Making more of primary traits by incorporating secondary traits

6th Annual - Nordic Plant Phenotyping Network Workshop

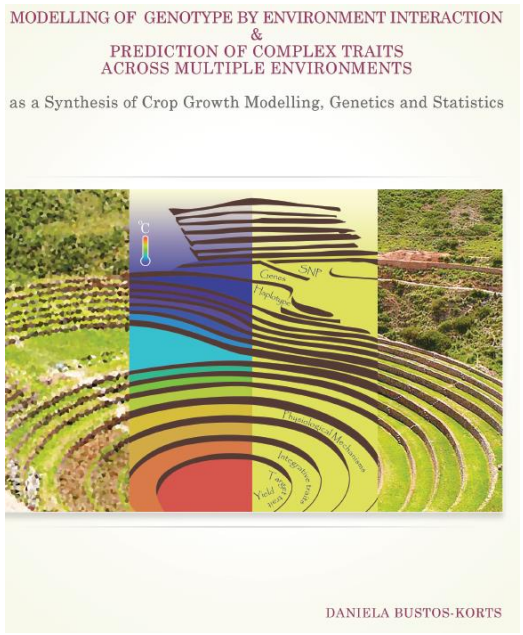
Fred van Eeuwijk

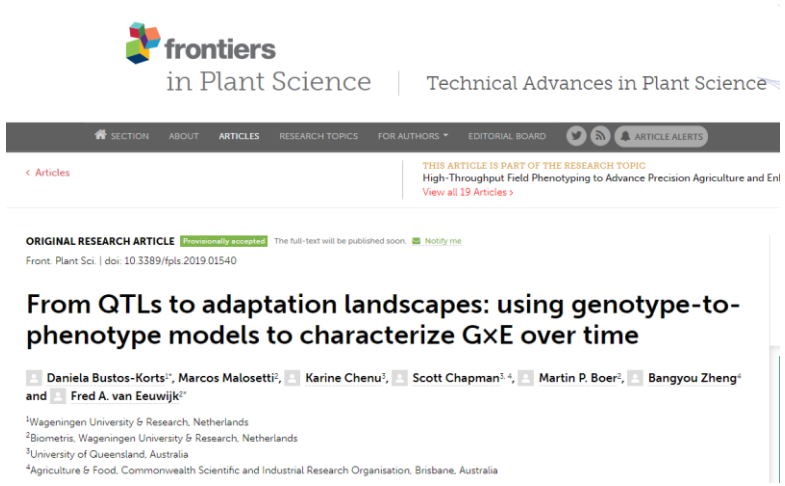
27 November, 2020

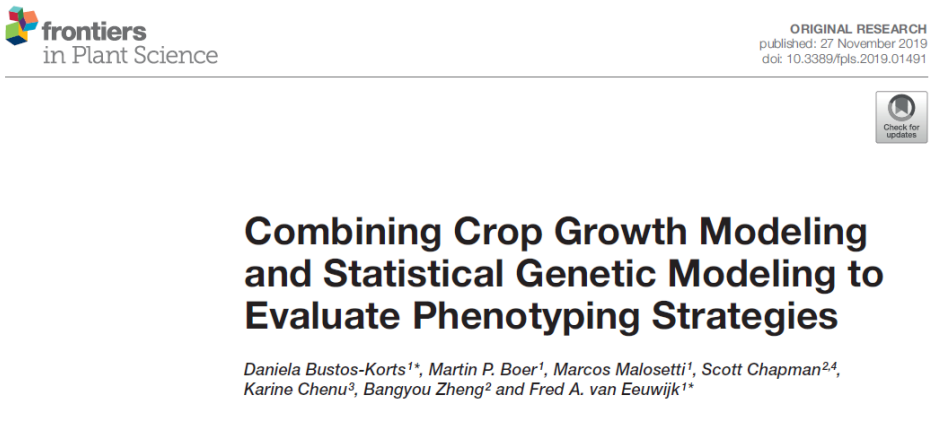
WAGENINGEN UR
For quality of life

Studying multi-trait dynamics by APSIM with genotype-specific parameters (Daniela Bustos Korts)

Role of secondary phenotyping & genotype specific crop growth simulations in environmental characterization







WAGENINGEN UNIVERSITY & RESEARCH

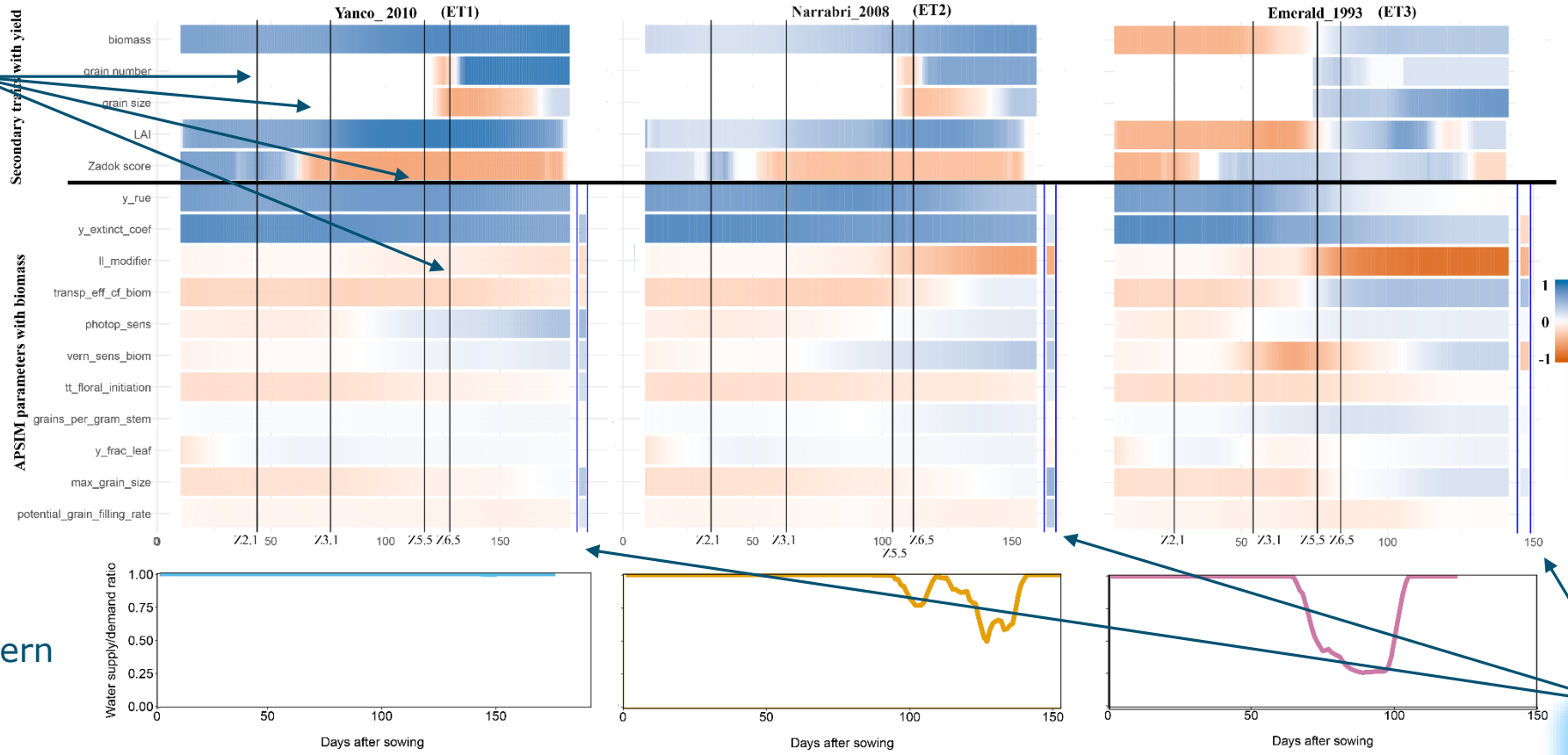
100years 1918 — 2018

Correlations between target (yield), intermediate and physiological parameters **over time** (Australian wheat)

Delimitation of developmental stages

trait x trait x time x environment interactions

Drought stress pattern



Correlations of **yield** with *intermediate traits*

Correlations of **biomass** with *physiological parameters*

Correlation of physiological parameters with **yield**

Note environment specificity of correlation patterns

Correlation patterns between yield and yield components over time define environmental types

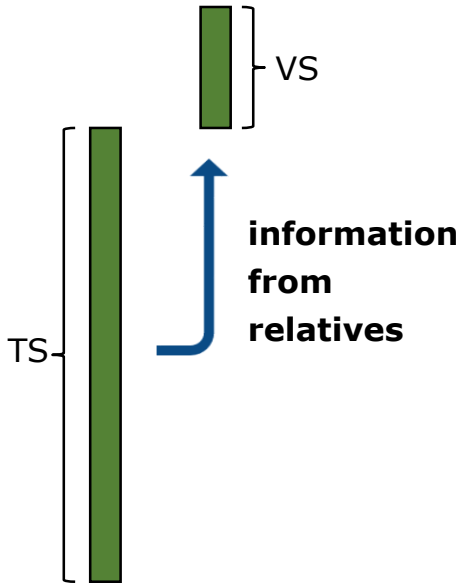
WAGENINGEN
UNIVERSITY & RESEARCH

Problem

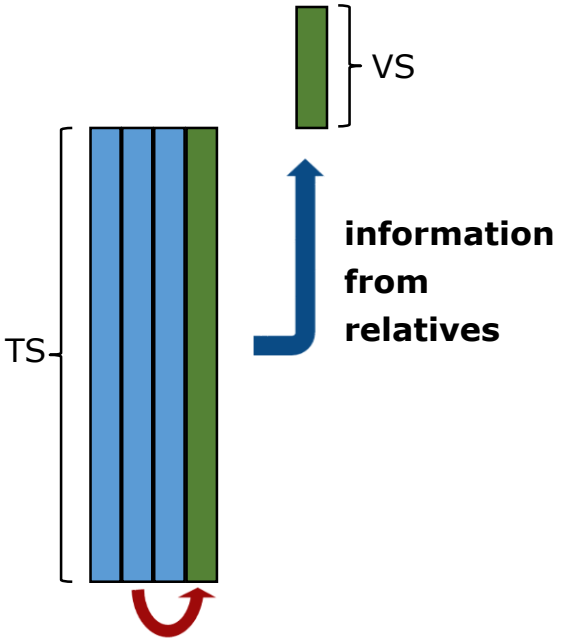
- Primary traits, like yield, can be predicted to a certain extent from molecular markers
- New phenotyping tools and devices produce large amounts of additional data on secondary phenotypes (spectra, root growth, biomass assessments)
- Can these secondary phenotyping data be useful to predict primary traits or made useful to better understand the genetic basis of primary traits (pleiotropic QTLs)?
    - Suggestion
        - Convert the secondary phenotyping data into genotypic covariables / responses and use them for prediction of primary traits / QTL mapping
    - Problems (secondary traits)
        - often come as time series
        - occur as discrete realizations of underlying continuous processes (wavelengths)
        - genetic signal requires adjustment as well as smoothing in multiple dimensions (spatial, temporal, wavelength)
        - no standard or default analysis methods available

Prediction scenarios

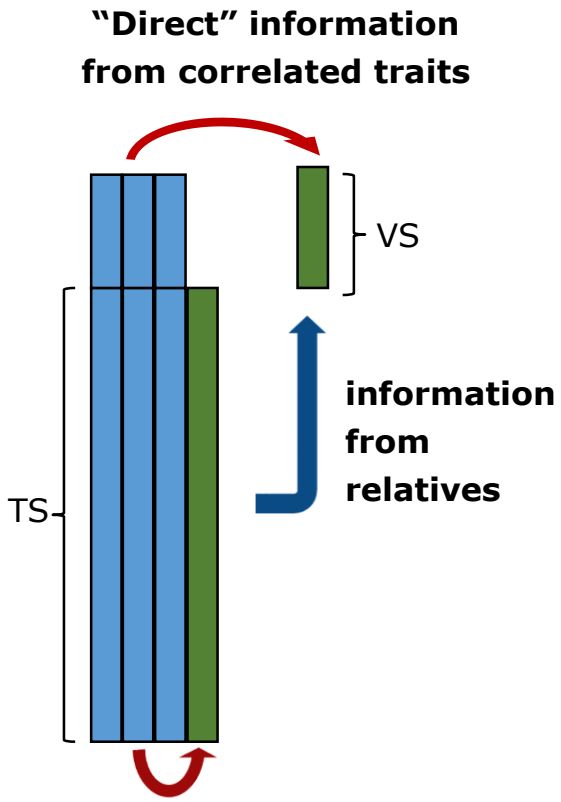
Single Trait



Multi Trait only in Training Set



Multi Trait in Training and Validation Set



Design & analysis of phenotyping experiments

Fred van Eeuwijk & Emilie Millet

Introduction JRA2: Design & analysis

- Develops tools for **statistical design and analysis of phenomics experiments** across platforms and scales of plant organization
- **Standardize statistical analyses** and **facilitate the combined analysis of data** coming from multiple installations and measurement scales (including field)
  - (i) outlier detection and data quality control
  - (ii) statistical design and analysis of single platform experiments
  - (iii) data integration between platforms and platform/field.
- Start EPPN^{2020} clear demand for a unified set of tools and methods to analyze platform data. Diversity phenotyping techniques and increasing amount of data prevented direct application of designs, models and analysis methods originally developed for field trials
- Main objectives (strategic):
  - Level 1
    - Describe installations in a statistically intelligible way and give a motivation for design and analysis choices for the current state
  - Level 2
    - Description of the properties of the platform
    - Documented routine use of automated statistical methods and tools
    - Traceability and reversibility of decisions

JRA2: Statistical design & analysis - Tasks

TJRA2.1 Outliers & data quality control

- Flagging different types of outliers
  - Point, time series, set of time series (plant)

TJRA2.2 Design & analysis of single platform experiments

- Row-column blocking to neutralize spatial variation
- Choosing standard genotypes and allocating them
- Spatial adjustment of phenotypic traits at single and multiple time points
- Simultaneous analysis of multiple growth curves

TJRA2.3 Data integration

- Analysis of multiple platform trials
- Trans platform EPPN^{2020} experiments
- Combining analysis of field trials with analysis of platform trials
- Analysis of genotype x environment interactions
- Multi-trait analyses (platform & field)

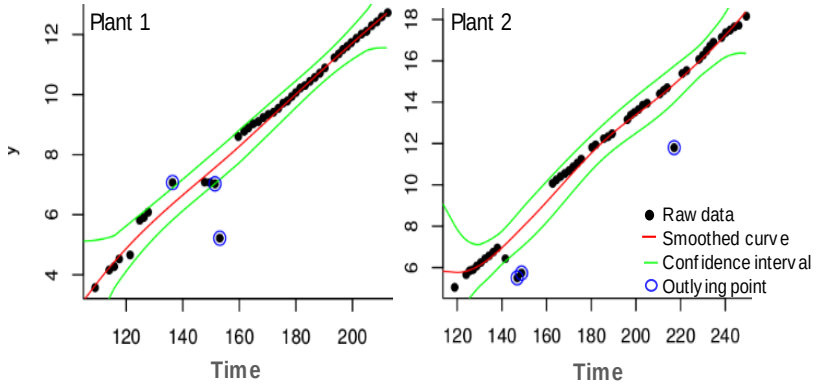
JRA2: Design & analysis outputs

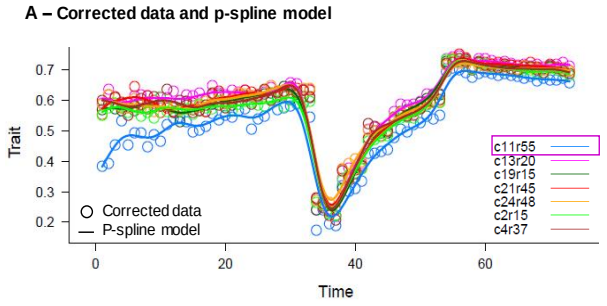
- Statistical methodology
- Software (R procedures)
- Web-applications (design generator)
- Instructions on how to apply the methodology
- Teaching material & courses
- Guidelines & recommendations
- Consultation & support

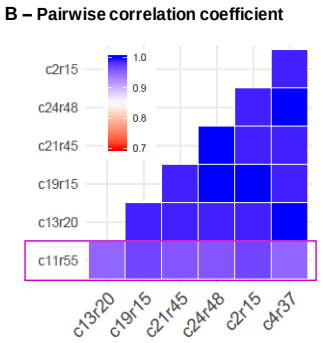
Outliers

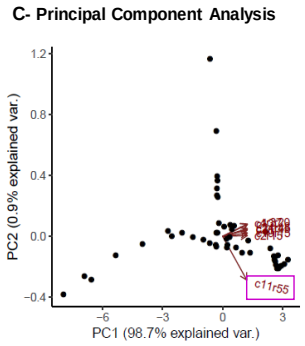
Nadine Hilgert, Isabel Sanchez,
Emilie Millet

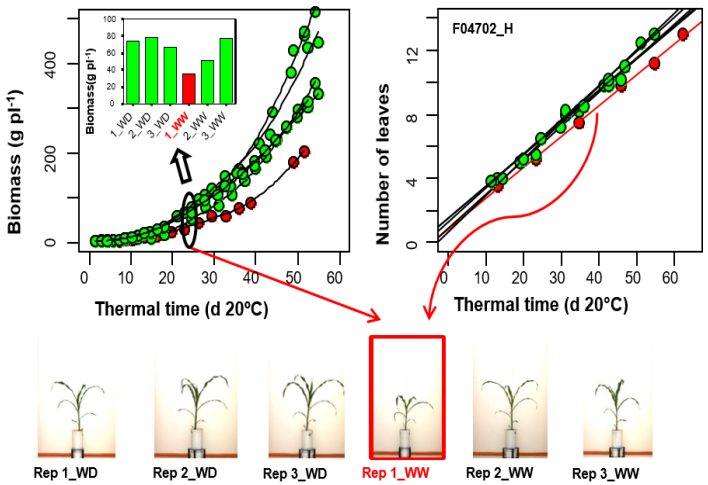
- Single points
- A time series
- A set of time series (plant)







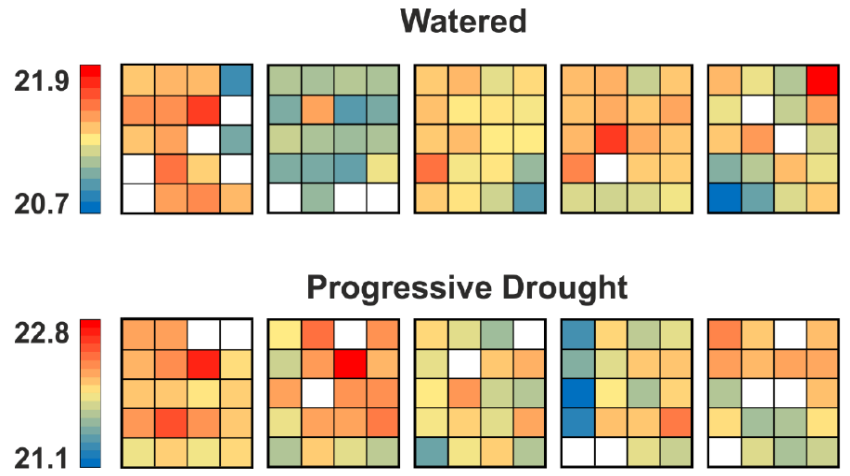


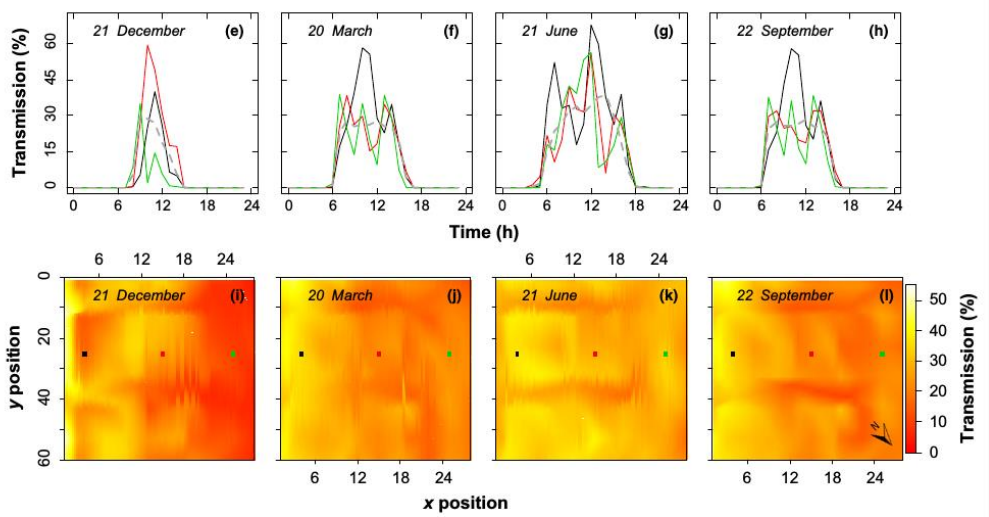


Spatial trends & Design

Observed trends

To counteract spatial trends on platforms, we can use row-column designs and perform spatial analyses

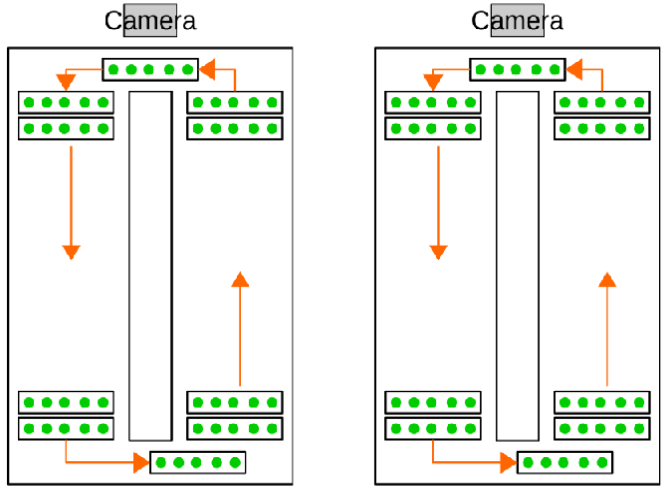




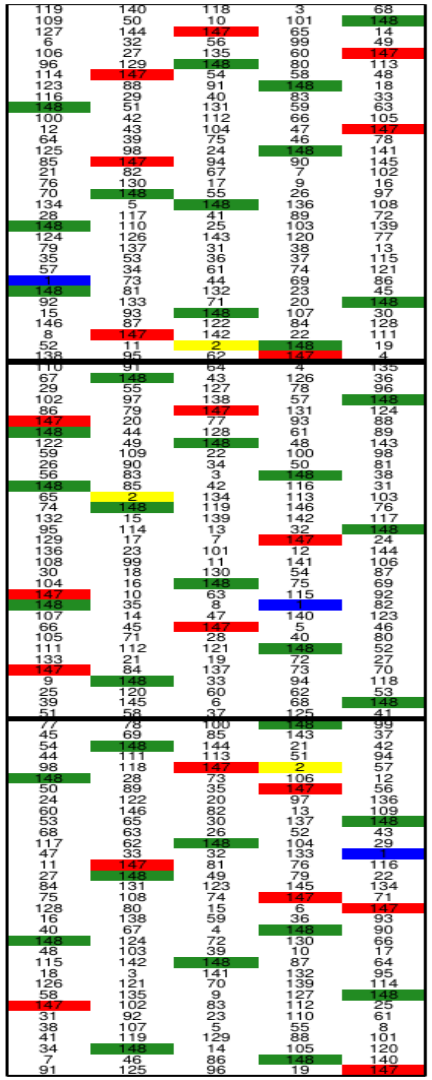
Rhizotubes Dijon Resolvable Row-Column **Design**

21	29	2	26	4	22	16	6	1	30	14	12	3	5	10
33	17	25	7	5	13	14	8	30	23	16	24	2	1	11
16	12	30	19	9	21	23	7	20	32	33	15	17	31	29
20	7	26	1	25	29	9	12	11	17	27	28	24	4	6
9	13	6	27	24	15	29	19	18	31	22	2	7	14	26
32	31	19	14	18	8	25	21	26	5	4	6	33	15	20
14	24	27	28	23	12	15	3	5	25	10	20	18	13	16
3	10	8	30	2	20	24	4	13	11	9	26	32	12	19
15	28	11	32	16	10	2	27	33	7	18	1	8	22	21
5	1	23	6	33	17	32	22	31	21	13	3	9	28	27
4	22	18	31	11	3	17	28	10	29	8	19	23	25	30
22	20	15	5	17	11	27	30	28	12	2	23	14	32	4
8	32	29	16	12	14	3	33	4	22	25	13	19	27	7
23	5	33	29	10	2	19	15	12	24	6	31	1	11	25
24	30	4	3	32	31	1	10	17	14	15	7	26	18	33
11	19	1	4	27	9	21	13	7	3	20	18	5	29	23
31	18	3	24	6	1	8	2	14	27	19	16	12	21	17
10	27	16	8	26	33	20	31	29	9	5	4	28	24	2
17	25	14	15	19	7	11	5	9	26	28	21	22	16	13
6	26	12	20	13	18	22	23	16	8	17	11	10	30	31
7	2	28	22	21	25	18	24	6	33	30	32	20	9	3
13	21	9	23	30	28	26	25	32	10	1	29	6	8	15

Rhizotron UCL Augmented Row Column **Design**

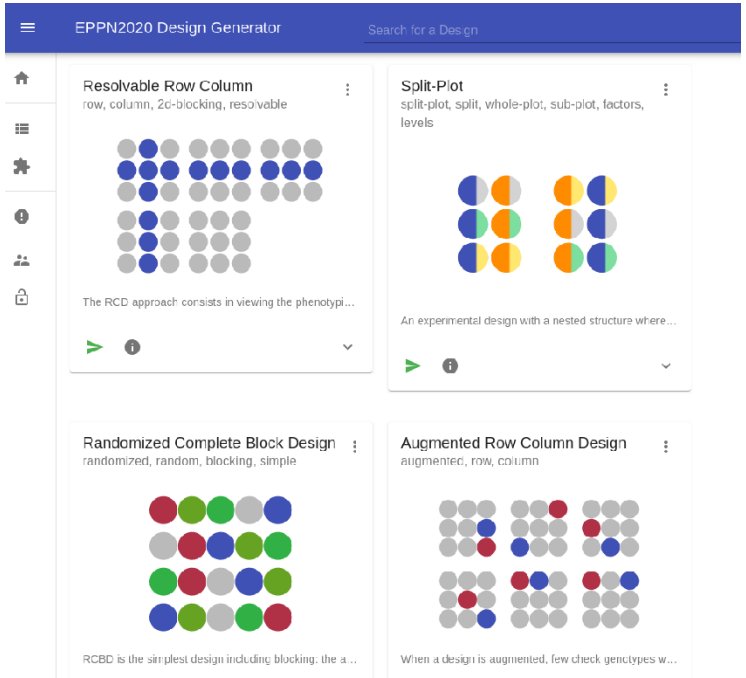


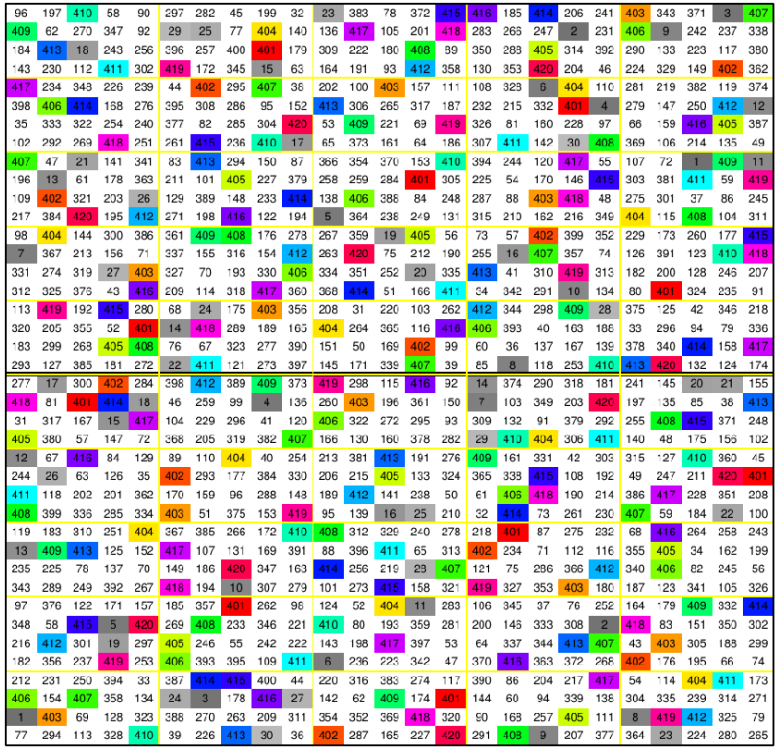
- Guidelines, software & support for choosing suitable statistical designs and corresponding statistical analyses
- Experimental design with row and column block structures to correct for spatial trends
- Choice and allocation of standard genotypes
- Unbiassed estimates of treatment differences with good precision at acceptable costs



Userfriendly Web App Design Generator

Robert Horne & Darren Murray VSNi





Spatial analysis by 2D spline surface

Theor Appl Genet
DOI 10.1007/s00122-017-2894-4

ORIGINAL ARTICLE

Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model

Julio G. Velazco[1,2] · María Xosé Rodríguez-Álvarez[3,4] · Martin P. Boer[1] · David R. Jordan[5] · Paul H. C. Eilers[6] · Marcos Malosetti[1] · Fred A. van Eeuwijk[1]

Spatial Statistics 23 (2018) 52–71

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Correcting for spatial heterogeneity in plant breeding experiments with P-splines

María Xosé Rodríguez-Álvarez [a,b,*], Martin P. Boer [c], Fred A. van Eeuwijk [c], Paul H.C. Eilers [d]

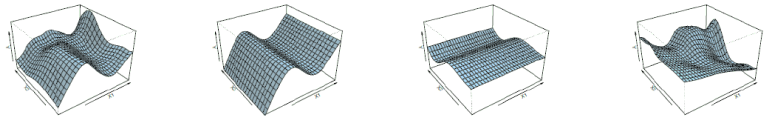
[a] BCAM – Basque Center for Applied Mathematics, Alameda de Mazarredo, 14. E-48009 Bilbao, Basque Country, Spain
[b] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[c] Biometris, Wageningen University & Research, Wageningen, The Netherlands
[d] Erasmus University Medical Centre, Rotterdam, The Netherlands

$$y = f(r, c) + G + e$$

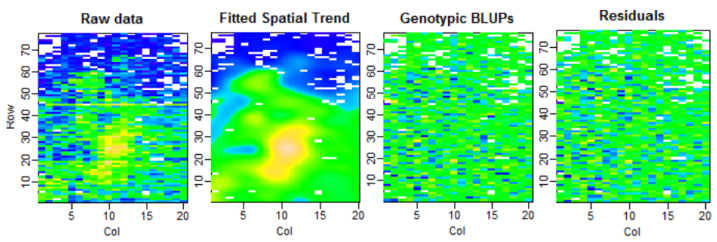
genetic effects

ANOVA-type decomposition of the smoothed surface:

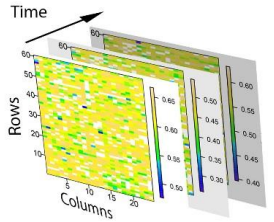
$$\hat{f}(r, c) \equiv f_r(r) + f_c(c) + f_{r|c}(r, c)$$



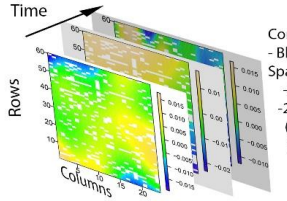
$$y = f(r, c) + G + e$$



A Raw phenotypic data

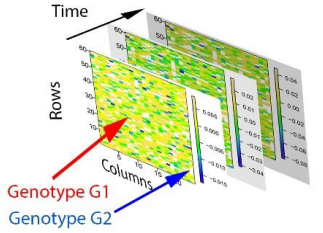


B Fitted spatial trend

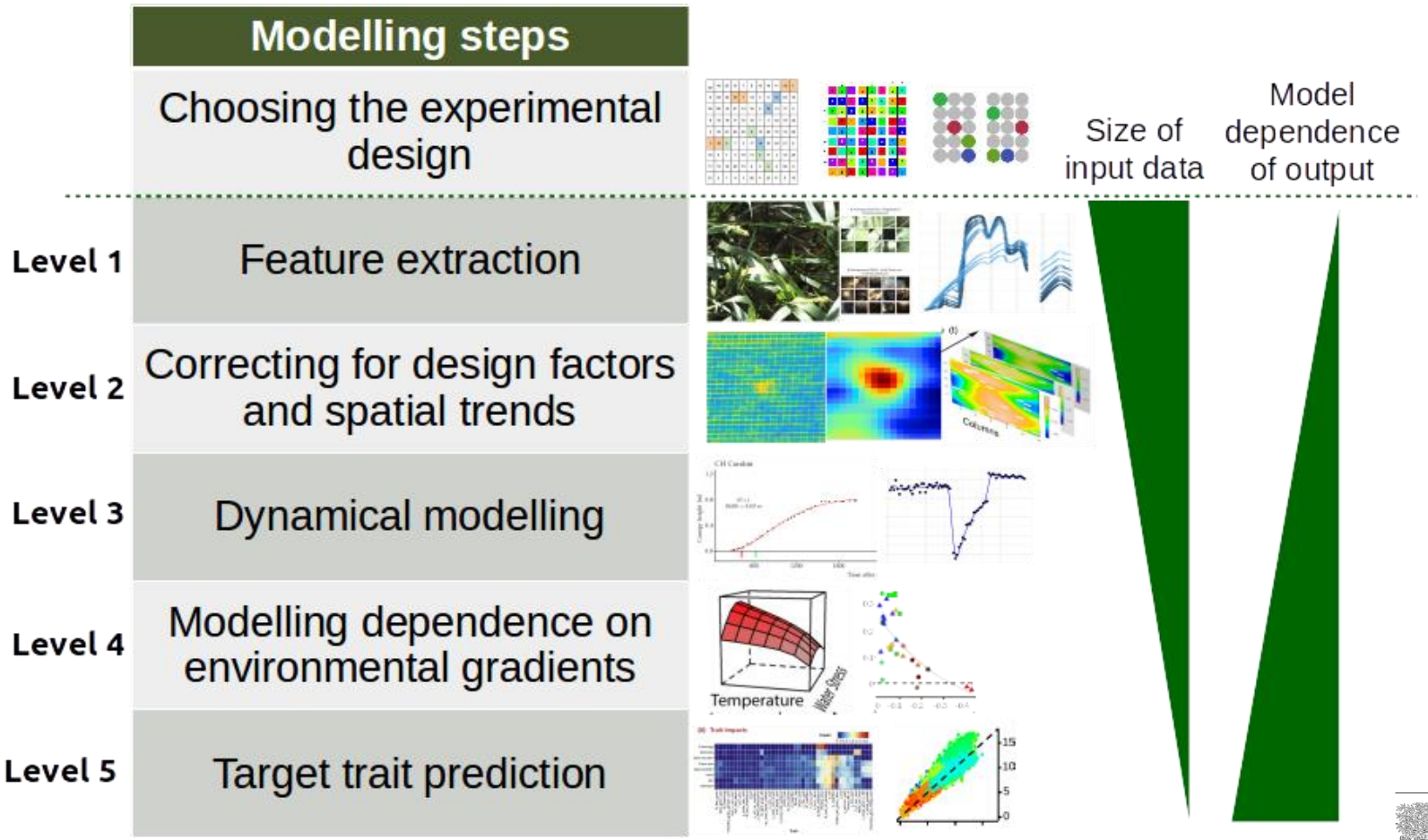


Correction fro spatial effects:
- Blocks
Spatially-dependent error
-AR1⊗AR1 (Gilmour et al., 1997)
-2D-splines
(Rodriguez-Álvarez et al., 2018, Velazco et al., 2017)

C Genotypic values (BLUEs)



Data integration



Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding

Fred A. van Eeuwijk[a,*], Daniela Bustos-Korts[a], Emilie J. Millet[a], Martin P. Boer[a], Willem Kruijer[a], Addie Thompson[b,1], Marcos Malosetti[a], Hiroyoshi Iwata[c], Roberto Quiroz[d], Christian Kuppe[e], Onno Muller[e], Konstantinos N. Blazakis[f], Kang Yu[g,h], Francois Tardieu[i], Scott C. Chapman[j,k]

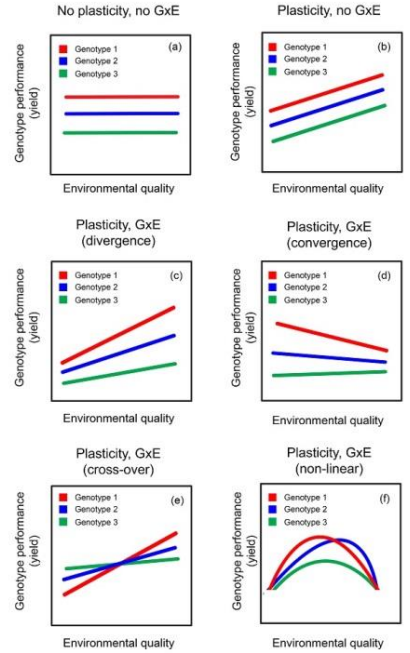
Baseline genotype to phenotype model includes GxE

Published August 30, 2016

RESEARCH

What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype × Environment Interactions?

Fred A. van Eeuwijk,* Daniela V. Bustos-Korts, and Marcos Malosetti



- Phenotype =
  - Genotype +
  - Environment +
  - Genotype by Environment Interaction +
  - Error
- Two-way ANOVA, fixed genotypes and environments, GxE fixed (lack of fit) term
  - $y_{ij} = \mu + g_i + e_j + ge_{ij} + \epsilon_{ij};\ \epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$
  - Subscript *i* for genotype, *j* for environment
- Mixed model formulation, random genotypes, GxE as heterogeneity of genetic variances and correlations
  - $y_{ij} = \mu_{ij} + G_{ij} + \epsilon_{ij};\ \mathrm{VCOV}(y_{ij}) = \boldsymbol{\Sigma}_{gge} + \boldsymbol{R}_\epsilon$

GxE and separability

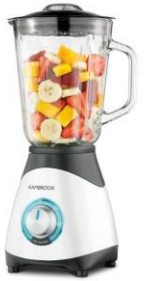
- $\underline{y_{ij}} = \mu_{ij} + \underline{G_{ij}} + \underline{\epsilon_{ij}}$
  - Multi-environment (multi-trait) model
- Mean μ_{ij}
  - $\mu_{ij} = \mu(\boldsymbol{\beta}_i^g, \boldsymbol{\beta}_j^e, \boldsymbol{x}_i, \boldsymbol{z}_j) = \sum_m a_{mi} b_{mj}$
    - $\boldsymbol{\beta}_i^g$: genotypic sensitivities
    - $\boldsymbol{x}_i$: genotypic covariables / molecular markers / groupings
    - $\boldsymbol{\beta}_j^e$: environment related quality / QTL
    - $\boldsymbol{z}_j$: environmental covariable / index / classification / management
- Variance-Covariance matrix $\underline{G_{ij}}$: $VCOV(\underline{G_{ij}})$;
  - $VCOV(\underline{G_{ij}}) = \sum_v \boldsymbol{A}_v^g \otimes \boldsymbol{B}_v^e$
    - $VCOV(\underline{G_{ij}}) = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\boldsymbol{g}} \otimes \boldsymbol{\Sigma}^e$
    - Structuring $\underline{G_{ij}}$ on genotypic dimension
      - Markers (GRM), pedigree, omics, phenotyping, grouping
    - Structuring $\underline{G_{ij}}$ on environmental dimension
      - Weather, soil, management, indices, classifications

Product models for GxE (separable)

- Finlay-Wilkinson / AMMI / GGE / SREG; implicit genotypic and environmental characterization
  - $\gamma_i \delta_j + \varepsilon_{ij}$
- Explicit genotypic characterization: QTL mapping (markers) & **secondary phenotyping** (platform / field)
  - $x_i \beta_j^e + \varepsilon_{ij}$
  - $y_i^* \beta_j^e + \varepsilon_{ij}$
  - $(\boldsymbol{y}_i^*)^t \boldsymbol{\beta}_j^e + \varepsilon_{ij}$
  - $(\boldsymbol{y}(\boldsymbol{x})_i^*)^t \boldsymbol{\beta}_j^e + \varepsilon_{ij}$
- Explicit environmental characterization (factorial regression)
  - $\beta_i^g z_j + \varepsilon_{ij}$
  - $\beta_i^g s(z_j) + \varepsilon_{ij}$
- Modelling using genotypic and environmental covariates
  - $\beta(\boldsymbol{x}_i) z_j + \varepsilon_{ij} = \kappa x_i z_j + \beta_i^* z_j + \varepsilon_{ij}$
  - $\beta(y_i^*) z_j + \varepsilon_{ij} = \kappa y_i^* z_j + \beta_i^* z_j + \varepsilon_{ij}$

Kernel methods

- Generalized GE-BLUP
  - $y_{ij} = \mu_{ij} + G_{1ij} + G_{2ij} + \epsilon_{ij}$
  - $VCOV(G_{1i}) = \boldsymbol{\Sigma}^{g_1} \otimes \boldsymbol{\Sigma}^{e}$ with $\boldsymbol{\Sigma}^{g_1} = \boldsymbol{K_1}\sigma^2_{G_1}$ and $\boldsymbol{K_1}$ a function of markers (genotypes)
  - $VCOV(G_{2i}) = \boldsymbol{\Sigma}^{g_2} \otimes \boldsymbol{\Sigma}^{e}$ with $\boldsymbol{\Sigma}^{g_2} = \boldsymbol{K_2}\sigma^2_{G_2}$ and $\boldsymbol{K_2}$ a function of omics / secondary phenotyping, defined on genotypes
  - $\boldsymbol{\Sigma}^{e}$ is a structuring matrix defined on the environments



Introducing a time component in GxE models

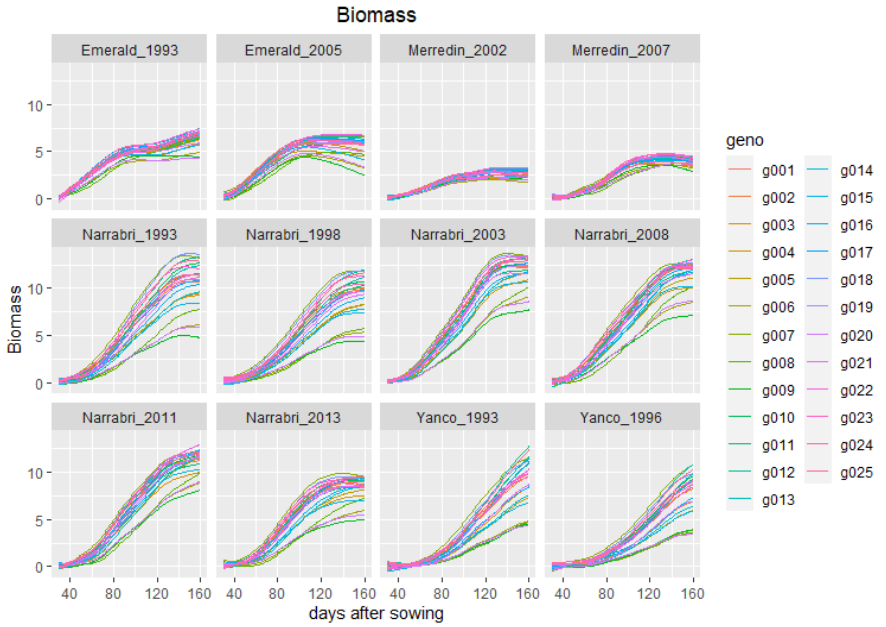
2-way P-spline ANOVA and 3-level hierarchical P-spline model

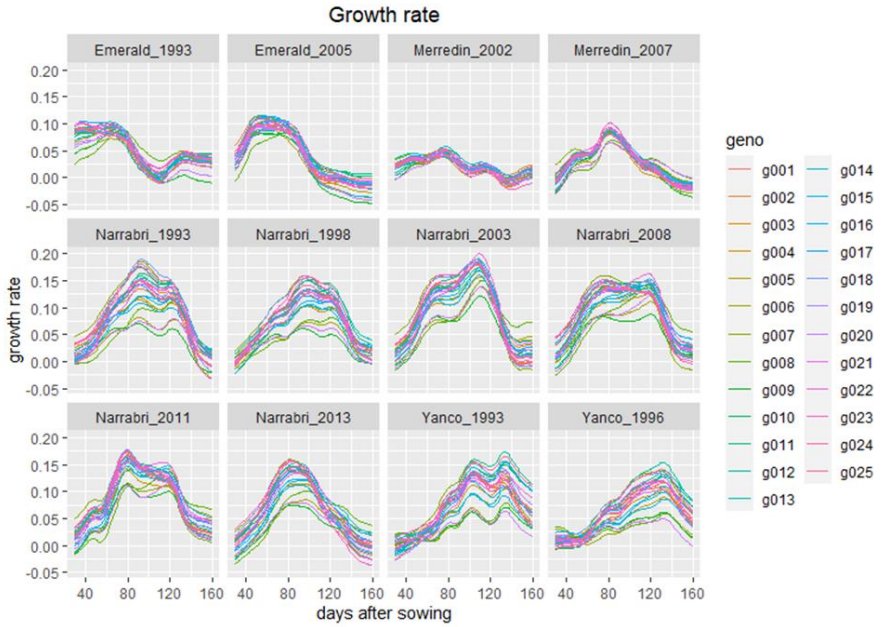
- Martin Boer: P-spline 2-way ANOVA
  - $y_{ij}(t) = \mu(t) + f_j^E(t) + f_i^G(t) + f_{ij}^{G.E}(t) + \varepsilon_{ij}(t)$
  - $y_{ij}(t) = \{\mu + lin(t) + spl(t)\} + \{g + e + g.e\}.\{lin(t) + spl(t)\} + \varepsilon_{ij}(t)$
    - Sparse mixed model equations, special choice of basis functions
    - Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.
- Diana M. Perez & Maria Xose Alvarez-Rodriguez
  - $y_{ijk}(t) = \mu(t) + f_j^E(t) + f_{ij}^{E.G}(t) + f_{ijk}^{E.G.P}(t) + \varepsilon_{ijk}(t)$
    - Choice of number of basis splines critical
    - Brumback and Rice. *JASA*, 1998; Durban, Harezlak, Wand and Carroll. *Stat Med*, 2005.

Dynamical predictions of biomass (G+E+G.E) and first derivatives

Data: APSIM wheat simulation – Daniela Bustos Korts

Analysis: Martin Boer



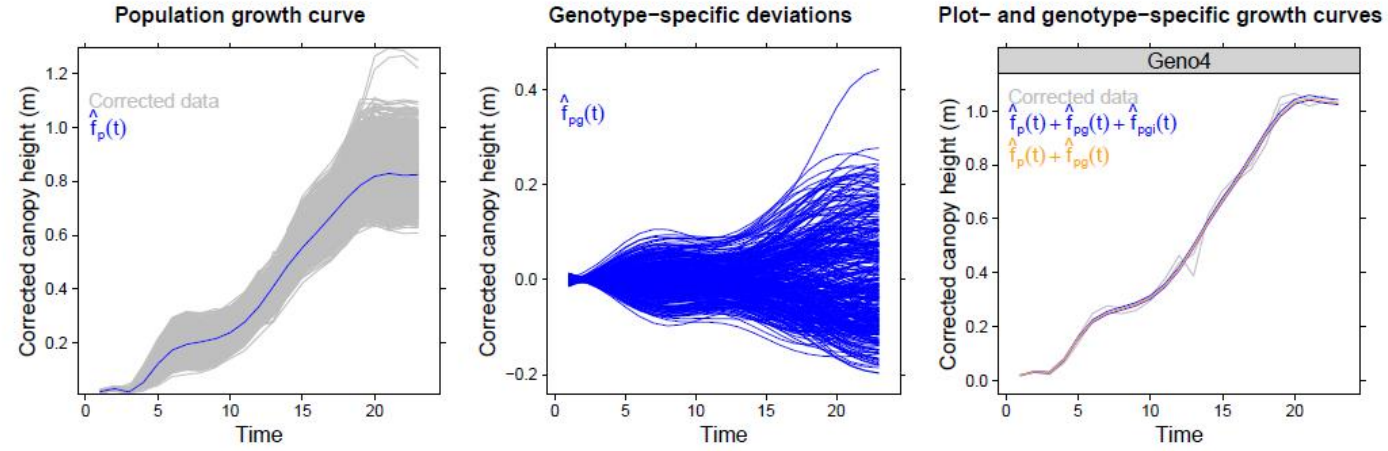


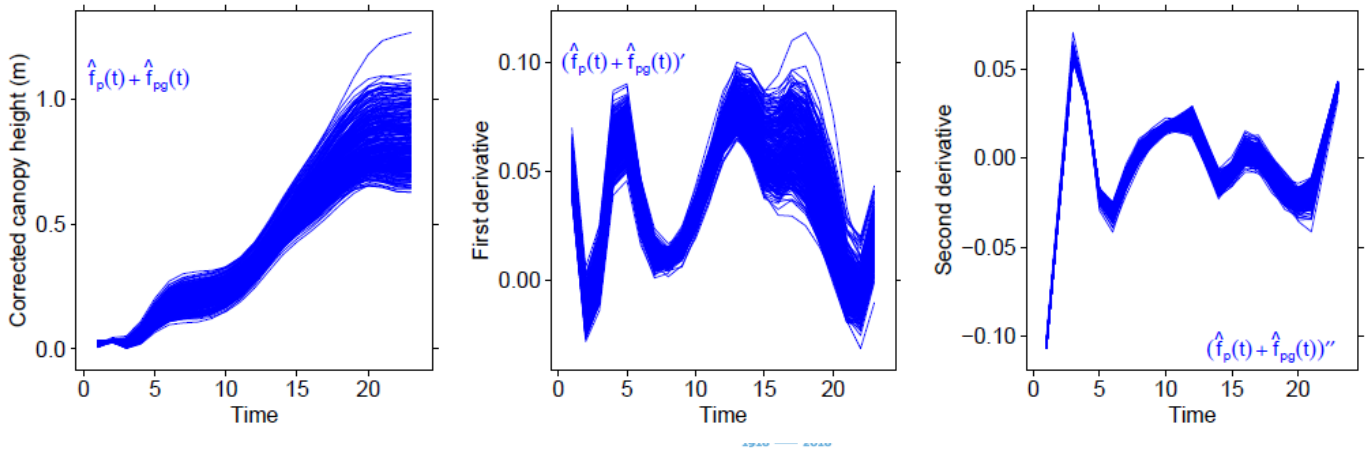
$$y_{ij}(t) = \mu(t) + f_G(t) + f_E(t) + f_{G\times E}(t) + \varepsilon_{ij}(t)$$

Illustration 3 level hierarchical splines and derivatives

Analysis: Diana M. Perez & Cote Alvarez Rodriguez (BCAM – Bilbao)

Data: wheat - group Andreas Hund (ETH Zurich)

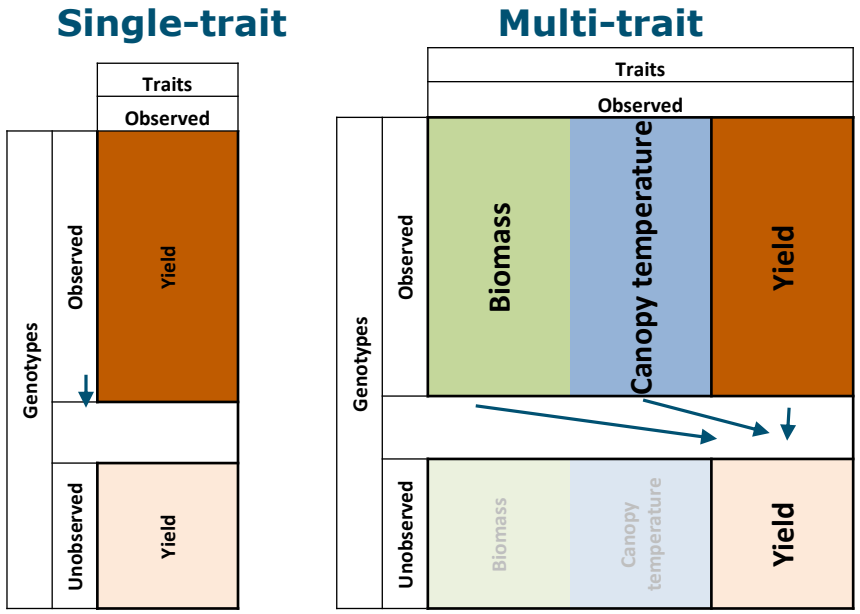


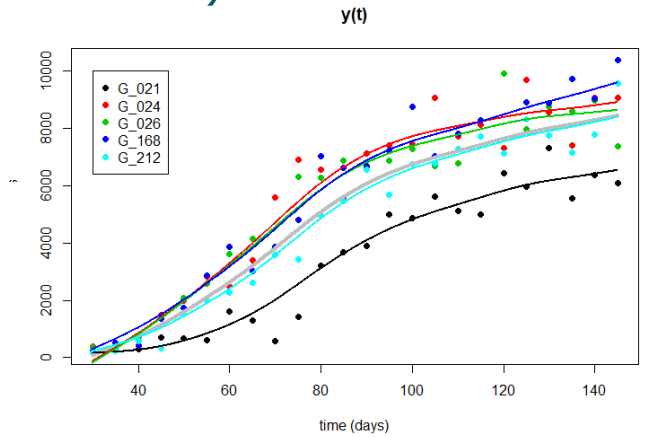


- Curves
  - population (prediction)
  - genotype (deviation)
  - plant / plot (prediction)
- Genotype prediction
  - first derivative
  - second derivative

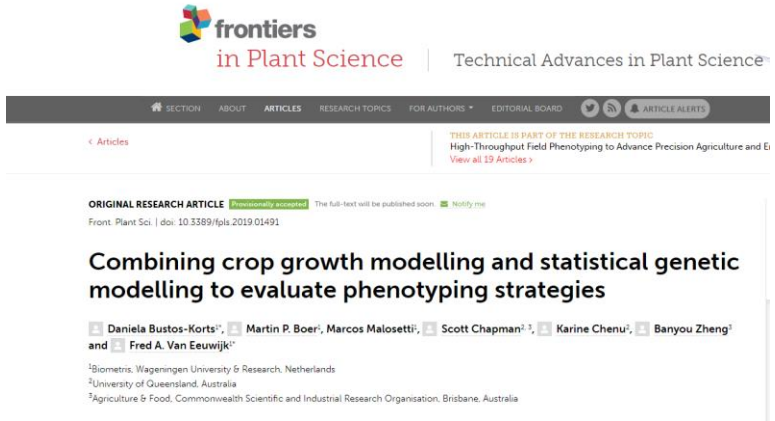
Data integration

Use dynamical parameters or full functional information of secondary phenotypes in a better prediction of target traits like yield (including GxE) (Daniela Bustos-Korts)





$$\underline{y}(t)_{ij} = \mu(t)_j + \underline{G}(t)_{ij} + \underline{\epsilon}(t)_{ij}$$

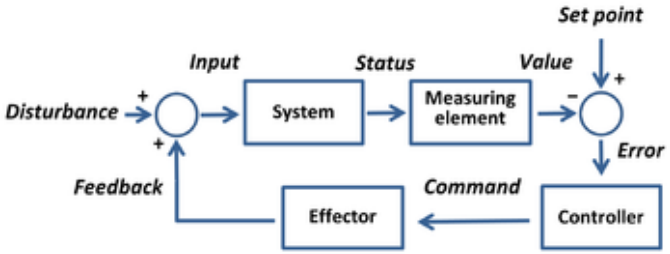


WAGENINGEN UNIVERSITY & RESEARCH

100years 1918 — 2018

Modelling principles: phenology and multiple traits

- Multiple kernel models
  - integrate genomic information with omics, phenotyping and environmental information
  - usually little attention for feature selection
- Feature selection in the form of characterization of genotypes and environments always pays off (mean and VCOVs)
  - adaptation & maturity groups (phenology)
  - eco zones, mega environments, environmental scenarios
- Yield = integrated output of processes triggered by environmental and internal stimuli following interactions between upstream traits starting from a DNA basis
  - Predicting GxE requires multi-trait modelling
  - Phenotyping platforms give access to many traits
- Prediction models for GxE need to include phenology/ timing of developmental events implicitly or explicitly
  - phenotyping platforms / devices provide temporal information
  - feature selection needs to respect phenology, scalability and granularity



WAGENINGEN
UNIVERSITY & RESEARCH

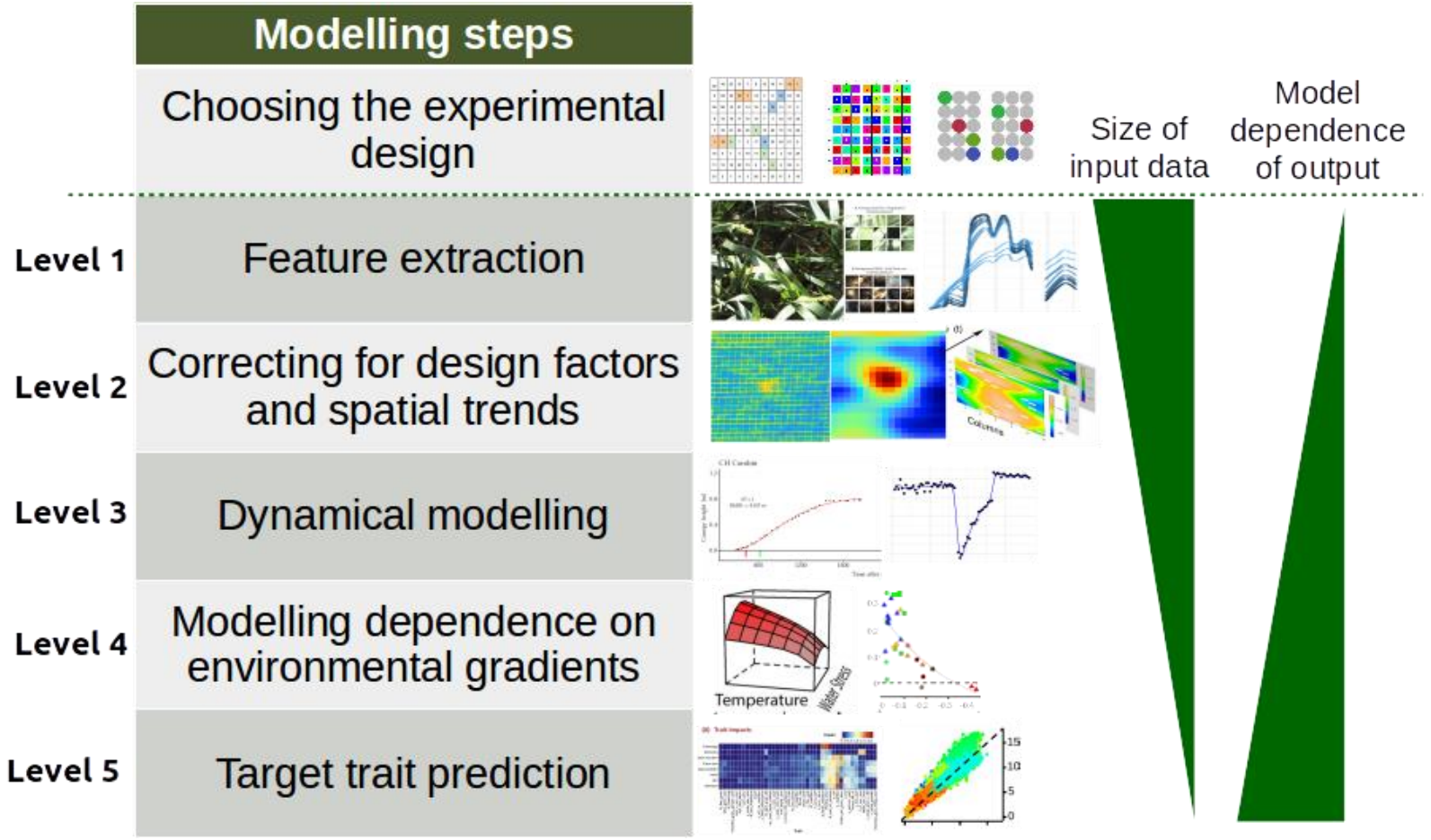
100years 1918 — 2018

Review TRENDS in Plant Science Vol.11 No.12
ScienceDirect

Models for navigating biological complexity in breeding improved crop plants

Graeme Hammer[1], Mark Cooper[2], François Tardieu[3], Stephen Welch[4], Bruce Walsh[5], Fred van Eeuwijk[6], Scott Chapman[7] and Dean Podlich[2]

Wrap up



Statistical approaches / tools

- Design
- Data quality control
- Single experiment analysis
  - Longitudinal data analysis
- Data integration
  - Primary trait prediction
  - Multi-experiment analysis
  - Multi-trait hierarchical models

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

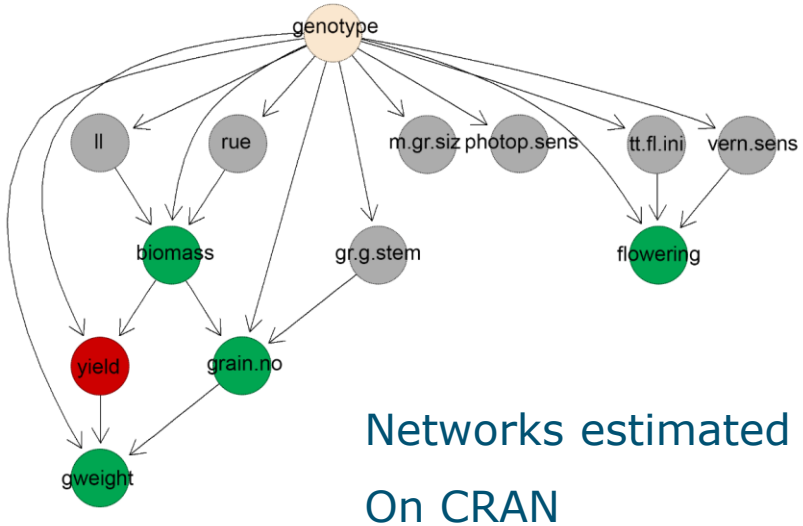
Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding

Fred A. van Eeuwijk[a,*], Daniela Bustos-Korts[a], Emilie J. Millet[a], Martin P. Boer[a], Willem Kruijer[a], Addie Thompson[b,1], Marcos Malosetti[a], Hiroyoshi Iwata[c], Roberto Quiroz[d], Christian Kuppe[e], Onno Muller[e], Konstantinos N. Blazakis[f], Kang Yu[g,h], Francois Tardieu[i], Scott C. Chapman[j,k]

Yield as part of a trait network with a genetic root

- Networks help visualizing genetic and phenotypic trait relationships (biomass plays a central role for yield)
- Topologies / adaptive mechanisms differ between environmental types with an important role for phenology
- Trait organisation shows '**modules**' that are **environment type** dependent, such modules can guide feature selection in prediction models (including kernel models)
- Elucidation of direct and indirect genetic effects on target traits

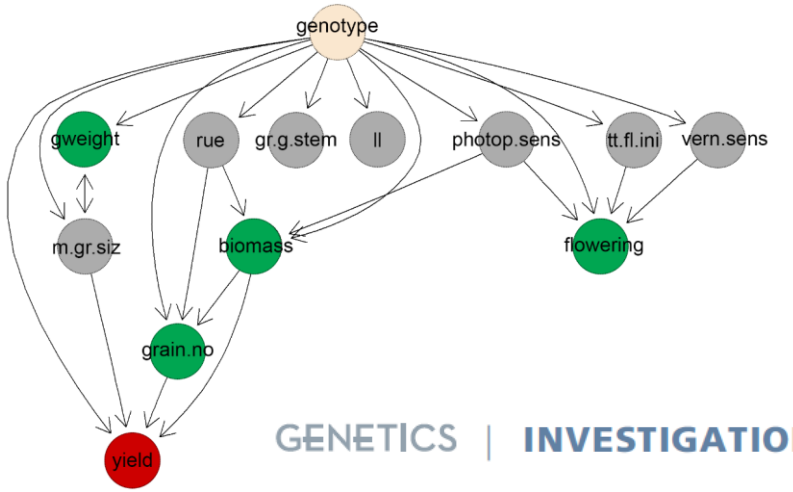
Drought



Networks estimated with **pcgen** algorithm

On CRAN

Non drought



GENETICS | INVESTIGATION

Reconstruction of Networks with Direct and Indirect Genetic Effects

Willem Kruijer,*,[1] Pariya Behrouzi,* Daniela Bustos-Korts,* María Xosé Rodríguez-Álvarez,[†,‡] Seyed Mahdi Mahmoudi,[§] Brian Yandell,** Ernst Wit,[††] and Fred A. van Eeuwijk*

WAGENINGEN UNIVERSITY & RESEARCH

100years 1918 — 2018

Data integration

$$GrainYield_{ij} = GrainNumber_{ij} \times GrainWeight_i$$

$$GrainYield_{ij} = \left(\mu + e_j + g_i + \sum_{l=1}^{s} \beta_{l,i} z_{l,j,i}\right) \times GrainWeight_i$$

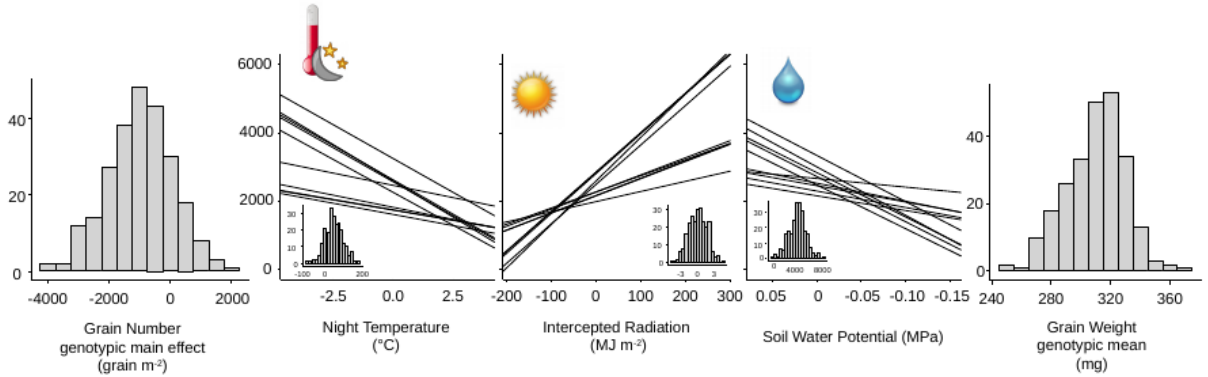
environmental main effect

genotypic main effect

genotypic sensitivities

environmental covariables

genotypic grain weight



⇒ Can be predicted using marker profile and/or environmental indices