Fostering sustainable legume-based farming systems and agri-feed and food chains in the EU

Deliverable D1.4

Map of achievable legume yields across studied EU-areas (Part A)

Planned delivery date: M36

Actual submission date: M37

Start date of the project: June 1st, 2017 **Duration:** 48 months

Workpackage: WP1

Workpackage leader: INRA **Deliverable leader:** UNIPI

Partners contributing to the deliverable: INRA, TERIN, WU, PGRO, WR, SSSA, INIAV, AICF, SEGES, LLKC, FH-SWF, CRAN, LAMMC, FiBL

Version: V1

Dissemination Level	
Public	x
Classified, as referred to Commission Decision 2001/844/EC	
Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Table of contents

1. Summary

In this report we summarized the major outcomes of the dataset on grain legumes yield built in Task 1.2 of WP1 of the LegValue project. According to the DoA, Task 1.2 might have delivered by month 36 a unique deliverable (D1.4) reporting on yield maps of grain legumes in different pedoclimatic conditions in the EU. Due to difficulties in gathering a significant amount of scientific data on grain legume yields and due to efforts higher than expected spent on validation of the modelling methodology adopted to generate the maps, the leader of task 1.2 proposed and the ExCom agreed upon splitting this deliverable in two parts (namely Part A and Part B).

This dataset, constituting the first part (Part A) of D1.4, was primarily intended for consolidating scientific data necessary to feed the modelling exercise adopted for the production of European maps of grain yield potential for the major pulse species (soybean, field pea, faba bean, chickpea, lentils), that will form the second part (Part B) of the D1.4. Additional objectives were:

- to allow more precise and realistic estimations of protein and starch yield from pulses in current and future scenarios of legumes included in cropping systems in the EU;
- to identify research gaps and technological lock-ins currently hampering a higher share of pulses in cropping systems;
- to identify sites with high productive potential for pulses that are currently unexplored in the EU;
- to set for a step up for science on legume crop management by highlighting ways to improve the existing knowledge on the topic.

The dataset that is made publicly available consists of data extracted from scientific papers published on international journals, that are complemented with data from LegValue Task 1.2 Partners own data, generated both on-station and on-farm but always under controlled conditions, and from the former EU-FP7 Project LEGATO.

Part of the unpublished data were made available by the LegValue partners only for the objectives of the project and consequently are not included in the public version of the dataset.

By outlining the 5244 entries of the dataset, we preliminarily explored the relationships between the grain yield of the five legumes and environmental (soil, climate, latitude) and agronomic (tillage, irrigation, fertilisation, weed control, organic farming practices) aspects. Grain yields varied a lot in function of pedoclimates but possibilities of cross-fertilisation among locations and contrasting conditions were identified in terms of knowledge transfer and establishment of future research.

Among management practices, special emphasis has been given in former studies on crop genotype, an evidence contrasting with poor resources allocated so far for the genetical improvement of these crops. The high presence also of combination between genotype and other management aspects reveals the high interest within the scientific community in developing new plant material adapted not only to a range of environmental conditions but also consistent with management practices.

In this “Part A” of the Deliverable 1.4 we also presented the preliminary results of the modelling activity performed on soybean yield data that were estimated by different algorithms based on former analysis on global dataset linking yield levels with climatic conditions. The preliminary tests performed on soybean yield estimations were very positive. The selected algorithm (Random Forest) allowed to predict in a reliable way soybean yields under historical climatic conditions also in the EU territory. According to preliminary analysis, although not considering the positive effect of irrigation on potential

yields, the internal demand of soybean grain seems to be likely met by EU production even if the legume would be present just 1 year over 6 of crop rotations.

This modelling approach will be then fed in the next 6 months with the yield data of all the 5 major pulse species considered and included in the dataset (Annex A to this report) in order to refine the validation of the model and to be able to produce more reliable yield estimations and more accurate yield maps (Part B of this Deliverable 1.4).

2. Introduction

Within the LegValue project, we aim at enhancing the presence of legume crops in cropping systems in the EU. Besides forage legumes, that are well known to deliver a number of different ecosystem services (e.g. higher availability of nitrogen in the soil through N_2 biological fixation, reduction of soil compaction, soil fertility restoration, weed suppression, stimulation of soil biological activity and biodiversity), also pulses (i.e. grain legumes) have potential to boost the agroecological transition towards more diversified cropping systems and viable value chains.

Despite these apparent benefits and subsidies that are dedicated in EU Rural Development Plans (RDPs) and in the Common Agricultural Policy (CAP), the area of farmland under legume production in the EU has been showed to steadily decline (except for soybean) in the last decades (Schreuder and de Visser, 2014). Pelzer et al. (2017) have recently estimated that grain legumes occupy only 1.8% of arable land in the EU. There are a number of factors causing this low percentage of land sharing of pulses, many of them being mostly related to socio-economic issues (Magrini et al., 2016), but also agronomic reasons are of paramount importance. Farmers do consider grain legumes so poorly attractive because of the uncertainty and instability of their yields, being normally affected by pedoclimates and biotic stress factors much more than other major arable crops like cereals.

So far, the research efforts paid on studying the relationships between environmental and management factors potentially affecting pulse yields have been not so extensive as for other groups of arable crops. That is why one of the objective of LegValue WP1 is to try to shed light on these aspects and to determine the actual yield potential of the most common pulse species grown in the EU (soybean, *Glycine max* (L.) Merr. ; field pea, *Pisum sativum* L. ; faba bean, *Vicia faba* L. ; chickpea, *Cicer arietinum* L. ; lentils, *Lens culinaria* Medik.).

In Task 1.2 of WP1, we intend to produce EU yield potential maps for the five major pulse species according to current soil and climate conditions with the following objectives:

- to allow more precise and realistic estimations of protein and starch yield from pulses in current and future scenarios of legume presence in cropping systems;
- to identify research gaps and technological lock-ins currently hampering a higher share of pulses in cropping systems;
- to identify sites with high productive potential for pulses that are currently unexplored in the EU;
- to set for a step up for science on legume crop management by highlighting ways to improve the existing knowledge on the topic.

To generate the maps, in Task 1.2 it was agreed on identifying a methodological approach based on simple mechanistic modelling able to link pedoclimatic conditions with grain legume performances, without the need to use highly detailed and specific information on crop physiology and crop phenology, but at the same time able to produce reliable and solid outcomes. Among the available models in the literature, it was decided to use the Ecological Niche Model approach, that revealed to be accurate enough in predicting yield levels of soybean worlwide according to climatic conditions.

In order to feed the model with science-based evidences on legume yields produced and collected in the EU for a finer model validation, we first needed to gather most of the existing knowledge and to structure it in a consolidated form, easy to use and to explore it also external to and after LegValue.

Deviation from the DoA

This first step (i.e. collection of scientific data on legume yields produced in the EU) was very time consuming and required extra efforts to integrate scientific paper results (that were very few compared to our expectations) with other sources of scientific information (e.g. former EU project results, unpublished experimental results owned by LegValue partners and data generated in LegValue On-Farm Networks -OFN-). Furthermore, the methodological approach for yield map generation (i.e. the Random Forest Ecological Niche Model) needed to be tested and validated on a test crop (i.e. soybean) with yield data produced within the EU before being extended to the other major pulses. This activity, led by INRA, required additional efforts respect to those foreseen in the DoA.

For all this reason, during the last annual meeting of the LegValue project, held virtually from April 28th to April 30th, 2020, **the task 1.2 leader proposed to the ExCom to split the Deliverable 1.4, expected by M36 according to the DoA, in two parts**. In this first part, named **"Part A",** delivered by **M37**, we included the dataset generated in Task 1.2 with explanation of major outcomes and the preliminary results of application of the modelling approach on soybean yields, estimated based on historical climatic data for the EU, based on the results obtained on global observed yield dataset. This preliminary analysis, presented in section 3.5 and 4.3, supports the strength and reliability of the adopted modelling approach and provides an anticipation of the yield maps that will be delivered in a second part **("Part B"**), expected by **M43** (December 2020).

In Part B we will present the final version of the yield maps generated by the Ecological Niche Model for all the 5 pulse species based on calibration and validation of the algorithm with data on yields and pedoclimates included in the dataset annexed to Part A. In Part B, an in-deep analysis of relationships between pulse yields and pedoclimates, as well as a science-based estimation of the potential of grain legume yields in the EU will be also provided.

The ExCom of LegValue agreed upon this deviation from the DoA, as it was considered that:

- the dataset produced with "Part A" of the deliverable, that was not expected according to the DoA, might represent a product with added value for the other tasks of the project (in particular for Tasks 1.3 and 1.4 that could benefit from more structured information on the link between pedoclimatic and agronomic variables and legume yields) and for the scientific community;
- the delay in the delivery of the yield maps should not affect anyhow the regular development of the Work Package 1 and of the other tasks linked to Task 1.2 (in particular, Task 1.3 and Task 1.4).

3. Materials and methods

3.1. Literature review

The collection of the data on legume yields produced under controlled conditions in the EU started by extracting the data on the 5 selected legume species (soybean, field pea, faba bean -including broad bean, pigeon bean and horse bean-, chickpea and lentils) from the global dataset published by Cernay et al. (2017). The paper includes 8,386 articles published globally from 1967 to 2016 and respecting the following 6 eligibility criteria on the dataset ISI-Web of Science:

1. 1/+ legume grown as sole crop in title/abstract;

2. >1 legume grown per each site (title/abstract);

3. at least one experiment for 1/several years from seeding to harvest;

4. peer reviewed journals;

5. written in English;

6. full text available

The dataset produced by Cernay et al. (2017) followed also other eligibility criteria:

- Studies conducted on grain legumes;
- Studies reporting grain/aerial biomass yields;
- Data coming from scientific experiments;
- Studies not including forage legumes, cover crops and intercrops.

By restricting the global dataset to the EU countries, we extracted 223 papers potentially eligible for the dataset. All of them were read in full to assess for their eligibility, according to our additional criteria (i.e., grain yield data reported as tables for each legume species, precise information on field site pedoclimatic conditions), and at the end only 25 papers were first considered eligible.

To expand the dataset we then proceeded as follows:

1. we updated the dataset to 2019 by launching the same search string used by Cernay et al. (2017) on Scopus and Web of Science databases;
2. we expanded the dataset to Turkey;
3. we modified the search string in order to:
    a. include also studies involving even only one legume crop at a given site;
    b. include also studies reporting on legumes grown in intercropping (in case also legume yield in pure stands was available);
    c. specify also the name of the 5 legume species considered;

    The final search string applied was as follows:

    “crop* AND (soybean OR soyabean OR pea OR faba bean OR chickpea OR lentil) AND (yield* OR ‘dry matter’ OR biomass) AND (compar* OR assessment OR product* OR performance*) AND (trial* OR factorial OR experiment* OR treatment* OR condition*)”;

4. we screened all the eligible papers and we also kept in papers reporting not only data in tables but also in charts, by using the WebPlotDigitizer-4.2.0 app (https://automeris.io/WebPlotDigitizer) to extract data from graphical forms;
5. we also gathered additional papers included in the literature review on ecosystem services related to legume crops performed in Task 1.1 and that were not identified according to our inclusion criteria;
6. we explored references cited in eligible papers identified as above.

3.2. Data from the EU-FP7 Legato Project

To expand the dataset, in a second moment it was agreed on including also open data coming from former EU project. The EU-FP7 LEGATO Project "LEGumes for the Agriculture of TOmorrow" (2014-2017) has delivered the results of the 2-yr experimental activities performed on field pea, faba bean, chickpea, lupins and chickling peas (*Lathyrus* spp.). The dataset was downloaded from https://intranet.iamz.ciheam.org/forms/Legato/WP6/index.php and eligible data on pea, faba bean and chickpea were extracted and added to the dataset.

3.3. Data from unpublished experiments carried out by LegValue partners

In order to expand the geographical coverage of the dataset, we also asked the Task 1.1 partners to deliver their own experimental data coming from LegValue Task 1.1 trials on legume yields (see Milestone MS3), from OFN trials as well as from other field experiments performed externally to LegValue and not already published.

The summary data reported in the Results chapter of this document referred to the whole dataset, including also data owned by LegValue partners that will remain not publicly accessible but that will enrich the production of yield maps that will be included in the Part B of this Deliverable 1.4.

3.4. Structure of the dataset

The dataset was structured as a worksheet with variables as columns and entries as rows.

Each combination of year x site x crop species x experimental treatment level, averaged over spatial or temporal replications, was reported as a single entry (i.e. a single row).

To facilitate browsing the dataset, we added also filters on each column.

An "Instructions" worksheet was also added to facilitate data entry operations as well as to inform the readers about the proper interpretation of the data.

The dataset includes 63 columns (i.e. variables), as described in details in Table 1.

Table 1 – Structure of the dataset and meaning of each column (variable)

COLUMN TITLE	**CONTENT**
ID	Entry ID
Source	Experiment/Paper/LEGATO Project
Publicly available	Y/N (Y if this data could be part of an open-access publication of the dataset on a Data Journal; N if for internal use in LegValue)
Experiment_ID	Experiment acronym_Surname of responsible
Site_Country	Name of the country in full
Site_Region	NUTS 3
Site_Name	Name of the site in full
Latitude	Decimal degrees of Latitude (XX.xx).
Latitude Cardinal	N/S
Longitude	Decimal degrees of Longitude (XX.xx).
Longitude Cardinal	W/E
Site_Soil_Classification_Name	Soil classification type (USDA)
Site_Soil_Texture_Name	Soil texture class (e.g. loam, sandy, sandy loam)
Site_Precipitation_mm	Total rainfall in the period considered. "NA" if not available
Site_Precipitation_Period	annual/growing season
Site_Precipitation_Period_Month	Initial Final month of the period for which precipitations are reported (e.g. Jan Dec)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

COLUMN TITLE	**CONTENT**
Site_Precipitation_Period_Year	Years of registration of the precipitations (e.g. 1993).
Site_Temperature_Celsius	Average temperature in the considered period. "NA" if not available
Site_Temperature_Period	annual/growing season
Site_Temperature_Period_Month	Initial Final month of the period for which temperature is reported (e.g. Jan Dec)
Site_Temperature_Period_Year	Years of registration of the temperature (e.g. 1993).
management evaluated	e.g. tillage, irrigation, variety
Crop_Sequence_Treatment_Name	Report the name of the treatment or (in case of factorial combination) the name of the combination
Scientific name	Latin name (without author initials) of the legume crop species (e.g. Glycine max)
Previous crop	Latin name (without author initials) of the crop species grown before the legume (e.g. Triticum aestivum)
Crop	Common name of the legume crop species (e.g. Soybean)
Cv	Name of the legume crop variety
Group of precocity	Only for soybean. Report here the precocity group (000 to 10)
GM	Genetically modified variety? Y/N
Sowing date	mm/dd/yyyy. "NA" if not available
Harvest date	mm/dd/yyyy. "NA" if not available
Tillage_no tillage	"tillage" if a tillage operation is performed before legume sowing, or "no-tillage" if sod-seeding legume
Days_crop_cycle	Length of crop cycle in the experimental year (nr. of days from sowing to harvest). "NA" if not available
Plant density (plant/m2)	nr of legume plants per m2 (alternative to sowing density). "NA" if not available
Sowing Density (seeds/m2)	nr of legume seeds per m2 (alternative to plant density). "NA" if not available
row spacing (m)	inter-row space in meters. "NA" if not available
N rate (kg N ha-1)	Total amount of N (kg ha$_{-1}$) supplied to the crop
N fertiliser type(s)	Name(s) of each N fertiliser applied to the crop with its level of N application rate (kg N ha$_{-1}$), separated by comma and in chhronological order of application (e.g. Poultry manure -30-, Urea -30-)
Nr. of applications (N)	Nr. of applications of N fertilisers
% N from organic fertiliser(s)	% of total N supplied to the crop coming from organic fertilisers or amendments
P rate (kg P ha-1)	Total amount of P (kg ha$_{-1}$) supplied to the crop
P fertiliser type(s)	Name(s) of each P fertiliser applied to the crop with its level of P application rate (kg P ha$_{-1}$), separated by comma and in chhronological order of application (e.g. Diammonium phosphate -30-, Superphosphate -30-)
Nr. of applications (P)	Nr. of applications of P fertilisers
% P from organic fertiliser(s)	% of total P supplied to the crop coming from organic fertilisers or amendments
K rate (kg K ha-1)	Total amount of K (kg ha$_{-1}$) supplied to the crop
K fertiliser type(s)	Name(s) of each K fertiliser applied to the crop with its level of K application rate (kg K ha$_{-1}$), separated by comma and in chhronological order of application (e.g. Potassium sulphate -30-, Sugarbeet liquid pulp -15-)
Nr. of applications (K)	Nr. of applications of K fertilisers
% K from organic fertiliser(s)	% of total K supplied to the crop coming from organic fertilisers or amendments
irrigation	Y/N_partial/full (Y if irrigation was applied or N if not; PARTIAL if irrigation did not cover the full water need of the crop or FULL if it did)
Mean irrigation quantity (mm)	Mean amount of irrigation water (mm) applied (exact amount or mean value if a range is reported). "NA" if not available
Herbicide application	Were chemical herbicides applied or not (Y/N)
Mechanical weed control	Was mechanical weeding applied or not (Y/N)
Crop protection	Were crop protection products, including natural or biocontrol agents, applied to the crop (Y/N)
number of replicates	Number of replicates concurring to the mean yield value reported in a single site x year combination (e.g. number of blocks or spatial replicates)
number of sites	Number of different sites considered as spatial replicates for computing the mean yield value reported, if mean yield values for each site are not available

COLUMN TITLE	**CONTENT**
number of years	Number of years concurring to the mean yield value reported, if single year mean yield values are not available
Moisture content harvest (%)	Moisture percentage of the marketable yield (e.g. "13" for 13%) as reported in the source material
Grain Yield (t d.m./ha)	Yield of the grain of the legume crop in t d.m. ha$_{-1}$ (the humidity reported in the previous column, when available, is removed from the grain yield reported)
SE crop yield	Value of the standard error of the mean of the yield, if available. If not available, "NA"
SD crop yield	Value of the standard deviation of the mean of the yield, if available. If not available, "NA"
CV crop yield	Value of the coefficient of variation of the mean of the yield, if available. If not available, "NA"
Var. crop yield	Value of the variance of the mean of the yield, if available. If not available, "NA"

3.5. Data analysis

To summarise the major outcomes of the full version of the dataset, graphical exploration of the data was performed by using the package ggplot2 (Wickham, 2009) of the statistical software R, version 136 3.3.1 (R Core Team, 2013).

A more detailed study on the relationships between legume yields and pedoclimatic conditions will be included in the Part B of the Deliverable. However, we briefly present below the method that we plan to use to generate the maps of achievable yield for each legume crop based on the relationships between legume yields and pedoclimatic conditions. The method has been developed for soybean in Europe. Building on two recently published global datasets including historical soybean yield and retrospective meteorological forcing (Iizumi et al. 2014a,b), we developed data-driven relationships between climate and soybean yield to estimate soybean suitable areas over Europe. Several machine learning algorithms were trained and tested at the global scale (Random Forest, Artificial Neural Networks, Generalized Additive Model, and Multiple Linear Regression) to predict soybean yield as a function of monthly climate inputs (solar radiation, minimum and maximum temperature, rainfall, and vapour pressure) calculated over the growing season (April to October). A large share of the training data was taken from major soybean-producing countries (Argentina, Brazil, Canada, China, India, Italy and the United States), and zero-yield data points were randomly sampled in climate zones known to be unsuitable for soybean production (e.g. deserts and arctic areas) and added to the dataset so that they represented about 20% of the final dataset. The most accurate algorithm was selected after running a cross-validation procedure assessing model transferability in time and space. The selected algorithm (Random Forest) was then run for the entire Europe to assess potential distribution of soybean suitable area in rainfed conditions under current and future climate. Projections of soybean suitability in Europe were performed for historical climate (1981-2010). The projections assume a growing season from April to October and no irrigation, although soybean is often irrigated in Europe. The no irrigation assumption prevents from making any hypothesis about available water for irrigation, which is a complex issue especially under climate change. We therefore acknowledge that the yield projections are probably a bit conservative from that point of view.

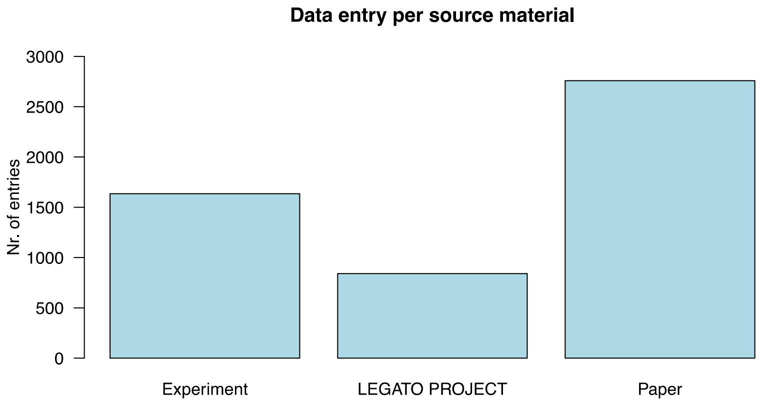
4. Results

4.1. Overview of the dataset

The data collected resulted in 5234 single entries related to the crop yield of the 5 major pulse species grown in the EU.

The distribution of the entries per type of source material (published papers, partners experimental data, LEGATO project) is depicted in Figure 1.

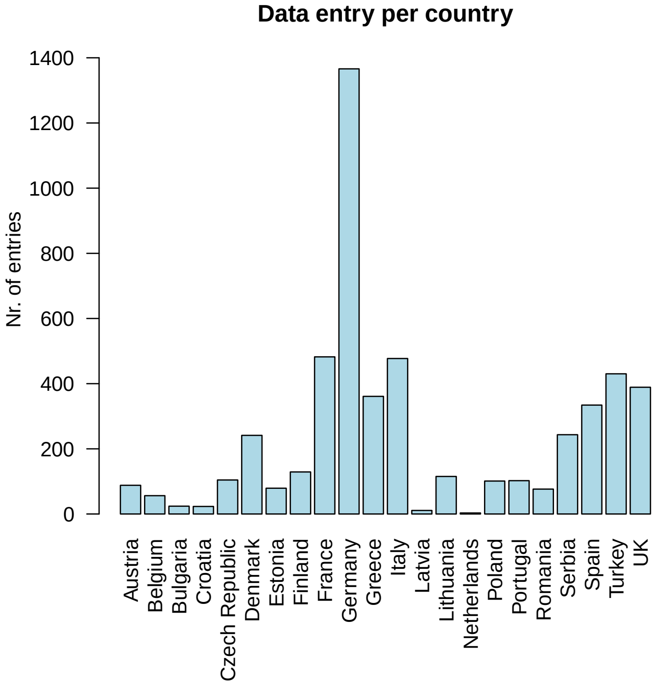
Figure 1 - Number of entries grouped per source material



Most of the entries (2759) comes from published scientific papers, whilst unpublished experimental data include 1635 entries and the results from the LEGATO project cover 840 entries.

As shown in Figure 2, Germany covers 26% of the entries. Also France (482), Italy (477) and Turkey (430) were well represented in the material. A total of 22 countries are represented in the dataset, covering a full range of different pedoclimates from Southern countries (e.g. Spain and Portugal) to Northern sites (e.g. Finland, Estonia, Lithuania) and from Eastern to Western countries.

Figure 2 - Number of entries grouped per country



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

The five pulse crops are not equally represented in the dataset (Figure 3). Pea (1553) and soybean (1511) are the most represented crops in the dataset.

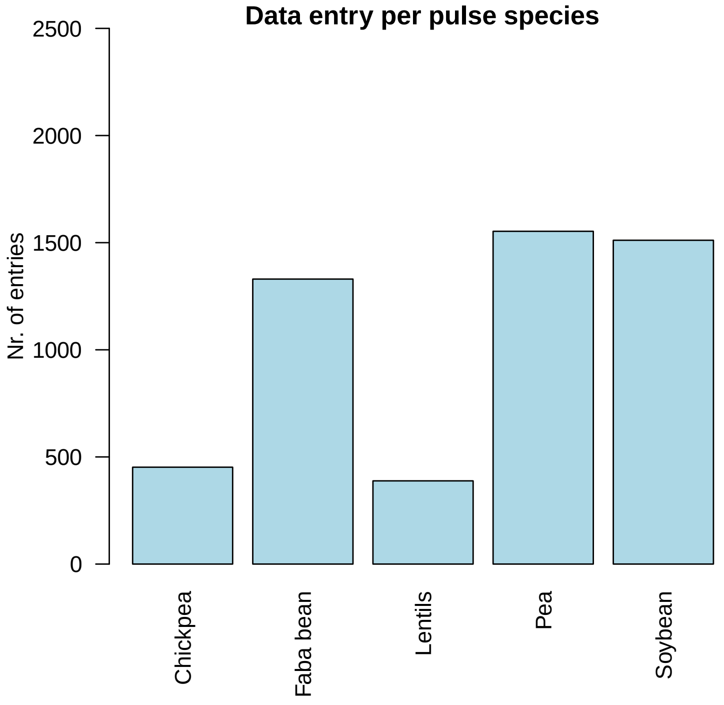
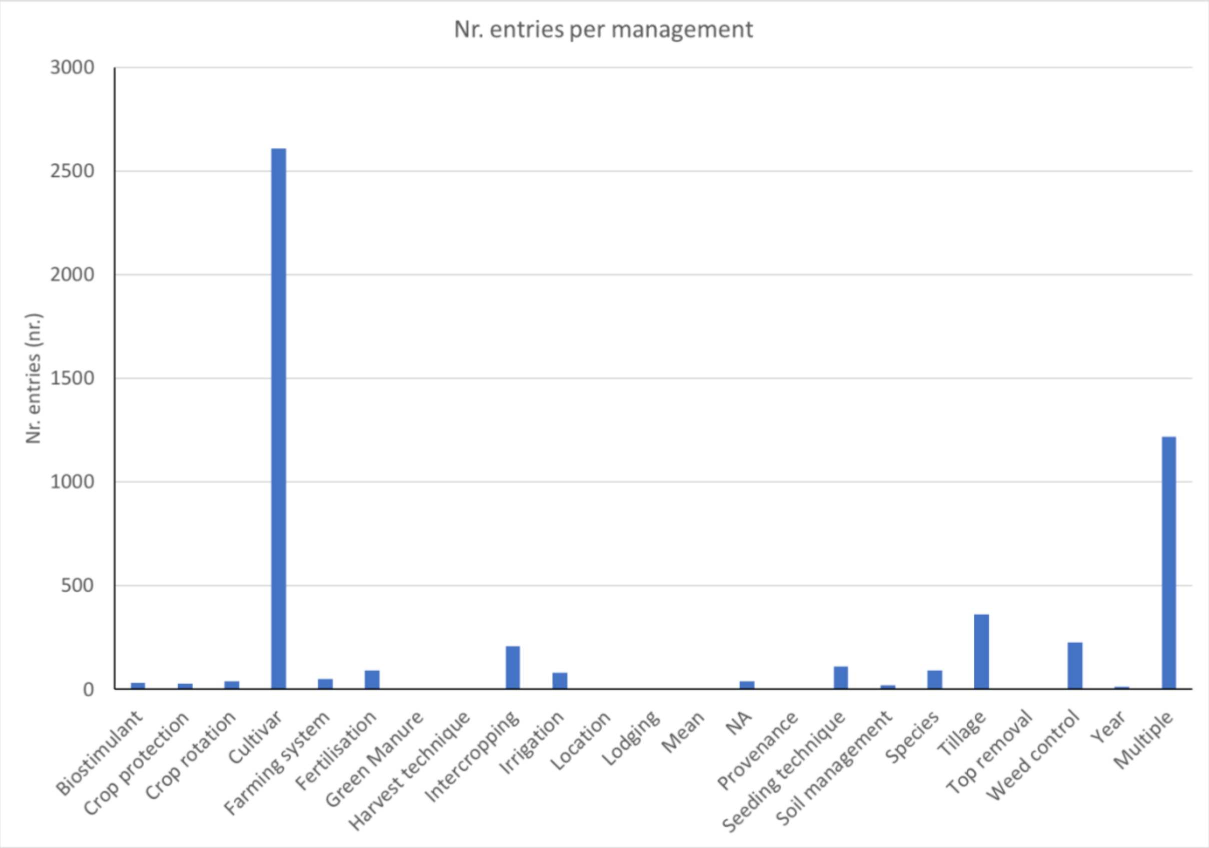


Figure 3 - Number of entries grouped per pulse species

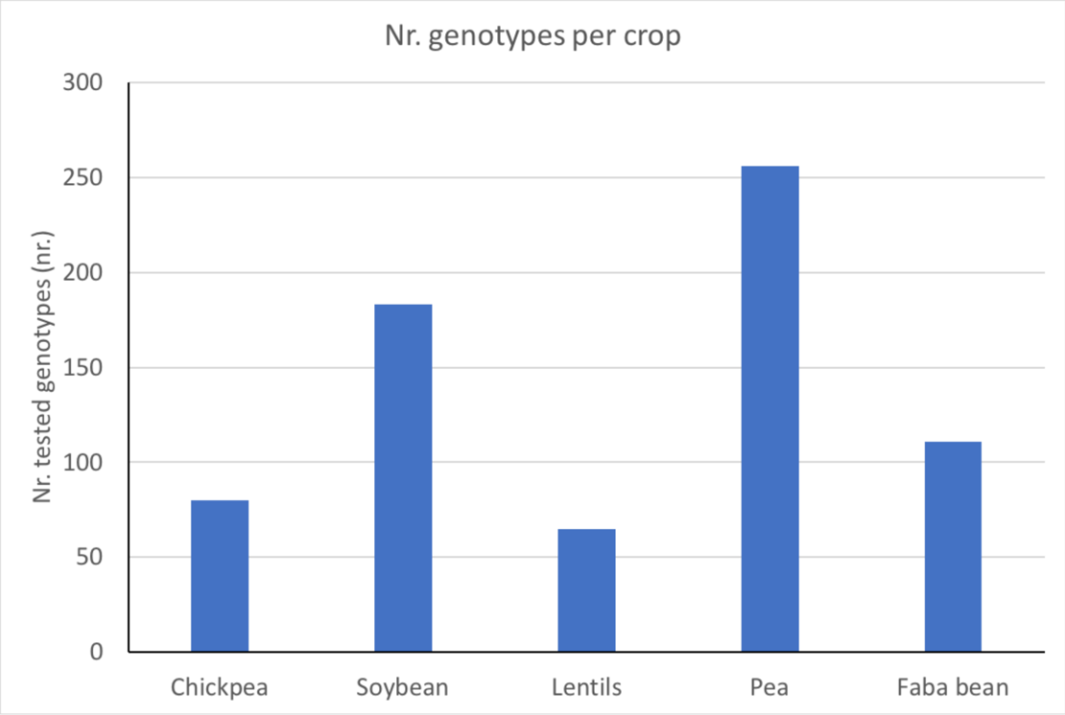
Genotype is the management aspect most represented, covering around 50% of all the entries of the dataset. Among the other agronomic management issues, interactions among several aspects, tillage, weed control and intercropping are the other more common elements (Figure 4).

Figure 4 - Number of entries grouped per management



The number of tested varieties grouped per crop is depicted in Figure 5. The highest number of different genotypes tested in the source material is observed for field pea (256), followed by soybean (183) and faba bean (111). For chickpea and lentils, only 80 and 65 genotypes are represented.

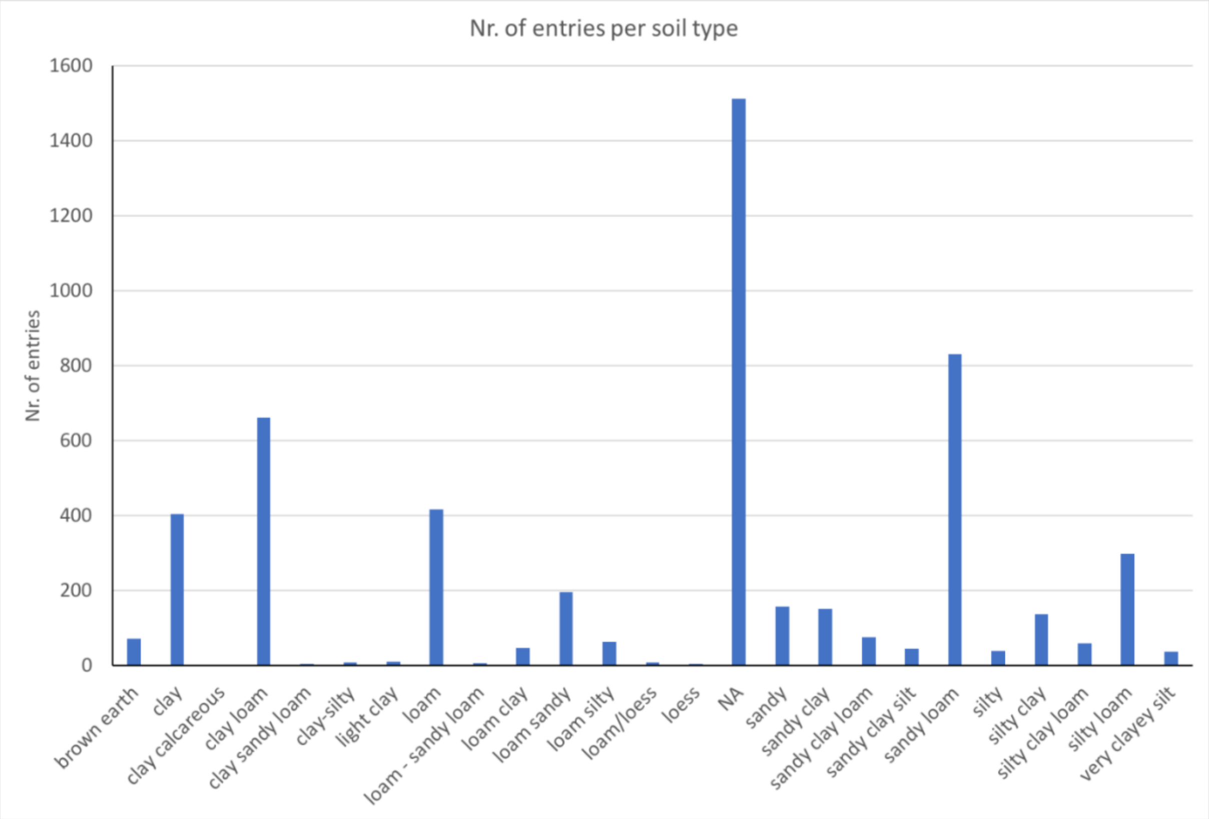
Figure 5 - Number of varieties grouped per legume species



This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°727672

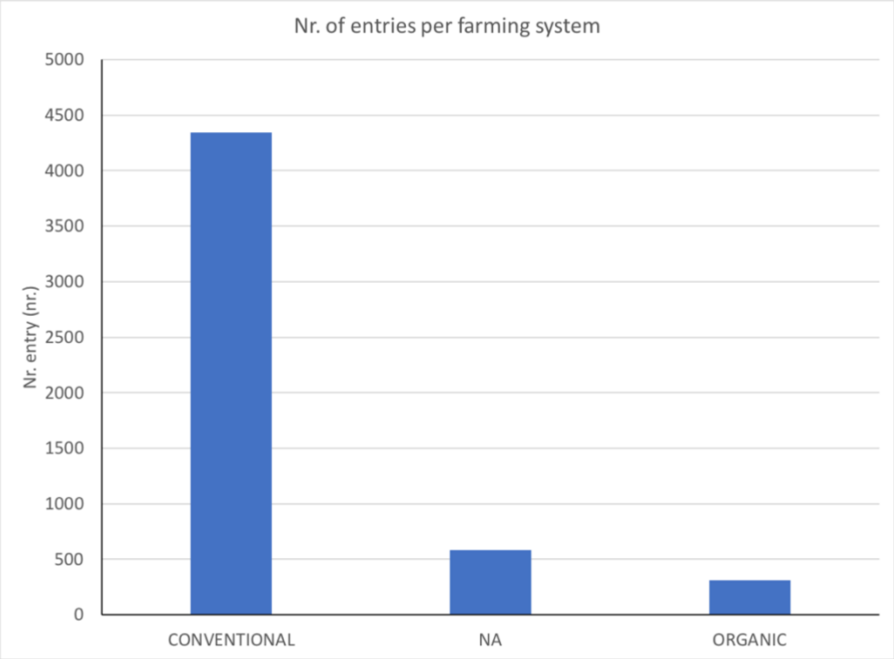
The dataset covers a wide range of different soil types and texture, from very sandy to very clay and calcareous soils, with highest presence of intermediate texture (Figure 6).

Figure 6 - Number of entries grouped per soil texture



For what concerns the farming system, most of the entries refers to studies or treatments managed according to non-organic management, whilst organic farming is represented only by 6% of the entries (Figure 7).

Figure 7 -Number of entries grouped per farming system



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

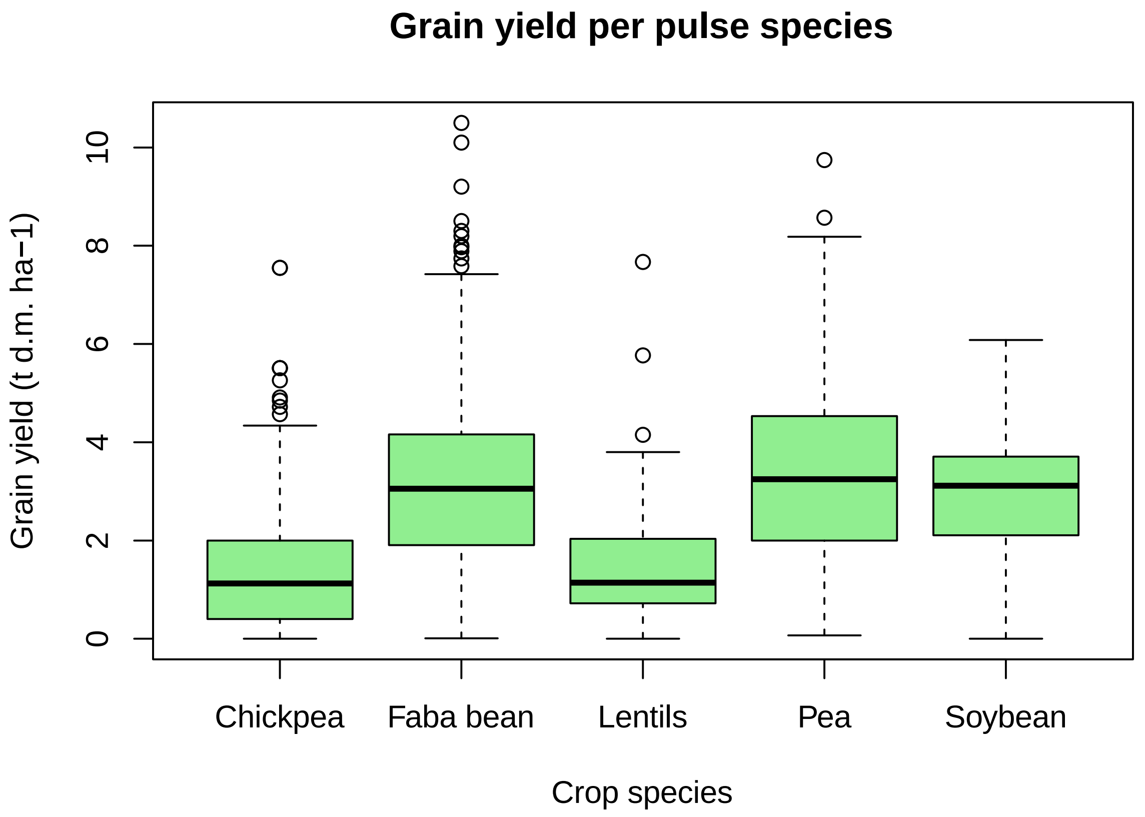
4.2. Effects of factors on crop grain yield

The grain yield in the dataset is normally expressed as t ha^{-1} of dry matter. Anyway, for many papers and experimental data the moisture level of the grain was not reported or not measured, affecting somehow the comparability of the results. In many cases, the mean values of grain yield were not accompanied by measures of variability (standard error, standard deviation, coefficient of variability, variance), thus hampering the possibility to fully exploit the dataset for meta-analyses.

Papers and experimental data not clearly stating the plant produce considered (e.g. grain or total aboveground biomass) were also removed from the dataset.

Concerning the trend of grain yield as affected by crop species, the dataset confirms that the three major pulses (i.e. soybean, field pea and faba bean) grown in Europe are the more productive ones, whilst a clear difference was observed for chickpea and lentils (Figure 8).

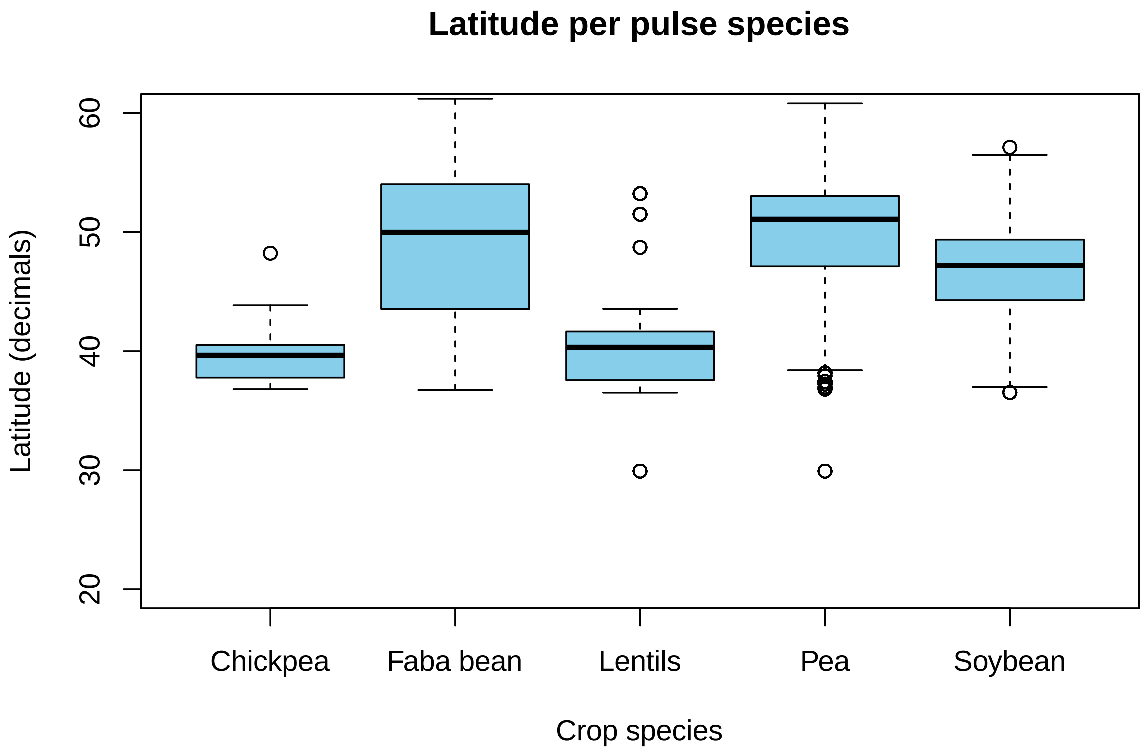
Figure 8 - Boxplot of grain yield (t d.m. ha^{-1}) as affected by legume species



As shown in Figure 9, the trend of grain yield reflects also the geographical distribution of the five crops, with chickpea and lentils mostly studied at lower latitudes.

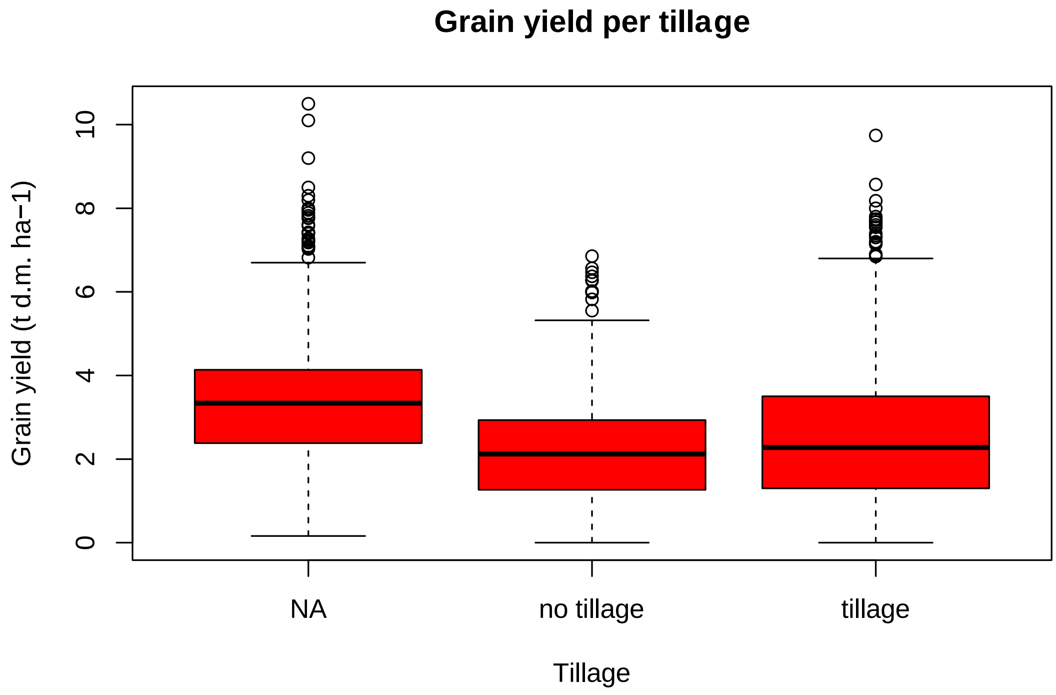
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Figure 9 - Boxplot of latitude decimals as affected by legume species



Among management practices, overall tillage systems (Figure 10), herbicide application (Figure 11), irrigation (Figure 12) and organic farming (Figure 13) did not clearly affected the levels of variability in legume yields.

Figure 10 - Boxplot of grain yield (t d.m. ha-1) as affected by tillage systems



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°727672

Figure 11 - Boxplot of grain yield (t d.m. ha-1) as affected by herbicide application

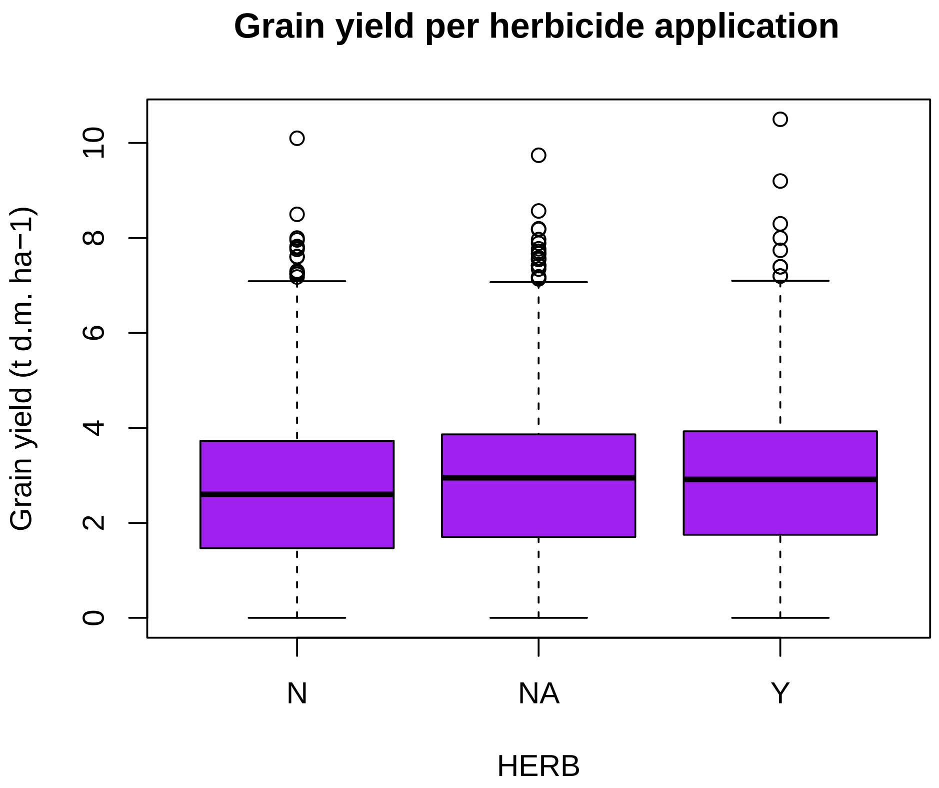


Figure 12 - Boxplot of grain yield (t d.m. ha-1) as affected by irrigation (covering part or full water needs of the crops)

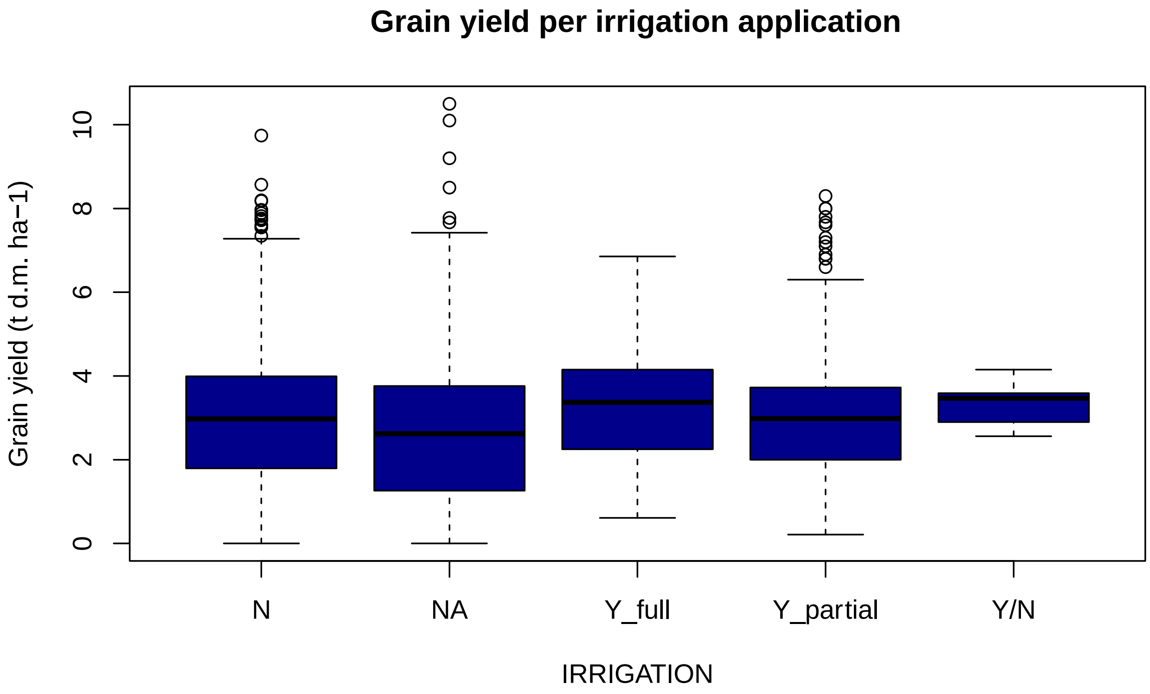
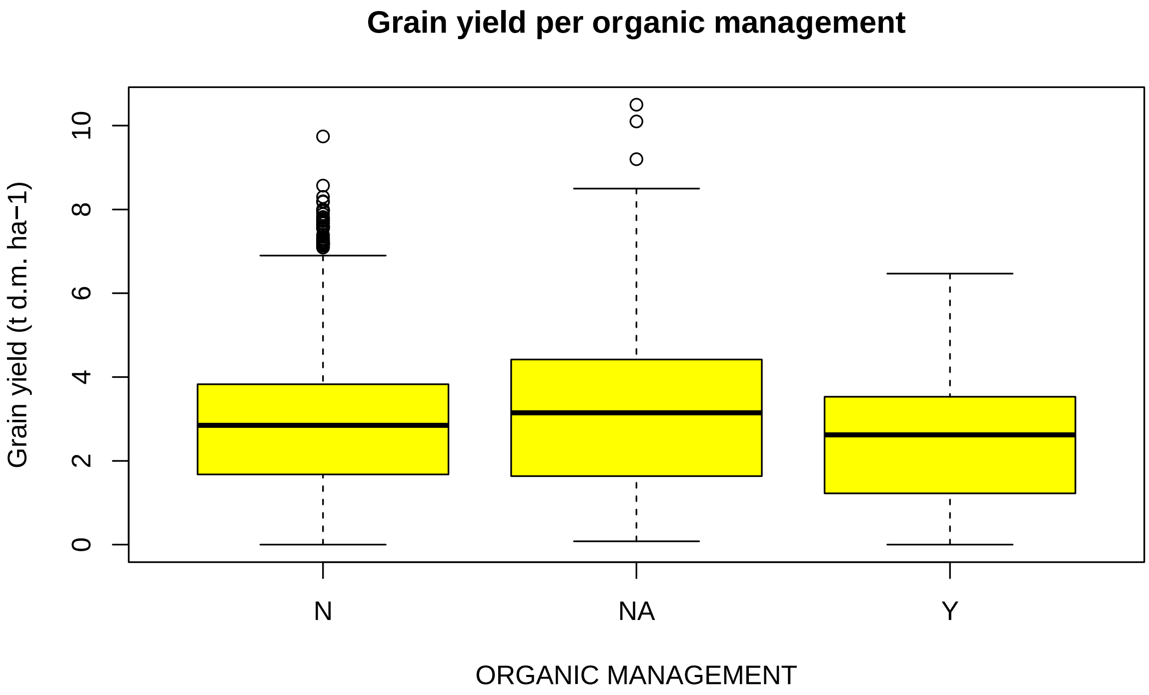
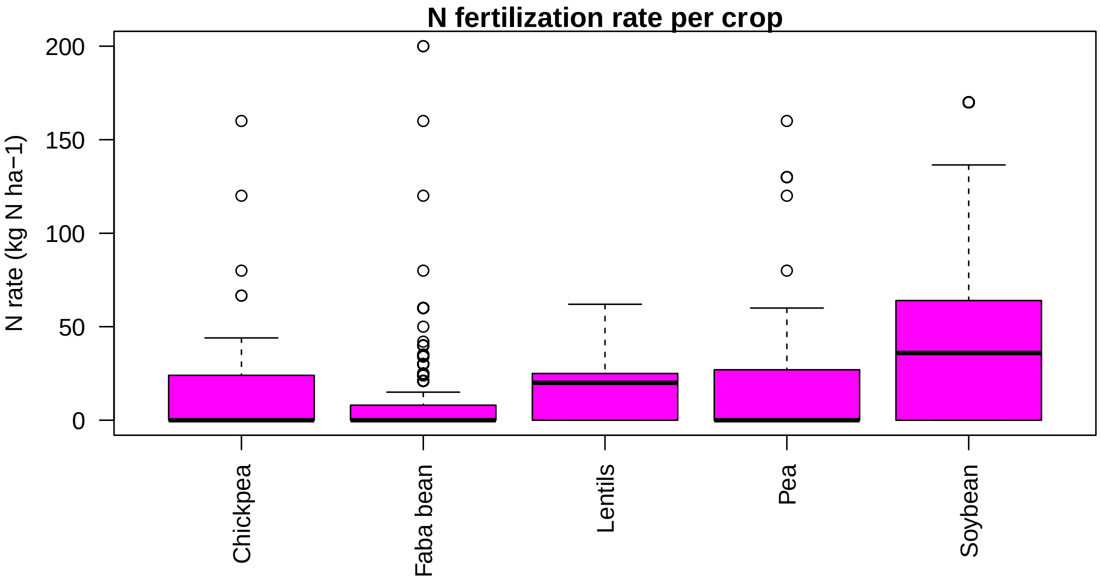


Figure 13 - Boxplot of grain yield (t d.m. ha-1) as affected by organic farming management



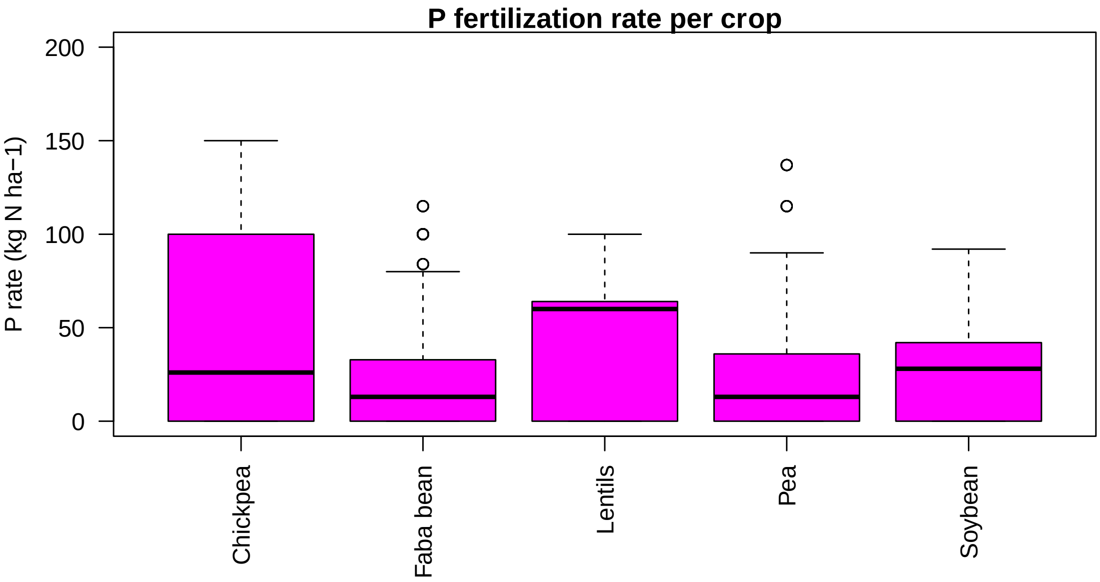
Concerning the fertilisation, nitrogen is normally applied to soybean at higher levels respect to the other legumes (Figure 14).

Figure 14 - Boxplot of N fertilisation rate (kg N ha^{-1}) as affected by legume species



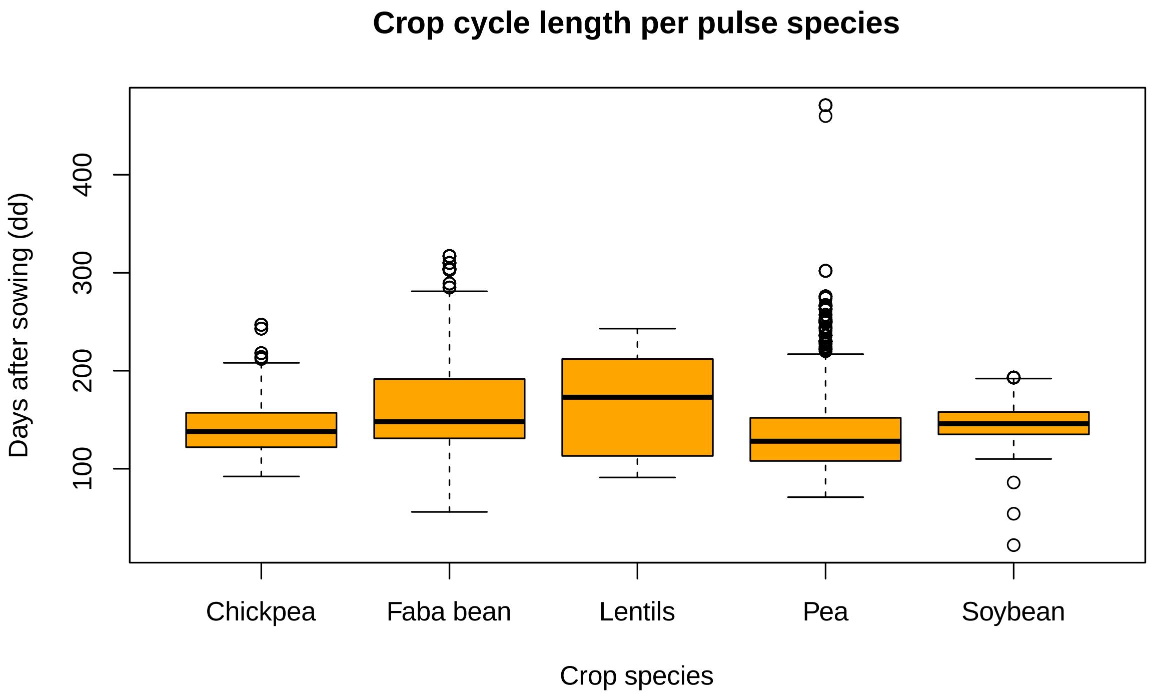
For phosphorus fertilisation, the data included in the dataset reveals an opposite trend, with higher levels often tested for less productive pulse species (chickpea and lentils).

Figure 15 - Boxplot of P fertilisation rate (kg P ha^{-1}) as affected by legume species



Concerning the length of the crop growing cycle, the data included in the dataset reveal narrower ranges for soybean and chickpea than for faba bean, lentils and pea (Figure 16).

Figure 16 - Boxplot of the length of growing cycle (expressed in terms of days from sowing to harvest)



As shown in Figure 17, the data included in the dataset do not cover evenly all the crops in terms of geographical distribution for chickpea and lentils, whereas for soybean, pea and faba bean a wider distribution can be observed. For all the crops we observed a high variability of the grain yield, likely related to interannual or spatial variation.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N°727672

For what concerns the soil type, the data in the dataset show a trend towards higher yield value for soybean in loam soils, and poor adaptation of all the legumes to silty soils (Figure 18).

Finally, for soil tillage we observed positive (soybean and lentils), neutral (faba bean and chickpea) and negative (field pea) effects of no tillage on the grain yield of the 5 legumes in terms of median and range of variability (Figure 19).

For the other management aspects tested for this report (i.e. irrigation, herbicide application, fertilisation), clear outcomes were not identified.

Figure 17 - Boxplot of grain yield (t d.m. ha-1) of the 5 legume species as affected by organic farming management

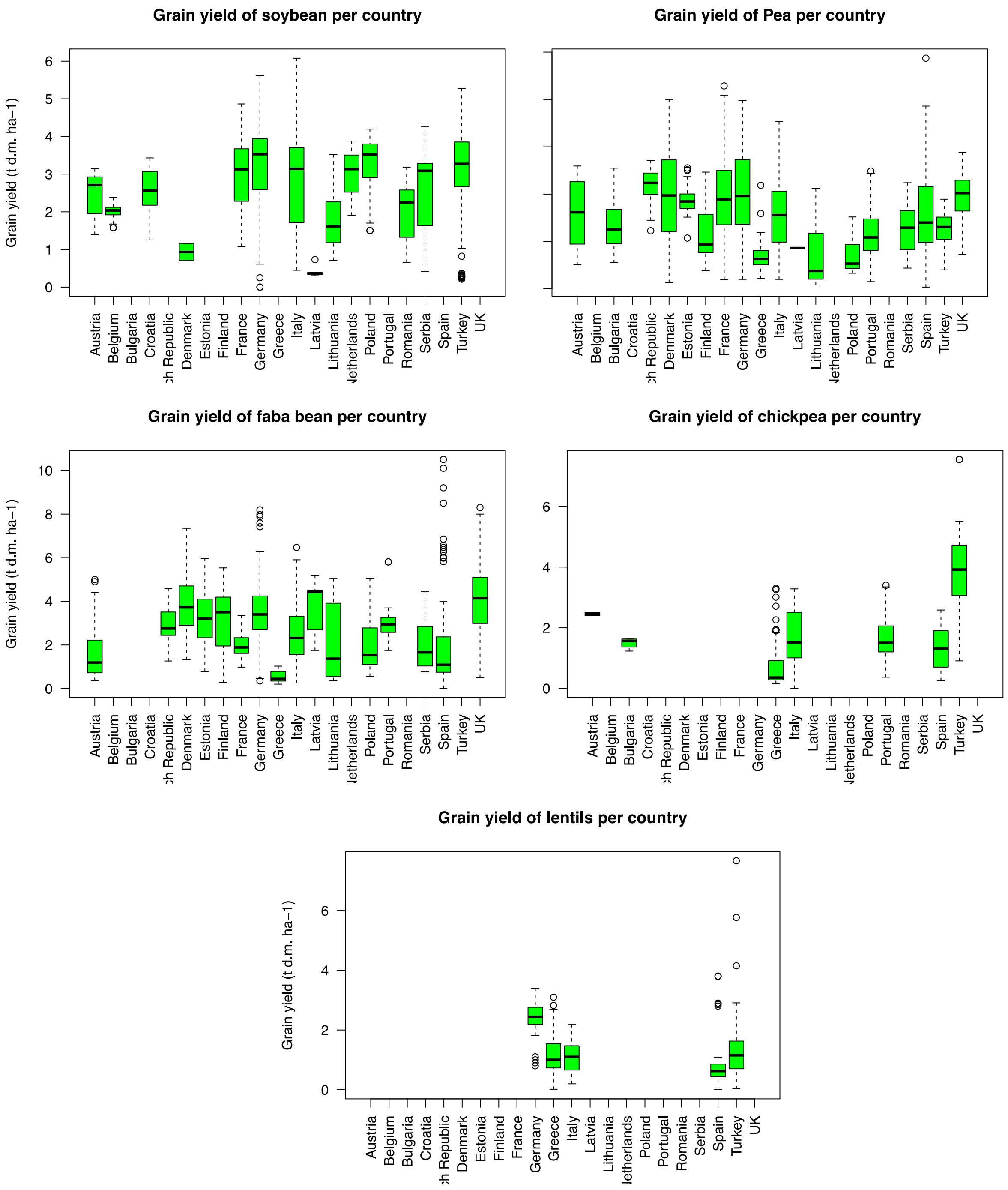


Figure 18 - Boxplot of grain yield (t d.m. ha^{-1}) of the 5 legume species as affected by soil texture

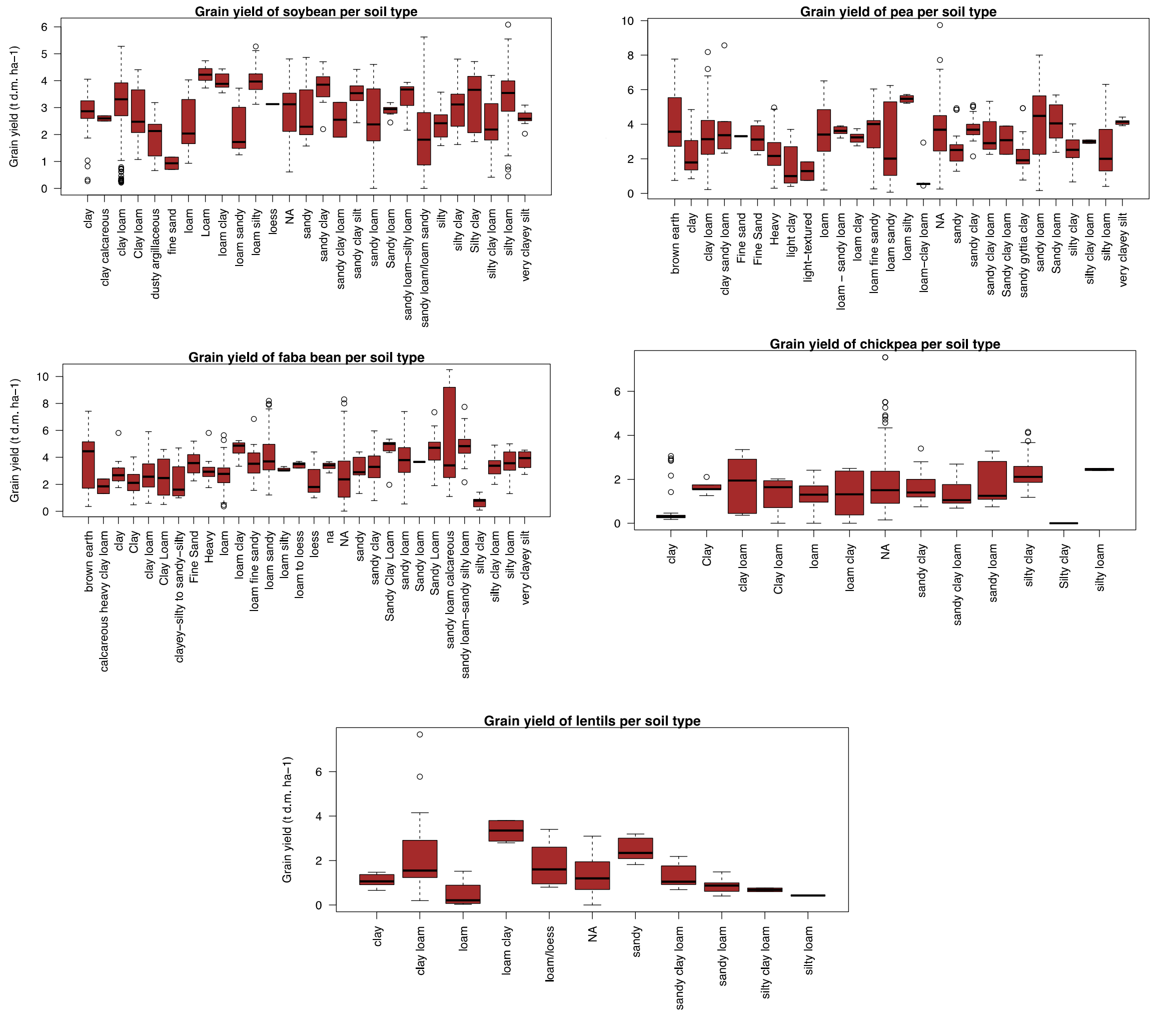
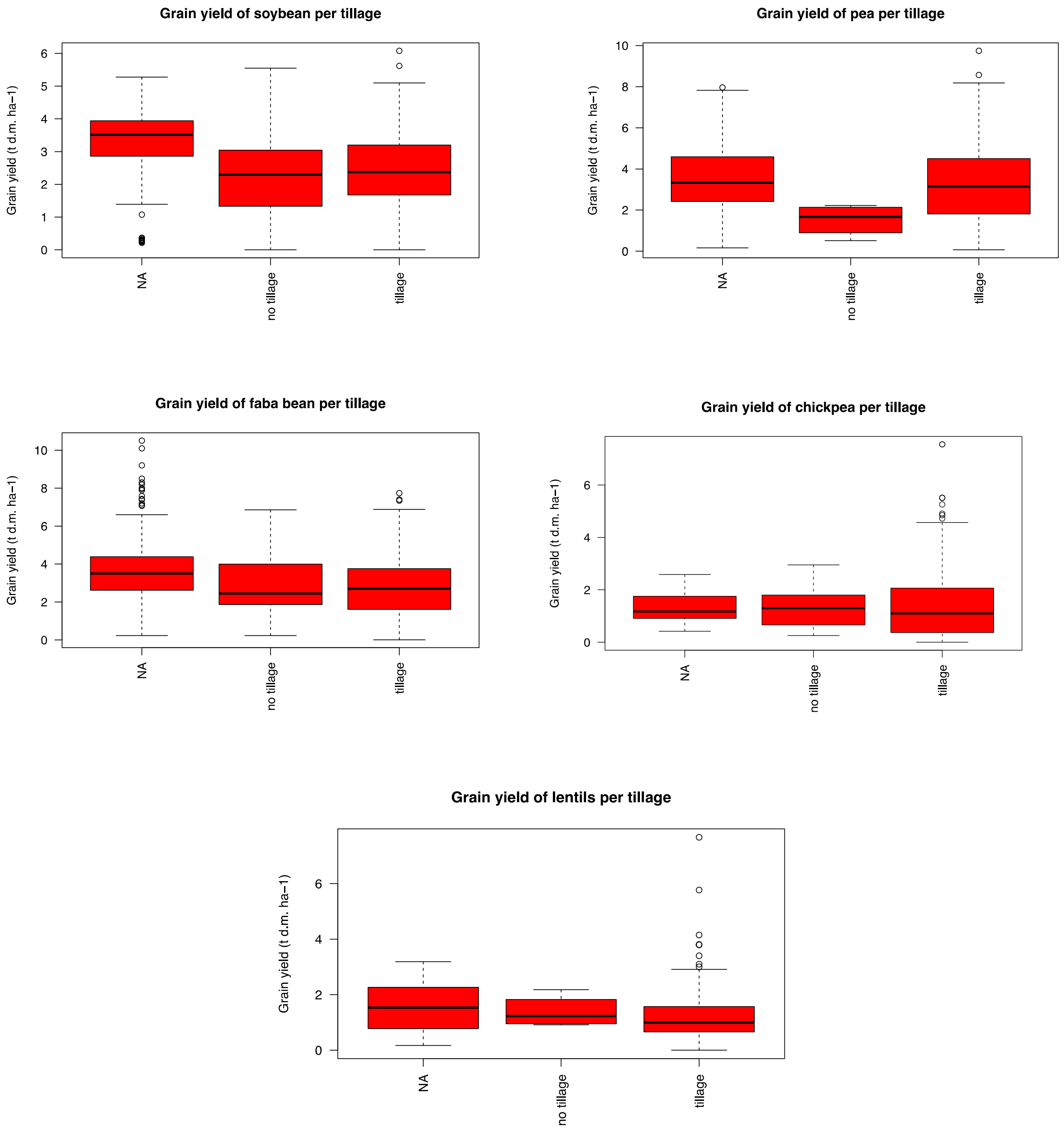


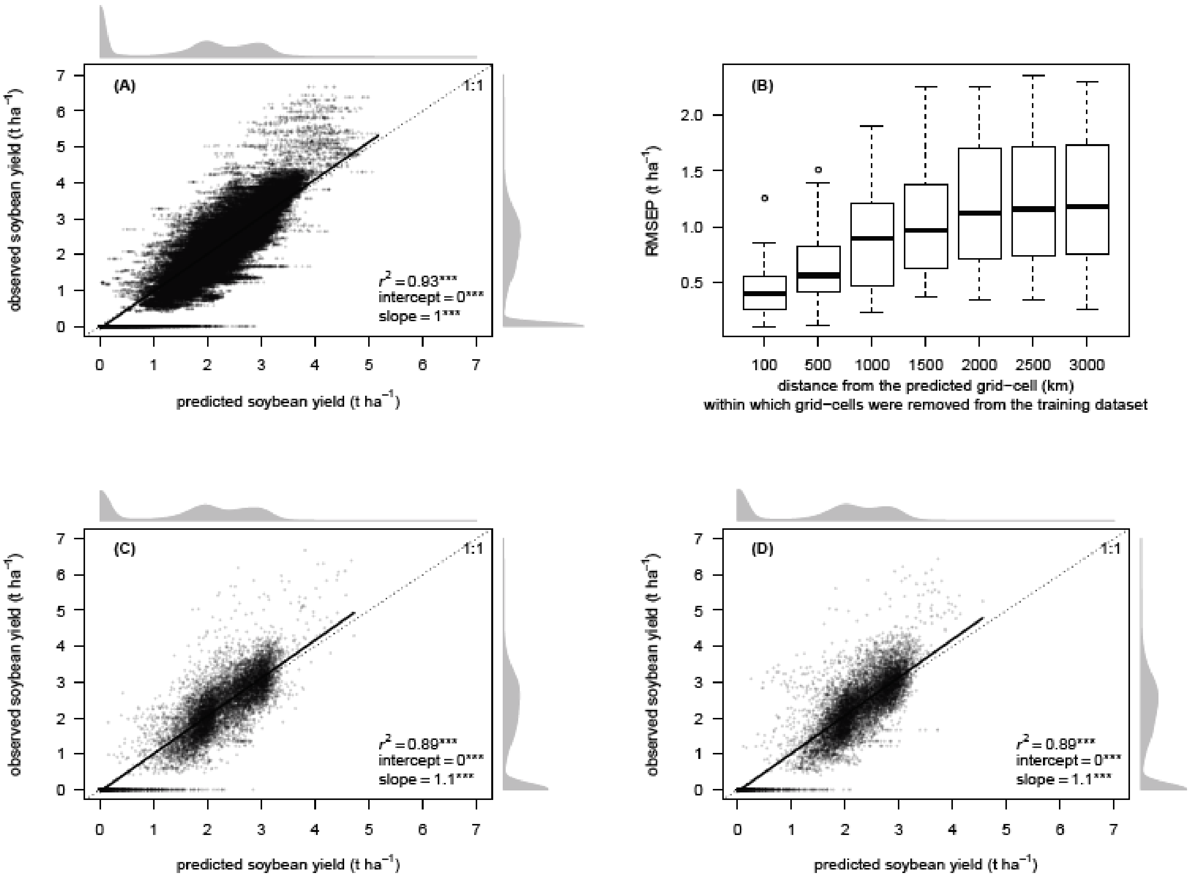
Figure 19 - Boxplot of grain yield (t d.m. ha^{-1}) of the 5 legume species as affected by soil tillage



4.3. A first example of achievable yield maps for soybean in Europe

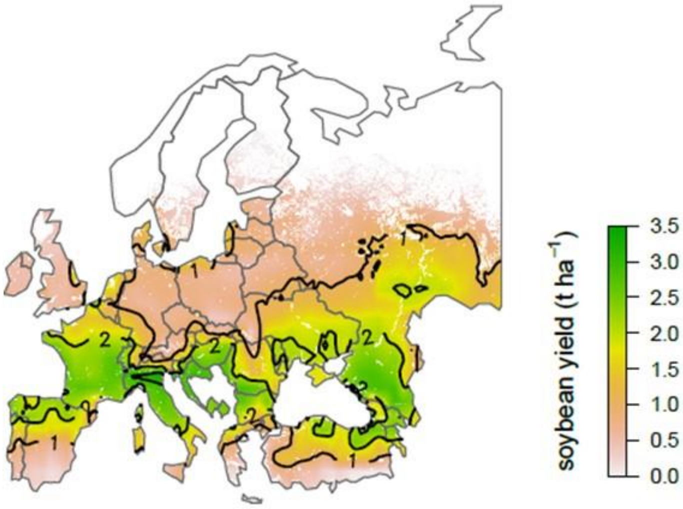
Model fitting. Here we present the results obtained for soybean in Europe using the global dataset of historical soybean yield (see section 3.5 for data and methods presentation). Among algorithms tested Random Forest appears to be the most accurate (Figure 20-A) in terms of root-mean-squared error of prediction (RMSEP = 0.35 t ha^{-1}) and Nash–Sutcliffe model efficiency coefficient (MEF = 0.93), as estimated with a classical (unstratified) cross-validation. It also displays the best transferability in time (RMSEP = 0.45 t ha^{-1} when applied to years different from those used for training, Figure 20- C,D) and space (RMSEP = 0.43 t ha^{-1} when applied in locations distant by 500 km, Figure 20-B). Our results reveal that transferability in space decreases with increasing distance between training and test datasets for all models, with a threshold of 1000 km above which the performance of the selected algorithm deteriorates markedly (Figure 20-B).

Figure 20 – Assessment of the Random Forest algorithm. (A) The model is first evaluated using a classical bootstrap approach with 25 resamplings. (B) Model transferability in space is then evaluated by ensuring a minimum spatial distance between training and test datasets. Finally, model transferability in time is assessed in (C) where model is fitted on 1981-1995 to predict 1996-2010, and in (D) where model is fitted on 1996-2010 to predict 1981-1995. RMSEP: root mean square error of prediction. Boxplot in panel (B) shows median (center line), 1st and 3rd quartiles (box limits), and 1.5 times the interquartile range (whiskers). Linear regression outputs are shown on panels (A), (C), and (D), as well as marginal distributions of observed and predicted soybean yields (in grey). Dotted lines represent the 1:1 line. In order to extend the range of climate conditions captured by the model and to capture climate conditions leading to zero yield, additional data points were randomly sampled in climate zones known to be unsuitable for soybean production (e.g. deserts and arctic areas) and added to the dataset with their yield value set to zero.



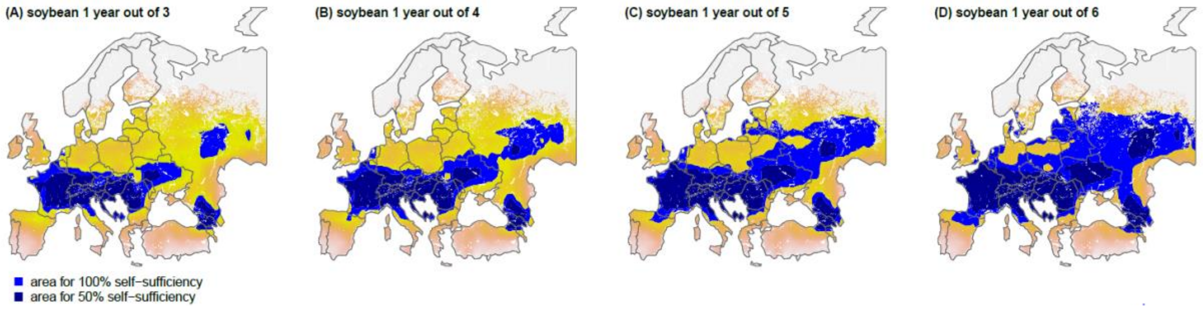
Projections of soybean yield. The projections of the Random Forest algorithm – which assume no irrigation and fixed growing period from April to October – suggest high suitability for soybean under historical climate (Figure 21). About 100 Mha show projected yield equal or higher than 2 t ha^{-1}, while in 2016 the soybean production area in Europe was only 5 Mha with 2 t ha^{-1} of average yield (FAOSTAT, 2019). Therefore, soybean suitable area appears to be much larger than current harvested area in Europe, which suggests that soybean production is not limited by climate conditions.

Figure 21 – Projected soybean yield in Europe under historical climate (1981-2010). Projections are shown only on agricultural area (cropland plus pasture), in the year 2000.



Area requirements for 50% and 100% soybean self-sufficiency in Europe. We estimate the soybean production area required to reach a self-sufficiency level of 50% and 100% based on yield projections presented in Figure 21. A three-step procedure was followed. First, we assumed that soybean could only be grown on current cropland. Under this assumption, soybean cannot be grown in place of permanent pastures, in line with the Common Agricultural Policy of the European Union aiming at their protection. Second, we considered four scenarios for the increase of soybean frequency in crop sequences. In these scenarios, soybean is grown in one year out of three, four, five, or six years, which correspond to 33%, 25%, 20%, and 16% cropland area in a grid-cell under soybean, respectively. These scenarios are consistent with observed and recommended soybean frequencies in crop sequences in Europe. Indeed, a 1-in-3 year or 1-in-4 year soybean cultivation is often recommended to limit the risk of disease occurrence (especially those caused by two fungal pathogens *Sclerotinia sclerotiorum* and *Rhizoctonia solani*), although higher frequencies are observed in Europe and other countries. Third we assumed that soybean is grown preferably in high-yielding grid-cells. Based on this assumption, soybean areas were allocated to grid-cells ranked in decreasing order of projected yield values until the cumulated production (calculated as the product of area and yield) reached 50% and 100% of current annual soybean consumption of Europe (58 Mt in average over 2009-2013[1]). Results suggest that a self-sufficiency level of 50% (100%) would be achievable in Europe under historical climate whatever the frequency of soybean in crop sequences, if ~5% (~10%) of the current European cropland is dedicated to soybean production (Figure 22).

Figure 22 – Area requirements for 50% and 100% soybean self-sufficiency in Europe under historical climate (1981-2000) based on soybean yield projections presented in Figure 21 and assuming various levels of soybean frequency in crop sequences (one year out for three, four, five and six years). Soybean areas were allocated to grid-cells ranked in decreasing order of projected yield values until the cumulated production (calculated as the product of area and yield) reached 50% (light blue) and 100% (dark blue) of the current annual soybean consumption of Europe (58 Mt, average 2009-2013). We assume that soybean can only be grown on current cropland, which excludes permanent pastures in line with the Common Agricultural Policy of the European Union aiming at their protection. Background colors indicate projected soybean yield in t ha^{-1} as in Figure 21.


(A) soybean 1 year out of 3
(B) soybean 1 year out of 4
(C) soybean 1 year out of 5
(D) soybean 1 year out of 6
■ area for 100% self–sufficiency
■ area for 50% self–sufficiency

5. Conclusions

The dataset built under Task 1.2 allowed to consolidate the scientific knowledge on the sources of variability in the yield of grain legumes in Europe, focusing on environmental and agronomical factors. This exercise allowed to confirm the evidence of poor representation of pulses in the existing scientific literature, but also highlighted the existence of grey literature and valuable unpublished results that could add value to the state of the art.

Among the five selected species, soybean and field pea confirmed to be the most studied by scientists, whilst chickpea and lentils are still much constrained by the market dimensions and are not so represented in the scientific production, especially in Northern countries.

Pulses are grown in very contrasting pedoclimatic conditions and this results in quite high variability of grain yields. Although the dataset contributes to identifying some relationships between different species and environmental conditions, additional efforts should be paid to better explore the interactional effects of soils and climates and management. For this aspect, the yield maps targeted to be produced by Task 1.2 will contribute substantially to identify more clear trends.

Among management options, crop genotype is definitively the most present in the studies considered for this activity and in many cases is tested in combination with other agronomic practices. The dataset could represent the baseline for identification of best performing varieties at the European level, thus boosting a wider adoption of cultivars adapted to similar conditions and characterized by desired traits (e.g. pest resistance, weed competitiveness, poor reliance on NP fertilisation, low water demand) and low variability. In this meaning, exploring the dataset addressing the issue of identifying genotypes already tested in different conditions could represent the first step to establish transnational cooperation for genetical improvement of pulses, taking into account not only environmental conditions but also agronomic factors.

Surprisingly, despite the high importance widely recognized to biological N_2 fixation of legumes, N fertilisation is still very present in studies involving also grain legumes. It should be remarked that, although many farmers still consider N fertilisation as a strategy to repair from instability of weather conditions resulting in variable effectiveness of root nodulation by Rhizobia, biological nitrogen fixation could be heavily depressed by availability of mineral forms of N in the soil. Variety trials, seed/root inoculation tests, bioactive compounds and innovation in soil tillage targeted to reduce soil disturbance and improve soil physical quality should be encouraged instead, in order to find out ways to support dinitrogen fixation in legume-based cropping systems.

Finally, the construction of the dataset clearly revealed poor scientific quality or clarity in the existing literature that is constraining a more effective exploitation of the background (e.g. to perform meta-analysis on selected traits). Future research efforts should accomplish for high scientific quality of the results, that is necessary to make significant steps further in the development of innovative and viable solutions to include grain legumes in cropping systems throughout the EU.

The preliminary tests performed on soybean yield estimations were very positive. The selected algorithm (Random Forest) allowed to predict in a reliable way soybean yield under historical climatic conditions. According to preliminary analysis, although not considering the positive effect of irrigation on potential yields, the internal demand of soybean grain seems to be likely met by EU production even if the legume would be present just 1 year over 6 of crop rotations.

6. Acknowledgements

We thank all the LEGVALUE partners that have contributed to this work, in particular by sharing their knowledge and experimental data: INRA, TERIN, WU, PGRO, WR, SSSA, INIAV, AICF, SEGES, LLKC, FH-SWF, CRAN, LAMMC, FiBL.

A special thanks goes to Nicola Guilpart (INRA) who took the lead of modelling activity and yield map generation.

We also acknowledge David Makowski and Nicolas Guilpart (INRA) for their intensive and clever support in the construction of the dataset.

The dataset was built at UNIPI thanks to the work of Silvia Pampana and Lorenzo Gabriele Tramacere who contributed significantly in literature review, data extraction and data entry.

7. References

Cernay, C., E. Pelzer, and D. Makowski. 2016. Data descriptor: A global experimental dataset for assessing grain legume production. Sci. Data 3: 1–20. doi: 10.1038/sdata.2016.84.

Iizumi, T. et al. Historical changes in global yields: Major cereal and legume crops from 1982 to 2006. Glob. Ecol. Biogeogr. 23, 346–357 (2014).

Iizumi, T., Okada, M. & Yokozawza, M. A meteorological forcing data set for global crop modeling: Development, evaluation, and intercomparison. J. Geophys. Res. Atmos. Res. 119, 363–384 (2014).

Magrini, M.-B., Anton, M., Cholez, C., Corre-Hellou, G., Duc, G., Jeuffroy, M.-H., Meynard, J.-M., Pelzer, E., Voisin, A.-S. & Walrand, S. 2016. Why are grain-legumes rarely present in cropping systems despite their environmental and nutritional benefits? Analyzing lock-in in the French agrifood system. Ecological Economics, 126, 152-162.

Pelzer, E., Bourlet, C., Carlsson, G., Lopez-Bellido, R., Jensen, E. & Jeuffroy, M.-H. 2017. Design, assessment and feasibility of legume-based cropping systems in three European regions. Crop and Pasture Science, 68, 902-914

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Schreuder, R. & De Visser, C. 2014. EIP AGRI Focus Group; Protein Crops: final report. European Commission, Brussels, Belgium.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

8. Annexes

The grain legume yield dataset (public version without data available only internally to LegValue).