European Plant Phenotyping Network

EPPN 2020

D2.2 Quality annotation protocols for phenotypic platform data

Nadine Hilgert, Isabelle Sanchez, Emilie Millet, Fred van Eeuwijk

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731013. This publication reflects only the view of the author, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

Document information

EU Project N°	731013	**Acronym**	EPPN2020
Full Title	European Plant Phenotyping Network 2020		
Project website	www.eppn2020.plant-phenotyping.eu		

Deliverable	**N°**	D2.2	**Title**	Quality annotation protocols for phenotyping platform data
Work Package	**N°**	WP2	**Title**	Design and analysis of phenotyping experiments across multiple platforms, scales of plant organisation, traits and management conditions

Date of delivery	**Contractual**	31/07/2019 (Month 27)	**Actual**	28/10/2019 (Month 30)	
Dissemination level	X	**PU Public, fully open, e.g. web**			
		CO Confidential, restricted under conditions set out in Model Grant Agreement			
		CI Classified, information as referred to in Commission Decision 2001/844/EC.			

Authors (Partner)	INRA			
Responsible Author	**Name**	Nadine Hilgert	**Email**	nadine.hilgert@inra.fr

Version log			
Issue Date	**Revision N°**	**Author**	**Change**
08/07/2019	0	Nadine Hilgert, Isabelle Sanchez	first version
18/07/2019	1	Emilie Millet	correction from co-authors
28/10/2019	2	Fred van Eeuwijk	first review by WP leader

Executive Summary

The present deliverable specifically addresses quality annotation protocols for phenotyping platform data. We first explain what cleaning phenotypic data is and why it is important to do it and keep track of how it was done. We then provide platform users with clearly described and defined rules for outliers identification and annotation in an automatic and traceable way.

An outlier is usually defined as an observation that appears to be inconsistent with the remainder of the dataset. After visiting a number of facilities and discussing with platform users, we have defined three types of outliers to annotate in the phenotypic data: (1) time points within a time course, (2) whole time courses of one or more variables and (3) a whole plant, defined here as a biological replicate deviating from the overall distribution of plants on a multi-criteria basis. This classification of outliers was proven relevant by the consortium partners. In this document, we propose procedures to identify them. For the first two types of outliers, statistical methods already exist and have been adapted and applied to datasets from different platform/species. The «plant outlier» type is new and a method has recently been published (Alvarez Prado et al., 2019). The common idea here is to provide annotated data to the user who, in the end, will decide whether or not to keep the annotated points, time course or plant for further analyses. The information will be stored in the information system, together with the rules for outlier detection as meta-data of the genetic analysis (see Neveu et al., 2019).

Authors/Teams involved:
INRA: Isabelle Sanchez and Nadine Hilgert (MISTEA), Llorenç Cabrera-Bosquet, Claude Welcker, Santiago Alvarez Prado and François Tardieu (LEPSE)
UCL: Xavier Dray
WUR: Emilie Millet, Fred van Eeuwijk

Table of contents

1. INTRODUCTION

1.1. Aim of the project

The European Plant Phenotyping Network 2020 (EPPN2020) project aims at providing European public and private plant scientists with access to a wide range of state-of-the-art plant phenotyping facilities, techniques and methods. It will aid the community in progressing towards excellence across the whole phenotyping pipeline, involving sensors and imaging techniques, data analysis in relation to environmental conditions, data organization and storage, data interpretation in a biological context and meta-analyses of experiments. It coordinates its activities with the future infrastructure EMPHASIS, listed in the ESFRI roadmap, and with national programs. EPPN2020 involves:

- access to 31 key installations in 15 infrastructures,
- a Work Package on sensors (WP1),
- a Work Package on data analysis (WP2),
- a Work Package about data management (WP3),
- networking activities for establishing cooperation and increasing integration between facilities both within and outside EPPN2020.

1.2. Scope and aim of the document

The Work Package 2 (WP2) develops tools for statistical analysis of phenomic experiments across platforms and scales of plant organization. These tools should be applied to data collected in the installations of the EPPN2020 consortium. They have been tested on data from previous projects and collaborations when this was deemed beneficial for the quality of the tools. The tools and methods should finally be applicable to the majority of phenotyping platforms. The activities in WP2 will allow the phenomic community to progress towards standardized statistical analyses and will facilitate the combined analysis of data coming from multiple platforms and measurement scales.

At the start of EPPN2020, there was a clear demand for a unified set of tools and methods to analyse platform data. The diversity of phenotyping techniques and the increasing amount of data points available made it difficult for platform users to directly apply designs, models and analysis methods originally developed for field trials. The objectives of the WP2 were defined to remedy the lack of appropriate statistical design and analysis tools for data from phenotyping platforms.
The aim of this document is twofold. We first explain what cleaning phenotypic data is and why it is important to do it and keep track of how it was done. In a second time, we provide platform users with clearly described and defined rules for data quality control by identifying and annotating outliers in an automatic and traceable way.

2. WHAT TO CLEAN IN PHENOTYPIC DATASETS AND WHY ?

An outlier is usually defined as an observation that appears to be inconsistent with the remainder of the dataset (Barnett and Lewis, 1994). Observations may be time points (Grubbs, 1950) or whole time courses of one or more variables (Hubert et al., 2015). An illustration is given in Figure 1, where we see the difficulty of deciding whether a biomass time course (red

curve at the bottom left) is atypical in a set of time courses from plants of the same genotype, or whether this slow growth is due to the plant's position in the platform.

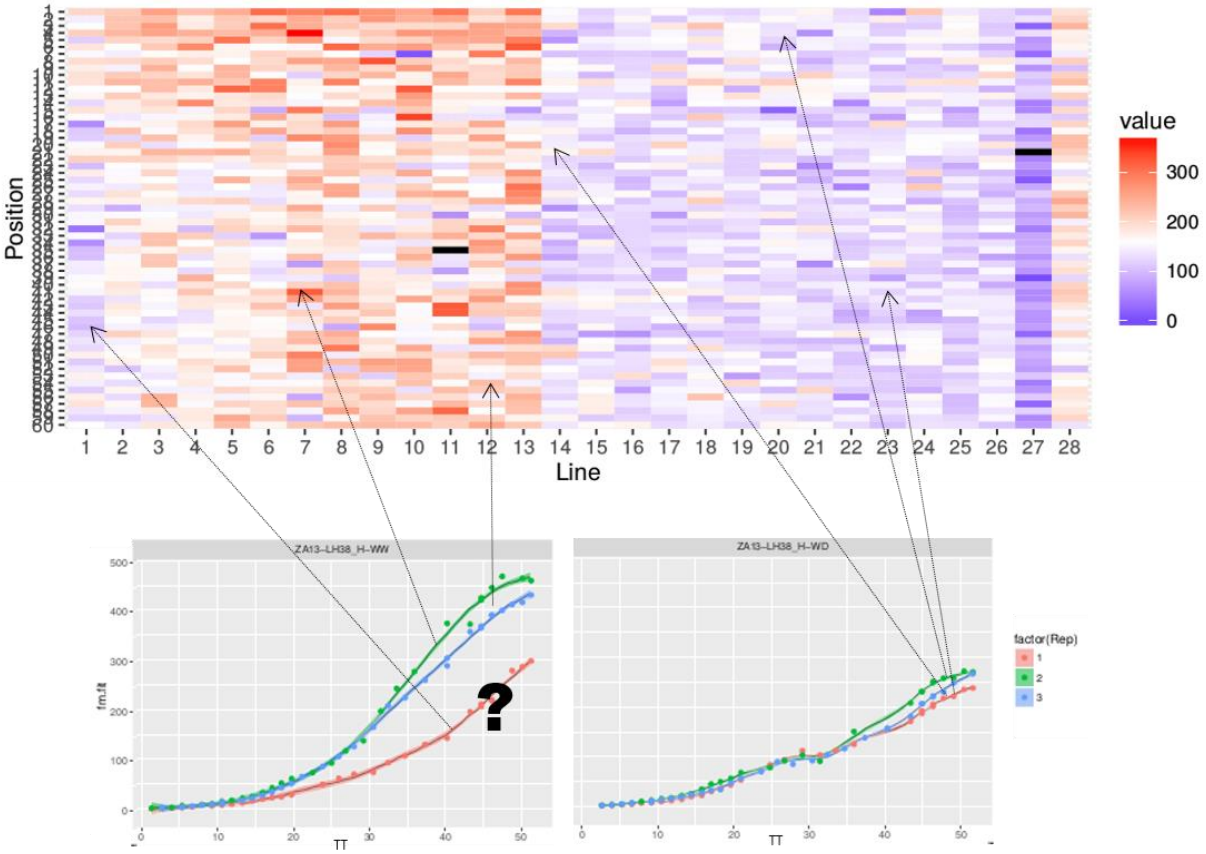

Figure 1. Heatmap of the biomass estimated at a specific time point (with a color gradient) of each plant according to its location in the platform (line, position). Well-watered (WW) treatment on the left, water deficit (WD) on the right. The graphs below show the biomass time courses for a given genotype (3 repetitions in each treatment). The question is whether the red growth curve on the left corresponds to an outlier plant (seed problem for example) or whether this slow growth is due to the plant's location in the platform (PenoArch Platform, INRA). *Courtesy: Llorenç Cabrera-Bosquet and Santiago Alvarez Prado.*

The concept of outlier can be extended to “outlier plants”, defined here as biological replicates deviating from the overall distribution of plants on a multi-criteria basis, regardless of the quality of measurements (Alvarez Prado et al., 2019). For example, outlier plants can originate from bad seed quality, from wrong genotype identification or from fertilization of ovaries by undesired pollen, e.g. generating a hybrid instead of an inbred line, with an important effect in the case for species grown as hybrids derived from lines with high consanguinity. In field experiments, outlier plants have a low impact on genotypic mean estimation, unless all seeds of the considered genotype have a low quality. This is because experimental units (defined as the smallest entity to which a treatment can be applied) are, in the field, microplots containing tens of plants. In phenotyping platforms with hundreds of genotypes, but also in many other experiments in controlled conditions, the experimental unit is frequently an individual plant with three to ten replicates per genotype, so the presence of one or more outlier plants may have a high impact on genotypic means (Estaghvirou et al., 2014).

Whereas numerous methods were developed for detecting outlier points or outlier time courses, the detection of outlier plants is still in its infancy. This is probably because the concept of outliers is less common in the case of individual plants, which are associated with a multiplicity of variables with different distributions. Statistical methods based on individual traits are reproducible for a given experiment, but they may exclude different plants depending

on the considered trait, resulting in different final trait-specific datasets for each variable, see Figure 2. Visually removing outlier plants based on expert intuition is the most used method, and can result in similar accuracy compared with statistical methods (Bernal-Vasquez et al., 2016). However, criteria for visual elimination can appreciably differ between experimenters. Moreover, whereas visual cleaning can be performed in small datasets, it becomes nearly impossible when thousands of time courses need to be analysed.

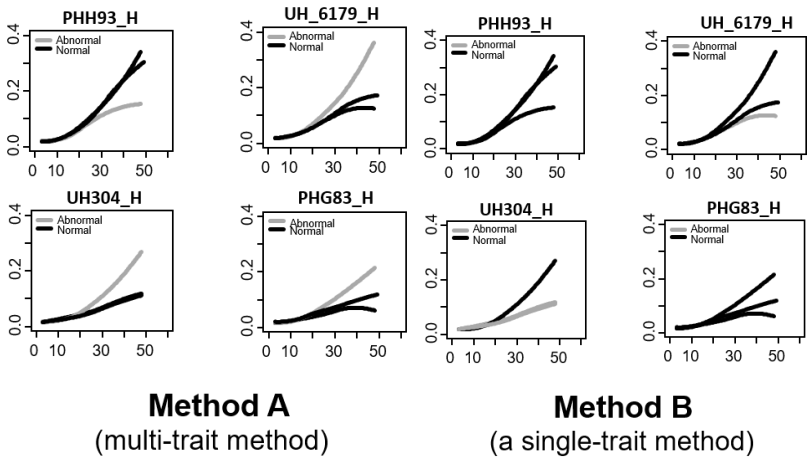

Figure 2. Detection of plant outliers with two approaches: a multi-trait method and a single trait one. Results are presented for 4 genotypes (3 replicates each) at the PenoArch platform (INRA, France). The grey curves correspond to replicates detected as outliers. They are not the same according to the methods. *Courtesy: Llorenç Cabrera-Bosquet and Santiago Alvarez Prado.*

In addition, issues other than the significance of statistical tests on variable of interest need to be considered for outlier plants. Indeed, the benchmark in this case is rather the degree to which one or another method affects the results of genetic analyses. In Alvarez-Prado *et al*, (2019), we showed that the results of genetic analyses largely depended on the cleaning method, as illustrated in Figure 3.

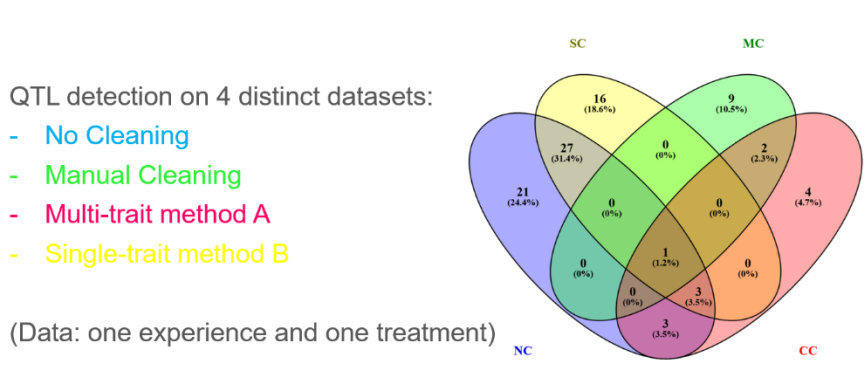

Figure 3. QTL detection using datasets originating from 4 different methods, as a function of the allelic frequency at the considered genomic position. The Venn diagram exemplifies the number of QTL for leaf area detected without cleaning (blue), with Statistical single-trait cleaning (yellow), with manual (visual) cleaning (green) or with a multi-trait method (red).

Removing outlier plants decreased the number of quantitative trait loci (QTLs), especially those at positions with highly unbalanced allelic frequencies. Consequently, outlier plants can generate false-positive QTLs and the cleaning method is an essential attribute of genetic analyses. The decision about cleaning or not cleaning depends on trades-off between the risk of false positive (with no-cleaning and/or a low threshold for minor allele frequency) and the risk of missing interesting rare alleles. Cleaning can lower the latter risk by making acceptable a higher threshold. Hence, datasets should be annotated, stored and organized in such a way that further re-analyses can be carried out with different methods for outlier detection.

3. METHODS FOR OUTLIER DETECTION

In the EPPN2020 project, a number of platforms have been visited in 2017/2018 to review both the experimental protocols implemented and the way the data is cleaned up. We tried to create a network of people interested in the process of data cleaning. We have worked with several (anonymized) datasets obtained on various platforms of the EPPN2020 project. The discussions with the experts of this data allowed us to refine both our questions (i.e. how to define an outlier plant?) and the statistical approaches and methods necessary to answer them.
In this section, we introduce the outlier detection methods identified as relevant for the platforms data, based on these discussions. It will help guiding the platform users in their work of cleaning up datasets. The methods presented here are the subject of a tutorial currently being finalised, that will be tested by EPPN2020 project partners next winter and presented at the next annual meeting.

3.1 Detection of outlying points

Time courses of phenotypic data are viewed as continuous time-related functions. The first cleaning step consists in checking the consistency of each point with respect to its neighbours within a time course. Outlying points are measurements at a given time that do not follow the expected behaviour. The detection requires fitting a model from the data, as a function of time. Two types of models are investigated below, one based on nonlinear parametric regression (Gompertz Model or sigmoidal function for example) and the other on nonparametric regression. Data annotation is based on the comparison of the experimental data with its estimated value from the model. If they significantly differ, the data will be annotated as suspect.

Parametric approach

Most of the data acquired in greenhouses are 'S-shaped' curves. The Gompertz or the sigmoid models are often used to fit such data. Once the model is fitted, a confidence interval is calculated. Points outside this interval will be declared outliers and annotated as such. The fitting requires the use of nonlinear least squares functions in R. We built a guide to facilitate the use of nonlinear fitting functions in R (with appropriate self-starting functions for example). The main advantage of this approach is that the fitted coefficients have a biological meaning. However, the models can be difficult to adjust if there is little data in the "exponential growth" part, or if the plateau is not reached.

Nonparametric approach: smoothing

The local regression is a well-known smoothing technique, which locally approximates an unknown function by parametric functions. It is a two-step procedure: 1/ evaluation at a set of points, then, 2/ interpolation to other points. A confidence interval can then be calculated. Points outside this interval will be declared outliers and annotated as such. The user can act on the smoothing parameter (the higher the parameter, the smoother the curve) and on the level with which the confidence interval is calculated (by increasing the level, there are fewer outliers detected). This nonparametric approach is well adapted when there is not enough data to fit a nonlinear model, or when no a priori shape can be distinguished from the times courses (Fig.4).

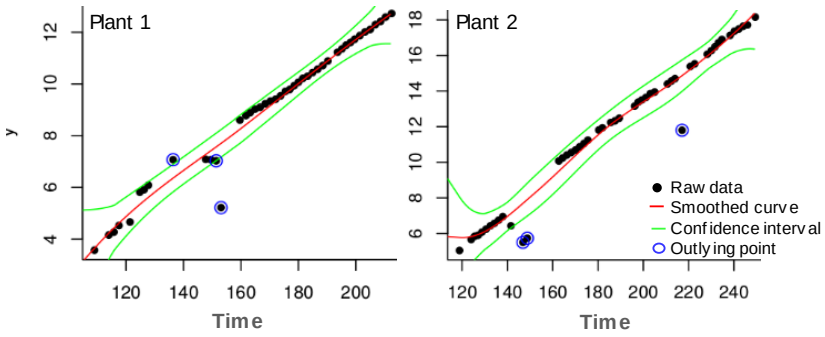

Figure 4. Detection of outlying points. The nonparametric approach applied on time courses of vertical coordinates (y) of the positions of the root tips (two plants from an experiment performed at the RootPhair platform, UCLouvain, Belgium). *Courtesy: Xavier Draye.*

The following Figure 5 shows an illustration of both parametric and nonparametric approaches on a time course of maize biovolume. The shape of the smoothed curves differs at the end of the experiment; the parametric curve flexes slightly to reach a plateau, while the non-parametric curve still increases.

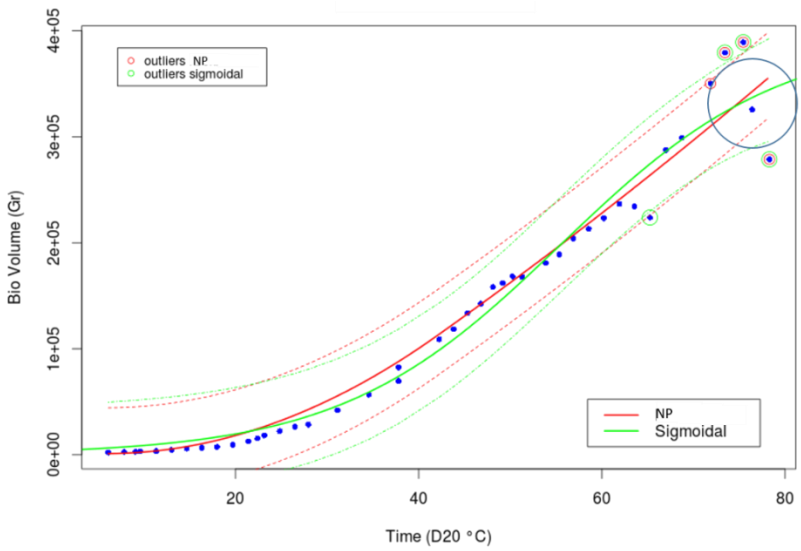

Figure 5. Detection of outlying points: The parametric (sigmoidal) and nonparametric (NP) approaches applied on time courses of Biovolume (experiment on maize performed in the PhenoArch platform, INRA). *Courtesy: Llorenç Cabrera-Bosquet*

Taking the spatial heterogeneity into account: Outlying points may also result from the spatial trends in the platform (due to spatial variability). In several installations, it is necessary to remove these effects (e.g. row/column, blocks, replicates) from the raw data. As the spatial trend of time 1 is not completely independent of time 0, an optimal solution is to simultaneously fit the spatial and temporal trends. However, so-called 3D methods are currently under development, so we will consider a two steps approach: 1/ Correcting for spatial trends at each time point, considering independent spatial patterns (when possible) 2/ Fitting a model as a function of time, using corrected data. The R-package SpATS makes it possible to carry out this approach successfully (Rodriguez-Alvarez *et al*, 2018).

3.2 Detection of outlier time courses

It consists in detecting outliers in a set of time courses resulting from the observation on one single trait. The detection procedure is applied to each trait individually without considering other traits, meaning that distinct outlier datasets are associated with each considered trait.

A nonparametric smoothing

Each time course is modelled by a nonparametric smoothing spline with a fixed number of knots. This is a piecewise cubic polynomial (Eubank, 1999, Eilers et al. 2015) fitted with the 'gam' function of the R-package 'mgcv'. The estimates for the spline coefficients are then extracted per time course (typically per plant) and correlations between those coefficient vectors are calculated to identify outlying time courses, i.e., plants. An outlying time course will have low correlation to the majority of time courses. To support the analysis by correlations, a principal component analysis can be done on the plant (time course) by spline coefficient matrix. A PCA plot of the plant scores will show the outlying plant. For example, when looking at time courses of plant replicates for a single genotype (Fig.6A), one can inspect correlation coefficients (Pearson's r, Fig.6B) and a PCA plot of plant time courses (Fig. 6C) to identify possible outliers.

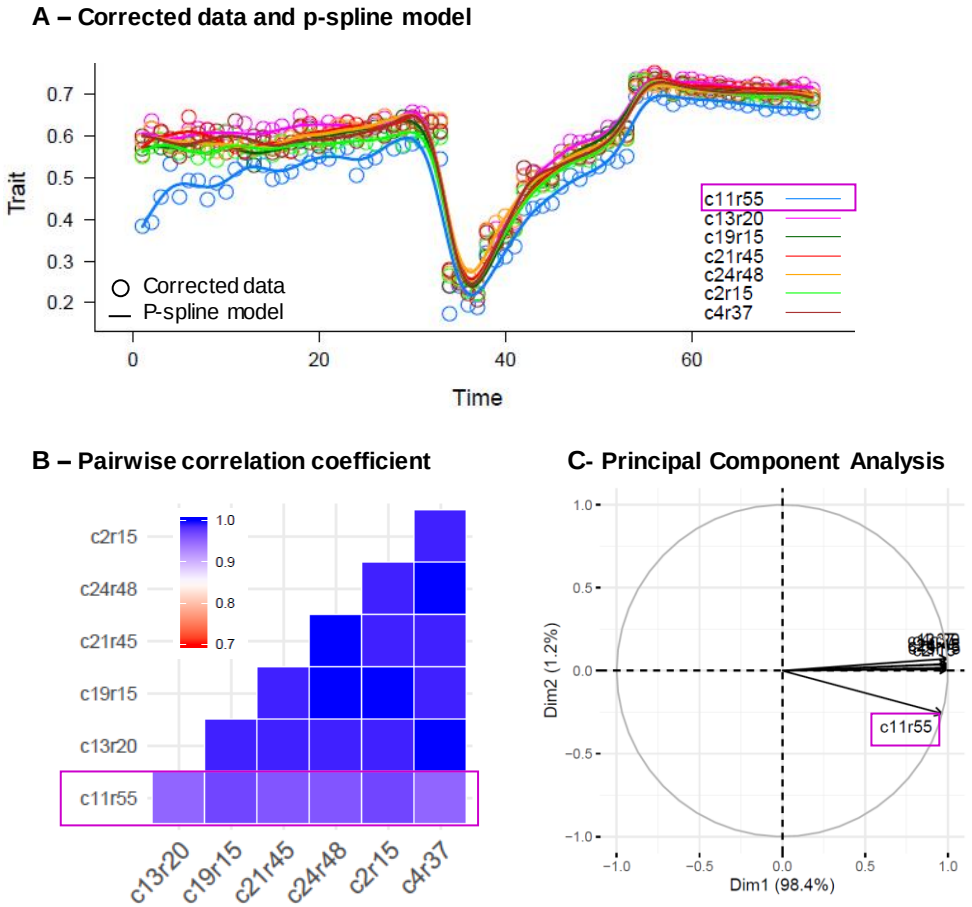

Figure 6. Annotation of an outlying time course for one variable and one genotype. A, Each time course of data corrected for spatial trends (using SpATS) is modelled using a p-spline with 50 knots. B, pairwise correlation of the p-spline coefficients. C, PCA of the p-spline coefficients. Data from the Phenovator platform (WUR, Netherlands). *Courtesy: Mark Aarts and René Boesten.*

In addition, a "functional" ANOVA decomposition (Gu, 2014) of the fitted splines can be done. The smoothing spline fitting and the functional ANOVA decompositions can be performed with the 'gss' R package. The outlier curves can be identified when a Kullback-Leibler distance surpasses a chosen threshold, see (Gu, 2014). This approach is easy to implement in R, and can be complemented by a first step of spatial correction at individual time points (using SpATS for example). The spline can thus be fitted either on the raw data or on the raw data corrected for spatial trends.

3.3 Detection of outlier plants

An outlier plant is defined as a biological replicate deviating from the overall distribution of plants on a multi-criteria basis, regardless of the quality of measurements. In a multi-trait approach, we consider several traits jointly, with rules set by experts depending on the species. We describe below the procedure we implemented to detect outlier plants in maize experiments conducted at the PHENOARCH phenotyping platform. The whole study is published in (Alvarez-Prado *et al.*, 2019).

Dataset: The dataset consisted of a diversity panel of 254 maize hybrids growing in three experiments (conducted in 2012 and 2013) with two irrigation treatments (WW: well-watered and WD: water deficit treatments). Each experiment involved 1680 plants in an image-based phenotyping platform located in a greenhouse. A randomized complete block design was used where each hybrid was replicated 3 times under both WW and WD treatments. Leaf area, total biomass and plant height were indirectly measured, among other variables, as presented in previous works (Alvarez Prado et al., 2018; for example).

Rules in the case of maize: We consider two categories of potentially outlier plants, namely apparently too small or too large plants. For the detection of unexpectedly small plants with likely physiological disorders, the progression of leaf stages was considered in addition to the time course of shoot biomass. Indeed, leaf appearance rate carries a non-redundant information compared with biomass (confirmed by standard correlation calculations). It usually presents a low plant-to-plant variability excepted in case of severe disorders, and is relatively insensitive to environmental cues other than temperature. Figure 7 illustrates the fact that a small plant, which would be difficult to classify as outlier based on biomass alone because of a continuous distribution of values, could be identified by combining the information on biomass with that on progression of leaf stages, for which one plant unambiguously differed from the others.

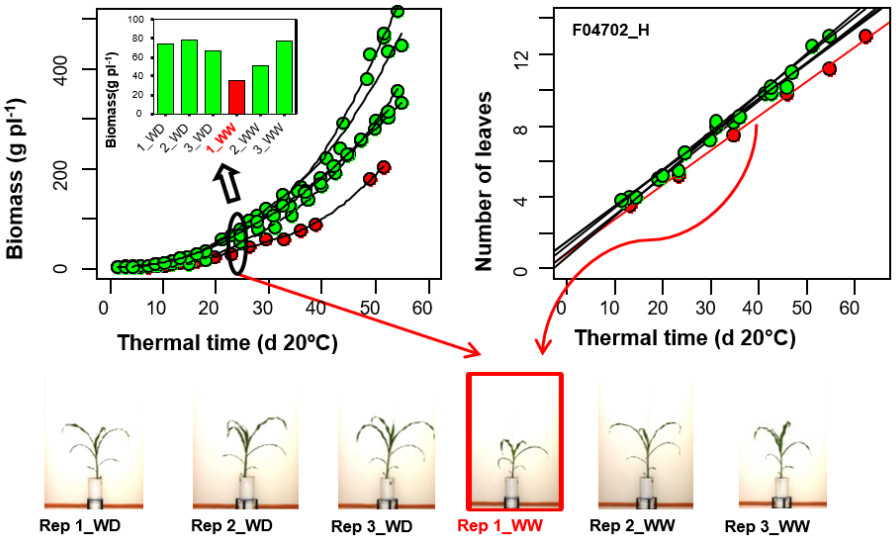

Figure 7. Example of multi-trait detection with expert rules ("small" maize plant) from Alvarez Prado et al. (2019).

For the detection of unexpectedly large plants, potentially associated with wrong genotype identification, combining plant height and biomass can result in an efficient identification.

Statistical modelling: Each of the selected traits was measured (or estimated) at a specific time (for example 24 $d_{20°C}$ for the Phenoarch dataset), time just before the beginning of the differentiation of the two watering treatments. This allows to have more replicates by genotype. It also reduces the dimensionality of the time courses to only one point, which will simplify the statistical models to be implemented later. As shown on Figure 8, traits are modelled with a mixed model that considers fixed experiment (ENV) effect and random genotypic (G), replicate (R) and spatial (C) effects. The model can be fitted with the SpATS R-package, (Rodriguez-Alvarez *et al*, 2018). Residuals (deviations) can be directly computed from the fitting, with a confidence interval. Plants, whose deviations for the criteria of leaf appearance rate and biomass are less than the lower bound of this interval, are considered as outlier small plants. Plants, whose deviations for the criteria of plant height and biomass are greater than the upper bound of the confidence interval, are considered as outlier large plants.

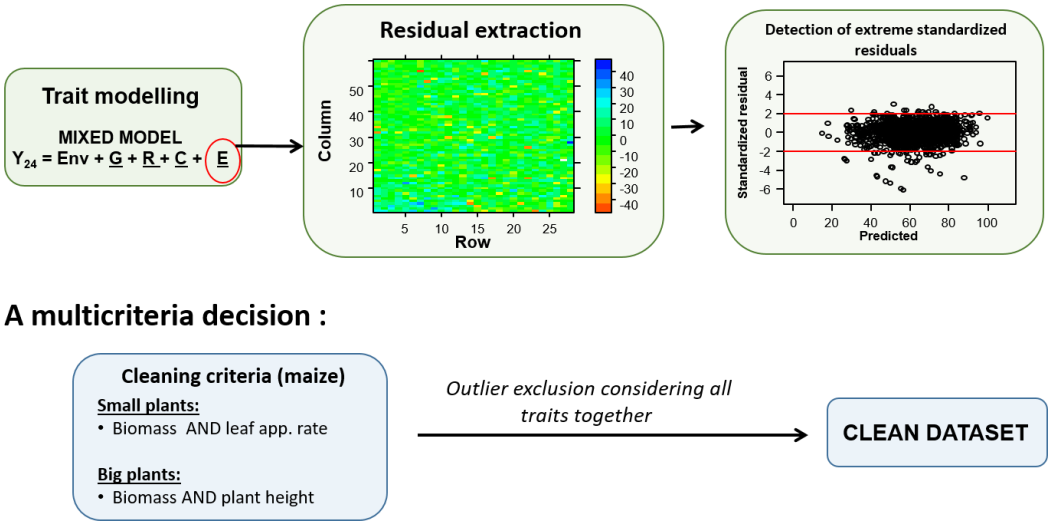

Figure 8. Schematic representation of a multi-trait approach for detection of outlier plants (case of maize), from Alvarez Prado et al. (2019).

4. CONCLUSION

Cleaning datasets coming from phenotyping platform is challenging. In this document, we have described some cleaning procedures that are suitable for platform data. The classification into three main types of outliers was proven relevant for the project partners and the procedures have been tested on various data provided by the EPPN2020 platforms. A tutorial will be developed, including codes to implement the methods with the R software. This tutorial will be tested in the first half of 2020.

To ensure data reuse, the annotated data should not be deleted in the information system but annotated as outliers with the rules of detection stored in the meta-data. Recent information systems for phenomic data allow these two conditions to be fulfilled, as the one promoted in EPPN2020-WP3 (Neveu et al., 2019). This topic has been the focus of a recently published paper (Alvarez Prado *et al*, 2019).

5. REFERENCES

Alvarez Prado S, Cabrera-Bosquet L, Grau A, Coupel-Ledru A, Millet EJ, Welcker C, Tardieu F. 2018. Phenomics allows identification of genomic regions affecting maize stomatal conductance with conditional effects of water deficit and evaporative demand. Plant, Cell & Environment 41, 314-326.

Alvarez Prado S, Sanchez I, Cabrera-Bosquet Ll, Grau A, Welcker C,Tardieu F and Hilgert N. 2019. Cleaning or not cleaning phenotypic datasets for outlier plants in genetic analyses? Journal of Experimental Botany, *to appear*.

Barnett V, Lewis T. 1994. Outliers in Statistical Data, John Wiley. New York.

Bernal-Vasquez AM, Utz HF, Piepho HP. 2016. Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. Theoretical and Applied Genetics **129**, 787-804

Eilers PH, Marx BD, Durbán M 2015. Twenty years of P-splines. SORT (Statistics and Operations Research Transactions). 39. 149-186.

Estaghvirou SBO, Ougutu JO, Piepho HP. 2014. Influence of Outliers on Accuracy Estimation in Genomic Prediction in Plant Breeding. G3: Genes|Genomes|Genetics **4**, 2317-2328.

Eubank, R. L. 1999. Nonparametric regression and spline smoothing. CRC press.

Grubbs FE. 1950. Sample Criteria for Testing Outlying Observations. 27-58.

Gu C. 2014. Smoothing spline anova models: R package gss. Journal of Statistical Software **58**, 1-25.

Hubert M, Rousseeuw PJ, Segaert P. 2015. Multivariate functional outlier detection. Statistical Methods & Applications **24**, 177-202.

Neveu P, Tireau A, Hilgert N, Nègre V, Mineau-Cesari J, Brichet N, Chapuis R, Sanchez I, Pommier C, Charnomordic B, Tardieu F, Cabrera-Bosquet L. 2019. Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. New Phytologist **221**, 588-601.

Rodriguez-Alvarez MX, Boer M, van Eeuwijk F, Eilers P. 2018. Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spatial Statistics **23**, 52 – 71. https://doi.org/10.1016/j.spasta.2017.10.003

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 160018.

Glossary

ANOVA – ANalysis Of VAriance
EPPN2020 - European Plant Phenotyping Network - 2020
QTLs - Quantitative Trait Loci
SpATS - Spatial Analysis of Trials using Splines
TNA – Trans-National Access
WP – Work Package