on relevant variables as well as on evaluation. As an example, the selection of environmental variables relevant for drought stress in *Beta* is described as an example in detail in Box 3.1.

Box 3.1 Drought resistance in *Beta* – results from the PGR Secure project

An important trait in *Beta* crops is drought resistance. To identify CWR occurring in sites where drought resistance traits are likely to develop, appropriate variables were needed to identify the sites. Simple precipitation values do not explain plant water availability. For instance, low annual precipitation in a particular location does not necessarily mean that plants may be subject to selective pressures for drought resistance. It will depend on how precipitation is distributed throughout the plant life cycle, the ability of the soil to retain water and on the distribution of temperatures throughout the period considered. Higher demand for water is expected during the growing stages and when temperatures are higher, i.e. when evapotranspiration values increase. Aridity indices are one way of bringing together precipitation and temperature (using temperature as a measure to estimate evapotranspiration) and the De Martonne index (Iar; De Martonne, 1926) was identified as the most appropriate for the ecogeographicalal filtering method. Following the De Martonne index, aridity can be classified as follows:

Classification of zones according to De Martonne index

Iar-DM value	Zone classification
0–5	Extremely arid (desert)
5–10	Arid (steppic)
10–20	Semi-arid (Mediterranean)
20–30	Sub-humid
30–60	Humid
> 60	Per-humid

Source: Adapted from Almorox Alonso, 2003.

For the 1596 available georeferenced *Beta* CWR occurrences in Europe the De Martonne aridity index was calculated using the temperature and precipitation of the three driest months (July, August and September).

Iar_i-DM = $12*P_i/(tm_i+10)$, where P_i is the mean precipitation of month *i* and tm_i is the mean temperature of month *i* in Celsius degrees.

Habitats for CWR likely to contain genetic diversity for resistance to drought would correspond to those with De Martonne aridity values below 10, including both the arid and extremely arid categories. Using the De Martonne aridity index and the identified threshold value when applying the ecogeographical filtering method to the *Beta* CWR data set, 31 occurrences were identified as growing or collected in habitats with an aridity index below 10.

Once the variable(s) have been identified, the values for each variable need to be extracted for each location and added to the occurrence data set, as they will be required for the R script. These variables can be geophysical, edaphic and bioclimatic variables and are extracted from relevant data sources such as those listed below. If you work at a national or sub-national scale, geographical or environmental national agencies, weather stations and meteorological services might provide additional, more detailed or more up-to-date layers.

Note that sometimes the most appropriate parameters helping in the selection of the subset might have to be calculated from other variables. That is the case, for example, for the De Martonne aridity index, which has been identified as more appropriate than precipitation alone to describe drought-prone environments (see Box 3.1). Once this new parameter is calculated, it has to be added to the occurrence data set with which you are working.

Geophysical data

Geophysical data such as elevation can be downloaded as layers from the Digital Elevation Model (DEM) (SRTM data version 4, http://srtm.csi.cgiar.org/). Data for each site can then be extracted from the layers with DIVA-GIS or other GIS software. Other variables, such as slope, aspect, northness and eastness[1], can be derived from elevation layer (DEMs) using aspect, slope, Cos and Sin functions from the ArcGIS / Arc toolbox.

Edaphic data

Edaphic data can be downloaded from the Harmonized World Soil Database at http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/HWSD_Data.html?sb=4

Bioclimatic data

Climate data specific to the locations in the occurrence data set can be extracted with the 'Extract data by Point from climate data' function in the DIVA-GIS Data menu. Global current climate data sets in DIVA-GIS-compatible format can be downloaded from the DIVA-GIS Web site at http://www.diva-gis.org/climate in 10, 5 and 2.5 minute resolution. If you require climate data at higher resolution, you can download climate data directly from WorldClim.

WorldClim – Global Climate Data (http://www.worldclim.org/) (Hijmans *et al.*, 2005) provides a set of global climate layers with a maximum spatial resolution of 30 seconds (about 1 square kilometre). It provides generic grids (raster files) which can be imported into most GIS applications, and ESRI grids which can be used in ArcMap, ArcInfo and ArcView. The site allows you to choose the appropriate resolution (30 seconds, 2.5 minutes, 5 minutes or 10 minutes) and the precipitation, temperature and BioClim variables you need (Figure 3.2).

[1] Northness and eastness are calculated from the aspect layers, which are obtained from the DEM or elevation layer: northness = cos(aspect); eastness = sin(aspect).

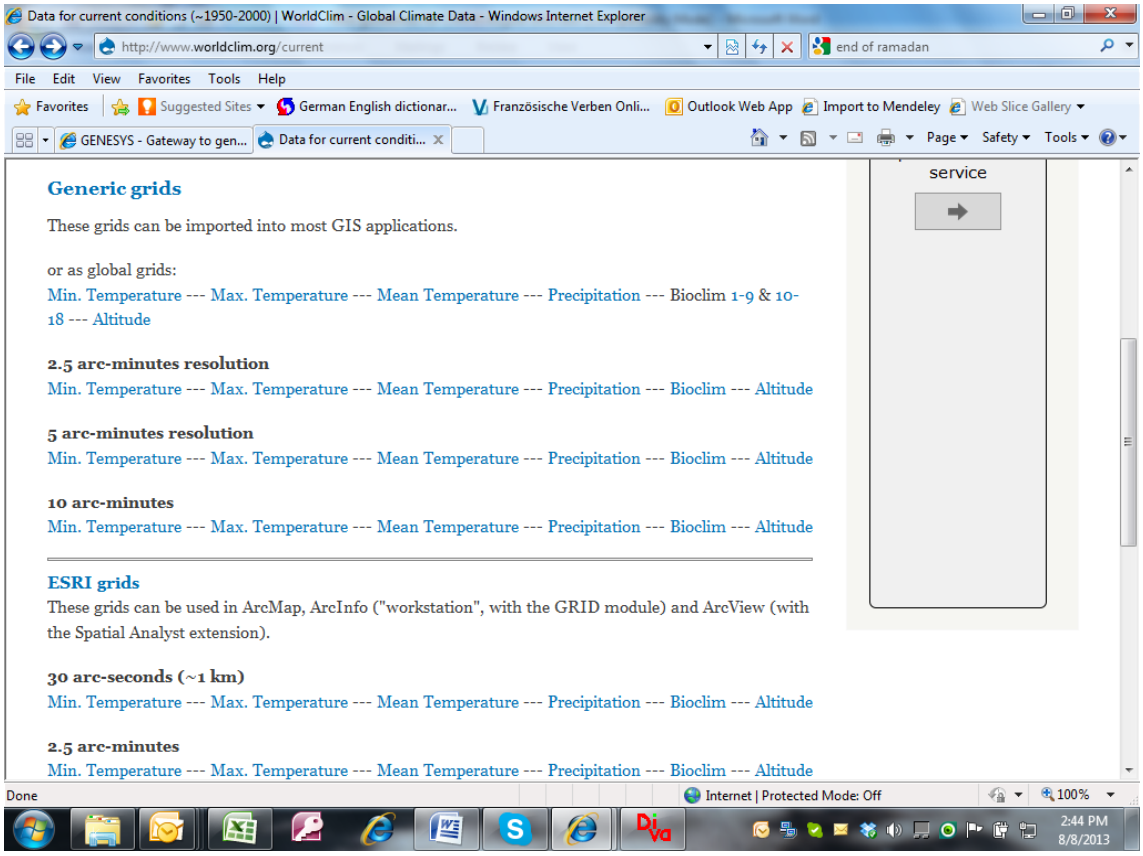


Figure 3.2. Data types available from WorldClim at http://www.worldclim.org/current, when global grids are downloaded.

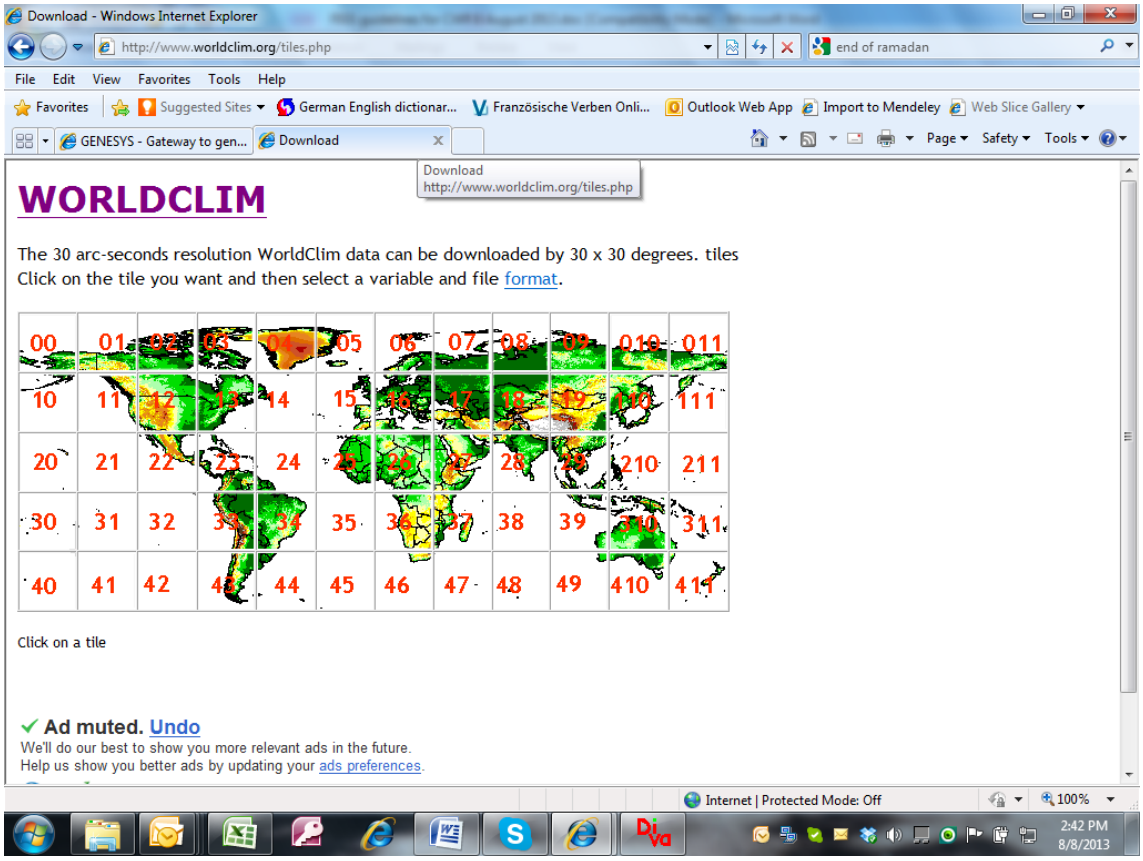


Figure 3.3. Data download by tile from WorldClim if high resolution data are requested.

If you require the highest resolution (30 arc-seconds [~1 km]), data are also available for download by tile (Figure 3.3).

See Section 4.2 for some examples on how to download and extract environment layers for your occurrence locations using the R software platform.

The identified variables and threshold values are included in the dataset to be used together with the ELC map to extract the final subsets for the target taxon from your data set.

3.3 Selection of occurrence set

The R environment is used to assign the ecogeographical category to each occurrence point and to generate the ecogeographical core set and subsequently the final occurrence subset, based on the ELC map and the threshold value of the critical environmental variable. An example R script, which uses *Avena* data and maps generated in the PGR Secure project, is provided below. The URLs from where the example *Avena* occurrence datasets, ELC map and raster file for the environmental variable can be downloaded are included in the script. You will need to adapt the working directory to the place where you download the files. Further adaption is required throughout the script to run it with your own occurrence data set, environmental variables and maps.

The following steps are executed within the R script:

- The ELC map is overlaid with the occurrence points of your dataset and the frequency of points that belong to each ecogeographical unit are obtained.
- Then, the number of samples to be taken from each ecogeographical unit is determined. The R script gives the choice to select a proportional allocation approach, a uniform allocation approach or a combination of both through the parameter *a* (see R-script), which ranges from 0 to 1 (0: for total uniform allocation; and 1: for total proportional allocation).
- The records belonging to each ecogeographical unit are ranked according to the environmental variable identified in Section 3.2 used to look for adaptation for the particular abiotic stress tolerance, and those records that better comply with this variable are selected, according to the allocation numbers generated in the previous step.
- The sum of these records produces an optimized ecogeographical core set.
- To obtain the final subset of interest, the ecogeographical core set is ranked based on the environmental profile variables and either all records above or below the previously determined threshold value are selected, or the 100 (or any other number defined by the user) highest ranking records are selected.

You should copy the example R script provided below into your R environment and adapt it to the respective file and folder names of your working environment as well as to your selected environmental variable(s). The example script (Code box 3.1) includes explanations and indications where you need to modify it.

The files you need to place in your R working folder are the following:

- Baseline data set as tab-delimited text file (.txt). Make sure that the values for the selected environmental profile variables are included.
- The ELC map or any existing ecogeographical map as an ASCII layer (.asc).
- The ASCII layer of your environmental profile variables, but only if you wish to plot the occurrences included in the final subset on the map showing this variable.

The R script generates a text file containing the identified final set of occurrences.

```
## CODE BOX 3.1

# Example R-script for the ecogeographical filtering method

setwd("G:/PredictiveCharacterization/Data") # sets the working directory
# Adjust this to the address of your working folder (“./your_path/data”)

Install.packages("raster") # install this package if it is not part of your installed
R-packages yet

library(raster)

# Download demo data
download.file("http://trait-mining.googlecode.com/svn/trunk/data/avena/AvenaCWR.txt",
"./AvenaCWR.txt")
download.file("http://trait-mining.googlecode.com/svn/trunk/data/avena/AvenaLR.txt",
"./AvenaLR.txt")
download.file("http://trait-mining.googlecode.com/svn/trunk/data/avena/elcmap_AvenaSativa.zip",
"./elcmap_AvenaSativa.zip")
download.file("http://trait-mining.googlecode.com/svn/trunk/data/avena/t_ph_w84.zip",
"./t_ph_w84.zip")



# Read Avena demo data set into R
baseline <- read.delim("./AvenaCWR.txt", header=TRUE, dec=".")

# AvenaCWR.txt contains the table with all occurrence data and the necessary ecogeo-
graphical information for Avena crop wild relatives. You can use AvenaLR.txt and
modify the code accordingly to run the script with the example data for Avena
landraces

# unzip the raster and read
unzip("./elcmap_AvenaSativa.zip")
elc<-raster("./elcmapAvena.asc")

# elcmapAvena.asc contains the ELC map for Avena

ecogeo <-extract(elc,baseline[,c("LONDEC","LATDEC")])
# ecogeo extracts from the ecogeographical map the ecogeographical units that
correspond to each occurrence point
baseline2<-cbind(baseline,ecogeo)
# baseline2 adds the column ecogeo with the ecogeographical units to the baseline
table,this will only work if all occurrence data contain coordinates

FAVENA<-table(baseline2$ecogeo)
FAVENA <- cbind(names(FAVENA),as.vector(FAVENA))
# The function 'table' obtains the frequencies of occurrences in each ecogeographical
unit
# cbind combines the names of the ecogeographical units with their corresponding
frequencies
# The ecogeographical units are usually numbered, e.g. from 1 to 78 in ELC map for
Avena created in the PGR-Secure project

# Here, aluminium toxicity is the target trait for Avena, and we use the following
variables as proxy for aluminum content: T_PH_H2O and T_OC. Values for these
variables are included in our occurrence data set
i <- order(baseline2$T_PH_H2O)
baseline2 <- baseline2 [i,]
# We order the subset according to the variable of our choice creating an index(i).
```

```
 n <- 1000
# n is the number of records we want to have in our optimized ecogeographical core
set
nc <- dim(FAVNEA)[1]
# nc is the number of ecogeographical units that have occurrences
# The dim function provides the dimension of the object (1:rows; 2:columns)


f <- as.numeric(FAVENA[,2])
# f provides the frequency of each ecogeographical unit

prop <- f/sum(f)
# Provides the proportional value of each frequency
even <- pmin(prop,rep(1/nc, times=nc))
even <- even/sum(even)
# even object provides the minimum value between the proportional value and the even
share
# In the second row it is adjusted to sum a total frequency of 1.

a <- .5
samples <- (a*prop + (1-a)*even) * n
# Samples is the number of samples that are to be selected from each ecogeographical
unit
# It combines the proportional and the even allocation approaches through the
parameter "a"
# a=1 provides a complete proportional allocation
# a=0 provides a truncated even allocation (even for those values where proportional
is greater than even)

samples[is.na(samples)] <- 0
# Provides 0 value to samples if na is obtained from log(0) in the previous step
samples <- pmax(round(samples),f>0)
# Rounds the values of samples and provides values of at least 1 for those
frequencies that are greater than zero

i <- order(baseline2$T_PH_H2O,decreasing=FALSE)
baseline2 <- baseline2 [i,]
# Reorders the whole subset according to the T_PH_H2O variable in ascending order.
# Change it to decreasing=TRUE if you wished to reverse the order

baseline2 <- baseline2 [!is.na(baseline2$ecogeo),]
# Eliminates the records that have na values for the field ecogeo. For T_PH_H2O na=na
and is not -9999, like in BIOCLIM variables.
baseline2 <- subset(baseline2, !is.na(baseline2$T_PH_H2O))
# Eliminates the records that have no data for the variable of interest for selection

egreg <- unique(baseline2$ecogeo)
egreg <- egreg[order(egreg)]
# egreg provides an ordered rank of the values of the ecogeographical units

bss <- NULL
# Creates a new object with empty values

for(j in 1:nc)
{
  bs <- subset(baseline2, ecogeo == egreg[j])
# Selects one by one (through the loop) the records of each ecogeographical unit
  bss <- rbind(bss, bs[1:samples[j],])
# Adds to bss the number of records assigned to each ecogeographical unit
}
```

```
i <- order(bss$T_PH_H2O,decreasing=FALSE)
bss <- bss[i,]
# It reorders bss (the ecogeographical core set) according to the variable of
interest


hist(bss$T_PH_H2O)
# Provides the results of the ecogeographical core set in a histogram according to
the variable of interest


semifinalsubset<-subset(bss,T_PH_H2O<=5.5)
# Selects a semifinal subset by setting a threshold in the variable of interest, here
topsoil pH below 5.5)


finalsubset<-subset(semifinalsubset, T_OC<=1.2)
# Selects from all the records in semi finalsubset (which are all those that meet the
first criteria of having a PH below the threshold)
# Those that have in addition an organic carbon content below 1.2%.
write.table(finalsubset, file = "taxon1_finalsubset_PHbelow55_TOCbelow12.txt", sep =
"\t", col.names = NA, qmethod = "double")


# If desired you can create a list of the 100 "best" records
Best100<-bss[1:100,]
# Selects a second final subset with the first best 100 records according to the
variable of interest T_PH_H2O
# This could include - in theory - also records where the variable is above our
threshold - if there were less than 100 occurrences with PH<5.5


# In the case of Avena we work with two variables and need to add to the selection of
the 100 best records the additional criteria t_OC < 1.2
# Therefore exclude from bss all records with T_OC > 1.2% and then reorder first on
PH and then on T_OC.
# Instead of best100<-bss[1:100,] you use the following script as we have two
variables:
bss2<-subset(bss, T_OC<=1.2)
i <- order(bss2$T_PH_H2O,decreasing=FALSE)
bss3 <- bss2[i,]
finalbest100<-bss3[1:100,]
write.table(finalbest100, file = "taxon1_100best_PHbelow55_TOCbelow12.txt", sep =
"\t", col.names = NA, qmethod = "double")


# Plotting the sub sets:


xmin<-min(finalsubset $LONDEC)-4
xmax<-max(finalsubset $LONDEC)+4
ymin<-min(finalsubset $LATDEC)-2
ymax<-max(finalsubset $LATDEC)+4
# Provides the margins for the extent of the territory to cover in the map taking
into account the distribution of points of the final subset
# The most extreme points of the distribution are provided and then we add some
degrees to make it a bit larger


# Adjust the margins of the map by changing the number you add or subtract
Extent<-extent(xmin,xmax,ymin,ymax)
# Defines the extent of the map


# unzip the raster and read
unzip("./t_ph_w84.zip")
pH_H2O <-raster("./t_ph_w84.asc")
# Obtains the map of the variable of interest.
```

```
pH_H2O <-raster("G:/PredictiveCharacterization/Data/t_ph_w84.asc")
# Obtains the map of the variable of interest. This needs to be placed in your
working folder


pH_H2O <-crop(pH_H2O,Extent)
# Crops the map of the variable of interest to the extent defined


par(mfrow =c(1,2))
# Prepares a figure with two maps left and right. If you prefer printing maps
separately, skip that step.


pH_H2O <-crop(pH_H2O,Extent)
plot(pH_H2O)
points(finalsubset[,c("LONDEC","LATDEC")],pch=20,cex=.1)
# Plots the map with the variable of interest and the distribution of the figs points


elc <-crop(elc,Extent)
plot(elc)
points(finalsubset[,c("LONDEC","LATDEC")],pch=20,cex=.1)
# Plots the map with the ecogeographical map and the distribution of final subset
points
```

3.4 Use with biotic resistance traits as variables

The ecogeographical filtering method can be used also with biotic resistance traits, e.g. pest resistances, if detailed occurrence data of a pest of interest is available, or if the environmental niche of a pest is known.

In the simplest case, taxon occurrences could be classified as co-occurring with the pest or not, and those occurrences co-occurring with the pest be selected for further evaluation and research.

If pest occurrence data are available with density information, the taxon occurrences could be classified as falling within a high, medium or low density area. This classification could be used as variable to generate an optimized ecogeographical core subset and the final subset.

Another possibility can be to calculate a predicted distribution for the pest based on ecological niche modelling methods, such as Maxent. The predicted distribution could be applied using the resulting habitat suitability index as variable to generate the optimized ecogeogrpahical core subset. Locations with habitat suitability values above a particular value would be selected as final subset.

If the environmental niche of the pest is known, variables delimiting this niche can be used to generate the optimized ecogeographical core set and define the final subset.

4. Application of the calibration method

The calibration method for predictive characterization can be used when you have access to a training set with characterization or evaluation data for your trait of interest. This training set (with known trait measurements) is required to calibrate a predictive model that can be used to calculate an estimate for the respective trait for other similar accessions not yet tested for this trait. Calibration of such predictive models can easily lead to very specific models that are over-fitted to the training set. This issue demands careful attention to find a balance for the appropriate model fitting. Even if the model has good predictive performance on the training set, it may not perform well on other similar accessions. The problem of over-fitting may often be that the model is too complex and too closely fitted to the training set. To find an appropriate model complexity the iterative model calibration should be stopped when a generic pattern is identified. To evaluate if the model is sufficiently generic to perform well on other similar accessions we need a test set of similar accessions with known trait data. It is highly recommended to use an independent test set and not just perform a simple data splitting on the training set. However, a separate and independent test set may not be available, and creating a test set by data splitting may for many experiments be the only practical option available.

This chapter provides example R code to perform a predictive characterization study using the calibration method for classification. For trait measurement values on a continuous scale you may want to use a regression algorithm (which is not covered in the demonstration example here). The demonstration example explores a prepared data set with occurrence and trait data for stem rust made available by USDA GRIN[2] (public domain), and environment data from the WorldClim data set[3] ("freely available for academic and other non-commercial use"). This is the same stem rust data and modelling approaches that was used by Endresen *et al.* (2011) and Bari *et al.* (2012). The example data set used in these examples can be downloaded from: http://goo.gl/Xp2dwq[4]. The R code presented in this Chapter 4 is from an R script made available at: http://goo.gl/FwZMZB[5].

For the examples presented in these guidelines, we use algorithms such as the Random Forest (Breiman, 2001) for calibration of the model, one of the methods that had performed well in previous FIGS studies (Bari *et al.,* 2012). The use of other algorithms, such as kNN and Boosted Regression Trees (BRT), should also be explored when conducting a practical study.

4.1 Installation of required R packages

In the examples below, some of the R code instructions require specific R-packages to be installed and loaded. The additional packages that are required are indicated using a comment at the end of the respective R code lines. To install the packages, use the graphical user interface (GUI) of RStudio or issue the command line as is described in Code box 4.1. When the package is installed, you only need to load it before using its functions and features in an R script.

[2] http://www.ars-grin.gov/cgi-bin/npgs/html/desc.pl?65049

[3] http://www.worldclim.org/current

[4] http://trait-mining.googlecode.com/svn/trunk/data/stemrust/stem_rust_set.txt

[5] http://trait-mining.googlecode.com/svn/trunk/R/stem_rust_example.R

```
## CODE BOX 4.1

install-packages(c("maps", "mapdata"))
install-packages("randomForest")
install-packages("raster")
install-packages("rgbif")

library(maps) # we will use this R-package to plot a world map
library(raster) # spatial raster data management
library(rgbif) # download species occurrences from GBIF
require(randomForest) # this example will use the Random Forest algorithm
```

4.2 Addition of climate data

The calibration method works preferably with climate data. Relatively fine-scale climate data at a 1 km (30 second) grid resolution would often improve accuracy of the models. However, it requires sufficient computing capacity and it is a large quantity of data to be downloaded. For data download of climate data and extraction of climate data for occurrence points see Section 3.2.

Climate data can also be extracted during the application process of the calibration method, including the necessary code in the R script.

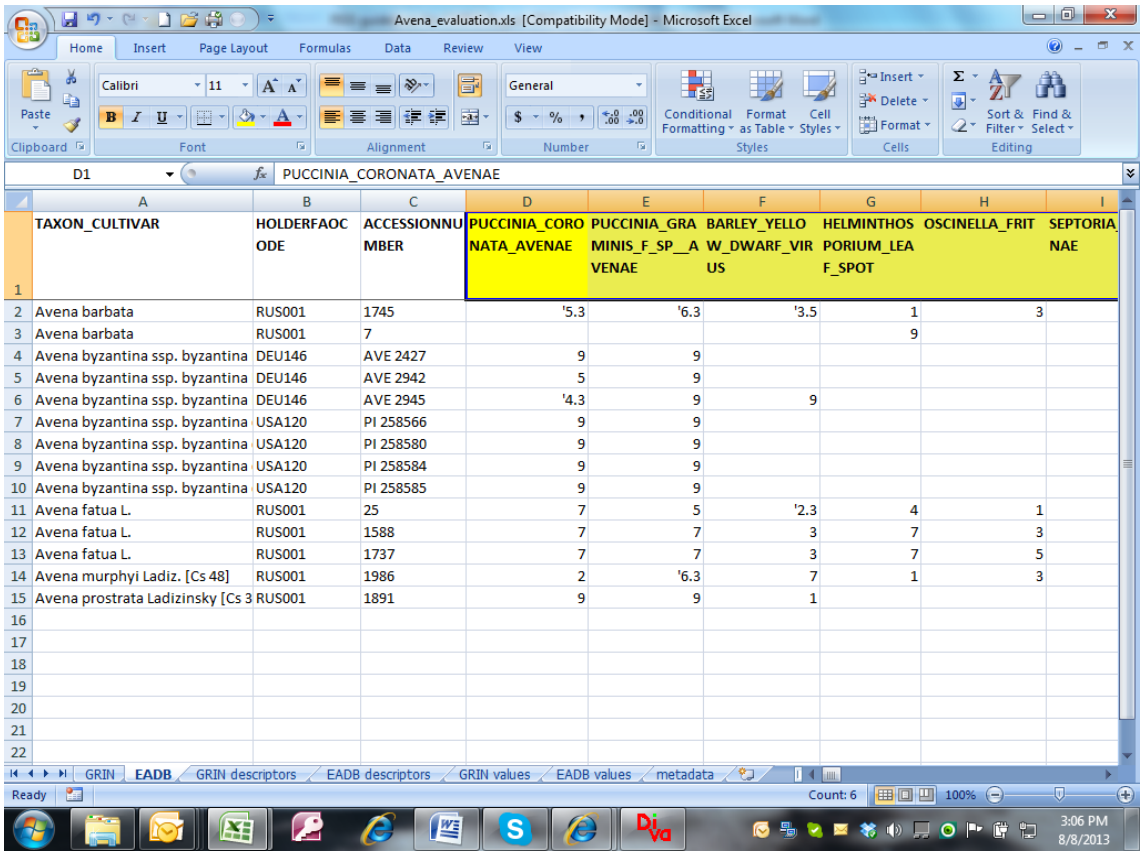
The GBIF portal provides occurrence data for most of the CWR species. The following example illustrates how you can download occurrence data from the GBIF portal (Chamberlain *et al*., 2014) and link these to environment layers from WorldClim using R (Hijmans, 2014). You can skip this example if you already have all the occurrence data you need.

```
## CODE BOX 4.2

# GBIF, http://www.gbif.org # rgbif, http://cran.r-project.org/package=rgbif
key <- name_backbone(name='Beta vulgaris')$speciesKey # taxonKey=5383920
# Example here is limited to 1000 (maximum limit is 1 million records per search)
bv <- occ_search(taxonKey=key, return='data', hasCoordinate=TRUE, limit=1000)
xy <- cbind('species'=bv$name, 'lon'=bv$decimalLongitude, 'lat'=bv$decimalLatitude);

# You may want to check the occurrence data downloaded from GBIF for duplicates
# and/or combine the occurrences with data from other sources.
# Uncomment the line below to write your occurrence data to a tab-delimited file
# write.table(xy, file="bv_set.txt", sep="\t", col.names=NA, qmethod="double")
# Uncomment the line below to read corrected occurrence data back into R
# xy <- read.delim("./YOUR_PATH/bv_set.txt", header=TRUE, dec=".")

map('world') # R-package: maps and mapdata
points (xy$lon, xy$lat, col='red') # plot points

# WorldClim, http://www.worldclim.org/
env <- getData('worldclim', var='bio', res=10) # (pkg raster)
Xbio <- extract(env, xy); # extract environment to points (pkg raster)

plot(env, 1) # plot the first bioclim layer
points (xy$lon, xy$lat, col='red') # plot points onto bioclim map
```

4.3 Trait data compilation

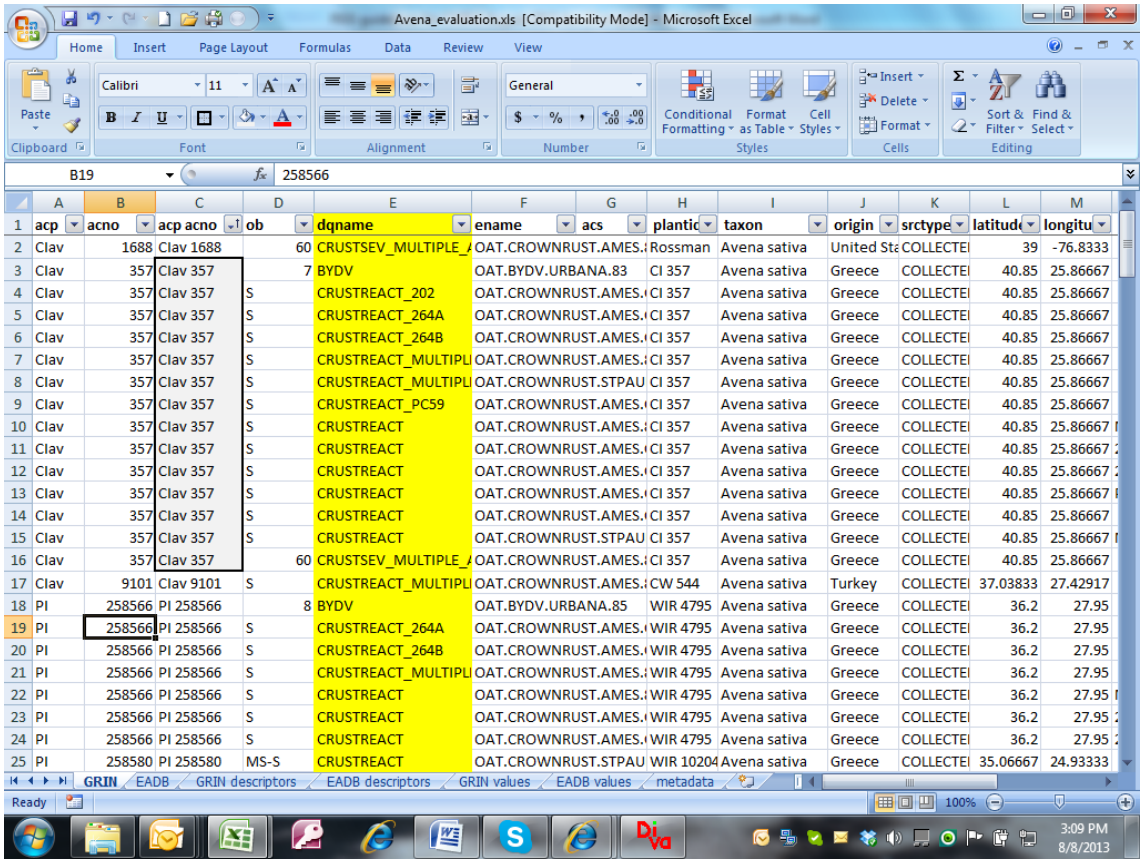
Characterization or evaluation data for traits of interest for the target taxon may be contained in the databases listed in Section 2.1, or may require contacting data curators, breeders or crop experts. The more evaluation data that are available, the better the calibration will work.

Trait evaluation data can have different formats in different databases. All trait scores for different variables can be listed in the same column, one after the other, as is used in GRIN, while other data sources have the variables distributed in different columns. It is recommended to follow the GRIN format when building a trait database, but you should prepare a cross-table with each variable in a separate column before using the calibration method.



	A	B	C	D	E	F	G	H	I
1	TAXON_CULTIVAR	HOLDERFAOCODE	ACCESSIONNUMBER	PUCCINIA_CORONATA_AVENAE	PUCCINIA_GRAMINIS_F_SP__AVENAE	BARLEY_YELLOW_DWARF_VIRUS	HELMINTHOSPORIUM_LEAF_SPOT	OSCINELLA_FRIT	SEPTORIANAE
2	Avena barbata	RUS001	1745	'5.3	'6.3	'3.5	1	3	
3	Avena barbata	RUS001	7				9		
4	Avena byzantina ssp. byzantina	DEU146	AVE 2427	9	9				
5	Avena byzantina ssp. byzantina	DEU146	AVE 2942	5	9				
6	Avena byzantina ssp. byzantina	DEU146	AVE 2945	'4.3	9	9			
7	Avena byzantina ssp. byzantina	USA120	PI 258566	9	9				
8	Avena byzantina ssp. byzantina	USA120	PI 258580	9	9				
9	Avena byzantina ssp. byzantina	USA120	PI 258584	9	9				
10	Avena byzantina ssp. byzantina	USA120	PI 258585	9	9				
11	Avena fatua L.	RUS001	25	7	5	'2.3	4	1	
12	Avena fatua L.	RUS001	1588	7	7	3	7	3	
13	Avena fatua L.	RUS001	1737	7	7	3	7	5	
14	Avena murphyi Ladiz. [Cs 48]	RUS001	1986	2	'6.3	7	1	3	
15	Avena prostrata Ladizinsky [Cs 3	RUS001	1891	9	9	1			
16									
17									
18									
19									
20									
21									
22									

Figure 4.1. Evaluation data for *Avena* from the European *Avena* Database. Evaluation data in different columns are for different traits of one accession.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	acp	acno	acp acno	ob	dqname	ename	acs	plantic	taxon	origin	srctype	latitud	longitu
2	Clav	1688	Clav 1688	60	CRUSTSEV_MULTIPLE_/	OAT.CROWNRUST.AMES.(		Rossman	Avena sativa	United Sta	COLLECTEI	39	-76.8333
3	Clav	357	Clav 357	7	BYDV	OAT.BYDV.URBANA.83		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
4	Clav	357	Clav 357	S	CRUSTREACT_202	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
5	Clav	357	Clav 357	S	CRUSTREACT_264A	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
6	Clav	357	Clav 357	S	CRUSTREACT_264B	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
7	Clav	357	Clav 357	S	CRUSTREACT_MULTIPLI	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
8	Clav	357	Clav 357	S	CRUSTREACT_MULTIPLI	OAT.CROWNRUST.STPAU		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
9	Clav	357	Clav 357	S	CRUSTREACT_PC59	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
10	Clav	357	Clav 357	S	CRUSTREACT	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
11	Clav	357	Clav 357	S	CRUSTREACT	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
12	Clav	357	Clav 357	S	CRUSTREACT	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
13	Clav	357	Clav 357	S	CRUSTREACT	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
14	Clav	357	Clav 357	S	CRUSTREACT	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
15	Clav	357	Clav 357	S	CRUSTREACT	OAT.CROWNRUST.STPAU		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
16	Clav	357	Clav 357	60	CRUSTSEV_MULTIPLE_/	OAT.CROWNRUST.AMES.(		CI 357	Avena sativa	Greece	COLLECTEI	40.85	25.86667
17	Clav	9101	Clav 9101	S	CRUSTREACT_MULTIPLI	OAT.CROWNRUST.AMES.(		CW 544	Avena sativa	Turkey	COLLECTEI	37.03833	27.42917
18	PI	258566	PI 258566	8	BYDV	OAT.BYDV.URBANA.85		WIR 4795	Avena sativa	Greece	COLLECTEI	36.2	27.95
19	PI	258566	PI 258566	S	CRUSTREACT_264A	OAT.CROWNRUST.AMES.(		WIR 4795	Avena sativa	Greece	COLLECTEI	36.2	27.95
20	PI	258566	PI 258566	S	CRUSTREACT_264B	OAT.CROWNRUST.AMES.(		WIR 4795	Avena sativa	Greece	COLLECTEI	36.2	27.95
21	PI	258566	PI 258566	S	CRUSTREACT_MULTIPLI	OAT.CROWNRUST.AMES.(		WIR 4795	Avena sativa	Greece	COLLECTEI	36.2	27.95
22	PI	258566	PI 258566	S	CRUSTREACT	OAT.CROWNRUST.AMES.(		WIR 4795	Avena sativa	Greece	COLLECTEI	36.2	27.95
23	PI	258566	PI 258566	S	CRUSTREACT	OAT.CROWNRUST.AMES.(		WIR 4795	Avena sativa	Greece	COLLECTEI	36.2	27.95
24	PI	258566	PI 258566	S	CRUSTREACT	OAT.CROWNRUST.AMES.(		WIR 4795	Avena sativa	Greece	COLLECTEI	36.2	27.95
25	PI	258580	PI 258580	MS-S	CRUSTREACT	OAT.CROWNRUST.STPAU		WIR 10204	Avena sativa	Greece	COLLECTEI	35.06667	24.93333

Figure 4.2. Evaluation data from GRIN. Evaluation data for different traits for one accession are in the same column.

It is important to understand how the evaluation data have been generated and which measurement protocols have been followed if evaluation data from different sources are used. Trait data compiled from different sources are often grouped into standardized categories following the Bioversity descriptor list recommendations (Gotor *et al.*, 2008) when used for computer modelling (for more details see Section 4.4 below), and values from different sources need to be correctly assigned to categories.

If you have evaluation data from different sources, the data source could be coded and included as one of the explanatory (independent) variables as input to the model. This will allow one to see to what extent the source itself influences the results of the model (by exploring the loading plot or similar statistics). Trait data compiled from different sources might follow slightly different measurement protocols. If these differences harm the analysis of these data together in one compiled data set, adding the experiment year and site as part of the explanatory variables might enable the computer model to incorporate these differences in the model and reduce the harmful effect from the different measurement practices.

Trait evaluation results are often a range within a scale of scores from several degrees of susceptibility to several degrees of resistance. When using classification algorithms, a better predictive performance can often be obtained by re-scaling the trait scores into fewer response category levels, e.g. into two or three category levels. These new categories can be included in the spreadsheet containing the data set or be calculated using R.

4.4 Preparation of training and test set

The trait data need to be combined with the occurrence data. Based on the accession number, the R script example in Code box 4.3 takes those records from the occurrence data set for which trait data exist, to create the training set for the calibration method.

```
## CODE BOX 4.3

# Write down "YOUR_PATH", "YOUR_GENUS_NAME" and "YOUR_SPECIES_NAME"

# Set the working directory
setwd("./YOUR_PATH/Data")
# Prepare traitfile.txt and masterfile.txt and copy to your working directory
trait<-read.delim("./YOUR_PATH/traitfile.txt")
# To correct the trait dataset for Beta we need to remove rows with missing ACCENUMB
trait <- subset(trait,ACCENUMB!="")
masterfile<-read.delim("./YOUR_PATH/masterfile.txt")
unitable<-merge(trait, masterfile, by.x = "ACCENUMB", by.y = "ACCENUMB", all.x=T)
# Creating an index and selecting only matching records
unitable <- unitable[!is.na(unitable$NUMCAT),]
# Deleting records from other genus (e.g. other than "Beta")
unitable<-subset(unitable,GENUS=="YOUR_GENUS_NAME")
# Deleting records from other species (e.g. other than "vulgare")
unitable<-subset(unitable,SPECIES=="YOUR SPECIES NAME")
# You may want to export the resulting merged table as tab-delimited text
write.table(unitable, file = "unitable.txt", sep = "\t", col.names = NA, qmethod =
"double")
```

The example script uses the accession or catalogue number as the link between the occurrence and the trait evaluation data set. If you work with occurrence data sets that include data for more than one taxon, the merge could return records for accessions with the same alphanumeric accession number, but from different taxa (i.e. the evaluation data of accession xyz of taxon A has been merged correctly with passport data of accession xyz of taxon A and wrongly with an accession of taxon B that carries the same accession number xyz). It is therefore good practice in these cases to check for double records and ensure that the evaluation data has been linked to the correct accession. The best practice is, of course, to always use globally unique and persistent identifiers for all the accessions and occurrence data. Unfortunately such persistent identifiers are not yet commonly available for germplasm accessions.

The resulting data set for the calibration method usually represents a subset of the available occurrence data set, as it includes only those records for which trait evaluation data are available.

As mentioned above, calibration and testing require two separate data sets. The ideal test set contains evaluation data for the same trait as the calibration set, but with the training and test sets respectively from two completely independent evaluation trials. If your data set is sufficiently large and a combined set that includes evaluation data from at least two independent and unrelated sources, it is advisable to use the data from one of the independent sources as an independent test set. Another option could be to split the trait data set into the respective training set and test set based on the trial year or trial site. This splitting approach assumes that different accessions are measured in different trial years or at different trial sites – as is common for C&E data sets (and as is the situation for the stem rust demonstration set). If the trait data set includes replications where the same accession is measured for each trial year and/or

trial location, this splitting approach will normally not be appropriate.

When you lack an independent test set, data splitting can be used as a second-best option. The data set is split in two parts, a training and a test set. Records can be assigned in a random way or in a systematic way to the two sets. The test set can, for example, be generated to include at least one-fifth to one-third of the samples in the working data set (Roy, Leonard and Roy, 2008). In our example we have split approximately half of the records to the training set and half to the test set. The training set is used to calibrate the model; the test set is used to test the model.

Two independent occurrence + trait data sets are created. Alternatively, a single occurrence + trait data set is available to be split into training and test sets.

4.5 Model calibration

The first step in the application of the calibration method is to load the data set into R. In our example (Code box 4.4) we have used a demonstration data set with accession occurrence and trait data from USDA GRIN, prepared and linked to environment data from WorldClim (Hijmans *et al.*, 2005).

```
## CODE BOX 4.4

# DEMO EXAMPLE : Download stem rust demo data set
# [the following command is one line in your script]
download.file("http://trait-mining.googlecode.com/svn/trunk/data/stemrust/stem_rust_set.txt",
"./stem_rust/stem_rust_set.txt")

# Read stem rust demo data set into R
sr <- read.delim("./stem_rust/stem_rust_set.txt", header=TRUE, dec=".")

# "s3" has the stem rust trait scores reclassified as three levels:
# 1 = resistant, 2 = intermediate, and 3 = susceptible germplasm accessions.
# [the following command is one line in your script]
Xbio <-
sr[c("s3","bio1","bio2","bio3","bio4","bio5","bio6","bio7","bio8","bio9","bio10","bio
11","bio12","bio13","bio14","bio15","bio16","bio17","bio18","bio19")]
```

You may want to preview the occurrences of the data set on a map. R provides numerous ways to open and display spatial map data. You may want to load a shapefile, such as by using the readShapePoly() function from the maptools R-package. Here we will just load a simple wireframe with country borders using the maps R-package (Code box 4.5).

```
## CODE BOX 4.5

map('world') # R-package: maps
points(sr[c("longitude","latitude")], col='red') # plot wheat set
```

The model ascertains an ecogeographical profile of the collecting sites of the training set and then – using statistical methods, i.e. random forest in this case – selects untested accessions from environments that are statistically similar to the training set environments.

When assessing the prediction performance for a categorical (or binary) response variable we can use a confusion table (also called confusion matrix) to calculate the performance metrics. The confusion table can be collapsed to a 2 by 2 table to calculate true positives, false positives, false negatives and true negatives (Table 4.1).

Table 4.1. Confusion matrix (2-by-2 contingency table)

	Observed resistant	Observed susceptible
Predicted resistant	True positive (TP)	False positive (FN)
Predicted susceptible	False negative (FP)	True negative (TN)

```
## CODE BOX 4.6

# Xbio [array] includes trait scores followed by the environment layers

# Autoscale = center around the mean, and divide by standard deviation.
scale(Xbio[,2:20], center=TRUE, scale=TRUE)

# Splitting into training and test sets
# An independent test set is highly recommended!!!
Xcal <- Xbio[1:3445,] # training set - for model calibration
Xtest <- Xbio[3446:6890,] # test set - to validate model performance

# Calibrate model using the training set
rf <- randomForest(as.factor(s3) ~ ., data=Xcal, ntr=50)
plot(rf) # preview

# Read the confusion table from the model object
conf <- rf$confusion
TP <- conf[1,1]
FP <- conf[2,1] + conf[3,1]
FN <- conf[1,2] + conf[1,3]
TN <- conf[2,2] + conf[2,3] + conf[3,2] + conf[3,3]

# Calculate calibration performance metrics
PO_cal <- (TP + TN)/sum(TP,FP,FN,TN) # proportion of observed agreement
PA_cal <- (2*TP)/(2*TP+FP+FN) # proportion of observed positive agreement
PPV_cal <- TP/(TP + FP) # positive predictive value
LRpos_cal <- (TP/(TP + FN)/(FP/(FP+TN))) # pos diagnostic likelihood ratio
Sensitivity_cal <- TP/(TP + FN)
Specificity_cal <- TN/(TN + FP)
```

The modelling of biotic resistance traits aims to detect germplasm with higher likelihood to contain adaptive genetic variation for resistances; i.e. it is more concerned with the identification of resistant samples than with an accurate classification of susceptible samples. Resistance to biotic stresses is generally a rare property with many more germplasm accessions susceptible to a given biotic stress. The distribution of the target trait property is often very skewed and this has implications for the prediction performance indicators we choose. If the distribution of the trait scores is indeed strongly skewed with few resistant samples, is

recommended to use a so-called “positive” prediction performance indicator in the calibration of the prediction models (Endresen *et al.*, 2011), as in these very skewed data sets it may provide a much better calibration approach (Table 4.2). This model calibration strategy of using so-called positive prediction metrics during model calibration will improve the tendency of the prediction model to correctly identify the target genetic diversity for biotic resistance. If a neutral prediction performance indicator is chosen, equal weight is given to correct identification of either the target resistant samples or the susceptible samples. With few resistant samples available, the model is not “punished” very much for prediction errors for these samples compared with the “reward” for predicting the susceptible samples correctly.

Some indicators that can be used to assess calibration performance of the model are listed in Table 4.2 (Endresen *et al.*, 2011). These are included in the R-script (Code box 4.6).

Table 4.2. Indicators to assess calibration performance

PO	Proportion of observed agreement	(TP + TN)/(TP+FP+FN+TN)
PA	Proportion of positive agreement	(2*TP)/(2*TP+FP+FN)
PPV	Positive predictive value	TP/(TP + FP)
Prevalence	Number or proportion of disease-resistant samples in relation to all samples.	TP//(TP+FP+FN+TN)
LR+	Positive predictive likelihood ratio (measures how much more likely it is for the model to predict a LR to be resistant in the group of LR observed to be resistant compared to making this prediction in the LR group observed to be susceptible).	(TP/(TP + FN)/(FP/(FP+TN)))
Gain	Improved predictive performance compared with a random selection.	PPV/prevalence

KEY: TP = True positive; TN = True negative; FP = false positive; FN = false negative

The different prediction performance indicators have some different properties that can be useful to keep in mind. The positive predictive value (PPV) provides a number between 0 and 1 making it easy to compare the numeric values where a higher number closer to 1 indicates a better prediction performance. However, it is important to remember that the PPV score is dependent on the prevalence, so great care should be taken when comparing PPV scores for trait data sets that can have different prevalence. Thus, if one data set includes 20% resistant samples, and another data set includes 10% resistant samples, care should be made when comparing the PPV scores. For these cases, the LR+ ratio is a useful indicator in that the numeric value is not inherently dependent on the prevalence in the same way. The LR+ provides numeric values not limited by 0 and 1 so it can be more difficult to directly assess the prediction performance for a single trait data set based on a single LR+ score. The LR+ indicator is useful when comparing the prediction performance between different trait data sets, and has particular merits when these trait data sets have different prevalence.

4.6 Model testing

Once the model has been calibrated with the training set, it can be validated by applying it to the independent data set before it is used to generate a subset from the accessions of interest (Code box 4.7). The model will predict the trait scores for the independent data set. As the real scores of the independent data set are already known, the predicted scores can be compared with the real scores to assess the quality of the predictions. Without availability of the independent test set, validation can still be carried out using the test set with results from the data splitting. However, the evidence for model performance generated in this way is weaker, because the general bias patterns in the training and the test set are the same.

The model prediction performance metrics can be evaluated against a statistical significance level. Statistical significance is the probability that the effect is not likely to be due to chance alone. It is common to choose either the 0.05, 0.01 or 0.001 levels. At the 0.05 significance level, the null hypothesis that the effect is due to chance alone is incorrectly rejected in 5% cases. This means that when repeating the experiment 20 times, a pure random effect will on average be statistically significant in one of the experiments. With scripting it is very easy to run the same modelling experiment many times. It is therefore important to be careful and critical when evaluating statistical significance for your predictive model. Remember to keep the test set separate during the entire model calibration phase, and choose the final model complexity before evaluating the prediction performance with the test set.

```
## CODE BOX 4.7

# Xbio [array] includes trait scores followed by the environment layers
# Recall the test set, Xtest was produced in code box 4.6 as:
Xtest <- Xbio[3445:6890,] # test set - to validate model performance

# Predict scores for the test set
prediction <- predict(rf, Xtest) # pkg stats
hist(Xtest$s3)) # preview classification histogram

# Read the confusion table from the model object
conf <- table(Xtest$s3, prediction)
TP <- conf[1,1]
FP <- conf[2,1] + conf[3,1]
FN <- conf[1,2] + conf[1,3]
TN <- conf[2,2] + conf[2,3] + conf[3,2] + conf[3,3]

# Calculate prediction performance metrics
PO_test <- (TP + TN)/sum(TP,FP,FN,TN) # proportion of observed agreement
PA_test <- (2*TP)/(2*TP+FP+FN) # proportion of observed positive agreement
PPV_test <- TP/(TP + FP) # positive predictive value
LRpos_test <- (TP/(TP + FN)/(FP/(FP+TN))) # pos diagnostic likelihood ratio
Sensitivity_test <- TP/(TP + FN)
Specificity_test <- TN/(TN + FP)
#LRpos_test<- Sensitivity_test / (1-Specificity_test)
```

The Statistics Calculator[6] from the Centre for Evidence-Based Medicine (CEBM, 2014) provides useful calculations including the 95% confidence intervals for the model prediction metrics in our case study presented here. The formulae used by this online calculator are captured in the R code below (Code box 4.8) – but be aware that this code has not yet been fully tested in practical modelling studies. The R code can be used as a guide, but the accuracy of the calculated confidence intervals should of course always be critically evaluated when making your own studies.

The value of the prediction can vary from situation to situation. For a modelling point of view, you can look at the 95% confidence interval (as calculated in the example R code in Code box 4.8 for the prediction performance metrics and say that the prediction is statistically significant at the 0.05 level when the prediction metric is between the estimated lower and upper boundaries of the confidence interval for the respective performance indicator.

[6] http://ktclearinghouse.ca/cebm/toolbox/statscalc

You may also create a completely random test set by re-ordering the response column (column s3 in the demonstration data set) randomly while keeping the environment layers constant (permutation). Such a permutation treatment should remove any signal between the target trait and the environment layers. If the confidence interval for the respective performance indicator is not overlapping between the test set and this modified randomized test set, then you have reasonable evidence that the model prediction metric for the test set is statistically significant. In a script you could repeat the permutation many times and use the average values. This approach simulates a random sampling from the test data set.

```
## CODE BOX 4.8

# Evaluation of prediction metrics
# Calculation of lower and upper 95% confidence interval limits
# TP = true positives, FP = false positives
# TN = true negatives, FN = false negatives

z = 1.959964

Sens <- TP/(TP + FN)
Sens_low <- ((2*TP)+z*2-z*sqrt((4*TP*FN/(TP+FN))+z*2))/((2*(TP+FN))+(2*z*2))
Sens_upp <- ((2*TP)+z*2+z*sqrt((4*TP*FN/(TP+FN))+z*2))/((2*(TP+FN))+(2*z*2))

Spec <- TN/(TN + FP)
Spec_low <- ((2*TN)+z*2-z*sqrt((4*TN*FP/(FP+TN))+z*2))/((2*(FP+TN))+(2*z*2))
Spec_upp <- ((2*TN)+z*2+z*sqrt((4*TN*FP/(FP+TN))+z*2))/((2*(FP+TN))+(2*z*2))

PPV <- TP/(TP + FP) # positive predictive value
PPV_low <- ((2*TP)+z*2-z*sqrt((4*TP*FP/(TP+FP))+z*2))/((2*(TP+FP))+(2*z*2))
PPV_upp <- ((2*TP)+z*2+z*sqrt((4*TP*FP/(TP+FP))+z*2))/((2*(TP+FP)) +(2*z*2))

LRpos <- (TP/(TP + FN)/(FP/(FP+TN))) # pos diagnostic likelihood ratio
LR_low <-exp(log(((FP+TN)*TP)/((TP+FN)*FP))-
z*sqrt((FN/(TP*(TP+FN)))+(TN/(FP*(FP+TN)))))
LR_upp <-exp(log(((FP+TN)*TP)/((TP+FN)*FP))+
z*sqrt((FN/(TP*(TP+FN)))+(TN/(FP*(FP+TN)))))
```

A similar approach can compare prediction performance between different models (such as when using different algorithms). When the confidence intervals for different models or algorithms are not overlapping, they have statistically significant different (better or worse) prediction performance. And when the confidence interval is overlapping, we can conclude that there is not sufficient statistical evidence (at the chosen significance level, e.g. 0.05) to claim that one model is performing better or worse than another model.

4.7 Selection of occurrence set

The final step is to apply the model to make predictions for accessions not yet tested for the respective trait. The objective is to identify a subset of accessions with a greater probability of having the target trait, and to bring these accessions further into a field screening experiment. The predictive calibration approach is successful if the field trials based on accessions in the subset give a higher hit-rate for accessions with the target trait scores than the prevalence of the target trait in the complete prediction set.

```
## CODE BOX 4.9

# Final step: predictions for accessions with unknown trait scores
# Extract the exact same environmental layers for these accessions
# Note, exact same environment columns as used for the training set
# Xpred <- CLIMATE_VARIABLES

# The example below will use the 19 BioClim variables

# Tab-delimited pred_accessions input with columns "longitude" and "latitude"
pred <- read.delim("./YOUR_PATH/pred_accessions.txt", header=TRUE, dec=".")
xy_pred <- pred[c("longitude","latitude")]

# WorldClim, http://www.worldclim.org/ (See also CODE BOX 4.1)
env <- getData('worldclim', var='bio', res=10)
Xpred <- extract(env, xy_pred); # extract environment to points (pkg raster)


# Predict scores for the new set
prediction <- predict(rf, Xpred) # pkg stats

table(prediction) # how many predicted to each class

# Note that the real trait scores are not available here, and prediction
# metrics can thus not be calculated (before after the final field trials are made).
```

5. Final remarks

The guidelines here presented are deliberately called Version 1. We hope that wider use of these guidelines for species with different breeding systems, different extents of distribution ranges, for a variety of traits and in different ecogeographical realitie, will allow the generation of a body of experience and results that will allow refinement and extension of the guidelines.

The calibration method may be said to have an advantage over the ecogeographical method in that the former is an evidence-based approach, the evidence being the explicit link to past C&E data, as it does take into account the genetic fixation of adaptive traits. So if there is only a partial relationship between site locality and presence of adaptive traits, it should be more successful at predicting sites with the desired adaptive traits than the ecogeographical method. Obviously where the basic assumption of predictive characterization is false and there is no correlation between site localities and patterns of genetic diversity for adaptive traits, both methods will be equally likely to fail in predicting localities where populations with desired adaptive traits might be found.

Some general important aspects of the two predictive characterization methods are the following:

- Accurate georeferenced information for all occurrences is important to allow proper extraction of climatic, edaphic and geophysical data.
- Increasing number and improved quality of environmental variables that are made available globally will make the methods more accurate.
- ELC maps and calibration models need to correctly reflect the assumption that is implemented in these methods (i.e. that different environmental conditions generate different selective pressures and genetic differentiation of adaptive value).
- The environmental profiles that promote target traits in LRs or CWRs need to be carefully described with environmental variables for which we have data in the territory.
- The methods are not appropriate for modern cultivars as they are not expected to show this association between traits and the environment. This is because their traits have not arisen as a result of natural selection but have been artificially selected to provide a high yield under a wide range of environmental conditions.
- The ecogeographical filtering method is the method better suited for CWR, as it is very unlikely that a sufficient number of C&E data records for a specific CWR species required to implement the calibration method will be available.

The predictive characterization methods here presented have not been tested with genetic data as variables. Further research is required to know how the inclusion of genetic diversity data might support predictive characterization. The relationship between neutral genetic marker diversity and population fitness, as well as heritability of quantitative traits, has found to be very weak (Reed and Frankham, 2001; McKay and Latta, 2002), and low differentiation at neutral markers does not necessarily mean corresponding lack of adaptive differentiation (Conner and Hartl, 2004). It therefore should not be used as predictor variable for adaptive genetic differentiation, but its use in the development of ELC maps or the generation of the ecogeographical core set might be investigated. Data about adaptive genetic diversity or diversity of loci linked to quantitative traits could probably serve as variable in the calibration method in a similar way in which C&E data are used, or support the description of environmental profiles. Considering the decreasing costs for genotyping and the continuous improvement of technologies, more data are expected to be generated that can be used in this research.

References

Alercia, A., Diulgheroff, S. and Mackay, M. 2012. FAO/Bioversity Multi-Crop Passport Descriptors (MCPD V.2). FAO (Food and Agriculture Organization of the United Nations), Rome, Italy, and Bioversity International, Rome, Italy. Available at: http://www.bioversityinternational.org/nc/publications/publication/issue/faobioversity_multi_crop_passport_descriptors_v2_mcpd_v2.html Accessed 12 Sep 2014.

Almorox Alonso, J. 2003. *Climatología aplicada al Medio Ambiente y Agricultura*. UPM. ETSI Agrónomos. 201 p.

Bari, A., Street, K., Mackay, M., Endresen, D.T.F., de Pauw, E. and Amri, A. 2012. Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables. *Genetic Resources and Crop Evolution*, 59(7): 1465-1481. doi:10.1007/s10722-011-9775-5

Bari, A., Amri, A., Street, K., Mackay, M., de Pauw, E., Sanders, R., Nazari, K., Humeid, B., Konopka, J. and Alo, F. 2014. Predicting resistance to stripe (yellow) rust (*Puccinia striiformis*) in wheat genetic resources using focused identification of germplasm strategy. *Journal of Agricultural Science*, 152: 906-916. doi:10.1017/S0021859613000543

Bhullar, N.K., Street, K., Mackay, M., Yahiaoui, N. and Keller, B. 2009. Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the Pm3 resistance locus. *Proceedings of the National Academy of Sciences of the United States of America,* 106: 9519–9524. doi:10.1073/pnas.0904152106

Breiman, L. 2001. Random forests. *Machine Learning*, 45(1): 5–32. doi:10.1023/A:1010933404324

Chamberlain, S., Boettiger, C., Ram, K., Barve, V. and Mcglinn, D. 2014. rgbif: Interface to the Global Biodiversity Information Facility API. R package version 0.7.0. https://github.com/ropensci/rgbif

Chapman, A.D. and Wieczorek, J. (eds). (2006). *BioGeomancer: Guide to best practices for georeferencing.* Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark. 80 p. ISBN: 87-92020-00-3. Available from http://www.gbif.org/orc/?doc_id=1288&l=en Verified 12 Sep. 2014.

Conner, J.K. and Hartl, D.L. 2004. *A primer of ecological genetics.* Sinauer Associates, Inc., USA.

CEBM. 2014. *Statistics calculator.* Statisticians: F. Khandwala, K. Thorpe; developers: D. Newton and P. Wong. Center for Evidence-Based Medicine (CEBM), University Health Network, Toronto, ON, Canada. Available at http://ktclearinghouse.ca/cebm/toolbox/statscalc Verified 12 Sep. 2014.

De Martonne, E. 1926. Une nouvelle fonction climatologique: L'indice d'aridité. *La Meteorologie*, 2: 449–458.

El Bouhssini, M.E., Street, K., Joubi, A., Ibrahim, Z. and Rihawi, F. 2009. Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. *Genetic Resources and Crop Evolution*, 56(8): 1065–1069. doi:10.1007/s10722-009-9427-1

El Bouhssini, M.E., Street, K., Amri, A., Mackay, M., Ogbonnaya, F.C., Omran, A., Abdalla, O., Baum, M., Dabbous, A. and Rihawi, F. 2011. Sources of resistance in bread wheat to Russian wheat aphid (*Diuraphis noxia*) in Syria identified using the focused identification of germplasm strategy (FIGS). *Plant Breeding,* 130(1): 96–97. doi:10.1111/j.1439-0523.2010.01814.x

Endresen, D.T.F. 2010. Predictive association between trait data and ecogeographical data for Nordic barley landraces. *Crop Science*, 50(6): 2418–2430. doi:10.2135/cropsci2010.03.0174

Endresen, D.T.F., Street, K., Mackay, M., Bari, A. and de Pauw, E. 2011. Predictive association between biotic stress traits and ecogeographical data for wheat and barley landraces. *Crop Science*, 51(5): 2036–2055. doi:10.2135/cropsci2010.12.0717

Endresen, D.T.F., Street, K., Mackay M., Bari, A., Amri, A., De Pauw, E., Nazari, K. and Yahyaoui, A. 2012. Sources of resistance to stem rust (Ug99) in bread wheat and durum wheat identified using focused identification of germplasm strategy. *Crop Science*, 52(2): 764–773. doi:10.2135/cropsci2011.08.0427

FAO [Food and Agriculture Organization of the United Nations]. 2010. *The second state of the world report for plant genetic resources for food and agriculture*. FAO, Rome, Italy. ISBN 978-92-5-106534-1.

GADM. 2012. GADM database of Global Administrative Areas. Version 2.0 January 2012. Available online from http://www.gadm.org/ Accessed 16 Sep. 2014.

Google Inc. 2014. Google Maps [Online]. Google Inc., TerraMetrics Inc. and Tele Atlas BV. Mountain View, CA. Available at http://maps.google.com Verified 12 Sep. 2014.

Gotor, E., Alercia, A., Ramantha Rao, V., Watts, J. and Carracciolo, F. 2008. The scientific information activity of Bioversity International: the descriptor lists. *Genetic Resources and Crop Evolution*, 55(5): 757–772. doi:10.1007/s10722-008-9342-x

Guarino, L., Jarvis, A., Hijmans, R.J. and Maxted, N. 2002. Geographical information systems (GIS) and the conservation and use of plant genetic resources. pp. 387–404, *in:* J.M.M. Engels, V. Ramanatha Rao, A.H.D. Brown and M.T. Jackson (eds). *Managing plant genetic diversity*. CABI Publishing, Wallingford, UK. ISBN: 9780851995229.

Hijmans, R.J. 2014. raster: Geographic data analysis and modeling. R package version 2.3-12. http://CRAN.R-project.org/package=raster

Hijmans, R.J., Schreuder, M., De la Cruz, J. and Guarino, L. 1999. Using GIS to check coordinates of genebank accessions. *Genetic Resources and Crop Evolution*, 46(3): 291–296. doi:10.1023/A:1008628005016

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. and Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology,* 25(15): 1965-1978. doi:10.1002/joc.1276 [WorldClim]

Kaufman, L. and Rousseeuw, P.J. 1987. Clustering by means of medoids. pp. 405–416, *in:* Y. Dodge (ed.). *Statistical data analysis based on the L1–norm and related methods*. North-Holland, Amsterdam, The Netherlands. ISBN: 9780444702739.

Kaur, N., Street, K., Mackay, M., Yahiaoui, N. and Keller, B. 2008. Allele mining and sequence diversity at the wheat powdery mildew resistance locus Pm3. *In:* R. Appels, R. Eastwood, E. Lagudah, P. Langridge, M. Mackay, L. Mcintyre and P. Sharp (eds). 11th International Wheat Genetics Symposium. Sydney University Press, Brisbane, Australia.

Ketchen, D.J. and Shook, C.L. 1996. The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic Management Journal*, 17(6): 441–458. doi:10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G

Khazaei, H., Street, K., Bari, A., Mackay, M. and Stoddard, F.L. 2013. The FIGS (focused identification of germplasm strategy) approach identifies traits related to drought adaptation in *Vicia faba* genetic resources. *PLoS ONE*, 8(5): e63107. doi:10.1371/journal.pone.0063107

Mackay, M.C. 1986. Utilizing wheat genetic resources in Australia. pp. 56–61, *in*: R. McLean (ed.). Proceedings of the 5th Assembly of the Wheat Breeders' Society in Australia. Merredin 18–22 August 1986. Western Australian Department of Agriculture, Perth, Australia. ISBN: 9780730913269.

Mackay, M.C. 1990. Strategic planning for effective evaluation of plant germplasm. pp. 21–25, *in:* J.P. Srivastava and A.B. Damania (eds). *Wheat genetic resources: Meeting diverse needs.* John Wiley & Sons, Chichester, UK. ISBN 0-471-92880-1.

Mackay, M.C. 1995. One core collection or many? pp. 199–210, *in:* T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum and A.A.V. Morales (eds). *Core collections of plant genetic resources.* Proceedings from the IBPGR/CGN/CENARGEN workshop on 'Core Collections: Improving the Management and Use of Plant Germplasm Collections', held in Brasilia, August 1992. John Wiley & Sons, Chichester, UK. 269 p. ISBN: 978-0-471-95545-0.

Mackay, M.C. and Street, K. 2004. Focused identification of germplasm strategy – FIGS. pp. 138–141, *in:* C.K. Black, J.F. Panozzo and G.J. Rebetzke (eds). *Cereals 2004.* Proceedings of the 54th Australian Cereal Chemistry Conference and the 11th Wheat Breeders' Assembly, 21–24 September 2004, Canberra, Australian Capital Territory (ACT). Cereal Chemistry Division, Royal Australian Chemical Institute, Melbourne, Australia.

McKay, J.K. and Latta, R.G. 2002. Adaptive population divergence: Markers, QTL and traits. *Trends in Ecology and Evolution*, 17(6): 285–291. doi:10.1016/S0169-5347(02)02478-3

Parra-Quijano, M., Iriondo, J.M. and Torres, E. 2012. Ecogeographical land characterization maps as a tool for assessing plant adaptation and their implications in agrobiodiversity studies. *Genetic Resources and Crop Evolution*, 59(2): 205–217. doi:10.1007/s10722-011-9676-7

Parra-Quijano, M., Iriondo, J.M., Frese, L. and Torres, E. 2012. Spatial and ecogeographical approaches for selecting genetic reserves in Europe. pp. 20–28, *in:* N. Maxted, M.E. Dulloo, B.V. Ford-Lloyd, L. Frese, J. Iriondo and M.A.A. Pinheiro de Carvalho (eds). *Agrobiodiversity Conservation: securing the diversity of crop wild relatives and landraces*. CABI, Wallingford, UK. ISBN: 9781845938529.

Parra-Quijano, M., Torres, E., Iriondo, J.M. and López, F. 2014. CAPFITOGEN Tools User Manual Version 1.2. International Treaty on Plant Genetic Resources for Food and Agriculture, FAO, Rome, Italy. 138 p. ISBN 978-92-5-108493-9. Available from: http://www.agrobiodiversidad.org/blog/?p=1189 Verified 12 Sep. 2014, and ftp://ftp.fao.org/ag/agp/planttreaty/publi/2014/capfitogen_ENG92014.pdf Verified 12 Sep. 2014.

Reed, D.H. and Frankham, R. 2001. How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. *Evolution*, 55(6): 1095–1103. doi:10.1554/0014-3820(2001)055[1095:HCCAMA]2.0.CO;2p

Rousseeuw, P.J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65. doi:10.1016/0377-0427(87)90125-7

Roy, P.P., Leonard, J.T. and Roy, K. 2008. Exploring the impact of size of training sets for the development of predictive QSAR models. Chemometrics and Intelligent Laboratory Systems, 90(1): 31–42. DOI: 10.1016/j.chemolab.2007.07.004

Street, K., Mackay, M., Zuev, E., Kaul, N., El Bouhssini, M., Konopka, J. and Mitrofanova, O. 2008. Diving into the genepool – a rational system to access specific traits from large germplasm collections. ISBN: 9781920899141. Available at http://hdl.handle.net/2123/3390 Verified 12 Sep. 2014.

Thormann, I. 2012. Applying FIGS to crop wild relatives and landraces in Europe. *Crop Wild Relative*, Issue 8: 14–16. ISSN 1742-3694 (Online). Available at: http://www.pgrsecure.bham.ac.uk/sites/default/files/documents/newsletters/CWR_Issue_8.pdf Verified 12 Sep. 2014.

Bioversity International is a member of the CGIAR Consortium. CGIAR is a global research partnership for a food-secure future.

Bioversity Headquarters
Via dei Tre Denari 472/a
00057 Maccarese, (Fiumicino)
Rome, Italy

www.bioversityinternational.org

Tel. (39-06) 61181
Fax. (39-06) 61979661
Email: bioversity@cgiar.org

ISBN 978-92-9255-004-2

Bioversity International

CGIAR
Science for a food-secure future