Project No. *283496*

transPLANT

Trans-national Infrastructure for Plant Genomic Science

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

D10.1
Statistical descriptors for genotype-phenotype map construction

Due date of deliverable: month 24
Actual submission date: month 24

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: IPG PAS

Project co-funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission	
RE	Restricted to a group specified by the consortium (including the Commission	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributor
IPG PAS (P. Krajewski, A. Markiewicz, H. Ćwiek)

Introduction

1 Introduction

The widely accepted systems biology approach to problems in plant science and the rapid development of fast and accurate high-throughput measurement techniques cause increase, both in size and complexity, of the sets of experimental results obtained in laboratory, greenhouse or field experiments. The observations of traditionally interesting phenotypic traits pertaining to yielding capacity or stress resistance of plants are now usually supplemented by the so-called phenotypic '-omics' traits, allowing to get insight into molecular, protein and metabolomic layers of plants. In addition to that, the number of data sets obtained by different research groups potentially interesting for building-up the knowledge about the plant organisms increases enormously. A proper integration of these data, both within and between experiments, is required to get new knowledge about plant systems. Such integration can not be achieved without tools that would be able to effectively store data and accompanying metadata, query appropriate databases, search for required information, and compare the obtained results. Such tools require building standards in the area of understanding of experimental designs, description of data complexity, data exchange formats and protocols, data compression, and should use generally acknowledged principles of statistical data processing; they must obey standard rules of metadata annotation and must be able to properly transfer these annotations to the computed results.

To achieve the goals described above we call upon the concept of 'sufficiency', widely used in mathematical statistics for theoretical considerations. To our knowledge, this concept has not been used so far in a coherent way to solve practical problems of experimental data management. We show that it can be applied to address several aspects, in particular data compression and data integration, and in consequence can be used for operations required for effective utilization of growing volumes of phenotypic data in systems biology.

Another application of the results produced by the presented approach is to use them as input to other computations. From our project's point of view, the most important case is related to building the phenotype-genotype maps by localization of the quantitative traits loci (QTL) or by genome-wide association studies (GWAS). For numerical optimization it is better to use in those procedures not the raw data, but properly computed parameters allowing e.g. to avoid lengthy computations when no phenotype-genotype relation is expected; sufficient statistics can play also this role.

The present report concerns description of the necessary theoretical background. The numerical implementation is described for the case of factorial experiments and linear mixed models only, and currently depends on functions that could be found in open-source R packages. Extensions of the methodology and numerical procedures to other experimental situations and models are under development, in particular in the area of repeated measurement experiments with application in image phenotyping. In this report we also present a first version of a web tool performing the computations, and describe possible applications of this service.

Methods
Literature and Internet studies Statistical modeling Data formatting Web service implementation

Results (if applicable, interactions with other workpackages)

2 Sufficiency

2.1 Definition, practical meaning, example

The concept of sufficiency introduced by R. A. Fisher allows to summarize data without any loss of information. Consider the problem of statistical inference about an unknown parameter θ based on a sample $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$. All the information about θ is of course contained in $\boldsymbol{y}$. However, we might wish to reduce data, especially when the sample size n is large. Data reduction can be expressed in terms of a particular statistic $T(\boldsymbol{y})$ that captures all information about θ from the sample, i.e., that is sufficient for θ. A sufficient statistic is formally defined as follows.

Definition 1. A real valued (or vector valued) statistic $T(\boldsymbol{y})$ is said to be sufficient for θ if the conditional distribution of the random sample $\boldsymbol{y}$, given $T = t$ does not depend on θ.

The practical meaning of the sufficiency is the following. If $T(\boldsymbol{y})$ is a sufficient statistic for θ, then any inference about θ should depend only on $T(\boldsymbol{y})$. Thus, in recording the experiment results it is sufficient to record T only, assuming model adequacy. As an example consider a random sample $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ from normal distribution $N(\theta, 1)$. The statistic $T = \sum_{i=1}^n y_i$ is sufficient for θ. In the case of normal distribution $N(\theta, \sigma^2)$ the statistic $T(\boldsymbol{y}) = (\sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2)^T$ is sufficient for the vector of parameters $(\theta, \sigma^2)^T$.

2.2 Sufficiency in linear models

If we restrict ourselves to linear models in which we estimate parameters by linear functions of observations, we can consider the property called 'linear sufficiency'.

Assume model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, where $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ and $Var(\boldsymbol{y}) = \boldsymbol{I}_n$. The best linear unbiased estimator (BLUE) of the expectation vector $\boldsymbol{X\beta}$ has a form

$$BLUE(\boldsymbol{X\beta}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y},$$

where $\boldsymbol{A}^-$ denotes a generalized inverse of the matrix $\boldsymbol{A}$. Baksalary and Kala (1981) defined a linear statistic preserving BLUE($\boldsymbol{X\beta}$), later on called linearly sufficient by Drygas (1983). It is defined as follows.

Definition 2. A linear statistic $\boldsymbol{Fy}$ is said to be linearly sufficient for $\boldsymbol{X\beta}$ if there exists a matrix $\boldsymbol{T}$ such that $\boldsymbol{TFy}$ is the BLUE of $\boldsymbol{X\beta}$.

An example of a quadratically sufficient statistic for $\boldsymbol{X\beta}$ is $\boldsymbol{Fy} = \boldsymbol{X}^T\boldsymbol{y}$. Drygas (1983) observed that under normality assumption, $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{I}_n)$, a linearly sufficient statistic is also sufficient in usual sense.

Mueller (1987) extended the concept of linear sufficiency to quadratic sufficiency in the context of linear model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, where $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ and $Var(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$

with $\sigma^2 > 0$ unknown. A statistic $(\boldsymbol{Ly}, \boldsymbol{y}^T\boldsymbol{Ty})$ is called quadratically sufficient if $\boldsymbol{Ly}$ is linearly sufficient and there exists a symmetric matrix $\boldsymbol{\Lambda}$ and a real α such that $\boldsymbol{y}^T\boldsymbol{L}^T\boldsymbol{\Lambda}\boldsymbol{Ly} + \alpha\boldsymbol{y}^T\boldsymbol{Ty}$ is the best quadratic unbiased estimator (BQUE) of σ^2.

An example of a linearly sufficient statistic for $(\boldsymbol{X\beta}, \sigma^2)^T$ in a model with $\boldsymbol{V} = \boldsymbol{I}_n$ is $\boldsymbol{Fy} = (\boldsymbol{X}^T\boldsymbol{y}, \boldsymbol{y}^T\boldsymbol{My})^T$, where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$. Under normality assumption a quadratically sufficient statistic is also sufficient in usual sense.

Let $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$ be a mixed linear model, where $\boldsymbol{u} = (\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_s^T)^T$ is a vector of uncorrelated random effects and $E(\boldsymbol{y}) = \boldsymbol{X\beta}$, $Var(\boldsymbol{y}) = \boldsymbol{V} = \sum_{i=1}^{s} \sigma_i^2 \boldsymbol{Z}_i\boldsymbol{Z}_i^T + \sigma_0^2\boldsymbol{I}_n$ with $\sigma_0^2 > 0, \sigma_i^2 \geq 0, i = 1, \ldots, s$, unknown. Under normality assumption $\boldsymbol{u} \sim N(0, \sigma_1^2\boldsymbol{I}_r)$, $\boldsymbol{e} \sim N(0, \sigma_0^2\boldsymbol{I}_n)$, an example of sufficient statistic for $(\boldsymbol{X\beta}, \sigma_0^2, \sigma_1^2)^T$ is $\boldsymbol{Fy} = [(\boldsymbol{X} : \boldsymbol{Z})^T\boldsymbol{y}, \boldsymbol{y}^T\boldsymbol{My}]^T$ with $\boldsymbol{M} = \boldsymbol{I}_n - (\boldsymbol{X} : \boldsymbol{Z})[(\boldsymbol{X} : \boldsymbol{Z})^T(\boldsymbol{X} : \boldsymbol{Z})]^-(\boldsymbol{X} : \boldsymbol{Z})^T$. For $s > 1$ the above statistic is sufficient for $(\boldsymbol{X\beta}, \sigma_0^2, \ldots, \sigma_s^2)^T$ under some balancedness conditions. In general this statistic is a basis of so-called Henderson mixed model equations leading to commonly used estimators and predictors in mixed linear model; cf. McLead et al. (1991) and Witkovský (2012).

3 The method for factorial experiments

3.1 Fixed effects model

Assume that an experiment is of the factorial type, that is, observations (samples) are classified by a number of factors A, B, C, ... , with replications forming a completely randomized design. The data is provided as a properly formatted set $\{\boldsymbol{y}, \boldsymbol{T}\}$ consisting of metadata concerning the treatment structure (in the minimal case, the columns of factor levels) - forming a text matrix $\boldsymbol{T}$, and data $\boldsymbol{y}$, a vector of observations of a trait (in practice we have many traits $\boldsymbol{y}$, so observations also form a matrix, but we consider here one trait for simplicity of the description). In the statistical model we want to consider effects of factors A, B, C, ..., and of their interactions AB, AC, ABC, ... It is possible that the experiment is not balanced (different number of replications in subclasses).

The model usually used for this situation is of the form $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with $\boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I})$, where

$$\boldsymbol{X} = [\boldsymbol{1} : \boldsymbol{X}_A : \boldsymbol{X}_B : \boldsymbol{X}_C : \boldsymbol{X}_{AB} : \ldots]$$

with each sub-matrix constructed 'in the usual way' of 0-1 indicator columns describing the allocation of samples to factor levels and to their combinations, constructed from $\boldsymbol{T}$. $\boldsymbol{X}$ is not of full rank, but each submatrix for a given factor or combination of factors is. The easiest way to estimate $\boldsymbol{X\beta}$ would be to use the previously introduced formula for BLUE$(\boldsymbol{X\beta})$ with any choice of the generalized inverse. However, as we want to extend the procedure to a mixed model later, our method is as follows.

We will follow the procedure applied in many statistical packages (Genstat, R) that consists of assuming a set of linear restrictions on model parameters in such a way that the remaining (unrestricted) parameters can be estimated. The numerical procedure in R assumes the value of zero for some parameters (technically, it removes some columns of matrix $\boldsymbol{X}$). In our notation, it corresponds to imposing on the parameters vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_R^T, \boldsymbol{\beta}_0^T)^T$ a linear restriction $\boldsymbol{\beta}_0 = \mathbf{0}$ and to solve the resulting system of normal equations $\boldsymbol{X}_R^T \boldsymbol{X}_R^T = \boldsymbol{X}_R^T \boldsymbol{y}$, with $\boldsymbol{X}_R$ of full column rank (equal to the rank of $\boldsymbol{X}$). Observing that the matrix

$$\begin{pmatrix} (\boldsymbol{X}_R^T \boldsymbol{X}_R)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

is a generalized inverse of the matrix

$$\boldsymbol{X}^T \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_R^T \boldsymbol{X}_R & \boldsymbol{X}_R^T \boldsymbol{X}_0 \\ \boldsymbol{X}_0^T \boldsymbol{X}_R & \boldsymbol{X}_0^T \boldsymbol{X}_0 \end{pmatrix},$$

where $\boldsymbol{X} = (\boldsymbol{X}_R : \boldsymbol{X}_0)$, we get that the vector $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_R^T, \mathbf{0}^T)^T$ is a solution of the original system of normal equations $\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{y}$. It is not a unique solution, yet for a given estimable function $\boldsymbol{p}^T \boldsymbol{\beta}$ the statistic $\boldsymbol{p}^T \hat{\boldsymbol{\beta}}$ is its unique best linear unbiased estimator. A vector $\boldsymbol{p}$ defining an estimable function $\boldsymbol{p}^T \boldsymbol{\beta}$ can be represented as $\boldsymbol{p} = \boldsymbol{X}^T \boldsymbol{l}$, for some vector $\boldsymbol{l}$; i.e., it is a column or a linear combination of columns of the matrix $\boldsymbol{X}^T$. To estimate $Var(\boldsymbol{p}^T \hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{p}^T (\boldsymbol{X}^T \boldsymbol{X})^- \boldsymbol{p}$ we can use the matrix $(\boldsymbol{X}_R^T \boldsymbol{X}_R)^{-1}$ and the usual estimate of σ^2. Note that for estimable function $\boldsymbol{p}^T \boldsymbol{\beta}$ the expression $\boldsymbol{p}^T (\boldsymbol{X}^T \boldsymbol{X})^- \boldsymbol{p}$ is invariant with respect to the choice of a generalized inverse of $\boldsymbol{X}^T \boldsymbol{X}$.

In particular, the BLUE$(\boldsymbol{X}\boldsymbol{\beta}) = \hat{\boldsymbol{\mu}} = \boldsymbol{X}_R \hat{\boldsymbol{\beta}}_R$ gives the estimated means (expected values) for all experimental combinations. The variance (and so the standard error) of this estimator can be also obtained by the general formula given above. Subsequently, for any $\boldsymbol{X}_F$ being a sub-matrix of $\boldsymbol{X}$ corresponding to a factor or combination of factors, the formula $(\boldsymbol{X}_F^T \boldsymbol{X}_F)^{-1} \boldsymbol{X}_F^T \hat{\boldsymbol{\mu}}$ provides the estimated (marginal) means for levels of individual factors or their combinations; the standard error of this estimator is obtained as above.

3.2 Mixed model

In a more general situation we consider a factorial experiment performed in an experimental design appropriate for the set of applied units (pots, plots, fields). In this case the data set is of the form $\{\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{B}\}$, with $\boldsymbol{B}$ denoting the matrix of meta-data describing the block structure of the experiment. In the experiments conducted in plant science the examples of designs which can be included here are: (incomplete) block designs, row-column designs, and latin square designs. In this case the commonly used model is the mixed model with the treatment structure $\boldsymbol{T}$, defining its fixed part, and the block structure $\boldsymbol{B}$ – the random part.

First we note that the procedure described previously for the fixed model can be used in the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{V})$, where $\boldsymbol{V}$ is known and nonsingular. We transform this model to $\tilde{\boldsymbol{y}} = \boldsymbol{V}^{-1/2}\boldsymbol{y}$ and then we use the same method taking $\tilde{\boldsymbol{X}}_R = \boldsymbol{V}^{-1/2}\boldsymbol{X}_R$ and $\tilde{\boldsymbol{X}}_0 = \boldsymbol{V}^{-1/2}\boldsymbol{X}_0$.

Then, let $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$ be a mixed linear model, where $\boldsymbol{u} = (\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_s^T)^T$ is a vector of uncorrelated random effects and $E(\boldsymbol{y}) = \boldsymbol{X\beta}$, $Var(\boldsymbol{y}) = \boldsymbol{V} = \sum_{i=1}^{s} \sigma_i^2 \boldsymbol{Z}_i \boldsymbol{Z}_i^T + \sigma_0^2 \boldsymbol{I}_n$ with $\sigma_0^2 > 0, \sigma_i^2 \geq 0$, $i = 1, \ldots, s$, unknown. The estimation procedure is as follows. First, variance components $\sigma_i, i = 0, \ldots, s$ are estimated by the residual maximum likelihood (REML) method; cf. McLean et al. (1991) and Wikovský (2012). Then linear estimable functions of $\boldsymbol{\beta}$ are estimated by empirical BLUE; i.e., using the approach of fixed model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with $Var(\boldsymbol{y})$ replaced by $\hat{\boldsymbol{V}} = \sum_{i=1}^{s} \hat{\sigma}_i^2 \boldsymbol{Z}_i \boldsymbol{Z}_i^T + \hat{\sigma}_0^2 \boldsymbol{I}_n$, where $\hat{\sigma}_i^2, i = 0, \ldots, s$ are REML estimates of $\sigma_i^2, i = 0, \ldots, s$.

3.3 Dimensionality of sufficient statistics and model reduction

The sufficient statistics pertaining to the fixed part of the model, $\boldsymbol{X}^T\boldsymbol{y}$, is easily calculated, but its dimension is large, equal to the number of columns of $\boldsymbol{X}$. As we see from the derivation a vector of smaller dimension, $\hat{\boldsymbol{\beta}}_R^T$, of independent parameters, is sufficient to reconstruct the estimators of expectations. This observation is useful when we consider application of the procedure for data compression.

Also, we should note that the dimension of the vector of sufficient statistics depends on the model – namely, on the number of considered interactions. The model can be reduced to a version with an acceptable fit to the data by ANOVA-based (or other) procedures of model selection by removing interactions which are not significant. We plan to use this possibility in numerical implementation of the procedure.

4 Numerical implementation

We implement the estimation and testing procedure in R environment using the *lme4* library and other necessary functions. For a given data set, REML computations are performed, and the results comprising:

- vector $\hat{\boldsymbol{\beta}}_R^T$, with its covariance matrix,

- vector $\boldsymbol{S}$ of variance components estimates, including the error variance,

are saved to proper structures. We consider the following versions of the reporting scenario for the input data set $\{\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{B}\}$:

a) A data set $\left\{\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{B}, \hat{\boldsymbol{\beta}}_R^T, \boldsymbol{S}\right\}$ is returned. This data set can be used for compressed storage of the data. It cannot be reported to the user, as the interpretation of the

estimates depends on the estimation procedure, and the inverse procedure based on the matrix $\boldsymbol{T}$ has to be applied for production of interpretable results.

b) A data set $\{\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{B}, \hat{\boldsymbol{\mu}}, \boldsymbol{S}\}$ is returned. This data set is a useful version of the results, but marginal means must be formed to get interpretable estimates.

b) A data set $\{\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{B}, \hat{\boldsymbol{\mu}}_A, \hat{\boldsymbol{\mu}}_B, \hat{\boldsymbol{\mu}}_{AB}, ..., \boldsymbol{S}\}$ is returned. This data set is the final version of the results which could be used for all aims that were indicated in the introduction.

NOTE: the algorithm for computation of the variances of variance components' estimates is under development.

5 Phenalyse tool

Phenalyse is a web tool calculating sufficient statistics for ISA-TAB formatted dataset. The program performs statistical analysis of data coming from phenotyping experiments. Data must be provided in zipped ISA-TAB (ISArchive) format, and should contain well annotated phenotyping assays. The application provides an interface (Fig.1) to upload a zipped dataset, run the analysis and download the results of the processing as

- a text file containing the computed sufficient statistics,

- updated ISArchive, including the newly created sufficient statistics file and references to it (dedicated column in the assay file),

- modified ISArchive, containing only the newly created statistics, with all the intermediate data and its reference removed from the set.

Statistical computations are performed in R environment. For each phenotyping assay-study pair from the uploaded dataset a separate statistical analysis is performed. For each observed trait a mixed linear model is constructed and evaluated. Based on ISA-TAB file annotations specific parts of the model are defined. All variable characteristics of a dataset are assumed to belong to the fixed part of the model. Random parameters are those describing the design of the experiment (e.g. plot, block, row, column, field, rank, replication); the remaining factors are also assigned to the fixed part. Parameters of the model are estimated using REML procedure from *lme4* library, based on which mean values of traits and their variances are computed. Results for each assay-study pair are saved in the 'sufficient statistics file', whose format is specified in ISA-TAB phenotyping configuration. After successful processing the assay file is updated to include references to statistics file in Sufficient Statistics File column, and an archive containing enriched dataset is constructed and set for download.

Phenalyse is an open source Java application using R environment for statistical computations, Spring framework and JPA2/Hibernate ORM layer for MySQL database; it runs on Jetty/Tomcat server.

6 Applications

Possible application of the Phenalyse tool can be as follows (to be developed in the project):

1) Production of a compressed phenotypic dataset in ISA-TAB format: the input ISArchive is transformed to an output containing the sufficient data file, but not the raw and derived data files; as an option, the set of sufficient statistics can be reduced to contain only the significant interactions.

2) Construction of compressed datasets for storage in a database: data sets with sufficient, or sufficient of minimal dimension, statistics are stored in a database instead of the raw data. In the latter case, the database must be equipped with an algorithm converting the minimal statistics to interpretable parameter estimates.

3) Data queries and comparisons: ISArchives with parameter estimates can be queried for meta-data or parameter values; different datasets can be compared.

4) Supplementary data publication: ISArchives with parameter estimates can be used as information supplementing published papers.

5) Data integration: data sets from different study/assays can be integrated at the level of samples or model parameters.

6) Production of input for another procedures: phenotypic ISArchive is processed to obtain the dataset for another analysis, e.g. a QTL/GWAS application. Within Phenalyse the model is reduced to the one with just significant interactions so that unnecessary computations to find non-existing QTL by environment interactions are avoided.

7 Example

The exemplary data is a set concerning 5 varieties of barley treated with drought in a block design. The design is taken from a real experiment, but observations of the traits are fictitious. The data set and results are presented in Fig. 2-7. Two situations are presented:

- in Fig. 6, the ISA-TAB formatted data set with parameter estimates as sufficient statistics in the model containing the interaction of two factors,

- in Fig. 7, the ISA-TAB formatted data set with sufficient statistics of minimal dimension for the model with main factor effects only.

Due to the property of sufficiency, the whole information concerning all the factor levels and their combinations under the model of no interaction can by recovered from the data shown in Fig. 7 using the methodology presented in this report. The difference in number of data corresponds to the gain of storage size (compression) obtained by application of the sufficiency principle.

8 Other models

The theory presented in Section 2 is a general one and can be applied to other situations arising in plant experiments. For example, sufficient statistics can be computed for regression models. In this case, the model matrix is usually of full column rank, so that computation of the sufficient statistics with minimal dimension is equivalent to usual parameter estimation. The methods reported here are applicable if the regression model is of a mixed type and includes fixed and random parameters pertaining to factors, and regression parameters pertaining to explanatory variables. Numerical implementation for this situation is under development.

The theory can also be used for the case of experiments with repeated measurements in time. For such cases, several models are described in the literature. For phenotyping the most interesting situation of this type is the 'image phenotyping' procedure carried out in modern installations. The experiments are of factorial type. Usually no blocking is involved because it can be assumed that the environmental conditions (apart from the differences resulting from applied treatments) are the same for all samples (pots). However, the measurements are done over time, usually once per day for a period required to record the plants' reaction. Possible modelling approaches for this data are:

a) mixed models with correlated errors,

b) models with (optional) factorial structure and time effects modelled as a regression function of a number of parameters, e.g. EMAX model,

c) models with (optional) factorial structure and time effects modelled by functional data analysis approaches (Ramsay and Silverman, 2005).

Numerical implementations for these situations are being developed in collaboration with project partners.

9 Conclusions

We have used the theory of sufficiency to set up the basis for standardised processing and management procedures for phenotypic data obtained in plant experiments. A basic numerical implementation for the case of linear mixed models have been described. A model of a web tool performing fundamental operations has been constructed. As planned in the project, the methodology is under development directed towards inclusion of other experimental situations and optimization for big data sets.

10 References

Baksalary, J.K. and Kala, R. (1981). Linear transformations preserving best linear unbiased estimators in a general Gauss–Markoff model. *The Annals of*

Statistics, 4, 913–916.

Drygas, H. (1983). Sufficiency and completeness in the general Gauss–Markov model. *Sankhyā, Series A*, 45, 88–98.

McLean, R.A., W. L. Sanders, and W. W. Stroup (1991). A unified approach to mixed linear models. *The American Statistician*, 45, 54–64.

Mueller, J. (1987. Sufficiency and completeness in the linear model. *Journal of Multivariate Analysis*, 21, 312–323.

Witkovský, V. (2012). Estimation, Testing, and Prediction Regions of the Fixed and Random Effects by Solving the Henderson's Mixed Model Equations. *Measurement Science Review*, 12, 234–248.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional data analysis.* 2nd ed., New York : Springer,

Interactions with other packages:
With WP3, on standards for data formatting

Publications
In preparation

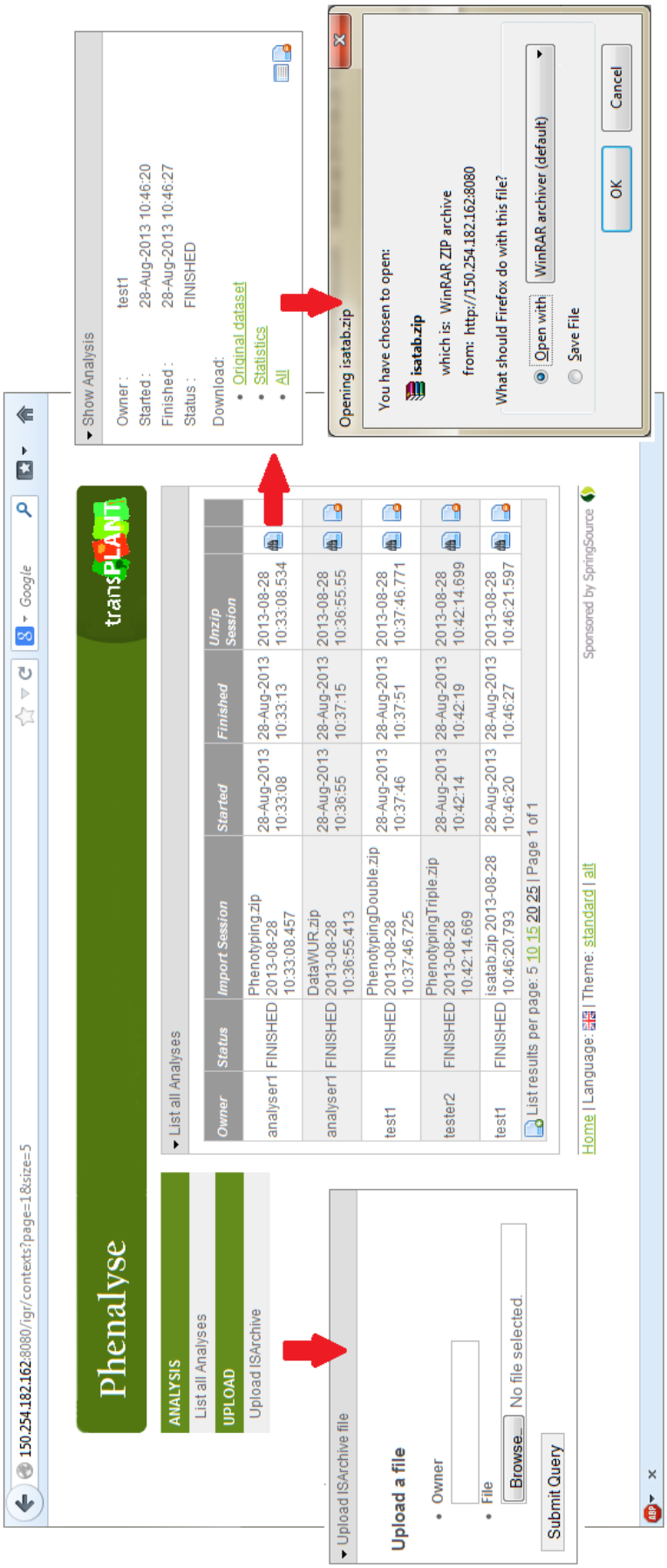

Fig.1 Phenalyse application

Source Name	Characteristics[Organism]	Term Source REF	Term Accession Number	Characteristics [Infra-specific name]	Term Source REF	Term Accession	Characteristics[Organism part]	Term Source REF	Term Accession Number	Protocol REF	Protocol REF	Factor Value[Treatment]	Term Source REF	Term Accession
source1	Hordeum vul	NCBITaxon	xon_112509	Sebastian	EURISCO		stem	PO	9047	drought ap	sample co	Control		
source2	Hordeum vul	NCBITaxon	xon_112509	Sebastian	EURISCO		stem	PO	9047	drought ap	sample co	Drought		
source3	Hordeum vul	NCBITaxon	xon_112509	Amarena	EURISCO		stem	PO	9047	drought ap	sample co	Control		
source4	Hordeum vul	NCBITaxon	xon_112509	Amarena	EURISCO		stem	PO	9047	drought ap	sample co	Drought		
source5	Hordeum vul	NCBITaxon	xon_112509	Nagradowicki	EURISCO		stem	PO	9047	drought ap	sample co	Control		
source6	Hordeum vul	NCBITaxon	xon_112509	Nagradowicki	EURISCO		stem	PO	9047	drought ap	sample co	Drought		
source7	Hordeum vul	NCBITaxon	xon_112509	HOR 198	EURISCO		stem	PO	9047	drought ap	sample co	Control		
source8	Hordeum vul	NCBITaxon	xon_112509	HOR 198	EURISCO		stem	PO	9047	drought ap	sample co	Drought		
source9	Hordeum vul	NCBITaxon	xon_112509	Basza	EURISCO		stem	PO	9047	drought ap	sample co	Control		
source10	Hordeum vul	NCBITaxon	xon_112509	Basza	EURISCO		stem	PO	9047	drought ap	sample co	Drought		

Fig. 2. The ISA-TAB study file for example data

Source Name	Sample Name	Factor Value[Block]	Term Source REF	Term Accession Number	Raw Data File	Protocol REF	Derived Data File	Trait Definition File
source1	sample1	1				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source1	sample2	2				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source1	sample3	3				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source1	sample4	4				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source1	sample5	5				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source2	sample6	1				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source2	sample7	2				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source2	sample10	5				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source3	sample11	1				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source3	sample12	2				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source3	sample13	3				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source3	sample14	4				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source3	sample15	5				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source4	sample16	1				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source4	sample17	2				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source4	sample18	3				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source4	sample19	4				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source4	sample20	5				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source5	sample21	1				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source5	sample22	2				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source5	sample23	3				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source5	sample24	4				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source5	sample25	5				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source6	sample26	1				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source6	sample27	2				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source6	sample28	3				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source6	sample29	4				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx
source6	sample30	5				data transfo	a_study1_processed_data.xlsx	a_study1_tdf.xlsx

Fig. 3. The ISA-TAB assay file for example data

Sample Name	Trait Value[len]	Trait Value[Co	Term Source	Term Accessio	Trait Value[Stem
sample1	0,7221	green	PATO	320	1,08
sample2	0,4288	yellow	PATO	324	1,15
sample3	0,5585	green	PATO	320	1,46
sample4	0,1394	yellow	PATO	324	1,23
sample5	0,5313	yellow	PATO	324	1,23
sample6	0,8398	green	PATO	320	1,31
sample7	0,8084	green	PATO	320	1,62
sample10	0,1138	green	PATO	320	1,46
sample11	0,9038	green	PATO	320	1,00
sample12	0,1341	yellow	PATO	324	1,00
sample13	0,2490	yellow	PATO	324	1,69

Fig. 4. The ISA-TAB processed data file for example data

Trait Name	**Trait Source REF**	**Trait Term Accession Number**	**Method Name**	**Method Source REF**	**Method Term Accession Number**	**Scale Name**	**Scale Source REF**	**Scale Term Accession Number**
len	TO	576	Stem length measuring method	BO	12	cm	UO	15
Colour	PATO	14	Color assessed visually by 2 specialists			colour scale	PATO	14
Stem diameter			Diameter measured in the middle			mm	UO	16

Fig. 5. The ISA-TAB trait definition file for example data

Parameter	Characteristics[Infra-specific name]	Factor Value[Treatment]	Factor Value[Block]	Estimate[len]	Standard Error[len]
Mean	Amarena	Control		0.3605	0.1875
Mean	Amarena	Drought		1.5759	0.1875
Mean	Basza	Control		0.4106	0.1875
Mean	Basza	Drought		0.9345	0.1875
Mean	HOR 198	Control		0.5525	0.1875
Mean	HOR 198	Drought		0.9441	0.1875
Mean	Nagradowicki	Control		0.6379	0.1875
Mean	Nagradowicki	Drought		1.3359	0.1875
Mean	Sebastian	Control		0.4760	0.1875
Mean	Sebastian	Drought		0.5609	0.2415
Mean		Control		0.4875	0.0906
Mean		Drought		1.0703	0.0956
Mean	Amarena			0.9682	0.1353
Mean	Basza			0.6725	0.1353
Mean	HOR 198			0.7483	0.1353
Mean	Nagradowicki			0.9869	0.1353
Mean	Sebastian			0.5185	0.1553
Mean				0.7789	0.0712
Variance			*	0.00733	0.0184
Error Variance				0.16849	0.0409

Fig. 6. The ISA-TAB sufficient data file for example data

Parameter	Characteristics[Infra-specific name]	Factor Value[Treatment]	Factor Value[Block]	Estimate[len]	Standard Error[len]
Mean	Amarena	Control			
Mean	Amarena	Drought			
Mean	Basza	Control			
Mean	Basza	Drought			
Mean	HOR 198	Control			
Mean	HOR 198	Drought			
Mean	Nagradowicki	Control			
Mean	Nagradowicki	Drought			
Mean	Sebastian	Control			
Mean	Sebastian	Drought			
Mean		Control			
Mean		Drought		0.6131	
Mean	Amarena				
Mean	Basza			-0.2957	
Mean	HOR 198			-0.2199	
Mean	Nagradowicki			0.0187	
Mean	Sebastian			-0.3738	
Mean				0.6616	
Variance			*	0.0000	0.0153
Error Variance				0.196	0.0449

Fig. 7. The ISA-TAB sufficient data file with minimal dimension for no-interaction model for example data