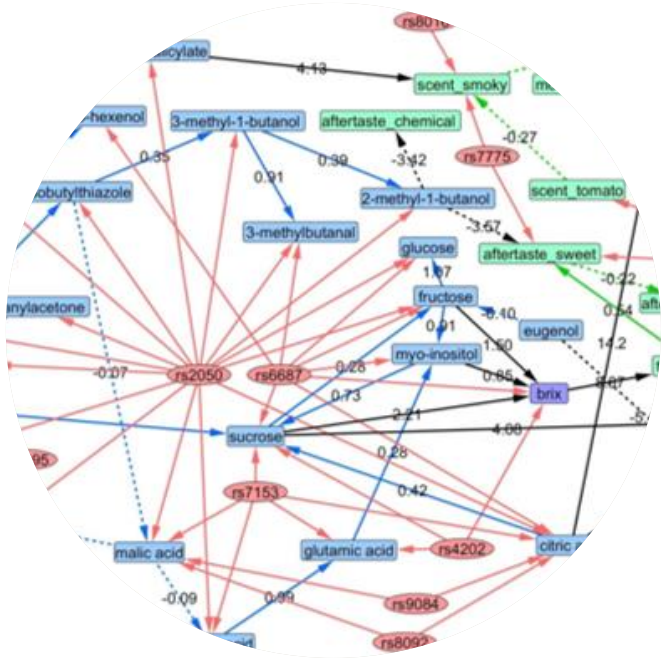
Introduction to Statistics and Data Science

Wageningen – Biometris, Friday 1 December, 2023

Fred van Eeuwijk



WAGENINGEN
UNIVERSITY & RESEARCH

Program

- 13.30-14.00 Fred van Eeuwijk
  - Introduction
- 14.00-14.30 Tom Theeuwen
  - The happy marriage between phenotyping and genetics: the key to unravel natural variation for photosynthesis
- 14.30-15.15 Felix Akens & Rick van de Zedde
  - Guided tour and introduction to the NPEC facilities
- 15.15-15.30 Coffee
- 15.30-16.00 Phillip Gillhausen
  - Digital Plant Phenotyping
- 16.00-16.45 Drinks

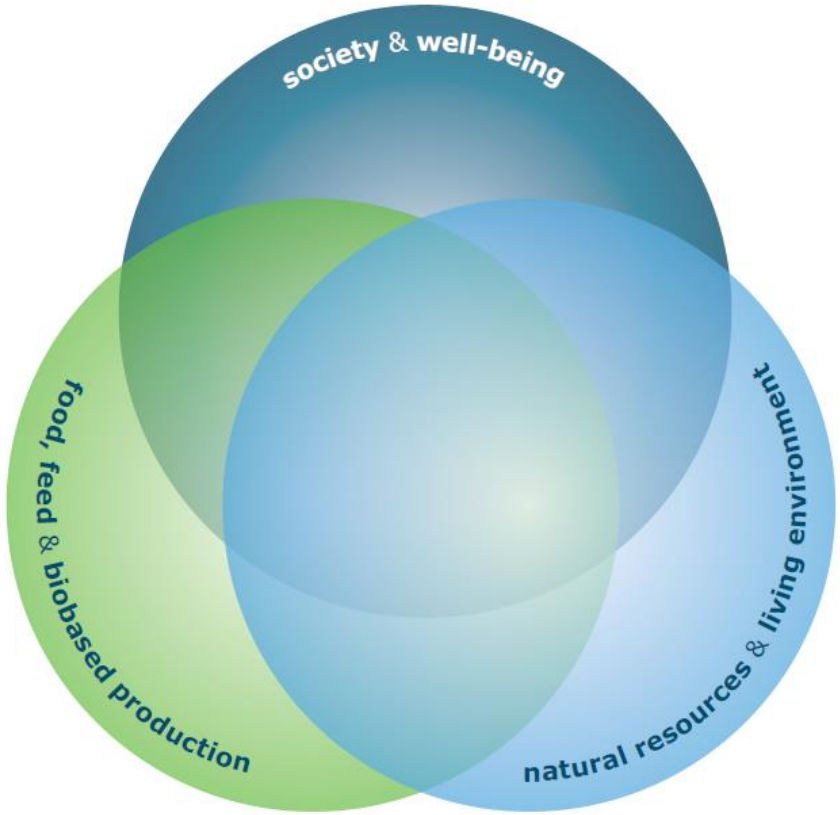
Participating groups in MSc Statistics and Data Science

- Leiden University Medical Centre (LUMC)
- Faculty of Social and Behavioural Science (FSW) Leiden
- Faculty of Science (FWN) / Mathematical Institute Leiden
- Biometris, Wageningen University and Research

- https://www.universiteitleiden.nl/en/education/study-programmes/master/statistics--data-science

WAGENINGEN
UNIVERSITY & RESEARCH

Our domain: healthy food and living environment



The university's 5 departments

Agrotechnology & Food Sciences

Animal Sciences

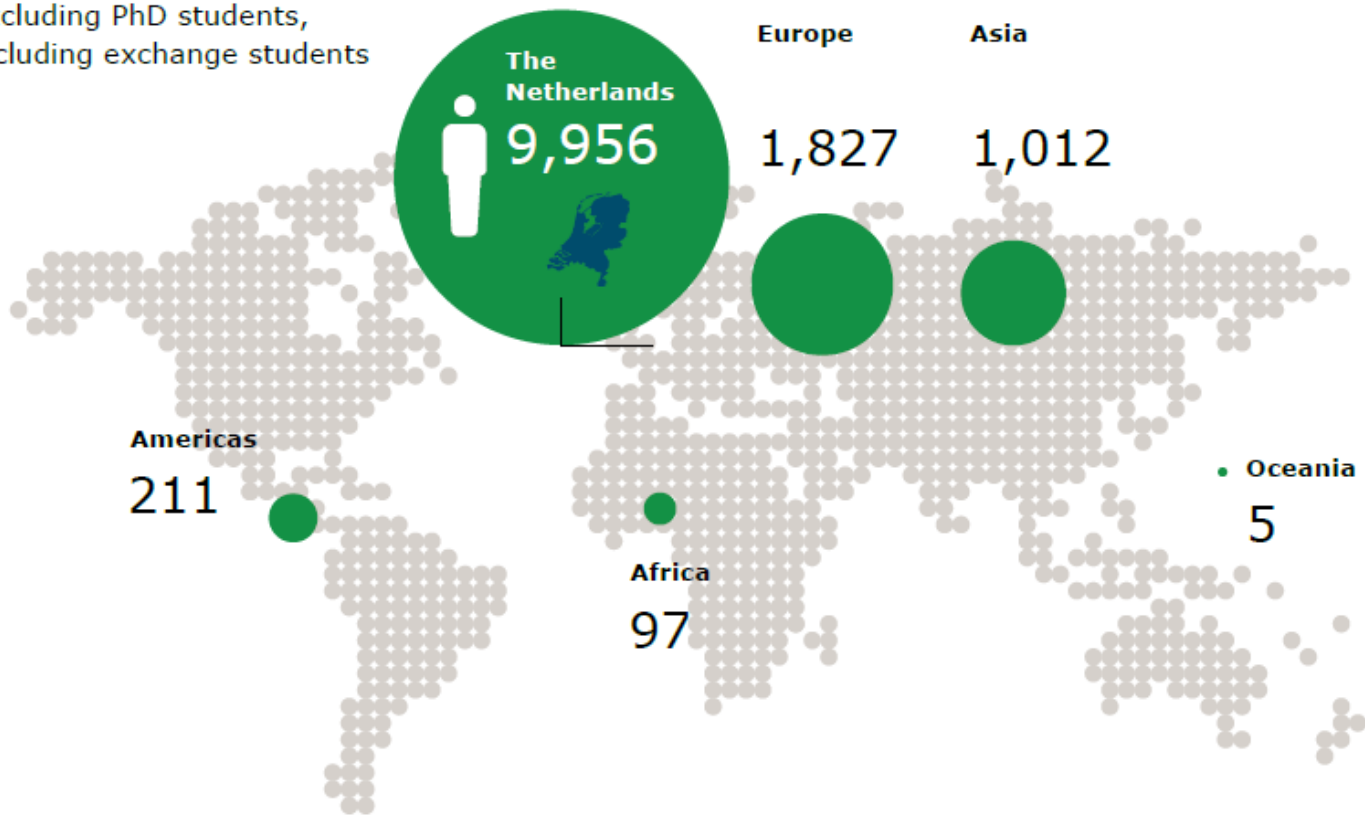
Environmental Sciences

Plant Sciences

Social Sciences

- Climate change
- Circular & Biobased Economy
- Nutrition & Health
- From hunger to food security
- Biodiversity

Origin of students
excluding PhD students,
including exchange students



109
nationalities

Albania, Australia, Argentina, Austria, Bangladesh, Belarus, Belgium, Bhutan, Bolivia, Brazil, Bulgaria, Cambodia, Cameroon, Canada, Chile, China, Colombia, Costa Rica, Croatia, Cyprus, Czech Republic, Congo, Denmark, Ecuador, Egypt, El Salvador, Estonia, Ethiopia, Finland, France, Georgia, Germany, Ghana, Greece, Guatemala, Guyana, Hungary, Iceland, India, Indonesia, Iran, Ireland, Israel, Italy, Japan, Jordan, Kazakhstan, Kenya, Kosovo, Latvia, Lebanon, Liberia, Lithuania, Luxembourg, Malaysia, Malta, Mauritius, Mexico, Mongolia, Morocco, Myanmar, Namibia, Nepal, New Zealand, Netherlands, Nicaragua, Nigeria, Norway, Pakistan, Panama, Peru, Philippines, Poland, Portugal, Romania, Russia, Rwanda, Saudi Arabia, Senegal, Sierra Leone, Singapore, Slovakia, Slovenia, Somalia, South Africa, South Korea, South Sudan, Spain, Sri Lanka, Sudan, Suriname, Sweden, Switzerland, Syria, Taiwan, Tanzania, Thailand, Tunesia, Turkey, Uganda, Ukraine, United Kingdom, United States of America, Venezuela, Vietnam, Yemen, Zambia, Zimbabwe

Locations

Locations in the Netherlands

Wageningen University & Research
- Wageningen, 1

Wageningen Academy
- Wageningen, 1

Agrotechnology &
Food Sciences Group
- Wageningen, 1

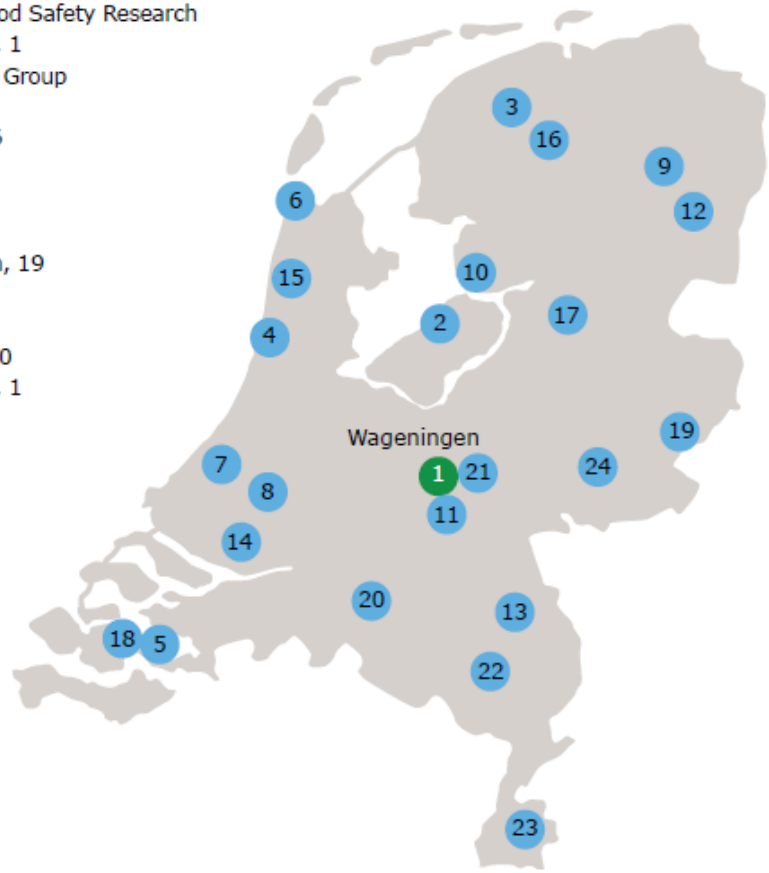
Animal Sciences Group
- Den Helder, 6
- Hengelo, 24
- IJmuiden, 4
- Leeuwarden, 3
- Lelystad, 2
- Wageningen, 1
- Yerseke, 5

Environmental Sciences Group
- Renkum, 21
- Wageningen, 1

Plant Sciences Group
- Bleiswijk, 8
- Lelystad, 2
- Marwijksoord, 9
- Nagele, 10
- Randwijk, 11
- Valthermond, 12
- Vredepeel, 13
- Wageningen, 1
- Westmaas, 14
- Wijnandsrade, 23

Wageningen Food Safety Research
- Wageningen, 1

Social Sciences Group
- Alkmaar, 15
- Drachten, 16
- Dalfsen, 17
- Den Haag, 7
- Goes, 18
- Haaksbergen, 19
- Lelystad, 2
- Meijel, 22
- Oisterwijk, 20
- Wageningen, 1



WAGENINGEN
UNIVERSITY & RESEARCH

Wageningen worldwide

Wageningen University & Research is active in many regions of the world. From China to Chile and from Ethiopia to the Arctic, we work together with partners in research programmes.

WAGENINGEN
UNIVERSITY & RESEARCH

Output/scientific prominence

Output 2022

PhD theses

320

Veni, Vidi, Vici in 2022

Veni	Vidi	Vici
9	10	1

ERC Grants

Starting Grant since 2007

9

Advanced Grant since 2007

8

Consolidator Grant since 2013

4

Rankings

WUR ranking in
QS World University Rankings 2023
Agriculture and Forestry

1 (8 years running)

WUR ranking in
National Taiwan University Ranking,
World Universities 2022
Agriculture

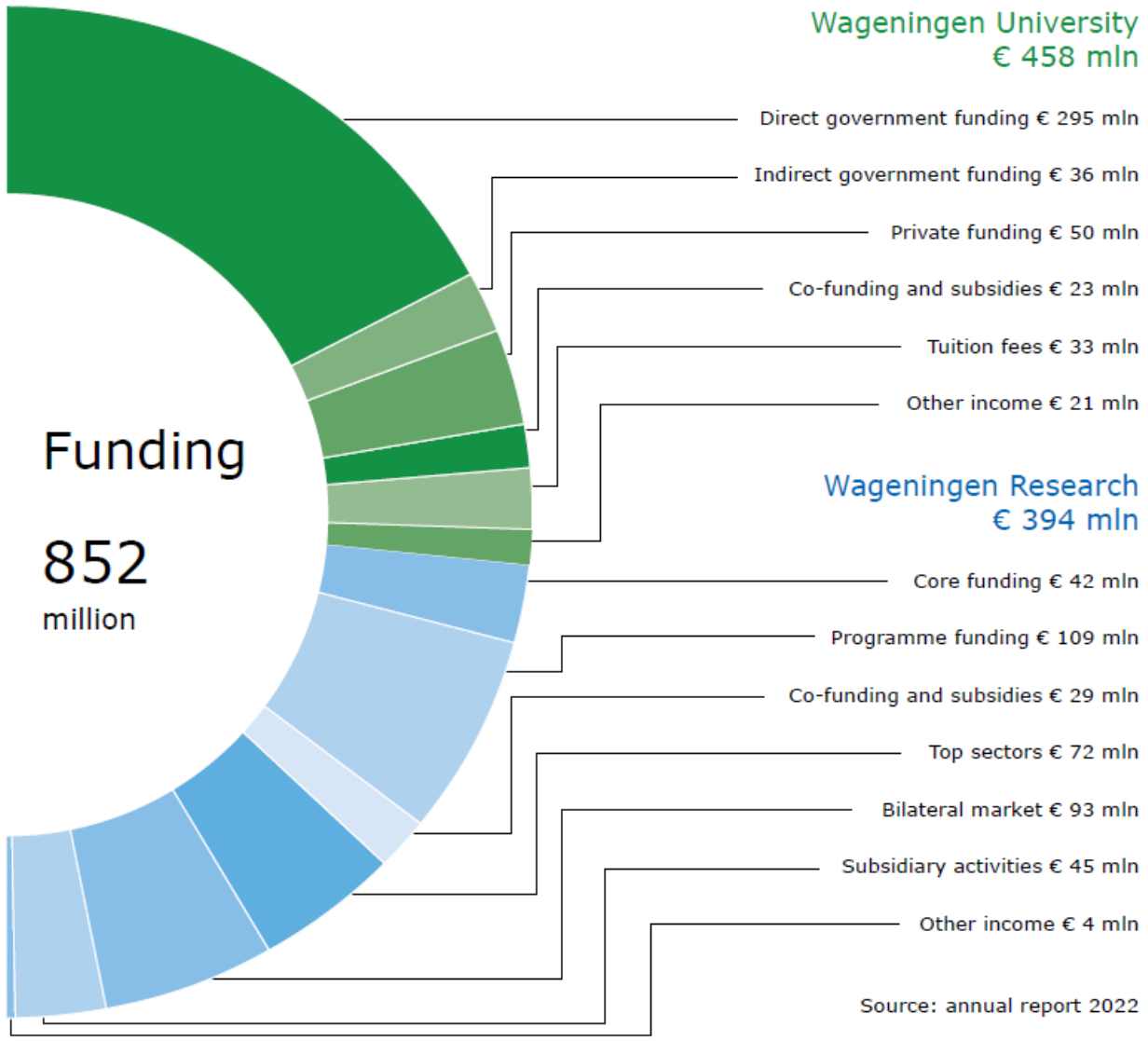
1

WUR ranking in QS World University Rankings 2023
Environmental Sciences

2

WUR ranking in Times Higher Education World University Rankings 2023

59



WAGENINGEN
UNIVERSITY & RESEARCH

Biometris

- Applied statistics (60 fte) and applied mathematics (45 fte)
- Quantitative methodology for life and environmental sciences
- Education
- Research
- Academic & commercial

Biometris education

https://www.wur.nl/en/research-results/research-institutes/plant-research/biometris/education.htm

Biometris provides education in applied mathematics and statistics within Wageningen University & Research. Biometris occupies a unique position within the university since quantitative methods are important for almost all Wageningen University study programmes. It is the largest supplier of education, and no other group is involved in so many of these programmes.

Focus on practical applicability

While theoretical knowledge of quantitative methods is highly relevant for students in Wageningen, practical applicability within the student's field of study is of greater importance. Our study material contains information on applications from many fields, putting much emphasis on the connection between quantitative methods and applications.

WAGENINGEN
UNIVERSITY & RESEARCH

Biometris research

https://www.wur.nl/en/research-results/research-institutes/plant-research/biometris.htm

Research areas of Biometris

The general expertise of Biometris includes mathematics and statistics applied to life sciences. Our focus is on the research themes below.

Phenotype, Genotype and Environment

- Phenotype
  - Properties you measure or observe on organism
- Genotype
  - Genetic constitution of organism
- Environment
  - Physical and biological conditions that the organism encounters

- Genetics, relating DNA variation to phenotypic variation and vice versa
  - phenotype = f(genotype, environment) + error
- Phenotyping
  - Measuring of plant properties by hand / eye or measurement devices in the field and under controlled conditions

Plant breeding

- Developing new plant varieties (cultivars, genotypes) that have higher yield and quality with lower environmental imprint
  - Create new genetic variation by crossing parents with interesting complementary properties (phenotypes) = production of offspring populations
  - Evaluate the offspring for desirable traits (phenotypes) under certain environmental conditions (for example, Dutch growing conditions)
  - Identify the genetic basis of those desirable phenotypes under certain conditions = locate the positions in the genome where the genes (alleles) are that produce the desired phenotype
  - Create or select genotypes with the genes that lead to the desired phenotypes

G2P models

- Genotype-to-phenotype (G2P) models describe phenotypes as functions of genetic and environmental parameters and inputs. They are essential for the identification of superior genotypes
- Phenotype =
  - Genotype +
  - Environment +
  - Genotype by Environment Interaction +
  - Error
- Two-way ANOVA, fixed genotypes and environments, GxE fixed (lack of fit) term
  - $y_{ij} = \mu + g_i + e_j + ge_{ij} + \epsilon_{ij}$; $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$
  - Subscript *i* for genotype, *j* for environment
- Mixed model formulation, random genotypes, GxE as heterogeneity of genetic variances and correlations
  - $y_{ij} = \mu_{ij} + G_{ij} + \epsilon_{ij}$; $\mathrm{VCOV}(y_{ij}) = \boldsymbol{\Sigma}_{gge} + \boldsymbol{R}_\epsilon$

Genotypic covariates, molecular markers & QTLs

- Single environment; partition main effect genotypic differences into part that can be explained by DNA differences and a residual genotypic effect
  - $\underline{y}_{ir} = \mu + \underline{G}_i + \underline{\epsilon}_{ir}$
  - $\underline{y}_{ir} = \mu + (x_i\beta^Q + \underline{G}_i^*) + \underline{\epsilon}_{ir}$
- Multiple environments; partition both main effect genotypic differences and GxE interaction
  - $\underline{y}_{ij} = \mu + E_j + \underline{G}_i + \underline{GE}_{ij} + \underline{\epsilon}_{ij}$
  - $\underline{y}_{ij} = \mu + E_j + (x_i\beta^Q + \underline{G}_i^*) + (x_i\beta_j^{QxE} + \underline{GE}_i^*) + \underline{\epsilon}_{ij}$

  - x_i: genotypic covariate, DNA variation at a particular genomic position (molecular marker / SNP)
  - β^Q: QTL main effect
  - β_j^{QxE}: QTLxE effect

Environmental covariates & genotypic sensitivity

- Introduction of environmental covariates to model GxE interaction
  - Which environmental covariates to include and how?
  - Genotypic sensitivities to env. covariates explain GxE
- Base model
  - $\underline{y_{ij}} = \mu + E_j + \underline{G_i} + \underline{GE_{ij}} + \underline{\epsilon_{ij}}$
- Model with environmental covariate
  - $\underline{y_{ij}} = \mu + E_j + \underline{G_i} + \left(\beta_i^{GxE} z_j + \underline{GE_i^*}\right) + \underline{\epsilon_{ij}}$
  - β_i^{GxE}: genotypic sensivity
  - z_j: environmental characterization
  - $\underline{GE_i^*}$: residual GxE interaction

Example: CIMMYT drought stress in maize

- Response
  - Yield
- Environments
  - 8 trials = 8 managed stress environments, intermediate and severe drought stress (IS, SS), low and high nitrogen (LN, HN), no stress
    - 1992, 1994, 1996
    - 2 locations (Tl, PR)
    - Winter and summer seasons
- Genotypes
  - 211 F2 derived F3 lines
- Covariables
  - Genotypes
    - 132 marker loci
  - Environments
    - Min. and max. temperature, radiation, rain and number or sun hours for vegetative, flowering and grain filling stages

frontiers in PHYSIOLOGY

METHODS ARTICLE
published: 12 March 2013
doi: 10.3389/fphys.2013.00044

The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis

Marcos Malosetti[1], Jean-Marcel Ribaut[2] and Fred A. van Eeuwijk[1]*

[1] Biometris - Applied Statistics, Department of Plant Science, Wageningen University, Wageningen, Netherlands
[2] Consultative Group on International Agricultural Research Generation Challenge Programme, México DF, Mexico

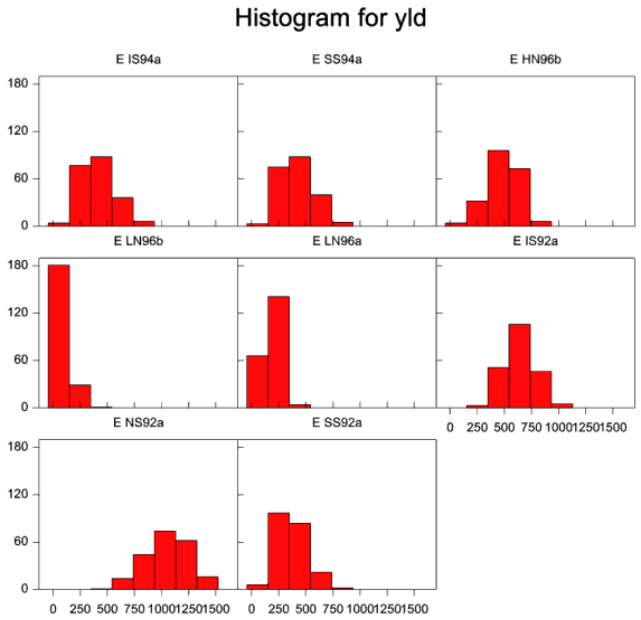
Theor Appl Genet (2006) 112: 1009–1023
DOI 10.1007/s00122-005-0204-z

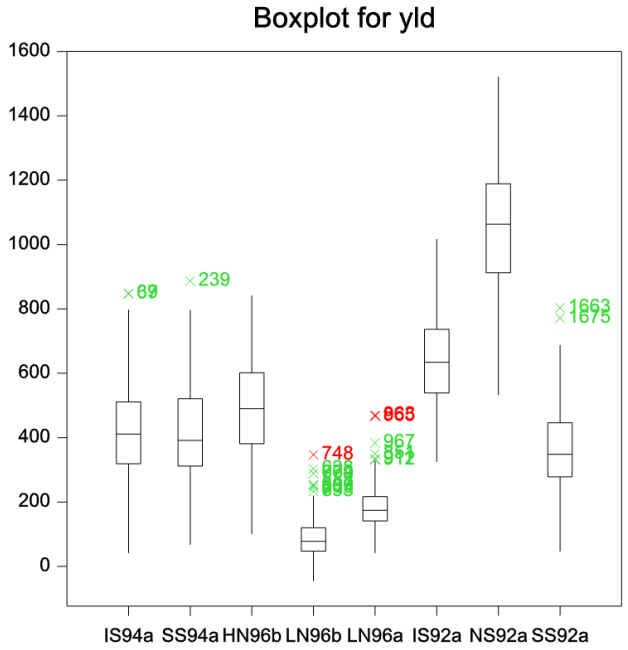
ORIGINAL PAPER

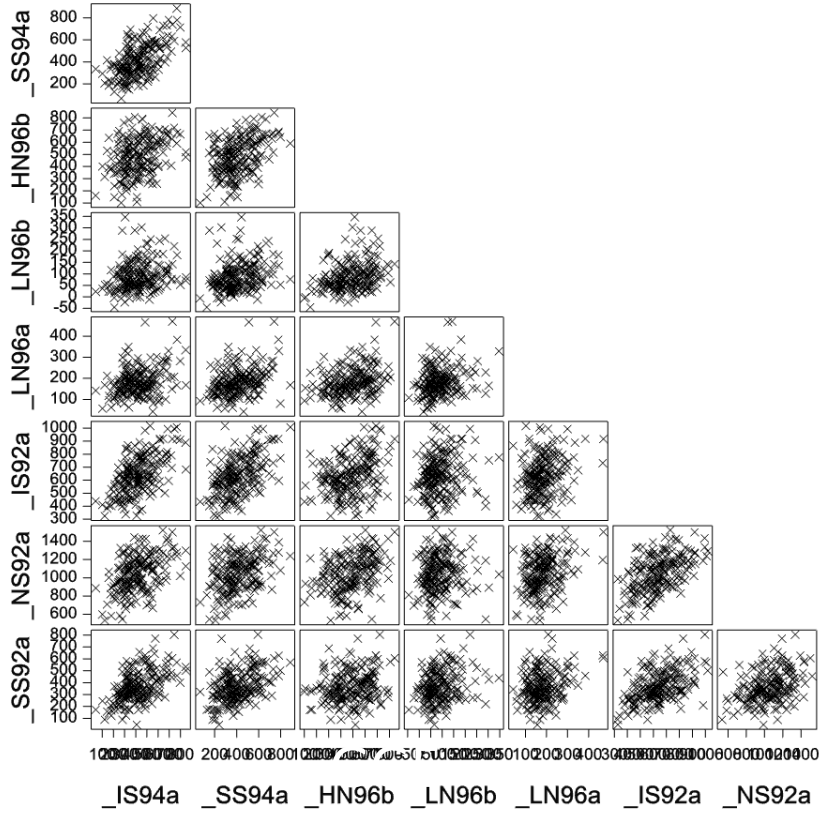
Mateo Vargas · Fred A. van Eeuwijk
Jose Crossa · Jean-Marcel Ribaut

Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods

GxE expresses itself in **heterogeneity of genetic variance and correlations** between trials (experiments)



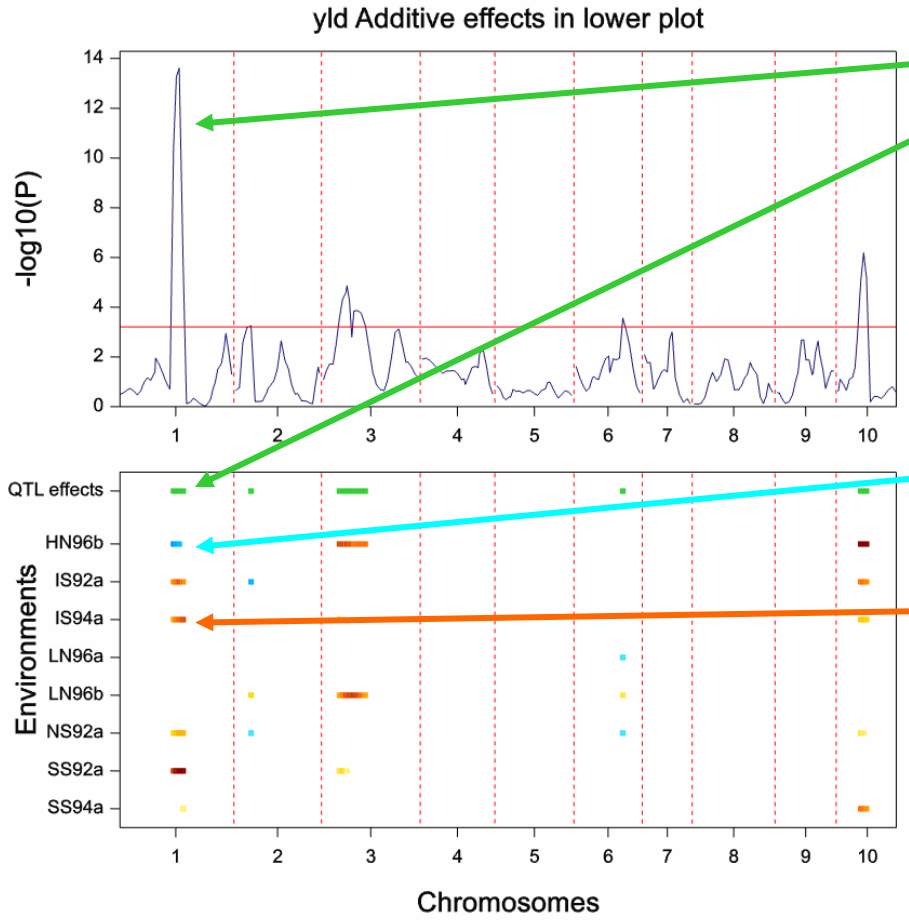




WAGENINGEN UNIVERSITY & RESEARCH

100years 1918 — 2018

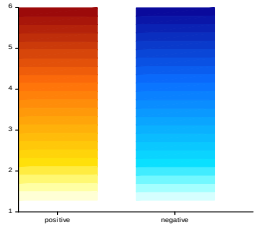
CIMMYT: QTL+QTLxE analysis for yield (CIM; VCOV = FA model)



Profile for environment specific QTLs

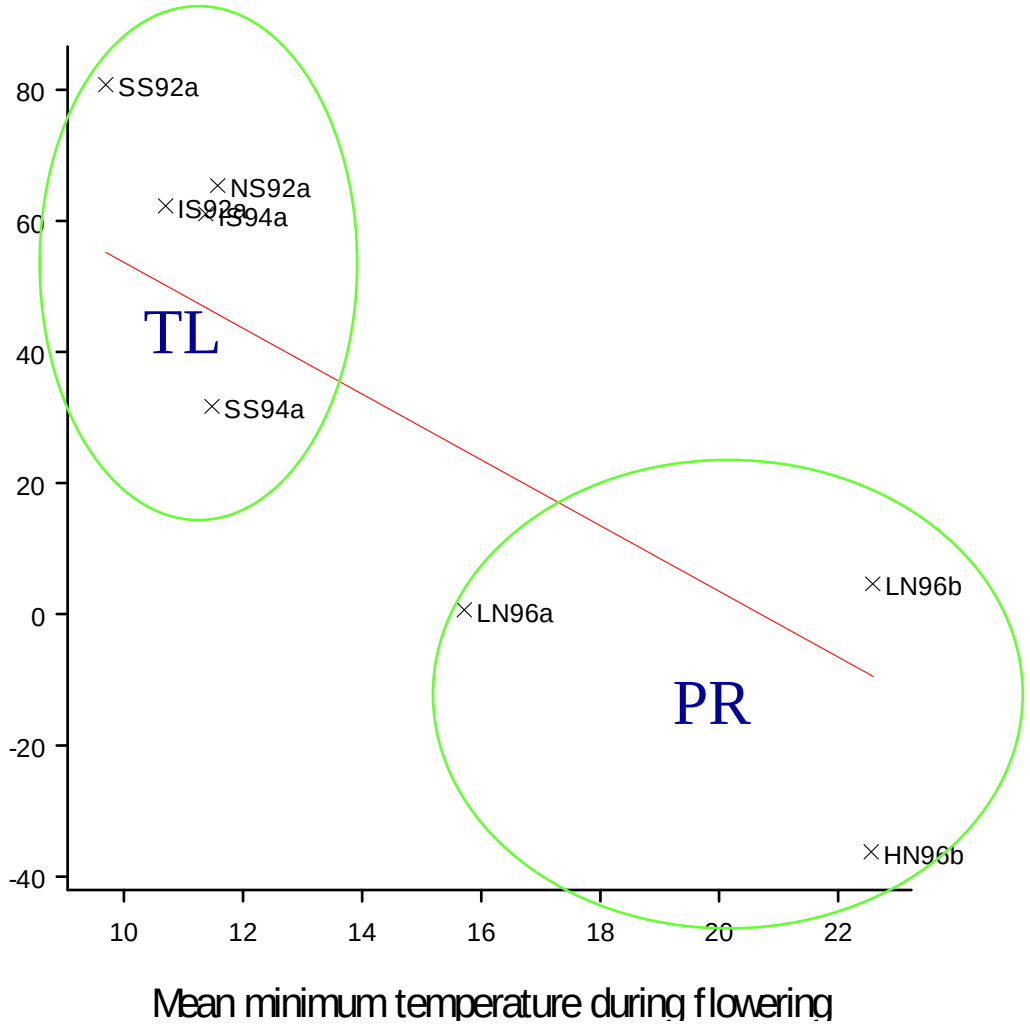
Positive effect P1 allele (dark/light blue)

Positive effect P2 allele (red/yellow)



Color code for p-values of QTL effects

Regression of QTLxE on min. temperature during flowering



$$\underline{y}_{ij} = \mu_j + \sum_{q=1}^{Q} x_{iq}\, \beta_{j,q}^{GGE} + \underline{G}_{ij} + \underline{\epsilon}_{ij}$$

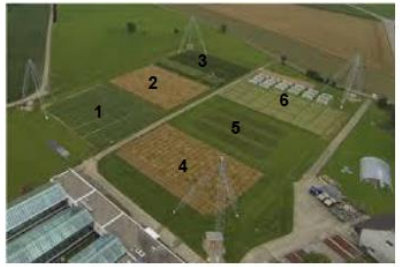
$$\underline{y}_{ij} = \mu_j + \sum_{q=1}^{Q} x_{iq}\, (\gamma_q + \delta_q z_j) + \underline{G}_{ij} + \underline{\epsilon}_{ij}$$

Types of phenotypic responses

- In plant biology and genetics, the most important traits (responses) are yield, biomass, and phenology (time to particular developmental stages like flowering time
- Yield and biomass are typically estimated at harvest, i.e., the end of the growing season, they are end-point traits
- Developmental stages are far more difficult to measure
- Because of a revolution in the availability of new measuring devices (unmanned aerial vehicles, drones, proximate sensing, remote sensing, field robots, phenotyping platforms) many new plant traits can be measured at high spatial and time resolution
- The collection and analysis of these new phenotypes is called phenomics
- These new traits themselves are often called secondary phenotypes

Longitudinal modelling of phenotyping data

Hierarchical splines for ETH FIP wheat



(b) FIP platform (ETH Zürich)

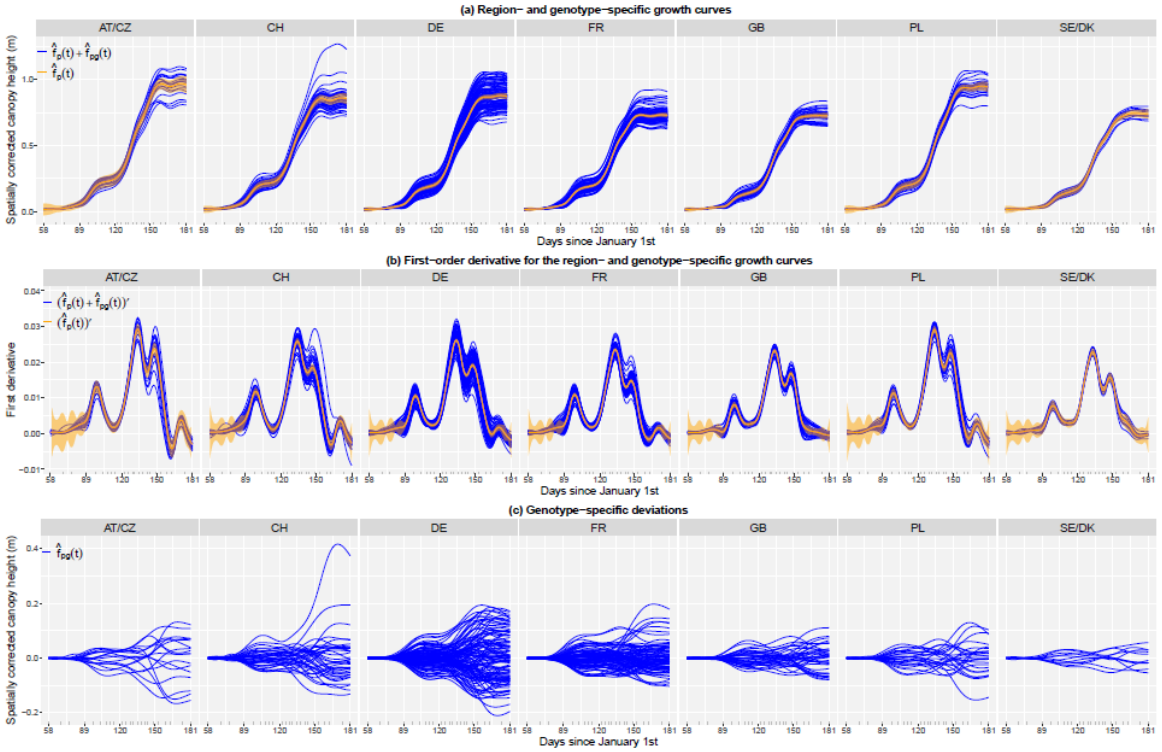


Figure 6. Results of the second stage of analysis for the ETH field phenotyping platform: (a) estimated region (orange) and genotype-specific (blue) growth curves, (b) estimated region (orange) and genotype-specific (blue) first-order derivatives, and (c) estimated genotype-specific deviations. In (a) and (b) the orange shaded areas denote 95% pointwise confidence intervals at the region level. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

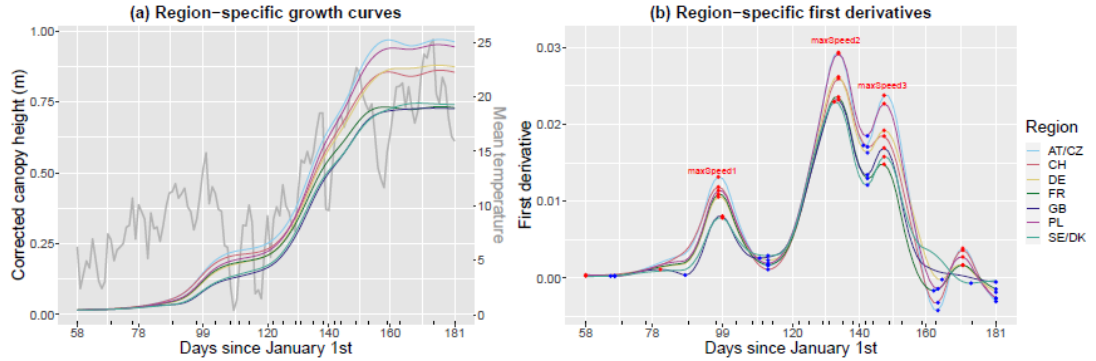


Figure 7. Results of the second stage for the ETH field phenotyping platform: (a) region-specific growth curves (coloured lines) vs. mean temperature (grey line), and (b) region-specific first-order derivatives; blue and red points indicate (local) minima and maxima, respectively. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

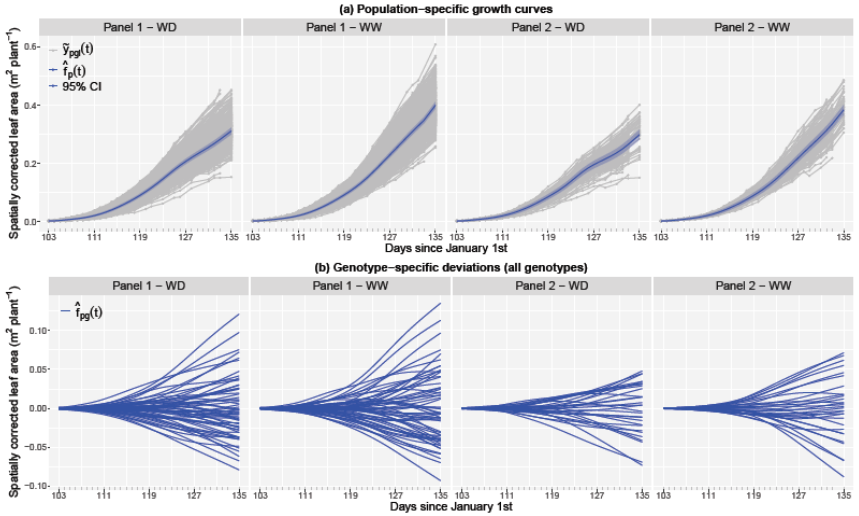
scientific reports

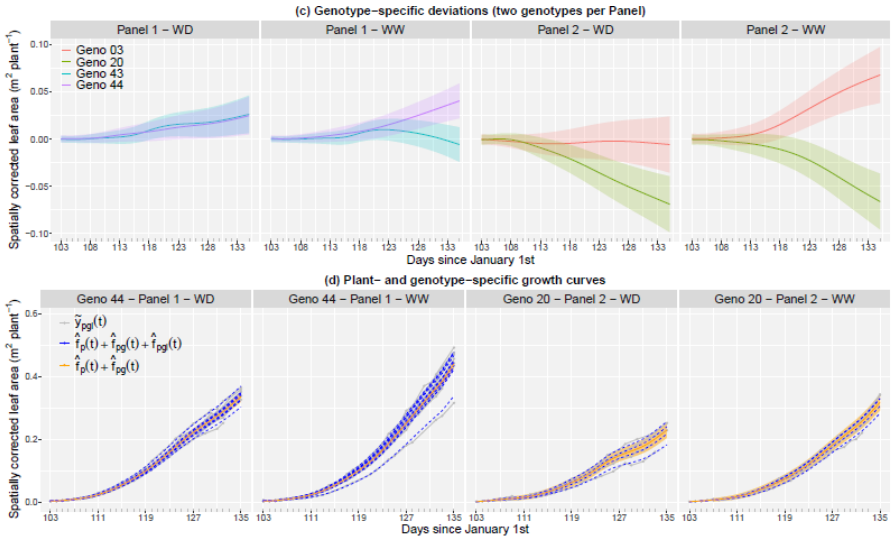
Check for updates

OPEN

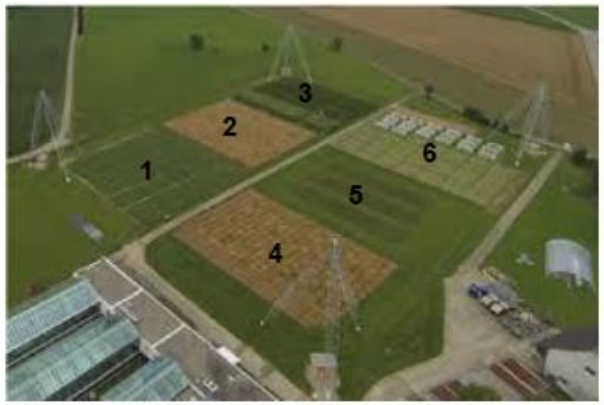
A two-stage approach for the spatio-temporal analysis of high-throughput phenotyping data

Diana M. Pérez-Valencia[1,2✉], María Xosé Rodríguez-Álvarez[1,3,7], Martin P. Boer[4], Lukas Kronenberg[5,6], Andreas Hund[5], Llorenç Cabrera-Bosquet[8], Emilie J. Millet[4,8] & Fred A. van Eeuwijk[4]





(a) PhenoArch platform (INRAE Montpellier)



(b) FIP platform (ETH Zürich)

$$y_{ijk}(t) = \mu(t) + f_j^{Management}(t) + f_{ij}^{Genotype}(t) + f_{ijk}^{Plant}(t) + \varepsilon_{ijk}(t)$$

$$\tilde{\mathbf{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon},\ \boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{G}),\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{W}),$$

where

$$\boldsymbol{X} = [\boldsymbol{Q}_{\text{pop}} \otimes \boldsymbol{X}_{\text{pop}}],$$
$$\boldsymbol{Z} = [\boldsymbol{Q}_{\text{pop}} \otimes \boldsymbol{Z}_{\text{pop}} \mid \boldsymbol{Q}_{\text{gen}} \otimes \boldsymbol{X}_{\text{gen}} \mid \boldsymbol{Q}_{\text{gen}} \otimes \boldsymbol{Z}_{\text{gen}} \mid \boldsymbol{I}_M \otimes \boldsymbol{X}_{\text{plant}} \mid \boldsymbol{I}_M \otimes \boldsymbol{Z}_{\text{plant}}] \text{ with } \boldsymbol{u} = (\boldsymbol{u}_{\text{pop}}^T, \boldsymbol{\beta}_{\text{gen}}^T, \boldsymbol{u}_{\text{gen}}^T, \boldsymbol{\beta}_{\text{plant}}^T, \boldsymbol{u}_{\text{plant}}^T)^T,$$

and

$$\boldsymbol{G} = \begin{pmatrix} \text{blockdiag}(\sigma_1^2 \boldsymbol{I}_{b_{\text{pop}}-2}, \ldots, \sigma_k^2 \boldsymbol{I}_{b_{\text{pop}}-2}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_L \otimes \begin{pmatrix} \boldsymbol{\Sigma}_{\text{gen}} & \mathbf{0} \\ \mathbf{0} & \sigma_{\text{gen}}^2 \boldsymbol{I}_{b_{\text{gen}}-2} \end{pmatrix} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_M \otimes \begin{pmatrix} \boldsymbol{\Sigma}_{\text{plant}} & \mathbf{0} \\ \mathbf{0} & \sigma_{\text{plant}}^2 \boldsymbol{I}_{b_{\text{plant}}-2} \end{pmatrix} \end{pmatrix}.$$

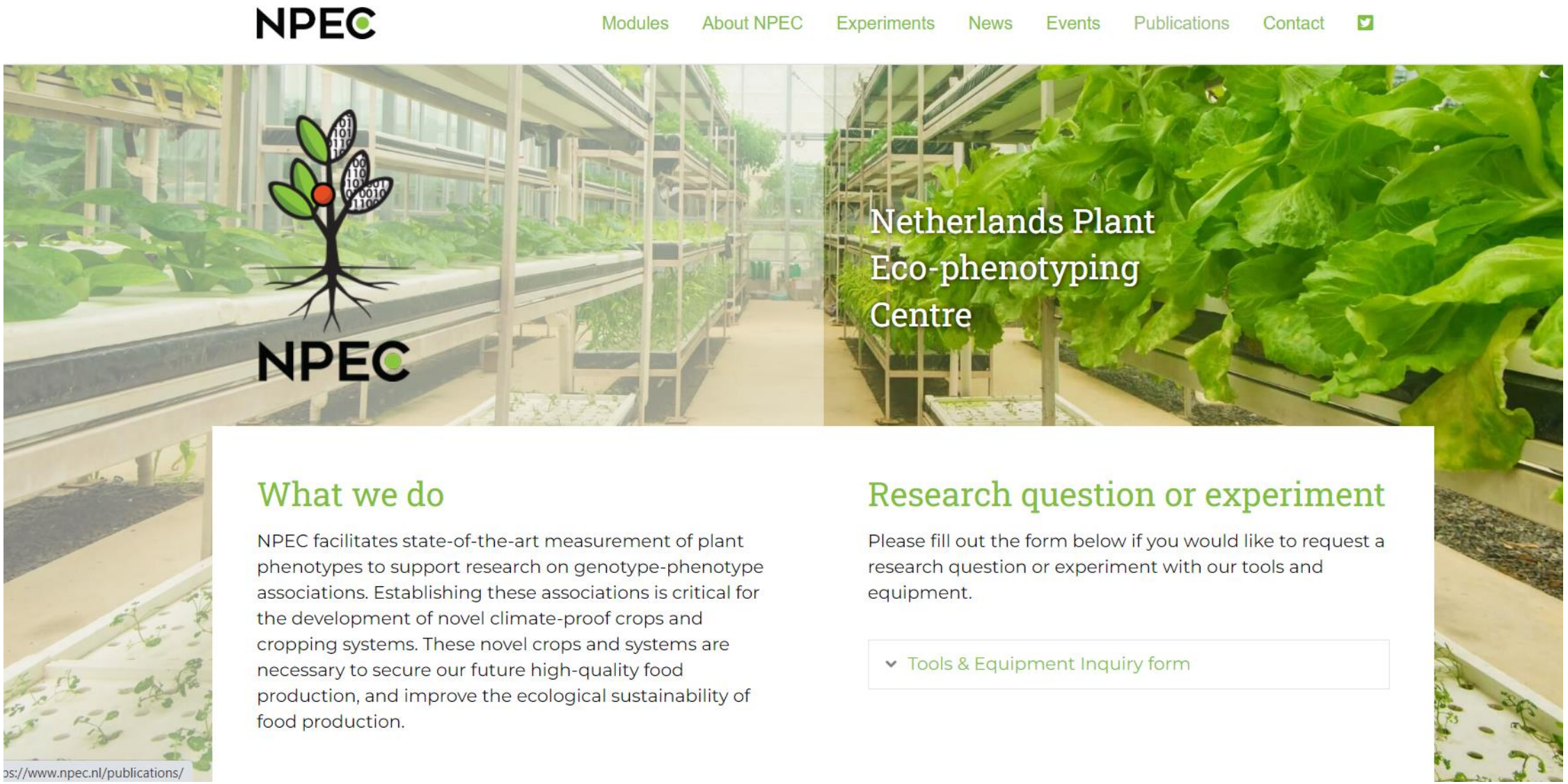
A two-stage approach for the spatio-temporal analysis of high-throughput phenotyping data

Diana M. Pérez-Valencia[1,2✉], María Xosé Rodríguez-Álvarez[1,3,7], Martin P. Boer[4], Lukas Kronenberg[5,6], Andreas Hund[5], Llorenç Cabrera-Bosquet[8], Emilie J. Millet[4,8] & Fred A. van Eeuwijk[4]

scientific reports

WAGENINGEN UR
For quality of life

https://www.npec.nl/



Summary

- Societal, environmental and consumer demands require development of new plant varieties with improved properties
- Genotype-to-phenotype models help identify the genetic basis of plant phenotypes and are instrumental in developing new plant varieties
- New phenotyping techniques make it possible to study plant development and behaviour in far more detail and with high time resolution
- These new phenotyping techniques can speed up the development of improved plant varieties
- Statistical methods are essential to support plant breeding efforts
- Popular methods are ANOVA, regression and linear mixed models. More advanced statistical models are necessary for modelling longitudinal trait information and genotype by environment interactions.
    - Splines, Generalized additive mixed models, differential equation models

References and links

To get an impression of phenotyping, you may have a look at the following web pages

- https://eppn2020.plant-phenotyping.eu/
- https://www.npec.nl/
- https://www.h2020-invite.eu/
- https://www.youtube.com/watch?v=2hOQPRb-z1A

To get an idea of which statistical methods are used look at:

- https://cran.r-project.org/web/packages/statgenHTP/vignettes/Overview_HTP.html

A review paper on statistical techniques in phenotyping can be found here (although not easy to read):

- https://doi.org/10.1016/j.plantsci.2018.06.018