Is the SpATS model as good as we would like it to be for the spatial analysis of field trials?

Lucas Peitton[1,2], Luciana Magnano[2], M. Valeria Paccapelo[1], M. Gabriela Borgognone[1]

[1]Department of Agriculture and Fisheries (DAF), QLD, Australia, [2]Universidad Nacional de Rosario (UNR), Santa Fe , Argentina

Background

Plant breeding programs aim to find genotypes with improved characteristics. Genotypes are tested in field trials where spatial variability is usually present (influencing their performance). Accounting for this variability in the data analysis is crucial for an efficient selection of the best genotypes. Generally this analysis is based on a laborious linear mixed model (LMM) approach [1]. A novel and much simpler LMM approach, Spatial Analysis of field Trials with Splines (SpATS) [2], has been developed and it has exhibited similar performance to the standard model for large partially replicated sorghum trials [3]. In this research, the SpATS model's performance is evaluated under different scenarios, such as **different crops**, **different trial sizes** and **different experimental designs.**

Materials

- 14 **mungbean** & 17 **chickpea** breeding trials spanning 12 years; 90-1128 **plots**; 30-768 **genotypes**; fully and partially replicated designs.

Standard LMM model

LMM under the standard approach is

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z}_g\boldsymbol{g} + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{\xi} + \boldsymbol{e}$$

Term	Meaning	Design matrix
$\boldsymbol{\tau}$	Overall mean and additional fixed effects	$\boldsymbol{X}$
$\boldsymbol{g}$	Random genotypic effects; $\sim\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$	$\boldsymbol{Z}_g$
$\boldsymbol{u}$	Other random effects (e.g. row or column); $\sim\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{U})$	$\boldsymbol{Z}_u$
$\boldsymbol{\xi}$	Correlated errors; $\sim\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R})$	–
$\boldsymbol{e}$	Independent errors; $\sim\boldsymbol{N}(\boldsymbol{0}, \sigma_e^2\boldsymbol{I})$	–

- Separable variance structure for an auto-regressive process across rows and columns in the field.
- Fixed and random effects to account for different spatial trends.
- Multi-step model fitting procedure based on residual diagnostic plots (*Fig. 1*) and formal tests to obtain the best spatial model.
- Different models likely to be fitted for different trials.
- Paid license (ASReml-R)[4].

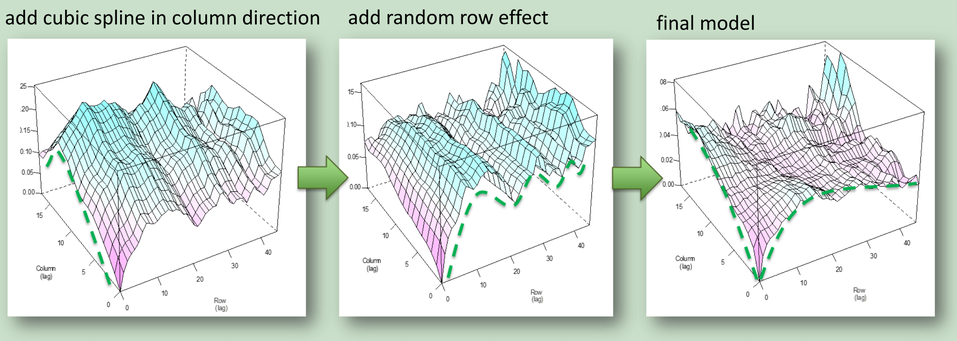

Fig. 1. Sample variograms for yield for a chickpea trial. In each step, the researcher adjusts an effect (fixed or random) according to the observed trends in the residuals and evaluates them through formal tests.

New SpATS model

LMM under the new approach is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \underbrace{\boldsymbol{X}_s\boldsymbol{\beta}_s + \boldsymbol{Z}_s\boldsymbol{s}}_{\text{fitted spatial surface}} + \boldsymbol{Z}_g\boldsymbol{g} + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{e}$$

- Penalised splines function fits the spatial surface under PS-ANOVA parametrization [5], decomposing the spatial effects, $\boldsymbol{s}$, into five mutually independent spatial components.
- Smoothing parameters automatically adjust degree of penalty.
- Single-step model fitting procedure (*Fig. 2*).
- A general model is fitted for different trials; the number of knots must be chosen.
- Free R-package (SpATS) [6].

Term	Meaning	Design matrix
$\boldsymbol{\beta}$	Overall mean and additional fixed effects	$\boldsymbol{X}$
$\boldsymbol{\beta}_s$	Fixed spatial effects **(unpenalised)**	$\boldsymbol{X}_s$
$\boldsymbol{s}$	Random spatial effects **(penalised)**; $\sim\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{S})$	$\boldsymbol{Z}_s$
$\boldsymbol{g}$	Random genotypic effects; $\sim\boldsymbol{N}(\boldsymbol{0}, \sigma_g^2\boldsymbol{I})$	$\boldsymbol{Z}_g$
$\boldsymbol{u}$	Random row, column & block effects; $\sim\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{U})$	$\boldsymbol{Z}_u$
$\boldsymbol{e}$	Independent errors; $\sim\boldsymbol{N}(\boldsymbol{0}, \sigma_e^2\boldsymbol{I})$	–

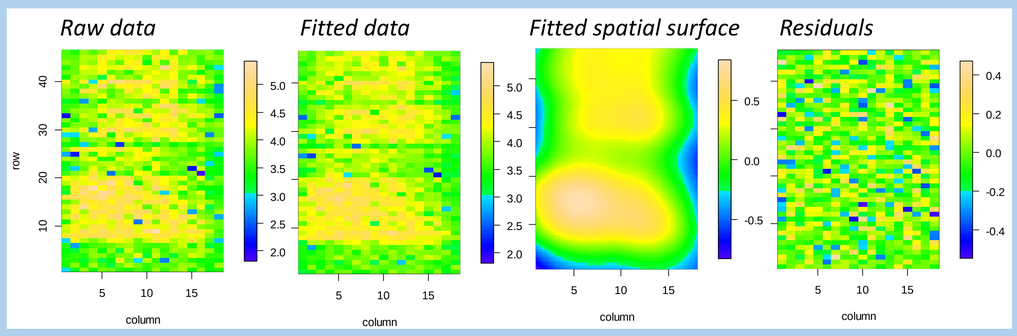

Fig. 2. Higher yield patches in the middle of the field and lower yield at the sides due to spatial trends (raw data). The fitted data is accurate due to the efficiently fitted spatial surface. No remaining trends can be seen in the residuals.

Number of parameters for the fitted spatial surface & penalty

- Under the PS-ANOVA parametrization, $\boldsymbol{s} = \text{vec}(\boldsymbol{s}_k)\ \ k = 1, \dots, 5$.
- Smoothing parameters tune the degree of smoothing and are estimated by REML [7];
  $\lambda_{s_k} = \sigma_e^2/\sigma_{s_k}^2 = \text{residual variance/the k}^{\text{th}}\text{ spatial effect variance}$
  - Complex spatial trends ($\sigma_{s_k}^2 \to \infty$), $\lambda_{s_k} \to 0 \Rightarrow$ lower penalty $\Rightarrow$ more parameters.
  - Smooth spatial trends ($\sigma_{s_k}^2 \to 0$), $\lambda_{s_k} \to \infty \Rightarrow$ higher penalty $\Rightarrow$ less parameters.
- The degree of flexibility of the function depends on the number of knots, which depend on the number of rows and columns of the trial;
  - If the number of knots > needed $\Rightarrow$ λ_{s_k}'s prevent overfitting.
  - If the number of knots < needed $\Rightarrow$ λ_{s_k}'s do not provide extra flexibility.
- It is preferable to choose a number of knots that exceeds the needed to provide the function enough flexibility [8]. This does not result in more computational time for the analysis of field trials.

Results

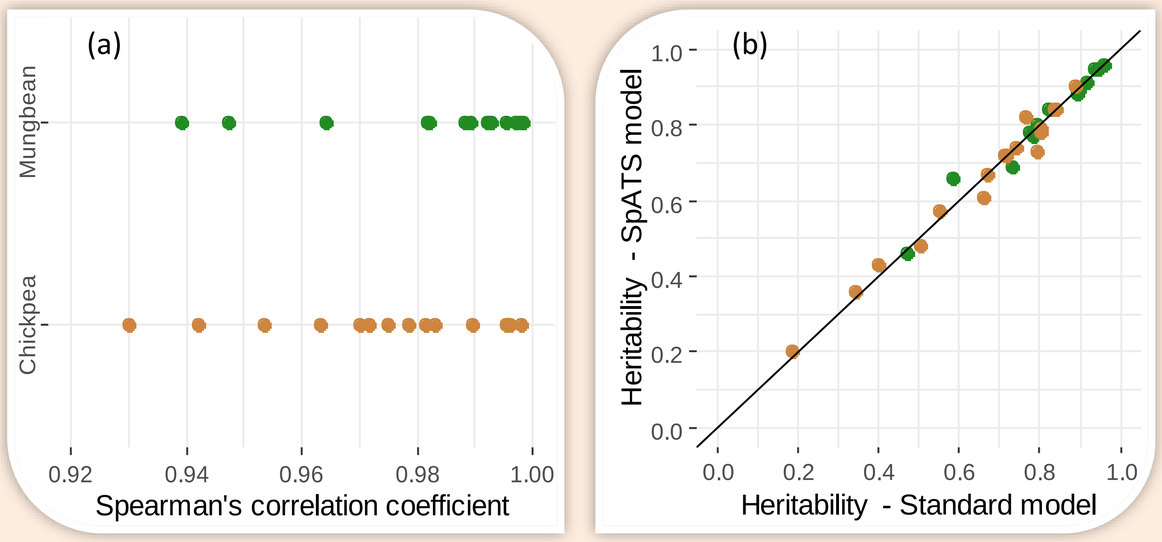

Fig 3. (a) The correlation between the ranking of predicted genotypic effects for both models ranged between 0.93 & 0.99 (for both crops), showing a strong agreement on ranking of genotypic predictions. **(b)** Similarity in computed heritability (the dots are very close to the identity diagonal) for both crops (green, mungbean; orange, chickpea), reflects consistency in trial precision.

Further research for the SpATS model

- Analysis of Multi-Environment Trial data.
- Incorporation of available genetic information (e.g. pedigree).

Conclusions

- SpATS model showed consistent results to the Standard approach in many different scenarios. No differences for different experimental designs or trial size.
- **More flexible:** Spearman correlations for chickpea trials were not as strong as correlations for mungbean trials (*Fig. 3a*) due to more complex trends, which SpATS likely accounted for more accurately than Standard model.
- **Greater time-efficiency:** A general model (given the number of knots) is obtained within seconds.
- **Non-subjective single-step process:** The fitted spatial surface is sufficiently flexible to account for the spatial variability regardless of its nature and the penalty is applied automatically.
- **Publicly available:** SpATS is freely available for researchers who cannot afford the license for the Standard approach package (ASReml-R [4]).
- **Anyone can use SpATS:** No LMM knowledge is required to fit the model and obtain genotypic predictions.

Acknowledgments The authors would like to gratefully acknowledge the Grains Research and Development Corporation of Australia for the financial support of the **DAQ1905-003RTX** project.

References

[1] Gilmour, Cullis & Verbyla (1997) J Agric Bio Environ Stat 2, 269-293. [2] Rodríguez-Álvarez, Boer, van Eeuwijk & Eilers (2016) arXiv:1607.08255 [stat.ME]. [3] Velazco, Rodríguez-Álvarez, Boer, Jordan, Eilers, Malosetti & van Eeuwijk (2017) Theor Appl Genet 130, 1375-1392. [4] Butler, Cullis, Gilmour and Gogel (2009) ASReml-R Reference Manual. Technical Report 3, Queensland Department of Agriculture, Fisheries & Forestry. [5] Lee, Durbán & Eilers (2013) Computational Statistics and Data Analysis 61, 22-37. [6] Rodríguez-Álvarez, Boer, Eilers & van Eeuwijk (2016) SpATS R-package. [7] Patterson & Thompson (1971) Biometrika 31, 100-109. [8] Eilers, Marx & Durbán (2015) SORT 30, 149-186.

Queensland Government

UNIVERSIDAD NACIONAL DE ROSARIO

GRDC GRAINS RESEARCH & DEVELOPMENT CORPORATION