in silico plants, 2023, **5**, 1–8,
https://doi.org/10.1093/insilicoplants/diad008
Advance Access Publication 27 June 2023
Original Article

in silico Plants

OXFORD

Original Article

Using copula graphical models to detect the impact of drought stress on maize and wheat yield

Sjoerd Hermes[1,2], Joost van Heerwaarden[1,2] and Pariya Behrouzi[1,*]

[1]Mathematical and Statistical Methods, Wageningen University, Wageningen, The Netherlands
[2]Plant Production Systems, Wageningen University, Wageningen, The Netherlands

*Corresponding author's e-mail address: Pariya Behrouzi, Mathematical and Statistical Methods, Wageningen University, Wageningen, The Netherlands; E-mail: pariya.behrouzi@wur.nl

Citation: Hermes S, van Heerwaarden J, Behrouzi P. 2023. Using copula graphical models to detect the impact of drought stress on maize and wheat yield. *In Silico Plants* **2023**: diad008; doi: 10.1093/insilicoplants/diad008

Handling editor: Daniela Bustos-Korts

ABSTRACT

Improving crop yields is one of the main goals of agronomy. However, yield is determined by a complex interplay between Genotypic, Environmental and Management factors (G×E×M), which varies across time and space. Therefore, identifying the fundamental relations underlying yield variation is a principal aim of agricultural research. A narrow, and not necessarily appropriate, set of statistical methods tends to be used in the study of such relations, which is why we aim to introduce a diverse audience of agronomists, production ecologists, plant breeders and others interested in explaining yield variation to the use of graphical models. More specifically, we wish to demonstrate the usefulness of copula graphical models for heterogeneous mixed data. This new statistical learning technique provides a graphical representation of conditional independence relationships within data that is not necessarily normally distributed and consists of multiple groups for environments, management decisions, genotypes or abiotic stresses such as drought. This article introduces some basic graphical model terminology and theory, followed by an application on Ethiopian maize and wheat yield undergoing drought stress. The proposed method is accompanied with the R package `heteromixgm` https://CRAN.R-project.org/package=heteromixgm.

KEYWORDS: Drought stress; heterogeneous data; mixed graphical models; on-farm data; yield.

1. INTRODUCTION

Crop yields are determined by the combination of Genotypes, Environments and Management techniques (G×E×M) that together define the production system and growing conditions. Part of the difficulty in optimizing the production of a certain crop is due to the strong variability of yield and its determinants across space and time (Landau *et al.* 2000; Lobell *et al.* 2009; Lischeid *et al.* 2022). Therefore, models fitted on certain conditions lack generalizability to other locations or seasons (Ronner *et al.* 2016). This is especially true for the statistical methods most commonly used to model yield variability, such as linear mixed or regression models and random forests (Bielders and Gérard 2015; Everingham *et al.* 2016; Niang *et al.* 2017), which generally incorporate large sets of marginally correlated explanatory variables, whose relationships to the response variable and to each other may be sensitive to changes in conditions. In practical terms, this limitation means that variables that are associated with yield variability under favourable conditions, may not show any relation under environmental stress. There is thus a clear need for methods that can identify both stable and condition-specific relations between potential explanatory variables and yields, while being less sensitive to spurious associations caused by co-linearity among explanatory variables. Such improved methods become increasingly pertinent, now that anthropogenic climate change (Dai 2013) may cause increasing stress and drought conditions (Dietz *et al.* 2021) that will call for changes to the production system using knowledge on what could enhancing yield under such novel conditions. Even though modern machine learning techniques such as random forests and neural networks have shown promising results in a predictive setting (Everingham *et al.* 2016; Khaki and Wang 2019), they lack interpretability, which is a large caveat to any method used to analyse data. The suitability of the method introduced in this article is derived from its interpretability and its ability to uncover direct dependence (fundamental) relationships that

underlay complex phenomena under contrasting conditions, by making use of partial rather than marginal correlations, and are therefore well suited to research questions involving the determinants of yield under abiotic stress conditions such as drought. Accordingly, uncovering such fundamental relationships that directly affect yield variability could provide insight into how better resistance against drought can be achieved.

This article introduces the copula graphical model for heterogeneous mixed data proposed in Hermes *et al.* 2022 to a diverse audience interested in the effects of drought stress on crop yield. To this end, in Section 2, we present the relevant statistical methodology. Section 3 presents an application of the model on maize and wheat data. Both of these contain non-drought and drought scenarios. Finally, the conclusion can be found in Section 4.

All functions pertaining to fitting and selecting copula graphical models for heterogeneous mixed data are presented in the R package `heteromixgm`, which is available on CRAN https://CRAN.R-project.org/package=heteromixgm.

2. METHODOLOGY

2.1. Gaussian graphical model

A Gaussian graphical model corresponds to a graph $G = (V, E)$ that represents the full conditional independence structure between variables represented by a set of vertices $V = \{1, 2, \dots, p\}$ through the use of a set of undirected edges E, and depends on a $n \times p$ data matrix with n observations and $\boldsymbol{X} = (X_1, X_2, \dots, X_p), X_j = (X_{1j}, X_{2j}, \dots, X_{nj})^T, j = 1, \dots, p$, where $\boldsymbol{X}$ is assumed to arise from a p-variate normal distribution $N_p(0, \Sigma)$. The inverse of the covariance matrix $\Sigma = \Theta^{-1}$ is known as the precision matrix, and contains the scaled partial correlations:

$$\rho_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}.$$

A partial correlation ρ_{ij} represents the dependence between X_i and X_j conditional on $X_{V\setminus ij}$ (all variables except for i and j). Therefore, $(i, j) \notin E$ if and only if $\Theta_{ij} = 0$.

2.2. Copula graphical models for heterogeneous mixed data

In typical agricultural datasets, the data are gathered under heterogeneous circumstances, that is, differing environmental, seasonal or management conditions which may affect the relations underlying yield. Consider, for example, on-farm trials where yields of a certain crop are measured at the end of a growing season across different locations (farms) in a country. Within a farm, we can assume that the observations arise from the same distribution, although not necessarily across farms. Fitting a Gaussian graphical model on the combined data would result in biased estimates, whereby farm-specific relations are not uncovered and might give the false impression that all discovered relations are the same across farms. In addition to the heterogeneous circumstances, the data are typically of mixed type, that is, a combination of Gaussian, non-Gaussian continuous, counts, ordinal or binomial data. Whereas variables like temperature or the amount of fertilizer applied might be normally distributed, the number of frost days during the growing season or the presence of livestock on the farm is not, as these are counts and binary variables, respectively. Applying a Gaussian graphical model to such mixed data violates the normal assumption, resulting in biased estimates.

To mitigate the problems mentioned above, we introduce the copula graphical model for heterogeneous mixed data (Hermes *et al.* 2022). The proposed model handles the heterogeneity by treating each differing condition as a different group, and in turn fitting a separate graphical model on each group to better account for the between-group heterogeneity in yield relations. The added value of applying the proposed method compared to fitting a different graphical model on each group 'by hand', is that the proposed method can borrow information across groups (through a penalty function) whenever the groups are similar to some extend. Therefore, when groups consist of only a few observations, borrowing information across groups potentially reduces bias in parameter estimates.

In order to draw valid inferences on mixed data, the Gaussian assumption underlying the Gaussian graphical model is bypassed by relying on the Gaussian copula, which treats each observed variable as if it represents some perturbation of a latent Gaussian variable. The latent variables are distributed as

$$Z^{(k)} \sim N_p(0, \Sigma^{(k)}),$$

where $\Sigma^{(k)} \in \mathbb{R}^{p\times p}$ represents the correlation matrix for group k, with $1 \leq k \leq K$. The j-th latent variable is linked to the observed data as

$$X_j^{(k)} = F_j^{(k)^{-1}}(\Phi(Z_j^{(k)})),$$

where the $F_j^{(k)}(\cdot)$ are non-decreasing marginal cumulative distribution functions, the $F_j^{(k)^{-1}}(\cdot)$ are quantile functions, $\Phi(\cdot)$ the standard normal cdf. Given that our interest lies in the dependencies encoded in the $\Theta^{(k)}$, we estimate the marginal cumulative distribution functions using a non-parametric approach, as the computational costs involved are substantial otherwise. These marginals are estimated as $\hat{F}_j^{(k)}(x) = \frac{1}{n_k+1}\sum_{i=1}^{n_k} \mathbb{I}(X_{ij}^{(k)} \leq x)$, where n_k is the sample size of group k.

By combining the multi-group setting with the Gaussian copula density, the following penalized log-likelihood function is obtained

$$\begin{aligned}\ell(\boldsymbol{\Theta}|\boldsymbol{Z}) = &\frac{1}{2}\sum_{k=1}^{K} n_k \log(\det(\Theta^{(k)})) \\ &- \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Z_i^{(k)^T}(\Theta^{(k)} - I)Z_i^{(k)} \\ &- \lambda_1 \sum_{k=1}^{K}\sum_{j\neq j'} \left|\theta_{jj'}^{(k)}\right| - \lambda_2 \sum_{k<k'}\sum_{j,j'} \left|\theta_{jj'}^{(k)} - \theta_{jj'}^{(k')}\right|\end{aligned} \quad (1)$$

for $1 \leq k \neq k' \leq K$ and $1 \leq j \neq j' \leq p$ and where $\boldsymbol{\Theta} = \{\Theta^{(1)}, \ldots, \Theta^{(K)}\}$. The log-likelihood function in Equation (1) appears similar to the typical log-likelihood for normally distributed data, but is made more complicated by the presence of latent variables and the two penalty terms, which will be discussed shortly. The method proposed in Hermes *et al.* (2022) is able to infer graph structures for a wide variety of data.

As direct maximization of the log likelihood is not feasible due to the non-existence of an analytic expression of Equation (1), an iterative method is needed to obtain the maximum-likelihood estimates for the various $\Theta^{(k)}$. Given that the Expectation Maximization (EM) algorithm tends to be the favoured method of maximizing a likelihood function containing latent variables, the proposed method also relies on this algorithm. The EM algorithm alternates between an E-step and an M-step. During the E-step, the expectation of the log-likelihood conditional on values for $\Theta^{(k)}$ and estimated intervals that discretize Z obtained during the previous M-step is computed. The algorithm requires the evaluation of a correlation matrix, which is not computed using its analytic expression, as this is computationally expensive. Instead, we resort to either an approximate method or a Gibbs method. The approximate method computes the diagonal (variance) elements through the second moment of the conditional $Z_{ij}|Y_i$, whilst the off-diagonal (covariance) elements are approximated using mean field theory (Guo *et al.* 2015; Behrouzi and Wit 2019). The model is fitted on data using the `heteromixgm()` function in the `heteromixgm` package with argument `method = 'Gibbs'` or `method = 'Approximate'` for either method, respectively. The former is slower and more accurate, whereas the latter is faster but slightly less accurate. We recommend researchers to stick to the Gibbs method whenever $p < 1000$, but resort to the approximate method when the number of variables exceeds this quantity. The rationale is that the computational cost of the Gibbs method becomes unfeasible when p exceeds 1000. These recommendations are conditional upon the grid over which penalty parameters are evaluated (see below). Denser grids require more computation time, making the approximate method more attractive.

The `heteromixgm()` function fits one model for each unique combination of λ_1 and λ_2 values supplied. These are penalty parameters, where higher values result in sparser (less edges) graphs and more similar graphs between groups, respectively, where the partial correlations are forced to equality. These two penalties are shown by the two rightmost terms in Equation (1). Typically, two vectors with values between 0 and 1 are supplied for `lambda1` and `lambda2` in the `heteromixgm()` function, as it is not a priori known which penalty value is best suited for the data. By supplying two vectors (possibly of length 1), the model fits $|\lambda_1| \times |\lambda_2|$ models, where $|\cdot|$ denotes the cardinality of a set.

Once these $|\lambda_1| \times |\lambda_2|$ models are obtained, it is time to select the 'best' model. Model selection is done using the `modselect()` function. Even though model selection can be done manually, based on prior assumptions (see Danaher *et al.* 2014)) or algorithmically (see Liu *et al.* 2010), the `heteromixgm` package only allows for the use of information criterion approaches. The rationale is that cross-validation-based approaches are computationally expensive, which would hinder the practical usage of the `heteromixgm` package, as the proposed method is computationally expensive for large p. In this article, we will use the Akaike information criterion (AIC) (Akaike 1973), which is of the following forms:

$$\begin{aligned}\text{AIC}(\lambda_1, \lambda_2) = \sum_{k=1}^{K} \Big[& n_k \text{tr}(S^{(k)} \hat{\Theta}^{(k)}_{\lambda_1,\lambda_2}) \\ & - n_k \log(\det(\hat{\Theta}^{(k)}_{\lambda_1,\lambda_2})) + 2\nu^{(k)}_{\lambda_1,\lambda_2} \Big],\end{aligned}$$

where the degrees of freedom $\nu^{(k)}$ represent the number of non-zero elements in the upper diagonal of the precision matrix, $\hat{\Theta}^{(k)}_{\lambda_1,\lambda_2}$ is the estimated precision matrix for group k using penalty parameters λ_1 and λ_2, and $S^{(k)}$ is the sample covariance matrix for group k. The values of λ_1 and λ_2 that minimize the AIC are the optimal penalty values and chosen as the values which are used to estimate the 'actual' graph. The `heteromixgm` package also allows for to the use of the Bayesian information criterion (BIC) (Schwarz 1978) and the extended Bayesian information criterion (EBIC) (Chen and Chen 2012). We chose for the AIC in this application due to the small sample sizes for each group, as, typically, using the BIC or EBIC would lead to a sparser and more stable network (Vujačić *et al.* 2015). However, we are interested in the true network and are not too worried about false negatives, especially given that we use a bootstrapping procedure to evaluate the stability of the graphs.

Due to the inherent uncertainty with respect to the graph structure, we suggest that a non-parametric bootstrap is used, in order to evaluate the uncertainty surrounding the estimated edges. The non-parametric bootstrap commences by permuting the data in the K groups B times, where random resampling of the data with replacement occurs, estimating the graph structure across all B resampled datasets, performing model selection and choosing a combination of λ_1 and λ_2 that minimizes the information criterion. This results in B graphs for all K groups. The last step consists of counting the number of times each estimated edge occurs across the B bootstraps for each group k and dividing this amount by B in order to obtain a probability or percentage that reflects the underlying uncertainty in the estimated graph, where a higher value indicates more certainty. In this analysis, we set $B = 120$.

Several R packages exist that allow for the application of graphical models. Three closely related packages will be mentioned here. The `huge` package (Jiang *et al.* 2021) offers a very efficient implementation of the Gaussian graphical model for a single group. The `netgwas` package by Behrouzi *et al.* (2023) moves beyond normality by implementing the copula transformation framework, but is again only suited for a single group. Finally, Danaher (2018)) developed the `JGL` package, which is suitable for fitting Gaussian graphical models on multi-group data. Yet none of these packages offers the advantages (lower bias for non-Gaussian variables and less variance for similar groups) of fitting graphical models on mixed multi-group data that the `heteromixgm` package offers.

3. RESULTS

The proposed copula graphical model was applied to maize and wheat data collected in Ethiopia by the Ethiopian Institute of Agricultural Research in collaboration with the International Maize and Wheat Improvement Center (Vasco Silva *et al.* ming). We use a subset of the data collected in Dugda (Oromia Region) which contains variables pertaining to soil properties, weather, management influences, external stresses and yield.

The data are grouped based on two variables: year and drought stress presence. We expect that these variables have a large impact on the underlying relationships, which required us to group samples together if we want to accurately infer the full conditional independence structure between the variables. Additionally, by grouping variables based on drought stress presence, we can evaluate what the influence is of drought stress on the underlying graph, granted all other conditions are equivalent to the situation without drought stress. Therefore, for each crop, four groups were created: data under drought stress in 2010 and 2013; data not under drought stress in 2010 and 2013 for maize; data under drought stress in 2009 and 2013; and data not under drought stress in 2009 and 2013 for wheat. Each group pertaining to maize yields consists of measurements across 61 variables with sample sizes 20, 66, 69 and 24, and each group pertaining to wheat yields consists of measurements across 58 variables with sample sizes 60, 9, 42 and 13. The model has shown excellent performance for low-dimensional ($n \gg p$) simulated data and real data, under varying network conditions, resulting in few false-positive and -negative edge discoveries for sample sizes of 100 and greater (Hermes *et al.* 2022). Whilst the copula graphical model outperforms competitors in high-dimensional ($p \gg n$) simulated settings, its performance on real data has not yet been assessed. As such, this article will be the first application of the copula graphical model on real-world high-dimensional data with unequal group sizes. Whilst typical agricultural datasets are not necessarily high dimensional, their effective dimensionality might increase whenever researchers expect local G × E × M interactions to be very specific to circumstances, in turn invalidating the typical i.i.d. assumption. The model is fitted on the data using the Gibbs method, with $\lambda_1 \in \{0, 0.1, \dots, 1\}$ and $\lambda_2 \in \{0.1, 0.2 \dots, 1\}$ to enforce (some degree of) similarity between graphs, as we do not expect the different seasons and drought circumstances to cause all edges to differ. In addition, forcing the proposed method to borrow information across graphs increases the stability of the estimated edges.

In this analysis, the results are limited to the yield graphs, that is, graphs where that only contain nodes that share an edge with yield. By the Local Markov Property (see Lauritzen 1996), given all neighbours of yield, yield is independent of all non-neighbours. Therefore, these yield graphs are valid subgraphs of the full graphs containing all variables. The maize yield graph is shown in Figure 1 and the wheat yield graph is shown in Figure 2. The estimated graph structures will be interpreted for each crop separately, as separate models were fitted on different crops, followed by a brief overview of some methodological considerations that hold for both crops. When interpreting the results, the main focus will be on the differences in graph structures for the varying drought conditions, rather than the overall graph structure.

3.1. Maize drought stress yields

None of the edges connected to yield can be considered stable, that is, having a bootstrap certainty of more than 70 % across all graphs. In fact, across all four graphs, there is only a a single edge, the edge between evaporation sum and daily temperature range, that has an associated bootstrap certainty of more than 70 %. When comparing the edge occurrence probabilities from the yield graph to those of the overall model [**see Supporting Information—Figures S1 and S2**], the stability of the maize yield graphs is higher compared to the overall model: the density of edges in the yield graphs with a very low bootstrap certainty is relatively low compared to the overall model, whilst substantial probability mass is allocated to edges with a high bootstrap certainty.

Concerning the estimated graphs, the differences observed between the yield graphs for drought and non-drought circumstances are of primary interest. We will only consider differing edges directly connected to yield. Two differences can be observed that are potentially affected by the presence of drought. First, a positive relationship between yield and weeding frequency is present for maize not undergoing drought stress, while this is not observed for maize under drought. One potential explanation is that weeds are more affected by drought than maize, leading to lower competition under drought conditions and a lower impact of multiple weedings. Second, a negative partial correlation is present between isothermality (ratio between annual mean temperature and mean diurnal range) and yield for maize in 2010 undergoing drought stress. A straightforward answer for this phenomenon cannot be given, but, using a random forests approach, Vasco Silva *et al.* (ming) found this variable to be an important predictor for yield using a superset of the data used here.

Even though the positive partial correlation between livestock ownership and yield can be found in all four graphs, its presence is of sufficient interest to warrant further discussion. Whilst the positive correlation might be due to the beneficial effects of livestock as draft animals or because of the positive effects on soil fertility, either directly through manure (nitrogen), or through greater resource endowment in general, for which livestock ownership is an indicator, the fact that yield is not conditionally independent of the amount of nitrogen fertilizer, the field size and the ox ploughing frequency, would suggest that this variable contains other information not captured by aforementioned variables. Depending on the type of livestock, manure may contain nutrients other than nitrogen such as phosphate, organic carbon or potash, all of which can have a positive effect on yield. Potentially, the livestock variable contains information regarding wealth not found in the field size variable. Wealth is typically associated with higher input and labour use, although this hypothesis cannot be confirmed due to the lack of presence for the relevant variables in the yield graph.

3.2. Wheat drought stress yields

Whilst overall, the estimated edges in the wheat graphs are more stable than those estimated in the maize graphs, there is no single edge connected to yield with a bootstrap certainty of more than 70 % present across all four graphs. Nonetheless, the edges between livestock and yield and the total labour use

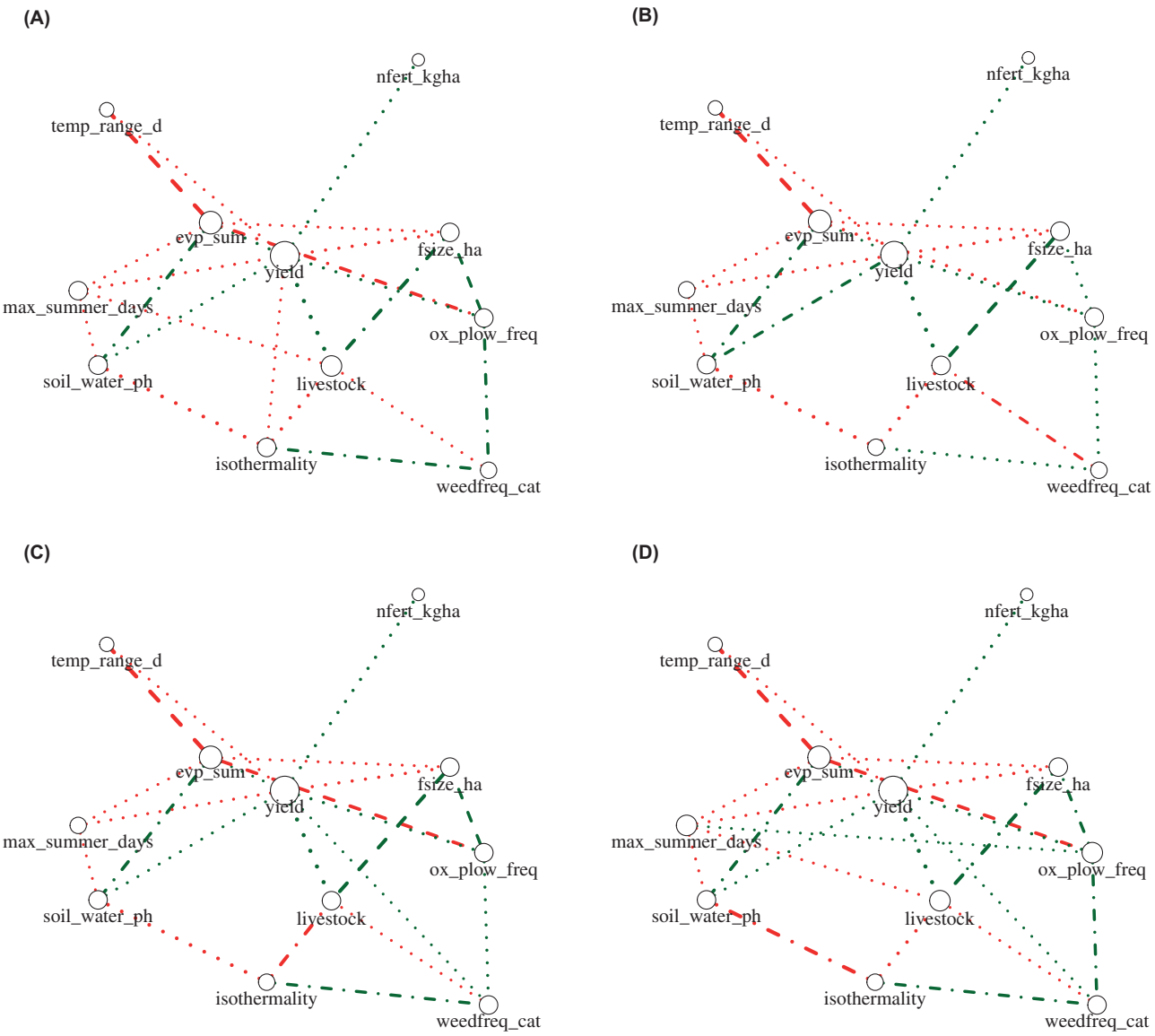

Figure 1. Results for the maize graphs, with estimated optimal penalty parameters $\lambda_1 = 0.1, \lambda_2 = 0.1$ as chosen by the AIC: (A) maize yield under drought stress 2010, (B) maize yield under drought stress 2013, (C) maize yield not under drought stress 2010, (D) maize yield not under drought stress 2013. Edge width reflects the absolute value of the partial correlation, color is used to make visual distinction between positive (green) and negative (red) partial correlations and line type represents the bootstrap certainty of a particular relationship: $\geq 90\%$_____, $90-70\%$_ _ _ _ _, $70-50\%$_._._._. and $\leq 50\%$ … … … …

and yield have a high certainty in three out of four graphs. The exception being the graph pertaining to wheat yields not undergoing drought stress in 2009. Moreover, as observed in the maize yield graphs, there is a single edge (not connected to yield) that does indicate high bootstrap certainty: the edge between the soil clay content and the maximum number of summer days. The wheat yield graphs, like the maize yield graphs exhibit more stability compared to the overall model, as can be seen in **Supporting Information—Figures S3 and S4**.

Contrary to the results obtained for maize yields, those of wheat do not show consistent differences for drought and non-drought conditions. Nevertheless, some differing edges connected to yield are observed. We find that the soil sand content has a negative partial correlation with yield for wheat not undergoing drought stress in 2013. Moreover, there is a positive partial correlation between the evaporation sum and yield. Given that these edges do not seem to indicate fundamental drought relations—they only hold for 2013 and not for 2009—they will not be discussed further.

What is striking is the low number of weather-related variables sharing an edge with yield, given that a substantial number (27 out of 58) of variables in the data are weather related. An explanation for this is that the variables that do share an edge with yield (arid days, evaporation sum and maximum summer days), contain all the information that the non-connected variables do, such that given arid days, evaporation sum and maximum summer days, yield is conditionally independent of all other weather variables, which is the definition of the partial

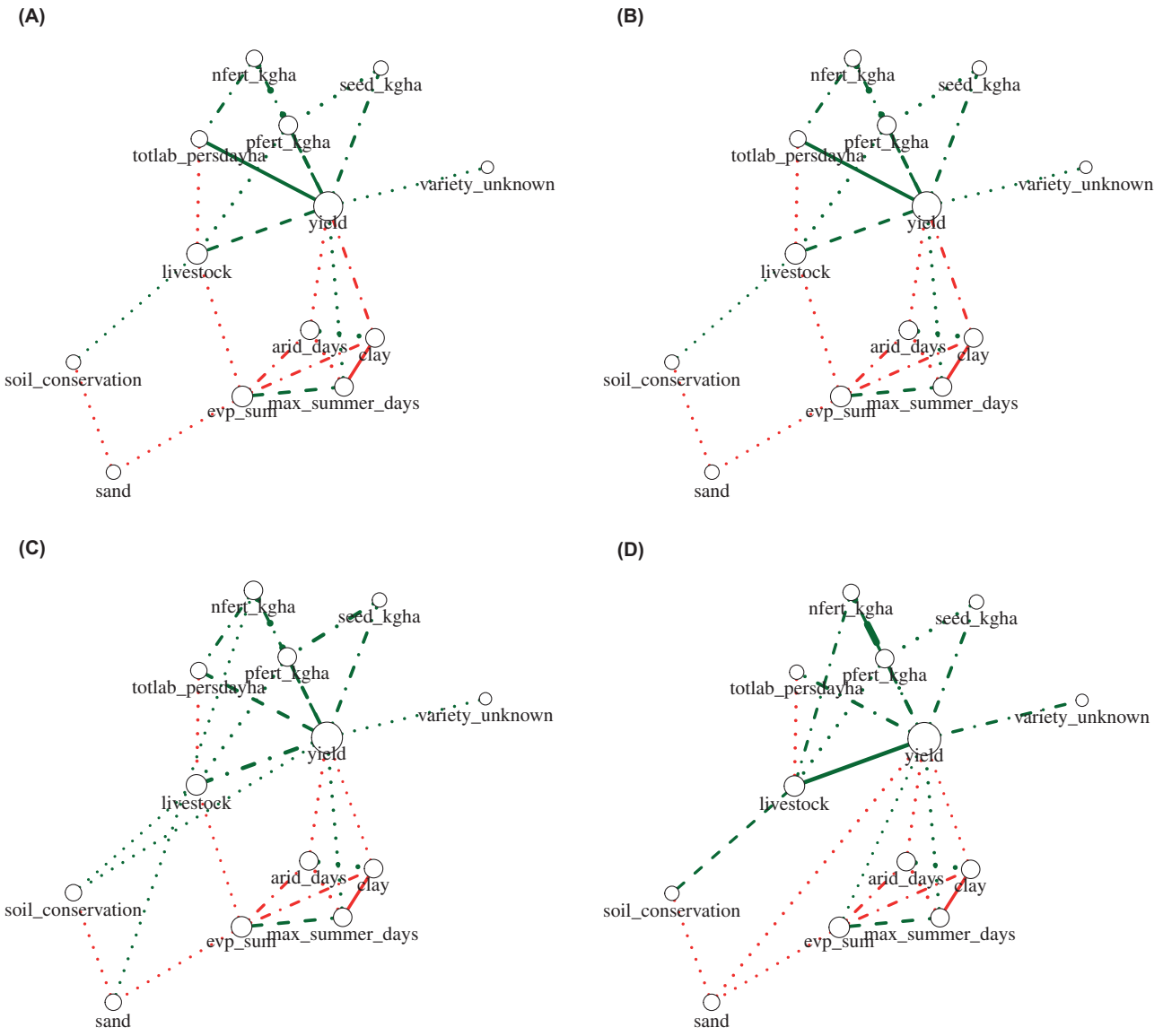

Figure 2. Results for the wheat graphs, with estimated optimal penalty parameters $\lambda_1 = 0.1, \lambda_2 = 0.1$ as chosen by the AIC: (A) wheat yield under drought stress 2009, (B) wheat yield under drought stress 2013, (C) wheat yield not under drought stress 2009, (D) wheat yield not under drought stress 2013. Edge width reflects the absolute value of the partial correlation, color is used to make visual distinction between positive (green) and negative (red) partial correlations and line type represents the bootstrap certainty of a particular relationship: $\geq 90\%$_____, $90-70\%$_ _ _ _ _, $70-50\%$_._._._. and $\leq 50\%$ ... ...

correlation. Conversely, management variables play a big role, indicating that, irrespective of the presence of drought, nitrogen and phosphorus fertilizer correlate positively with yield. Similarly, more labour also correlates positively with yield. In fact, this is one of the more potent and stable relations as evidenced by the edge width and the line type, respectively. Interestingly enough, the partial correlation between the amount of seed used (planting density) and yield is invariant across drought conditions. Whilst during droughts competition for water between plants might intensify, this is not reflected here. In fact, even during droughts, increasing the plant density has a positive relationship with yield. If a type of wheat was used that has a high density resistance, this relationship across drought conditions can be explained.

3.3. Methodological considerations

Due to the penalty parameters chosen by the AIC, $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$ for both crops, the graphs exhibit similar structure across groups, see Figures 1 and 2. This is not only evident by the presence of the edges, but also by magnitude and sign of the partial correlations. The phenomenon of forcing edge values to equality across graphs for high values of λ_2 is a special feature of the penalty imposed by the proposed method that is not typically found in other group-based graphical models. In addition, as $\lambda_1 = 0.1$, the graphs themselves are sparse only to a limited extend, but remain sufficiently sparse to be well interpretable. Evaluating the AIC over a denser grid $|\lambda_1| \times |\lambda_2|$ might have resulted in sparser graphs. Alternatively, evaluating values of λ_2 in a dense grid between 0 and 0.1 could have resulted in

estimated graphs exhibiting more structural differences. If time is not an issue, it is, therefore, recommended to use dense grids to evaluate the information criterion, in order to obtain optimal values for λ_1 and λ_2. Moreover, the researcher can restrict model selection to only similar, dense, sparse of dissimilar graphs, depending on prior knowledge. Needless to say, model selection remains a tricky part of most statistical analysis. Finally, it should be noted that the small sample size also hinders the effectiveness of the model selection step.

Even though, for both crops, some edges are estimated to be fairly certain ($> 70\%$), the majority of the edges are not. This can be attributed to the small sample size leading to variable estimates. When sample sizes are small, imposing sparsity is recommended. However, given that the optimal estimated penalty value for λ_1 was 0.1, many edges required estimation, inherently increasing the uncertainty associated with the graphs. Moreover, when a higher value for λ_2 was selected, variability would be lower and certainty higher, due to the increased borrowing of information across graphs.

4. CONCLUSION

The aim of this article is to introduce copula graphical models for heterogeneous mixed data to a diverse audience interested in the effects of drought stress on crop yield, thereby letting researchers expand their statistical toolkit in order to better analyse agricultural data. By introducing some basic theory, an overview concerning model fitting and selection and an R package, we aim to encourage researchers to use the proposed statistical tool into their analysis. With this method, we gain deeper insights into the complex mechanisms underlying crop response to abiotic stresses and develop more effective strategies to promote agricultural sustainability. The applications in this article served as examples of how to interpret results obtained by the proposed method, and illustrate that the method is able to capture both fundamental relationships (edges present in all graphs) as well as differing relationships between groups. Even though the present article introduced a grouping-based method to take into account differing spatial circumstances, strong spatial effects can be present on a smaller scale (within fields), which require further methodological developments. We are currently working on a new methodology, that takes these local spatial dependencies into account. Moreover, we are aiming to extend the proposed method to causal inference and prediction.

SOURCES OF FUNDING

No public funding sources were involved in the preparation of this manuscript.

CONTRIBUTIONS BY THE AUTHORS

All authors contributed to editing the manuscript. S.H. led the general structuring of the manuscript and contributed to the writing of the manuscript, the initialization of the research, data analysis and the methodology, J.vH. contributed to interpreting the results and P.B. contributed to the initialization of the research, the data analysis and the methodology.

CONFLICT OF INTEREST STATEMENT

None delcared.

ACKNOWLEDGEMENTS

The authors thank the reviewers for their useful comments that improved the readability of this article.

SUPPORTING INFORMATION

The following additional information is available in the online version of this article –

Figure S1. Maize edge density plots for the overall model. The figures show the edge occurrence probabilities for the data under drought stress in 2010 and 2013, and for the data not under drought stress in 2010 and 2013 respectively.

Figure S2. Maize edge density plots for the yield graph. The figures show the edge occurrence probabilities for the data under drought stress in 2010 and 2013, and for the data not under drought stress in 2010 and 2013 respectively.

Figure S3. Wheat edge density plots for the overall model. The figures show the edge occurrence probabilities for the data under drought stress in 2009 and 2013, and for the data not under drought stress in 2009 and 2013 respectively.

Figure S4. Wheat edge density plots for the yield graph. The figures show the edge occurrence probabilities for the data under drought stress in 2009 and 2013, and for the data not under drought stress in 2009 and 2013 respectively.

LITERATURE CITED

Akaike H. 1973. Second international symposium on information theory. *2nd International Symposium on Information Theory*.

Behrouzi P, Arends D, Wit EC. 2023. The R journal: netgwas: An R package for network-based genome wide association studies. *The R Journal 14*, 18–37. https://doi.org/10.32614/RJ-2023-011.

Behrouzi P, Wit EC. 2019. Detecting epistatic selection with partially observed genotype data by using copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 68*, 141–160.

Bielders CL Gérard B. 2015. Millet response to microdose fertilization in south–western niger: Effect of antecedent fertility management and environmental factors. *Field Crops Research 171*, 165–175.

Chen J, Chen Z. 2012. Extended BIC for small-n-large-p sparse glm. *Statistica Sinica*, 555–574.

Dai A. 2013. Increasing drought under global warming in observations and models. *Nature Climate Change 3*, 52–58.

Danaher P. 2018. *JGL: performs the joint graphical lasso for sparse inverse covariance estimation on multiple classes*. R package version 2.3.1.

Danaher P, Wang P, Witten DM. 2014. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*, 373–397.

Dietz K-J, örb CZ, Geilfus C-M. 2021. Drought and crop yield. *Plant Biology 23*, 881–893.

Everingham Y, Sexton J, Skocaj D, Inman-Bamber G. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development 36*, 1–9.

Guo J, Levina E, Michailidis G, Zhu J. 2015. Graphical models for ordinal data. *Journal of Computational and Graphical Statistics 24*, 183–204.

Hermes S, van Heerwaarden J, Behrouzi P. 2022. Copula graphical models for heterogeneous mixed data. *arXiv preprint arXiv:2210.13140*.

Jiang H, Fei X, Liu H, Roeder K, Lafferty J, Wasserman L, Li X, Zhao T. 2021. *huge: high-dimensional undirected graph estimation*. R package version 1.3.5.

Khaki S, Wang L. 2019. Crop yield prediction using deep neural networks. *Frontiers in Plant Science 10*, 621.

Landau S, Mitchell R, Barnett V, Colls J, Craigon J, Payne R. 2000. A parsimonious, multiple-regression model of wheat yield response to environment. *Agricultural and Forest Meteorology 101*, 151–166.

Lauritzen SL. 1996. *Graphical models*, Vol. 17. Oxford: Clarendon Press, United Kingdom.

Lischeid, G., H. Webber, M. Sommer, C. Nendel, and F. Ewert (2022). Machine learning in crop yield modelling: a powerful tool, but no surrogate for science. *Agricultural and Forest Meteorology 312*, 108698.

Liu H, Roeder K, Wasserman L. 2010. Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in Neural Information Processing Systems 23*.

Lobell DB, Cassman KG, Field CB. 2009. Crop yield gaps: their importance, magnitudes, and causes. *Annual Review of Environment and Resources 34*, 179–204.

Niang A, Becker M, Ewert F, Dieng I, Gaiser T, Tanaka A, Senthilkumar K, Rodenburg J, Johnson J-M, Akakpo C, Segda Z. (2017). Variability and determinants of yields in rice production systems of West Africa. *Field Crops Research 207*, 1–12.

Ronner E, Franke A, Vanlauwe B, Dianda M, Edeh E, Ukem B, Bala A, Van Heerwaarden J, Giller KV. 2016. Understanding variability in soybean yield and response to p-fertilizer and rhizobium inoculants on farmers' fields in northern Nigeria. *Field Crops Research 186*, 133–145.

Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Vasco Silva J, van Heerwaarden J, Pytrik R, Laborte AG, Tesfaye K, van Ittersum MK. Forthcoming. Big data, small explanatory power? Lessons learnt with random forest predictive modeling of crop yield in contrasting farming systems.

Vujačić I, Abbruzzo A, Wit E. 2015. A computationally fast alternative to cross-validation in penalized Gaussian graphical models. *Journal of Statistical Computation and Simulation 85*, 3628–3640.