Instance-based domain adaptation for Multi-Environment Trials data

François VICTOR

UMR MIA Paris-Saclay, INRAE, AgroParisTech

Supervisors:

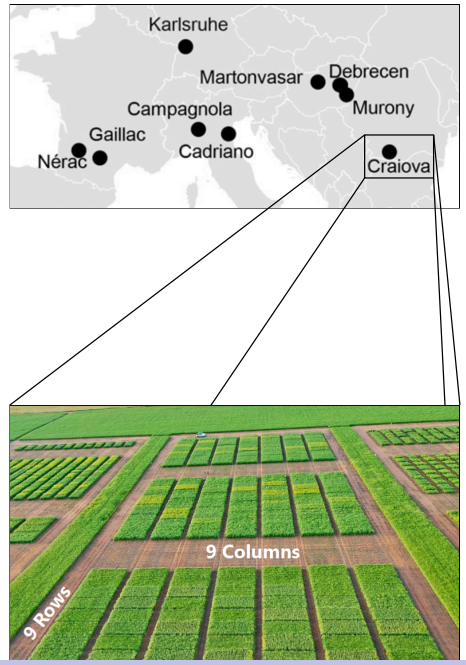
A.Charcosset, J.Chiquet, J-B.Léger, T.Mary-Huard

Context

Challenges in crop science: climate change and agro-ecology :

- Yield instability due to climate variability, biotic and abiotic stress.
- Agro-ecological practices : low inputs, pesticides,
- Need to identify adapted varieties for future climates and practices.

Multi-Environment Trials (MET): Evaluate genotypes across diverse locations and years.



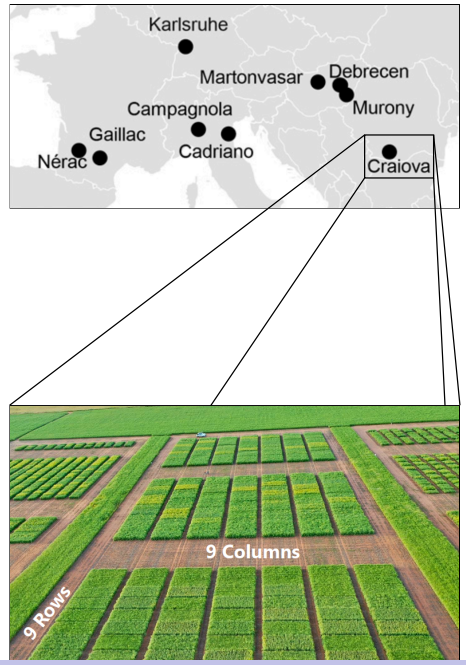
- **DROPS** (DROught-tolerant yielding PlantS)
- 256 maize hybrids
- 29 environments (Europe)
- Genotype $\times$ Environment ($G \times E$) interactions

Context

Challenges in crop science: climate change and agro-ecology :

- Yield instability due to climate variability, biotic and abiotic stress.
- Agro-ecological practices : low inputs, pesticides,
- Need to identify adapted varieties for future climates and practices.

Multi-Environment Trials (MET): Evaluate genotypes across diverse locations and years.



Field layout

Rep1						Rep2						Rep3					
G22	G02	G18	G08	G01	G09	G04	G23	G17	G01	G22	G16	G19	G08	G06	G23	G24	G07
G05	G20	G16	G19	G07	G24	G14	G03	G21	G10	G13	G06	G14	G09	G04	G10	G16	G20
G04	G10	G14	G03	G15	G12	G20	G15	G11	G09	G18	G07	G01	G15	G18	G13	G22	G05
G11	G21	G23	G13	G17	G06	G08	G24	G12	G05	G02	G19	G11	G02	G17	G12	G21	G03

Modeling Yield in MET

Linear Mixed Model (LMM) for a Randomized Complete Block (RCB) design:

$$y = X\beta + Zu + \epsilon$$

- y: vector of phenotypic observations ($n \times 1$).
- X, Z: incidence matrices for fixed and random effects.
- β: vector of fixed effects coefficients.
- u: vector of random genetic values, with $u \sim \mathcal{N}(0, \sigma_g^2 A)$.
- ϵ: vector of residual errors, with $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I)$.

Henderson's Mixed Model Equations (MME):

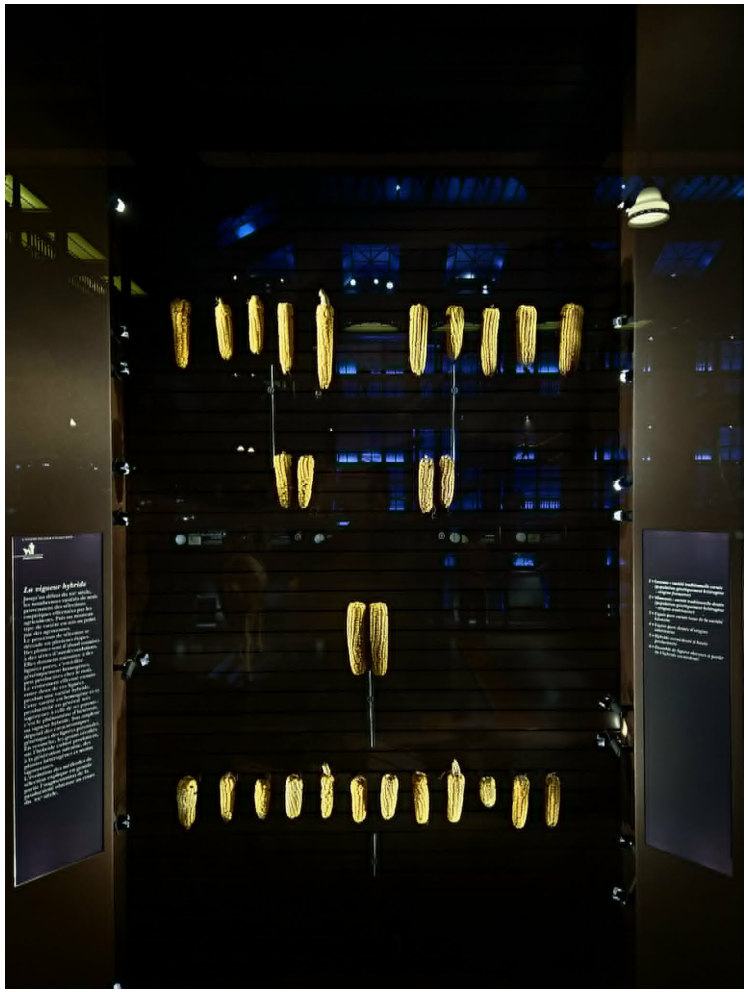
$$\begin{bmatrix} X^\top X & X^\top Z \\ Z^\top X & Z^\top Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^\top y \\ Z^\top y \end{bmatrix}$$

- λ: ratio of variance components σ_e^2/σ_g^2 (shrinkage factor).
- $\hat{\beta}, \hat{u}$: Best Linear Unbiased Estimators (BLUEs) and Predictions (BLUPs).
- A: relationship matrix (pedigree).

a small tour in the evolution gallery

a small tour in the evolution gallery

Variété *Lacaune* -
cornée
origine française



Variété *Minnesota* -
dentée
origine américaine

Genomic Relationship Matrix (GRM) aka Kinship matrix

Genotyping via Single Nucleotide Polymorphisms (SNPs): $W \in \{0, 1, 2\}^{n \times m}$. **Construction (VanRaden, 2008):**

$$K = \frac{ZZ^\top}{2\sum_{j=1}^{m} p_j(1 - p_j)}$$

where $Z_{ij} = W_{ij} - 2p_j$.

- K: genomic relationship matrix of size $n \times n$.
- W: matrix of observed allelic dosages.
- p_j: reference allele frequency at locus j.
- Z: centered marker matrix adjusted for allele frequencies.
- m: total number of marker loci.

Dual Interpretation of Genomic Prediction

Ridge Regression (RR-BLUP): $y = X\beta + Za + \epsilon$ with $a \sim \mathcal{N}(0, \sigma_a^2 I)$.

- a: vector of marker effects ($m \times 1$).
- σ_a^2: variance component associated with marker effects.
- I: identity matrix.

Equivalence: $y = X\beta + u + \epsilon$ where $u = Za$ and $\mathbb{V}[u] = ZZ^\top \sigma_a^2$.

- u: vector of total genetic values.
- $\mathbb{V}$: variance-covariance operator.

Motivations for Domain Adaptation

- MET are extremely expensive and time-consuming.
- Large datasets exist (e.g., G2F in the US), but they observe US hybrids in US environments.
- **Challenge**: Are US-trained models relevant for predicting European "ideotypes" or adapting to new climate scenarios?

Issue: Potential violation of the IID assumption.

$$P_{\text{train}}(G, E, Y) \neq P_{\text{test}}(G, E, Y)$$

- $P_{\text{train}}, P_{\text{test}}$: probability measures of the training and test domains.
- $G \in \mathcal{G}$: genotype space and $E \in \mathcal{E}$: environment space.
- Y: random variable representing the phenotypic response (yield).

Domain Adaptation Framework

$$\mathcal{D} = \{X \in \mathcal{X}, P(X)\}\text{: domain}$$
$$\mathcal{T} = \{Y \in \mathcal{Y}, P(Y|X = x)\}\text{: task}$$

Problem: Domain Shift.

- Source Domain $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s} \sim P_S(X, Y)$
- Target Domain $\mathcal{D}_T = \{x_j^t\}_{j=1}^{n_t} \sim P_T(X)$

Covariate Shift Hypothesis: $p_S(x) \neq p_T(x)$ but $p_S(y|x) = p_T(y|x)$.

Goal: Minimize the target risk $\mathcal{R}_T(h) = \mathbb{E}_{P_T}[\ell(y, h(x))]$.

- $\mathcal{R}_T$: risk function associated with the target distribution.
- h: predictive hypothesis or decision function.
- ℓ: loss function quantifying the discrepancy between prediction and observation.
- $\mathbb{E}_{P_T}$: expectation operator with respect to the target distribution.

Target Risk Minimization

$$
\begin{aligned}
\mathcal{R}_T(h) &= \iint \ell(y, h(x)) p_T(x, y) dx dy \\
&= \iint \ell(y, h(x)) \frac{p_T(x, y)}{p_S(x, y)} p_S(x, y) dx dy \\
&= \mathbb{E}_{P_S}\left[\frac{p_T(x, y)}{p_S(x, y)} \ell(y, h(x))\right] \\
&\overset{\text{cov. shift}}{=} \mathbb{E}_{P_S}\left[\frac{p_T(x)\cancel{p_T(y|x)}}{p_S(x)\cancel{p_S(y|x)}} \ell(y, h(x))\right] \\
&= \mathbb{E}_{P_S}\left[w(x)\ell(y, h(x))\right] \quad \text{where } w(x) = \frac{p_T(x)}{p_S(x)}
\end{aligned}
$$

- $w(x)$: importance weight.
- p_T, p_S: marginal density functions for the target and source domains.
- dx, dy: infinitesimal elements in the feature and label spaces.

$w(x)$ are the **Importance Weights**.

Importance Estimation: KLIEP

KLIEP (Kullback-Leibler Importance Estimation Procedure) (Sugiyama et al., 2007)

$$\widehat{w}(x) = \sum_{i=1}^{n_t} \alpha_i k(x, x_i^t)$$

- $\widehat{w}(x)$: importance weighting function estimator.
- α_i: non-negative expansion coefficients.
- k: kernel basis function (typically Gaussian RBF).
- x_i^t: i-th sample from the target domain.

Optimization Problem:

$$\max_{\alpha} \sum_{x_j \in X_T} \log \widehat{w}(x_j)$$

subject to $\mathbb{E}_{P_S}\left[\widehat{w}(x)\right] = 1$ and $\alpha_i \geq 0$.

Application to Genomic Data

We focus on the **Genotype Space** $\mathcal{X} = \mathcal{G}$.

Kinship-based Kernel:

$$d_K^2(x_i, x_j) = K_{ii} + K_{jj} - 2K_{ij}$$

$$k(x_i, x_j) = \exp(-\gamma d_K^2(x_i, x_j))$$

- d_K^2: squared genetic distance between genotypes i and j.
- K_{ij}: genomic relationship coefficient between individuals i and j.
- γ: kernel bandwidth parameter.
- k: kernel similarity measure used for importance estimation.

Experimental Strategy: Source/Target Split

Validation Strategy:

1. **Spectral Clustering** on the GRM laplacian to identify structured groups (sub-populations).
2. Select a cluster as the target domain.
3. **Split** the target cluster:
   - 1/2 labeled target (for testing).
   - 1/2 "proxy" source (to simulate the gap).

$\Rightarrow$ Mimics a realistic scenario where we want to predict a specific, genetically distinct sub-population.

Weighted G-BLUP Model

$$y = X\beta + Zu + e, \quad e \sim \mathcal{N}(0, W_{diag}^{-1}\boldsymbol{\sigma}_e^2)$$

- $W_{diag} = \text{diag}(w(x_1), \ldots, w(x_n))$: diagonal matrix of importance weights.
- $\boldsymbol{\sigma}_e^2$: residual error variance per environment.
- e: vector of weighted residuals.

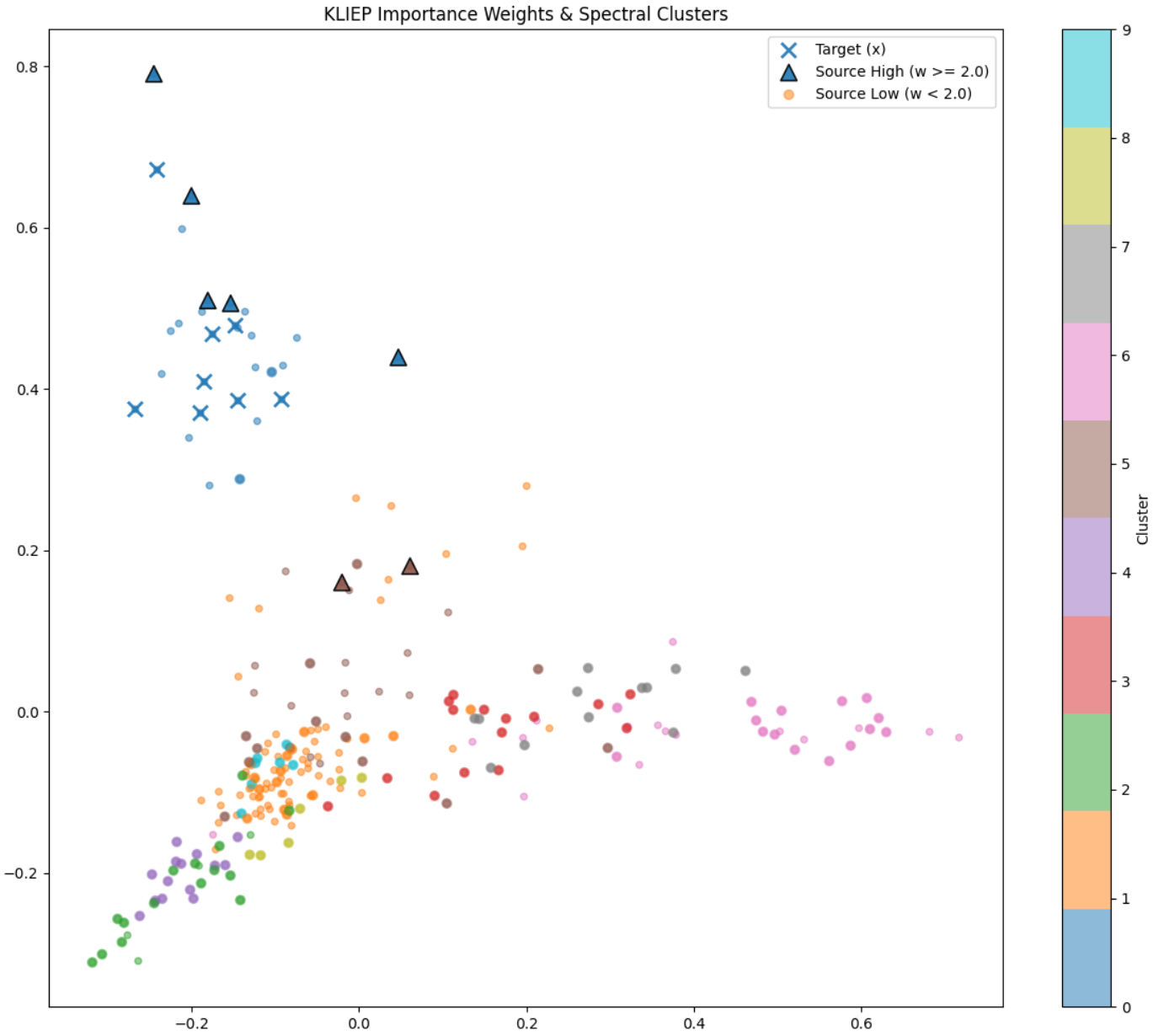
Weighted MME:

$$\begin{bmatrix} X^\top WX & X^\top WZ \\ Z^\top WX & Z^\top WZ + \lambda K^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^\top Wy \\ Z^\top Wy \end{bmatrix}$$

where $W = W_{diag}$. The weights W up-weight source observations similar to the target.

Results: Clustering and Weighting

Results: Visualisation of estimated importance weights

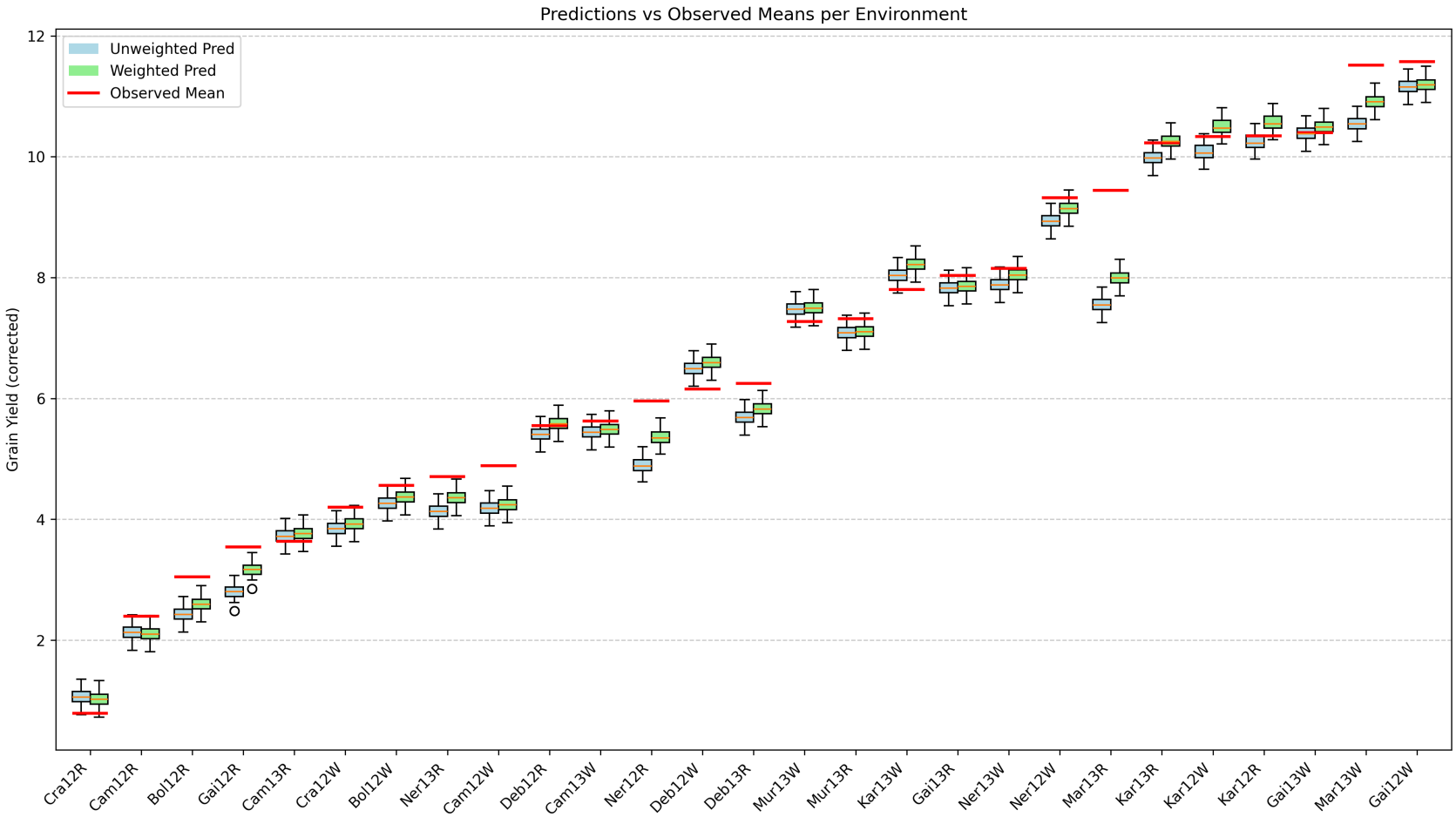


Results per environment

Experiment	Observed	Unweighted	(SE)	Weighted	(SE)
Bol12R	3.047	2.373	(0.069)	**2.548**	(0.080)
Bol12W	4.559	4.209	(0.077)	**4.321**	(0.086)
Cam12R	2.397	**2.073**	(0.057)	2.054	(0.071)
Cam12W	4.889	4.128	(0.087)	**4.193**	(0.100)
Cam13R	3.636	**3.666**	(0.057)	3.716	(0.070)
Cam13W	5.624	5.387	(0.066)	**5.440**	(0.083)
Cra12R	0.790	1.007	(0.060)	**0.975**	(0.071)
Cra12W	4.198	3.791	(0.075)	**3.878**	(0.082)
Deb12R	5.549	5.351	(0.057)	**5.536**	(0.071)
Deb12W	6.153	**6.439**	(0.065)	6.548	(0.077)
Deb13R	6.246	5.631	(0.060)	**5.781**	(0.074)
Gai12R	3.542	2.721	(0.076)	**3.098**	(0.089)
Gai12W	11.572	11.102	(0.073)	**11.143**	(0.082)
Gai13R	8.035	7.773	(0.058)	**7.809**	(0.072)
Gai13W	10.396	10.329	(0.077)	**10.444**	(0.082)
Kar12R	10.347	**10.200**	(0.078)	10.526	(0.083)
Kar12W	10.331	10.033	(0.085)	**10.457**	(0.093)
Kar13R	10.232	9.925	(0.078)	**10.209**	(0.084)
Kar13W	7.804	**7.981**	(0.092)	8.170	(0.100)
Mar13R	9.445	7.495	(0.081)	**7.947**	(0.091)
Mar13W	11.517	10.488	(0.116)	**10.861**	(0.122)
Mur13R	7.317	7.031	(0.070)	**7.057**	(0.082)
Mur13W	7.271	**7.420**	(0.089)	7.450	(0.098)
Ner12R	5.955	4.855	(0.059)	**5.327**	(0.079)
Ner12W	9.319	8.880	(0.066)	**9.097**	(0.077)
Ner13R	4.704	4.075	(0.058)	**4.310**	(0.073)
Ner13W	8.151	7.824	(0.064)	**7.996**	(0.075)

Table: Average grain yield predictions (BLUEs) vs Observed mean per environment.

Results per environment



Conclusion & Perspectives

- **Conclusion**:
  - Gain in prediction performance : positive transfer learning.
  - Proof of concept that validates the covariate shift assumption.

- **Perspectives**:
  - Application to more structured population.
  - Extend to $G \times E$.
  - Combine with Feature-based DA (representation learning).
  - Application to large-scale cross-continental datasets (US $\rightarrow$ EU).

Thank you for your attention !