Estimation of Missing Values Affects Important Aspects of GGE Biplot Analysis

Article *in* Crop Science · November 2016

DOI: 10.2135/cropsci2016.02.0100

8 authors, including:

Leomar Guilherme Woyann
Federal University of Technology - Paraná/Br…
24 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Giovani Benin
Federal Technological University of Parana C…
138 PUBLICATIONS **592** CITATIONS
SEE PROFILE

Lindolfo Storck
Federal University of Technology - Paraná/Br…
319 PUBLICATIONS **1,831** CITATIONS
SEE PROFILE

Volmir Sergio Marchioro
Universidade Federal de Santa Maria
98 PUBLICATIONS **216** CITATIONS
SEE PROFILE

Published November 28, 2016

RESEARCH

Estimation of Missing Values Affects Important Aspects of GGE Biplot Analysis

Leomar Guilherme Woyann, Giovani Benin,* Lindolfo Storck, Diego Maciel Trevizan, Cátia Meneguzzi, Volmir Sergio Marchioro, Matheus Tonatto, and Alana Madureira

ABSTRACT

Multi-environment trials often yield unbalanced datasets, thus necessitating the estimation of missing values. It is unknown whether this estimation affects the graphic characteristics of genotype plus genotype-by-environment interaction (GGE) biplots. Therefore, our objectives were to investigate the effects of different percentages of missing values on the number of significant principal components (PCs) and on mega environments, "winner" (highest-performing) genotypes, and the amount of variation explained by the PCs. Two complete sets of two-way data from wheat (*Triticum aestivum* L.) were used. The first set consisted of the original data (Data1, from which we created scenarios with 0, 30, and 60% missing data. For the second dataset (Data2), we removed 50% data from the original dataset, estimated missing values to make it a new complete dataset, and created scenarios like those for Data1. Missing values were estimated via expectation-maximization–GGE (EM–GGE) and EM–additive main effects and multiplicative interaction (EM–AMMI) methods. The percentage of variation explained by the PCs was affected by the percentage of missing data; a large percentage of missing values considerably increased the amount of variation explained by PC 1 and PC 2 and reduced the complexity of the genotype-by-environment interaction because two PCs accounted for more than 80% of the variation, instead of the three PCs that were required to explain the variation in the original dataset. The EM–GGE estimation method was able to maintain the original conformation of the 'which-won-where' biplot when ≤30% of estimated data were used. The EM–GGE was superior to the EM–AMMI method for estimating missing data. The estimation of more than 30% of the data should be avoided because it can lead to significant changes in mega environment conformation and the identification of "winner" genotypes.

L.G. Woyann, G. Benin, L. Storck, D.M. Trevizan, C. Meneguzzi, M. Tonatto, and A. Madureira, Univ. Tecnológica Federal do Paraná, UTFPR–Câmpus Pato Branco, Via do Conhecimento, Km 1, Pato Branco, PR 85503-390, Brazil; V.S. Marchioro, Cooperativa Central de Pesquisa Agrícola, COODETEC, Rodovia BR 467, Km 98, Cascavel, PR 85813-450, Brazil. Received 15 Feb. 2016. Accepted 15 Oct. 2016. *Corresponding author (benin@utfpr.edu.br). Assigned to Editor Shawn M. Kaeppler.

Abbreviations: ABL, Abelardo Luz; AMMI, additive main effects and multiplicative interaction; CMR, Campo Mourão; CSC, Cascavel; EM, expectation-maximization algorithm; GEI, genotype-by-environment interaction; GGE, genotype plus genotype-by-environment interaction; GVA, Guarapuava; ME, mega environment; NMT, Não-Me-Toque; PC, principal component; SS, sum of squares; SVD, singular value decomposition; VCU, value for cultivation and use

The presence of genotype-by-environment interaction (GEI) is common in multi-environment trials in wheat (*Triticum aestivum* L.) and other crops; it leads to changes in the performance of cultivars in different environments. A GEI demands that trials be conducted at multiple locations for several years to obtain reliable data for the possible release of a new cultivar.

Trials conducted across several years and locations increase the possibility of obtaining datasets with unbalanced or incomplete data. This can occur because of planned or unplanned actions by the breeder. The planned actions might be the removal of genotypes that did not perform well in a given year, introducing new genotypes developed by the breeding program and the introduction of newly released cultivars. On the other hand, unplanned actions, which are unrelated to the breeder, might be attributable to human or environmental causes (Yan, 2013).

There are three primary strategies for analyzing unbalanced or incomplete datasets. The first strategy consists in the removal of genotypes or environments with missing values. The second

Published in Crop Sci. 57:1–13 (2017).
doi: 10.2135/cropsci2016.02.0100

strategy involves replacing missing values by environmental means. The third strategy is the estimation of missing values using statistical tools that enable the analysis of such data (Yan, 2013; Arciniegas-Alarcón et al., 2014).

Several methods have been developed to estimate missing values in multi-environment trials (Gauch and Zobel, 1990; Yan, 2013; Arciniegas-Alarcón et al., 2014; Paderewski and Rodrigues, 2014) and are based on the expectation-maximization algorithm (EM) (Dempster et al., 1977). The EM involves a natural generalization of the maximum likelihood estimation of missing data (Do and Batzoglou, 2008). The EM–additive main effects and multiplicative interaction (EM–AMMI) method (Gauch, 1992) and EM–genotype plus genotype by environment interaction (EM–GGE) method (Yan, 2013) are the commonly used methods to estimate missing values for a biplot analysis.

A biplot analysis (Gabriel, 1971) summarizes a large amount of information within the main axes (typically one or two), which are designated as principal components (PCs). The GGE methodology (Yan, 2001) makes it possible to identify the "winner" genotype, i.e., the highest-performing genotype at a location or environment, and to verify the formation of mega environments (MEs). This methodology uses the additive effect of genotype plus the multiplicative effect of genotype and environment for the analysis (Yan and Kang, 2003; Yan et al., 2007; Gauch et al., 2008; Yan, 2014).

One of the criticisms of GGE biplot methodology relates to the absence of statistical tests for testing the significance of PCs (Yang et al., 2009; Forkman and Piepho, 2014). In the GGE biplot methodology, the first two PCs are usually used to explain the GEI data. However, it is necessary to verify whether this procedure is appropriate (Yang et al., 2009). A lesser or greater number of PCs may be necessary to better describe the interactions present in the data (Forkman and Piepho, 2014).

To our knowledge, no reports exist in the literature about the implications of using estimated data in the analysis of multi-environment trial data. In this context, we tested the hypothesis that the use of estimated data in GGE graphical analysis can affect the identification of "winner" genotypes, the formation of MEs, genotype allocation to a specific location, the adaptability and stability of genotypes, and the selection of an ideal environment. With respect to the effects of data estimation on the significance of PCs and respective changes in a GGE analysis, the objectives of this study were to investigate the effects of different percentages of missing values on the number of significant PCs, to ascertain changes (arising from the presence of missing data) in the identification of the "winner" genotypes for the environments evaluated, and to determine the effects of missing data estimation on the variation explained by the PCs in the GGE biplot graphs.

MATERIALS AND METHODS

We used wheat grain yield data from 17 cultivars tested at five locations in 2010, 2011, and 2012. The locations were Abelardo Luz (ABL), Campo Mourão (CMR), Cascavel (CSC), Guarapuava (GVA), and Não-Me-Toque (NMT) in southern Brazil. These sites are representative of the two most important regions of "value for cultivation and use" (VCU) for wheat in Brazil. The sites GVA and NMT are located in the VCU 1 region and ABL, CMR, and CSC are located in the VCU 2 region (Fig. 1). The experimental design was a randomized complete block with three replications. The cultivars evaluated were 'CD 104', 'CD 105', 'CD 114', 'CD 116', 'CD 117', 'CD 119', 'CD 120', 'CD 121', 'CD 122', 'CD 123', 'CD 124', 'CD 1440', 'CD 150', 'CD 1550', 'Fundacep Raízes', 'Guamirim', and 'Quartzo'. Additional information on cultivars, such as year of release, plant height, quality parameters, and breeding company, is provided in Table 1.

We used two datasets for the analysis. The first one was the original complete dataset, which was designated as Data1. For the second dataset (Data2), we removed 50% data from the original dataset. A flowchart of the procedures used in this study is shown in Fig. 2. Scenarios were created from the complete datasets by the random withdrawal of 30% of the values using the default sample function in R software (R Core Team, 2014). We further randomly removed another 30% of the data to generate a dataset with 60% missing values. We constructed scenarios with 0, 30, and 60% for each dataset and applied three independent replications for each scenario to each dataset. For each replication, we returned to the complete file and restarted the process, so the first 30% missing values obtained for scenario 1, replication 1, were not necessarily the same values in scenario 1, replication 2 (i.e., cells that had their values removed were randomly selected). Cells with missing data were determined randomly for the complete dataset (Data1). For the scenarios using Data2, the missing cells were the same as those for Data1. We used this procedure to compare the results obtained from the two datasets.

The estimation of missing data using the EM algorithm can be performed either by AMMI (Gauch and Zobel, 1990) or by GGE (Yan, 2013). The method using GGE (Yan, 2013), designated here as EM–GGE, only considers the multiplicative effects. Thus, the allocation is performed only on the p_{ij} term. This method does not consider the parameters μ_j and s_j in the iteration process (i.e., it neglects the additive effects of the genotype and the environment). In a different manner, the EM–AMMI estimation (Gauch and Zobel, 1990) represents an assessment of the value of Y_{ij}, which considers both the main effects and the multiplicative effects. Estimation methods based on EM, EM–AMMI, and EM–GGE consist of the performance of successive singular-value decomposition (SVD) cycles, until the values estimated in two successive cycles are sufficiently similar to the previous results (Yan, 2013).

All of the scenarios were estimated using GGEbiplot software (Yan, 2001). According to Yan and Holland (2010), the general model for two-dimensional GGE biplots is:

$$p_{ij} = \frac{\left(\overline{y}_{ij} - \mu_j\right)}{s_j} = \sum_{k=1}^{2} \lambda_k \alpha_{ij} \gamma_k + \overline{\varepsilon}_{ij}$$

where p_{ij} denotes a genotype-by-environment two-way table of GGE effects with $i = 1,..., g$ genotypes and $j = 1,..., e$ environments that is decomposed into two PCs, with singular values λ_k, genotype eigenvalues α_{ij}, and environmental eigenvalues γ_k for every

kth PC. $\bar{\varepsilon}_{ij}$ is the residue for genotype i in environment j, $\bar{y}_{ij}$ denotes the cell mean of genotype i in environment j, and μ_j is the mean value in environment j. The operation $\bar{y}_{ij} - \mu_j$ is assigned as "environment centering," in which the environment main effects are removed from the original dataset. The parameter s_j is the scaling factor, and the operation of dividing $(\bar{y}_{ij} - \mu_j)/s_j$ is referred to as "data scaling."

An ANOVA for each scenario for Data1 and Data2 was conducted by use of the GGEbiplot software (Yan, 2001) before the data were estimated. Subsequently, which-won-where GGE biplots were generated to identify the performance and the grouping of the genotypes in the environments evaluated. The following settings were used: untransformed data (transform = 0), standard deviation scale (scaling = 1), data centralized by the average of environments (centering = 2), and singular-value partition with focus on the environment (singular value partitioning = 2).

The missing values were also estimated using the EM–AMMI method, as described by Paderewski (2013), via the R software (R

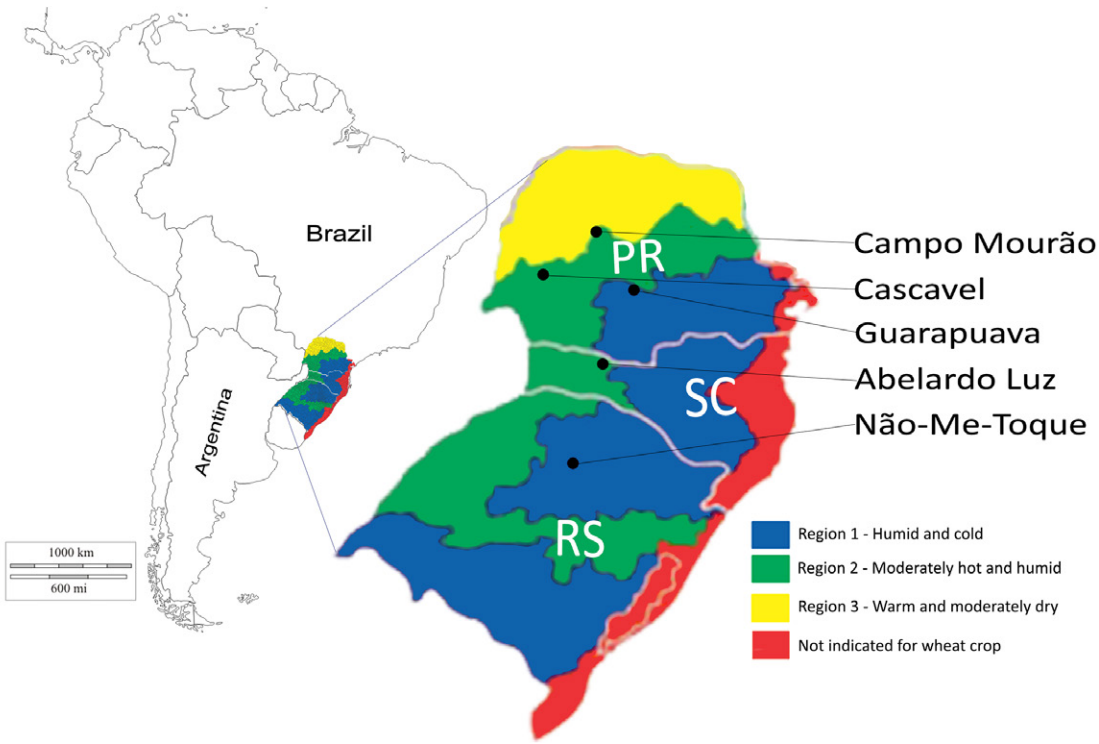

Fig. 1. Test locations used in this study and a description of the regions of "value for cultivation and use" (VCU) for wheat in southern Brazil. RS, Rio Grande do Sul State; SC, Santa Catarina State; PR, Paraná State.

Table 1. List of cultivars used in the current study with information about year of release, cycle length, height, quality parameters, and breeding company.

Cultivars	Year of release	Breeding program	Agronomic characters		Quality parameters	
			Plant height	Cycle length	Gluten strength	Stability
			cm	d	10^{-4} J	min
CD 104	1999	Coodetec	81	123	375.0	15.3
CD 105	1999	Coodetec	82	125	150.0	4.9
CD 114	2004	Coodetec	73	123	293.0	12.8
CD 116	2006	Coodetec	77	125	350.0	15.0
CD 117	2007	Coodetec	77	123	248.0	12.3
CD 119	2009	Coodetec	75	123	140.0	6.8
CD 120	2009	Coodetec	77	125	137.8	6.3
CD 121	2010	Coodetec	73	124	184.0	5.6
CD 122	2010	Coodetec	74	124	281.8	12.0
CD 123	2010	Coodetec	73	125	275.4	11.5
CD 124	2012	Coodetec	73	123	280.0	10.9
CD 150	2009	Coodetec	74	125	370.5	16.6
CD 1440	2013	Coodetec	79	124	324.0	14.9
CD 1550	2012	Coodetec	78	125	316.8	14.6
Fundacep Raízes	2006	Fundacep	77	123	222.0	19.0
BRS Guamirim	2005	Embrapa	74	125	271.5	8.0
Quartzo	2007	OR/Biotrigo	78	124	234.3	14.8

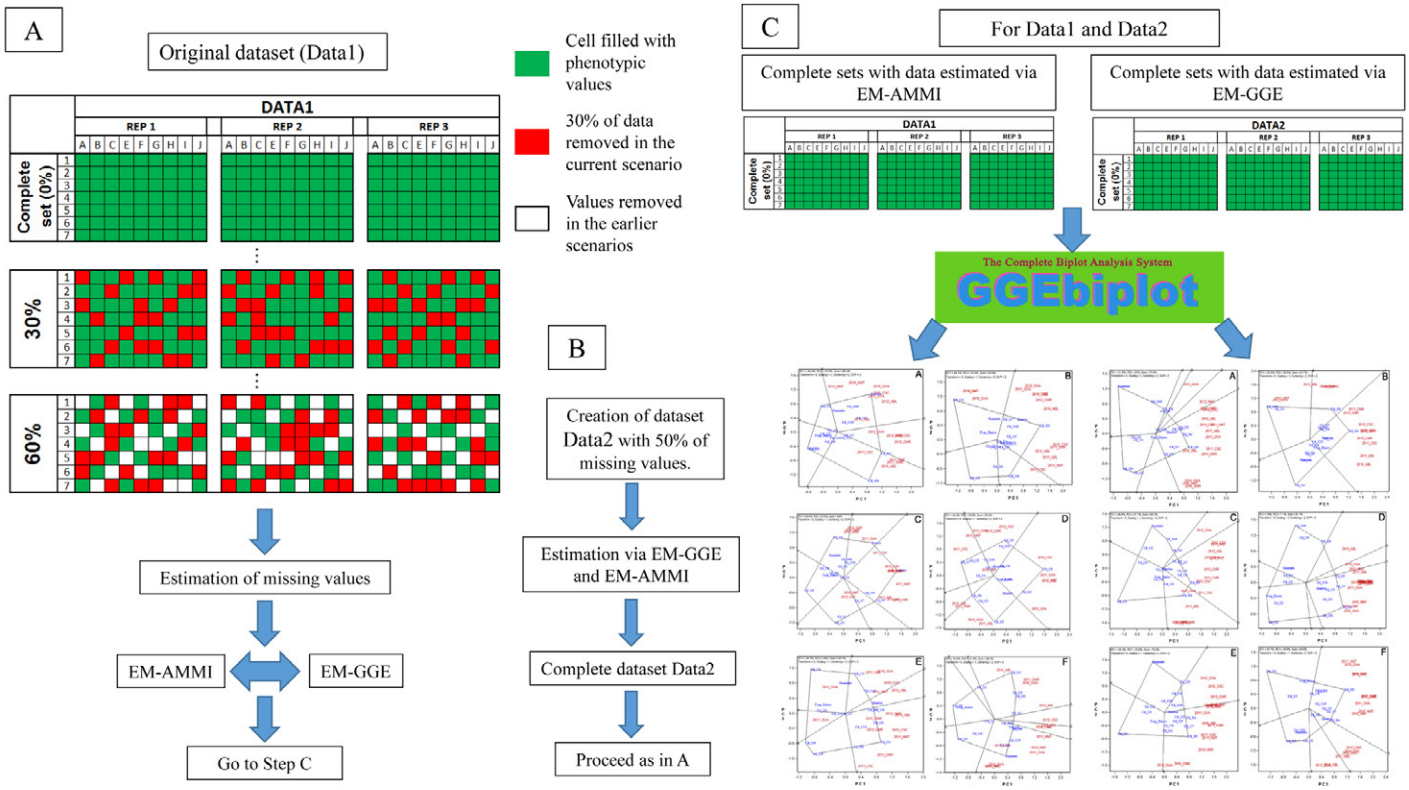

Fig. 2. Flowchart of the procedures performed in this study. (A) Process used to get the repetitions and scenarios with missing values from the Data1 set. Furthermore, it shows that the missing data were estimated via EM–GGE (expectation-maximization–genotype plus genotype-by-environment interaction) and EM–AMMI (EM– additive main effects and multiplicative interaction) methods. (B) Indicates that the Data2 set was obtained removing 50% of data (repetition 1) from the Data1. Later, missing values were estimated via EM–GGE to have a complete set of data. After that, the procedures adopted in (A) were repeated. (C) Scenarios and repetitions obtained for the datasets Data1 and Data2, after the missing values were estimated by EM–AMMI and EM–GGE; the GGEbiplot software (Yan, 2001) was used to obtain the graphic biplots.

Core Team, 2014). In the main function, the number of PCs in the AMMI model was defined as 2 (PC.nb = 2). The default number of PCs for AMMI is 1. The parameter PC.nb = 2 was used because the values obtained were used in a GGE biplot analysis, in which two PCs were normally used. The other parameters were used as default. The complete dataset was analyzed with the GGEbiplot software to obtain the graphs. For EM–GGE, the missing values were estimated with the GGEbiplot software (Yan, 2001). Both methods used SVD to estimate the missing values. Successive SVD cycles were performed until the values estimated in two successive cycles were sufficiently similar (Yan, 2013). The initial values used by the EM methods do not affect the final estimated values. In the same way, we ran the new complete sets to obtain which-won-where graphs based on missing values estimated via EM–AMMI.

The significance of the PCs in the GGE biplot analysis was verified by the methodology proposed by Forkman and Piepho (2014). The statistical test used was the *t* test and the probability value (*p*-value) was verified by simple bootstrap, with resampling 100,000 times. The scripts to perform these analyses for AMMI-type biplots with the R software (R Core Team, 2014) are in the supplementary data of Forkman and Piepho (2014). The following changes were implemented to adapt these scripts to GGE biplot types: $M = \min(I, J - 1)$; $D = \max(I, J - 1)$; $E = Y -$ RowMeans; Eb < −matrix{rnorm[$(D - K)(J - 1 - K)$], nrow = $I - K$, ncol = $J - 1 - K$}.

RESULTS AND DISCUSSION

The ANOVA obtained via GGEbiplot software are shown in Table 2. Changes in the level of significance of the genotype factor (entry) and the environment (tester) were present in only one replication in Data2. The significance probability decreased from 1 to 5%, but none ceased to be significant, even with 60% of missing data. Significant effects of both the genotypes and environments are sufficient to determine the formation of ME and identify "winner" genotypes in the GGE biplot analysis (Yan et al., 2007). However, Yang et al. (2009) indicated that statistical tests were necessary to verify whether genotypes and environments could form subgroups or whether all of them should be analyzed together. In this context, we report a method using confidence intervals to determine whether a significant difference among genotypes and among environments was present, allowing the discrimination of ME. In this regard, Yang et al. (2009) analyzed a dataset widely used in literature (Ontario winter wheat genotypes from the year 1993, with 18 genotypes and nine locations). They suggested that the environments in these trials should not be divided into MEs. However, it is unclear whether this methodology allows the generation of MEs and the designation

of "winner" genotypes via the which-won-where graph when other datasets are analyzed. So, this statistical analysis approach could restrict genotype grouping and the formation of MEs. This does not help breeders in selecting new genotypes with specific adaptations and hinders the elimination of sites that generate redundant information.

A methodology to verify the significance of the PCs, proposed by Forkman and Piepho (2014), was not sensitive to changes in explaining the PC 1 and PC 2 in different scenarios arising from the Data1 and Data2 datasets (Tables 3 and 4). The significance of the PCs was also affected by the percentage of missing data (Table 3). For the original dataset (Data1), replications indicated the presence of two significant PCs. When 30% of the data were estimated, discrepancies began to occur in the number of significant PCs, and three PCs were statistically significant ($p < 0.05$) in one of the replications. Similar results were obtained for the scenarios derived from the Data2 dataset.

The sums of squares (SS) differed in the datasets employed (Tables 3 and 4). In this sense, the randomization of missing data and their estimation can significantly affect the results. This is especially evident from Table 3,

Table 2. Analysis of variance for the original dataset (Data1) for grain yield (kg ha^{-1}) in wheat, where the first two principal components explained 42.3% of genotype-by-environment variation, and for a new dataset (Data2), wherein 50% data were estimated and the first two principal components explained 81.2% of genotype-by-environment variation.

		42.3% explanation						81.2% explanation					
		0% MV†		30% MV		60% MV		0% MV		30% MV		60% MV	
Replication	Variation sources	df	MS‡	df	MS	df	MS	df	MS	df	MS	df	MS
1	Entry	16	823718**	16	706230**	16	646837**	16	753157**	16	533006**	16	646836**
	Tester	14	5408203**	14	4481657**	14	3027002**	14	5484156**	14	4316492**	14	3027002**
	Interaction or error	224	221244	147	221075	93	198923	224	196347	147	189355	93	198923
2	Entry	16	823718**	16	905445**	16	888026**	16	753157**	16	894855**	16	1118704**
	Tester	14	5408203**	14	3833037**	14	2659020**	14	5484156**	14	3481900**	14	2638560**
	Interaction or error	224	221244	147	219446	93	231410	224	196347	147	169599	93	93416
3	Entry	16	823718**	16	542171**	16	677142**	16	753157**	16	351411*	16	498865*
	Tester	14	5408203**	14	3161860**	14	2157695**	14	5484156**	14	3091843**	14	2161529**
	Interaction or error	224	221244	147	240479	93	231103	224	196347	147	239411	93	242070

* Significant at the 0.05 probability level.

** Significant at the 0.01 probability level.

† MV, percentage of missing values

‡ MS, mean square

Table 3. Percentage of missing data (MV), sum of squares (SS), percentage of variation explained by each of the first five principal components, *t* test, and significance of the principal components for the genotype plus genotype-by-environment interaction (GGE) biplot analysis via simple bootstrap (SB) of the original dataset (Data1) that explained 42.3% of the genotype-by-environment interaction variation for grain yield (kg ha^{-1}) in wheat.

		Replication 1				Replication 2				Replication 3			
	Term†	SS‡				SS				SS			
MV	$K + 1$	$Y^2(k+1)$	%	Test *t*	SB	$Y^2(k+1)$	%	Test *t*	SB	$Y^2(k+1)$	%	Test *t*	SB
0%	1	54953894	40.09	0.401	0.000	54953894	40.09	0.401	0.000	54953894	40.09	0.401	0.000
	2	27019852	19.71	0.329	0.001	27019852	19.71	0.329	0.002	27019852	19.71	0.329	0.001
	3	14530720	10.60	0.264	0.277	14530720	10.60	0.264	0.280	14530720	10.60	0.264	0.276
	4	9318303	6.80	0.230	0.893	9318303	6.80	0.230	0.897	9318303	6.80	0.230	0.898
	5	8699475	6.35	0.278	0.550	8699475	6.35	0.278	0.558	8699475	6.35	0.278	0.546
30%	1	62591951	38.77	0.388	0.000	51268067	36.09	0.361	0.000	54930512	41.16	0.412	0.000
	2	30621289	18.97	0.310	0.006	30733435	21.64	0.339	0.001	32242137	24.16	0.411	0.000
	3	16371707	10.14	0.240	0.588	17828821	12.55	0.297	0.052	14376683	10.77	0.311	0.024
	4	12916022	8.00	0.249	0.682	12943587	9.11	0.307	0.103	7118195	5.33	0.223	0.939
	5	11056881	6.85	0.284	0.480	10064848	7.09	0.344	0.059	5762894	4.32	0.233	0.960
60%	1	55283287	34.35	0.346	0.000	73498078	32.54	0.325	0.000	57883994	36.73	0.367	0.000
	2	27216830	16.91	0.261	0.146	46787403	20.71	0.307	0.007	35660041	22.63	0.358	0.000
	3	17615053	10.95	0.228	0.759	27460026	12.16	0.260	0.314	23689053	15.03	0.370	0.001
	4	17226991	10.71	0.289	0.213	19831350	8.78	0.254	0.622	11021873	6.99	0.273	0.378
	5	13961459	8.68	0.330	0.101	16538829	7.32	0.284	0.485	7400890	4.70	0.252	0.832

† Sequential testing procedure for all terms (Marasinghe, 1985; Schott, 1986). This procedure tests the (K + 1)th multiplicative term as if it were the first term in a problem with the numbers of rows and columns reduced by K (Forkman and Piepho, 2014).

‡ Sum of squares for $\hat{\lambda}^2_{k+1}$ and proportions (%) of the additive model error sum of squares, $\sum_{k=1}^{M} \hat{\lambda}^2_k$ (Forkman and Piepho, 2014).

Table 4. Percentage of missing data (MV), sum of squares (SS), percent variation explained by each of the first five principal components, *t* test, and significance of the principal components for the genotype plus genotype-by-environment interaction (GGE) biplot analysis via simple bootstrap (SB) of the complete dataset (Data2) for which 50% of the values were estimated, which explained 81.2% of the genotype-by-environment interaction for grain yield (kg ha^{-1}) in wheat.

		Replication 1				Replication 2				Replication 3			
	Term†	SS‡				SS				SS			
MV	$K + 1$	$Y^2(k+1)$	%	Test *t*	SB	$Y^2(k+1)$	%	Test *t*	SB	$Y^2(k+1)$	%	Test *t*	SB
0%	1	55254874	42.47	0.425	0.000	55254874	42.47	0.425	0.000	55254874	42.47	0.425	0.000
	2	29083861	22.35	0.389	0.000	29083861	22.35	0.389	0.000	29083861	22.35	0.389	0.000
	3	14068526	10.81	0.307	0.031	14068526	10.81	0.307	0.028	14068526	10.81	0.307	0.030
	4	7370721	5.67	0.233	0.864	7370721	5.67	0.233	0.873	7370721	5.67	0.233	0.874
	5	5569658	4.28	0.229	0.974	5569658	4.28	0.229	0.970	5569658	4.28	0.229	0.974
30%	1	55435892	42.69	0.427	0.000	53231669	42.97	0.430	0.000	54211076	41.71	0.417	0.000
	2	28788355	22.17	0.387	0.000	27700923	22.36	0.392	0.000	29113131	22.40	0.384	0.000
	3	14252503	10.97	0.312	0.024	12087586	9.76	0.282	0.119	15103039	11.62	0.324	0.011
	4	7294630	5.62	0.232	0.867	7535794	6.08	0.244	0.740	7423791	5.71	0.235	0.845
	5	5480247	4.22	0.227	0.977	5683632	4.59	0.244	0.902	5428735	4.18	0.225	0.983
60%	1	55283309	34.62	0.346	0.000	63327154	41.96	0.420	0.000	61141180	42.84	0.428	0.000
	2	27216852	17.04	0.261	0.140	32280534	21.39	0.369	0.000	26307561	18.43	0.322	0.001
	3	17615058	11.03	0.228	0.754	14827637	9.82	0.268	0.224	17976451	12.60	0.325	0.010
	4	17227009	10.79	0.289	0.211	9443667	6.26	0.233	0.861	9539215	6.68	0.256	0.593
	5	13961422	8.74	0.330	0.105	7375481	4.89	0.238	0.937	7836515	5.49	0.282	0.498

† Sequential testing procedure for all terms (Marasinghe, 1985; Schott, 1986). This procedure tests the (K + 1)th multiplicative term as if it were the first term in a problem with the numbers of rows and columns reduced by K (Forkman and Piepho, 2014).

‡ Sum of squares for $\hat{\lambda}^2_{k+1}$ and proportions (%) of the additive model error sum of squares, $\sum_{k=1}^{M} \hat{\lambda}^2_k$ (Forkman and Piepho, 2014).

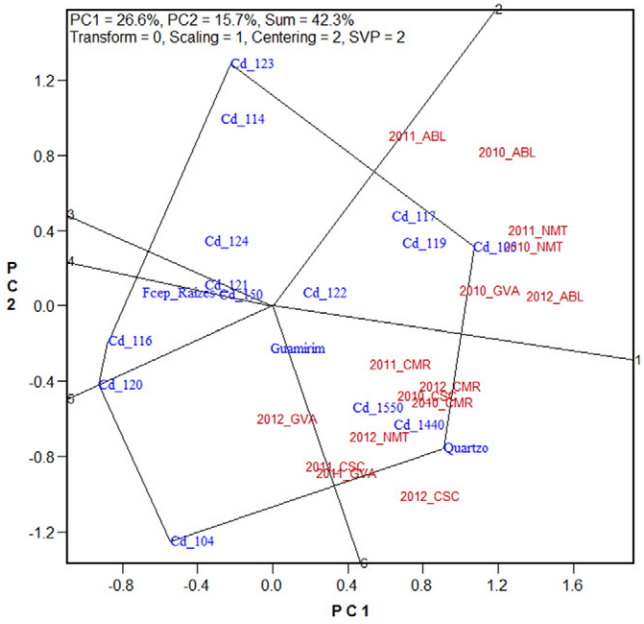

Fig. 3. Which-won-where biplot for a complete dataset of 17 Brazilian spring wheat genotypes and 15 environments (five locations from southern Brazil over the course of 3 yr). PC, principal component; SVP, singular value partitioning method; ABL, Abelardo Luz; CMR, Campo Mourão; CSC, Cascavel; GVA, Guarapuava; NMT, Não-Me-Toque.

where the SS values were higher in the second replication than in the other replications, which further supports the hypothesis that these parameters are strongly affected by the dataset employed (Yan, 2013). In Table 4, differences in the SS between replications were smaller than in Table 3. Through this method, the participation of PC 1 and PC 2 in the SS showed no significant changes when different percentages of missing data were evaluated, indicating that this method did not efficiently identify the effects of data estimation. However, this method is very useful to justify the use of a certain number of PCs in a biplot analysis.

In the original complete dataset (Data1), the sum of PC 1 and PC 2 explained only 42.3% of the variation (Fig. 3), which is considered low. Yan (2002) indicated that a satisfactory explanation should correspond to more than 80% of the interactions present in the dataset.

The necessity of conducting trials for several years and locations increases the probability of an incomplete dataset for various reasons. Different percentages of missing values could have different effects on the percentage of variation explained by the PCs. As the percentage of missing data increased in the scenarios derived from Data1, the variation explained by PC 1 and PC 2 greatly increased for data estimated via EM–GGE and via EM–AMMI. When 30% of the data were estimated via EM–GGE, the amount of variation explained varied between 58 and 66.2% for the three replications (Fig. 4). However, when 60% of the values were estimated, the percentage of variation explained by the first two PCs varied between 83.6 and 86.4% in the three random replications. When the missing data were estimated by EM–AMMI (Fig. 5), the variation explained by the estimation of 30% of the data ranged between 68.3 and 74.8%. When 60% of the data were estimated, the amount of variation explained varied between 81.1 and 92.6% in the three random replications. Therefore, the two methods of estimating missing data led to similar increases in the amount of variation explained by PC 1 and PC 2.

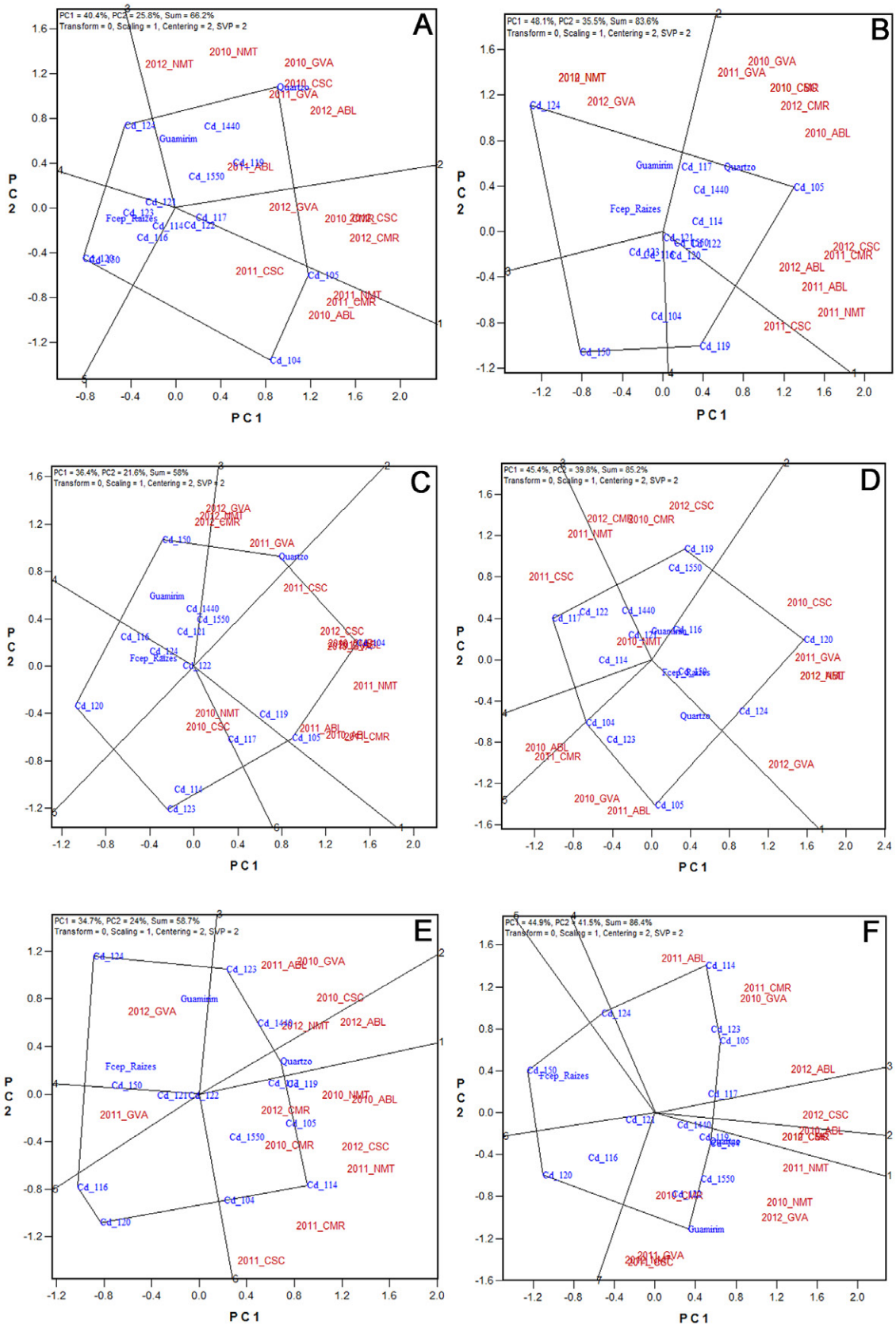

Fig. 4. Which-won-where biplot derived from a complete dataset of double entry for 17 Brazilian spring wheat genotypes and 15 environments (five locations from southern Brazil over the course of 3 yr). The missing values were estimated via the expectation-maximization algorithm considering only the multiplicative term genotype-by-environment (EM–GGE). (A), (C), and (E) consist of three independent replications when 30% of the data were estimated. (B), (D), and (F) consist of three independent replications when 60% of the data were estimated. PC, principal component; SVP, singular value partitioning method; ABL, Abelardo Luz; CMR, Campo Mourão; CSC, Cascavel; GVA, Guarapuava; NMT, Não-Me-Toque.

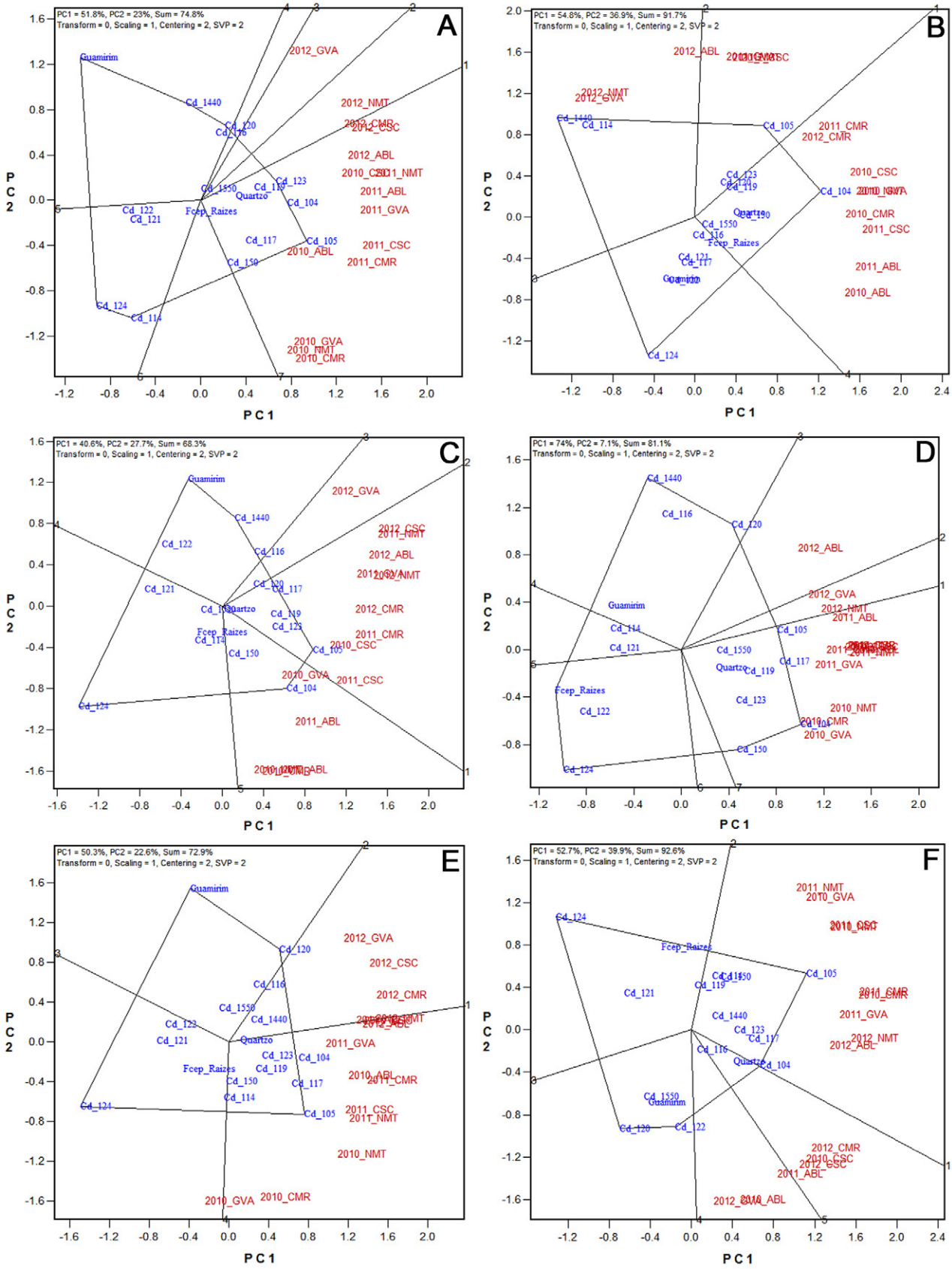

Fig. 5. Which-won-where biplot derived from a complete dataset of double entry for 17 Brazilian spring wheat genotypes and 15 environments (five locations from southern Brazil over the course of 3 yr). The missing values were estimated via the expectation-maximization algorithm considering the additive main effect of genotype, environment, and their multiplicative interaction (EM–AMMI). (A), (C), and (E) consist of three independent replications when 30% of the data were estimated. (B), (D), and (F) consist of three independent replications when 60% of the data were estimated. PC, principal component; SVP, singular value partitioning method; ABL, Abelardo Luz; CMR, Campo Mourão; CSC, Cascavel; GVA, Guarapuava; NMT, Não-Me-Toque.

Changes in the variation explained by the biplot PCs can alter the breeder's perception of the dataset and the inferences made. In a scenario where only 42.3% of the interaction was represented by PC 1 and PC 2, the subsequent estimation of missing data via EM–GGE or EM–AMMI increased the amount of variation explained to greater than 80% when 60% of the data were estimated. Accordingly, Yan (2015) indicated that a breeder can use datasets with up to 60% estimated data without an excessive increase in predicted errors attributable to data estimation.

The generation of MEs is one of the most important pieces of information obtained via a biplot analysis, particularly in the GGE biplot method (Yan et al., 2007). For this analysis, it is necessary to conduct tests at various locations. The formation of MEs has many consequences because the analysis of each ME must be performed separately. Furthermore, these results must be repeatable across years (i.e., repeatable location groups need to be found) (Gupta et al., 2013; Munaro et al., 2014; Xu et al., 2014). The first method used to verify the occurrence of MEs is through the strategy designated as "analyze yearly and summarize across years" (Yan, 2014). The ME formation is facilitated when the variance is low for the years but high for the locations (Luo et al., 2015). Yan (2015) developed a new approach to verify the formation of MEs, designated as the GGL + GGE biplot (GGL being the genotype main effect plus genotype-by-location interaction). This strategy uses two datasets, one for the ME analysis and test location evaluation (data from 5 yr) and the other for validation (data from 2 yr). This strategy allows the breeder to more reliably divide the target area into MEs.

A which-won-where biplot of the original complete Data1 dataset indicated the formation of the following two MEs: the first comprised locations ABL and NMT, and the second comprised CSC, CMR, and GVA (Fig. 3). The "winner" cultivars for the two groups were CD 105 and Quartzo, respectively. When 30% of the data were estimated via EM–GGE, this same pattern was retained in two replications (Fig. 4). However, one of the replications failed to show this pattern. When 60% of the data were estimated via EM–GGE, the original pattern was not present in any of the three replications. When the data were estimated by EM–AMMI (Fig. 5), the original pattern was absent in both datasets, i.e., the datasets with 30 and 60% estimated data.

The complete Data2 dataset showed that when 50% of data were estimated, no clear formation of MEs was evident because the same locations in different years were not grouped to characterize such a separation (Fig. 6). When 30% of the data were estimated via EM–GGE (Fig. 7), the pattern present in the complete set was maintained. However, this pattern was different from the pattern in the original complete dataset (Data1), which indicated the formation of the MEs. When data were estimated via EM–AMMI, the pattern found in Fig. 6 was not present

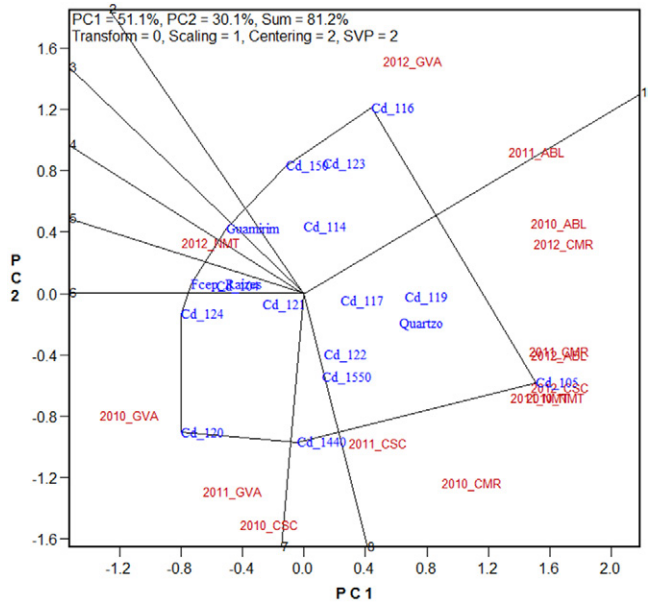

Fig. 6. Which-won-where biplot for a complete dataset obtained via an estimation of 50% of the original complete dataset for 17 Brazilian spring wheat genotypes and 15 environments (five locations from southern Brazil over the course of 3 yr). PC, principal component; SVP, singular value partitioning method; ABL, Abelardo Luz; CMR, Campo Mourão; CSC, Cascavel; GVA, Guarapuava; NMT, Não-Me-Toque.

in the dataset with 30 or 60% of estimated data (Fig. 8). This indicates that the EM–GGE method was able to maintain the pattern shown in the complete Data2 dataset when 30% of the data were estimated. In contrast, the EM–AMMI method did not preserve the pattern when 30% of the data were estimated. The patterns were not maintained when 60% of the data were estimated; this was true when data were estimated by EM–AMMI and when data were estimated by EM–GGE.

A thorough evaluation of the which-won-where analysis must be performed, especially for VCU tests, because a breeder would generally release only one or a few cultivars, so it is necessary to analyze the changes caused by data estimation. The which-won-where analysis for the complete original Data1 dataset indicated that the genotype CD 105 was the "winner" in ABL 2010 and 2012, NMT 2010 and 2011, and GVA 2010. The cultivar CD 123 had the best performance in ABL 2011, and the genotype CD 104 was the "winner" genotype in GVA 2011 and 2012 and in CSC 2011. The cultivar Quartzo had the best performance in the other environments (CMR 2010, 2011, and 2012; CSC 2010 and 2012; NMT 2012).

The estimation of missing data led to significant changes in the which-won-where pattern (Fig. 4), both for the different percentages of missing data and for the estimation of data by the independent removal of data in each of the three replicates. The cultivar Quartzo provides an interesting example. This genotype was the "winner" in six environments in the first replication, where 30% of the data were estimated (Fig. 4A).

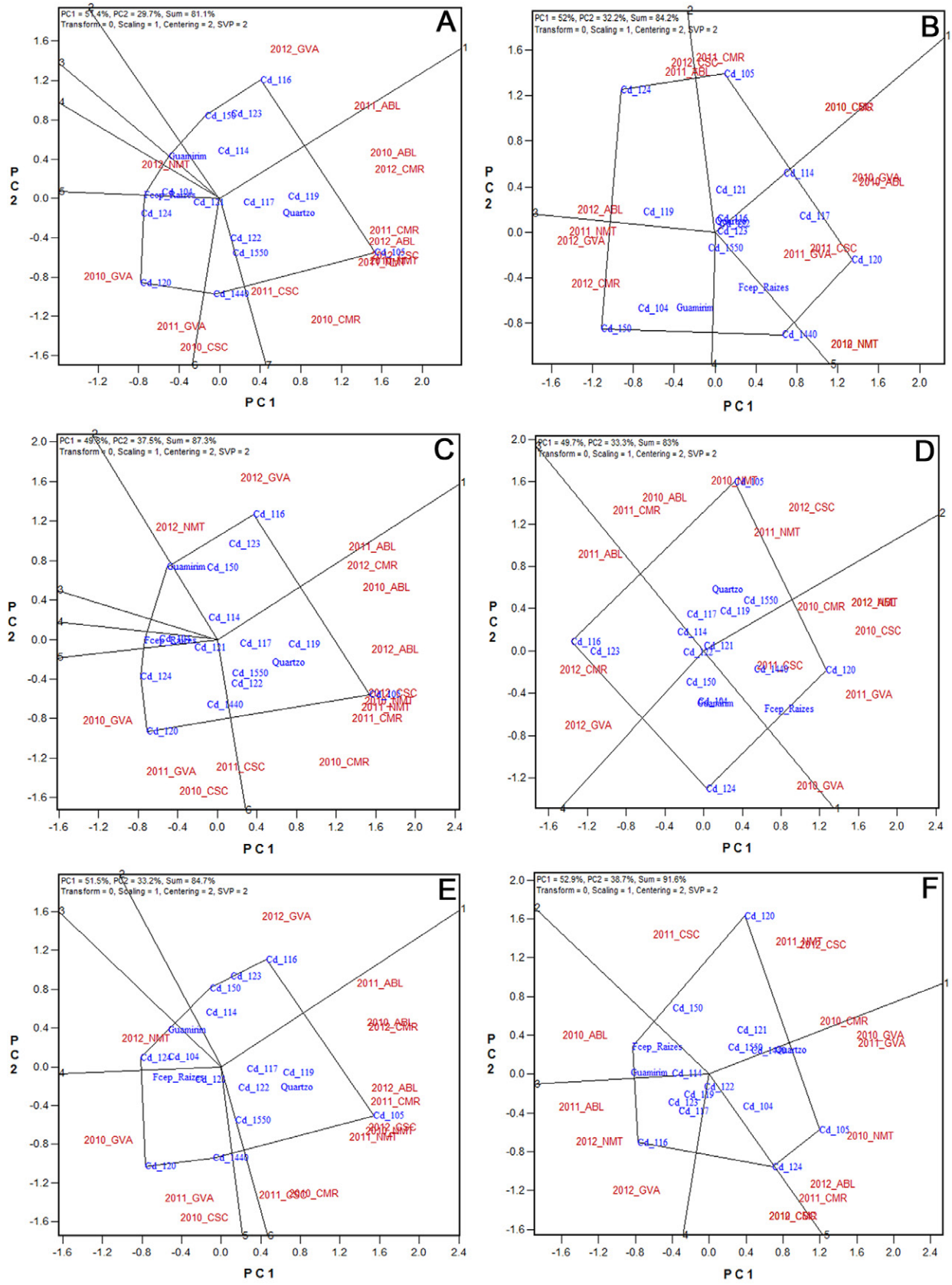

Fig. 7. Which-won-where biplot for a complete dataset obtained via the estimation of 50% of the original complete dataset for 17 Brazilian spring wheat genotypes and 15 environments (five locations from southern Brazil over the course of 3 yr). The missing values were estimated via the expectation-maximization algorithm considering only the multiplicative term genotype-by-environment (EM–GGE). (A), (C), and (E) consist of three independent replications when 30% of the data were estimated. (B), (D), and (F) consist of three independent replications when 60% of the data are estimated. PC, principal component; SVP, singular value partitioning method; ABL, Abelardo Luz; CMR, Campo Mourão; CSC, Cascavel; GVA, Guarapuava; NMT, Não-Me-Toque.

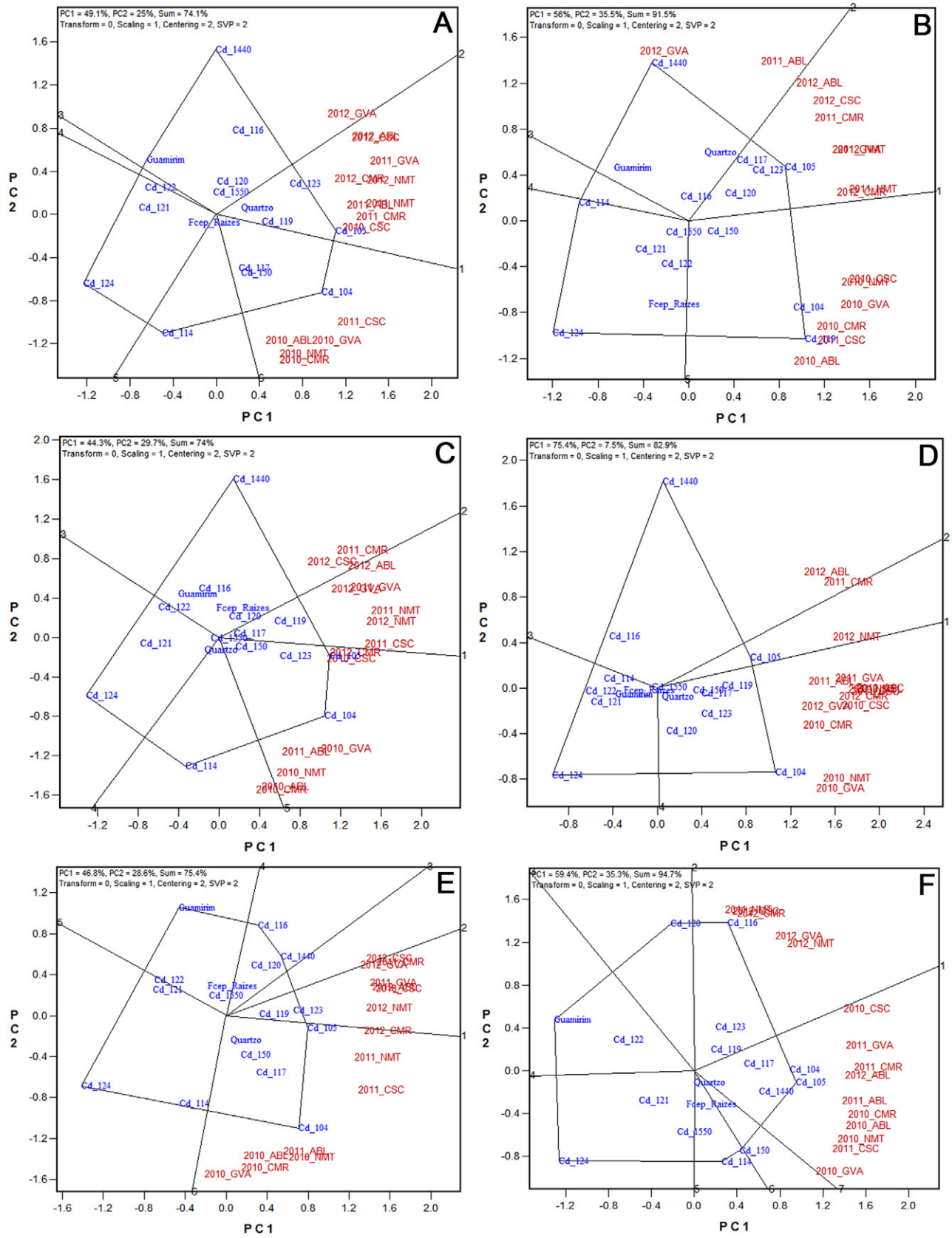

Fig. 8. Which-won-where biplot for a complete dataset obtained via the estimation of 50% of the original complete dataset for 17 Brazilian spring wheat genotypes and 15 environments (five locations from southern Brazil over the course of 3 yr). The missing values were estimated via the expectation-maximization algorithm considering the additive main effect of genotype, environment, and their multiplicative interaction (EM–AMMI). (A), (C), and (E) consist of three independent replications when 30% of the data were estimated. (B), (D), and (F) consist of three independent replications when 60% of the data were estimated. PC, principal component; SVP, singular value partitioning method; ABL, Abelardo Luz; CMR, Campo Mourão; CSC, Cascavel; GVA, Guarapuava; NMT, Não-Me-Toque.

Compared with the complete original dataset, this genotype won in only three environments (NMT 2012, CSC 2010, and GVA 2011) and became the "winner" in three other environments (NMT 2010, GVA 2010, and ABL 2012), where it had not won in the complete dataset (Fig. 4A). In both the second and third replication, Quartzo was the "winner" genotype in only one environment (GVA 2011 and ABL 2012, respectively) (Fig. 4C and 4E). This genotype did not win in any of the three scenarios where 60% of the data were estimated (Fig. 4B, 4D, and 4F). Similar results were found for the biplots obtained via EM–AMMI data estimation (Fig. 5).

The cultivar Quartzo was released in 2007 and was widely cultivated in Brazil between 2008 and 2013. It proved highly productive, disease resistant, and had a highly stable yield in several wheat-producing areas of Brazil.

The complete Data2 dataset obtained by the estimation of 50% of the original Data1 dataset showed a less pronounced increase in the amount of variation explained by the first two PCs when 30 and 60% of estimated data were employed (Fig. 6 and 7). The percentage of variation explained in the complete Data2 dataset was 81.2% (Fig. 6). When 30% of the data were withdrawn and estimated via EM–GGE, the percentage of variation explained ranged between 81.1 and 87.3%, and when 60% of the data were re-estimated via EM–GGE, the percentage of variation explained by the biplots ranged from 83 to 91.6% (Fig. 7). When 30 and 60% of data were withdrawn and estimated via EM–AMMI (Fig. 8), the percentage of variation explained ranged between 74 and 75.4% and between 82.9 and 94.7%, respectively. The difference in magnitude in variation explained by Data2 (81.1–91.6%) was smaller when compared with Data1 (42.3–92.6%). It is important to point out that this outcome occurred in any of the estimated percentages. This behavior happened similarly both for the estimation via EM–GGE and EM–AMMI. One possible explanation for this result may be the greater variation explained by PC 1 and PC 2 in the complete Data2 (81.2%) in comparison with the complete Data1 dataset (42.3%). This small increase in the amount of variation explained by the biplot was attributable to the reduction of environmental effects on the estimated values via SVD. Moreover, it can be noted that when 60% of the data were estimated, fewer values from the complete original dataset were present.

Cultivar CD 105 remained the "winner" genotype in most of the environments evaluated, when the full dataset was employed and also when 30 and 60% of the data were estimated, which indicated that, when two PCs could account for a higher proportion of the GEI, the presence of missing values and subsequent estimation did not affect the identification of "winner" genotypes via the which-won-where analysis. However, when the percentage of variation explained is lower, as it was in the case of the complete Data1 dataset, the "winner" genotypes varied considerably for the original complete dataset relative to datasets containing estimated data.

The variation in PCs showed a similar trend for different percentages of missing values, in which the missing data were estimated by EM–GGE or EM–AMMI. This comparison was very important to understand whether the EM–GGE and EM–AMMI estimation methods provided significantly different results. It is evident that EM–GGE and EM–AMMI yielded similar estimates, even though EM–AMMI uses the main additive effects plus the multiplicative term (genotype + environment + GEI), whereas EM–GGE uses only the multiplicative effects (GEI).

The formation and detection of MEs is crucial in plant breeding because it allows the identification of locations in which the behavior of a genotype is similar across years, and it especially favors the possibility of optimizing the finite resources of a breeding program. Eliminating locations that are not representative or those that do not produce redundant information allows for a more efficient operation of a breeding program. In addition, sites that provide little information or locations that produce redundant information should be replaced with new test sites; this, in turn, not only favors the expansion of the operating area but also improves representation of the region of interest in a breeding program.

The estimation of missing data is an essential tool for the breeder for optimizing the information from a dataset and the inferences made from it. The deletion of information obtained in the field is hampered by the need for versatility in the breeding program and by the high cost of conducting and evaluating the trials that produce such information. However, the accuracy of inferences and its effects on the patterns obtained must be determined so that the breeder does not make wrong decisions about the maintenance or withdrawal of VCU testing genotypes and the release of new cultivars.

CONCLUSIONS

The percentage of variation explained by the PCs is affected by the percentage of missing data, such that a large percentage of missing data considerably increases the amount of variation explained by PC 1 and PC 2 and reduces the complexity of the GEIs. This phenomenon can lead to erroneous inferences about the performance of genotypes and environments.

The EM–GGE estimation method maintains the original conformation of the which-won-where graph when $\leq$30% of the data are estimated. This method is superior to the EM–AMMI method for estimating missing data in multi-environment trials.

The number of significant PCs depends on the percentage of missing values. However, the method proposed by Forkman and Piepho (2014) did not effectively detect

changes in the percentage of variation explained by PC 1 and PC 2 obtained via a GGE biplot analysis.

The which-won-where pattern can become distorted and lead to erroneous inferences when a significant percentage of the data are estimated and cause breeders to commit errors in the selection of new cultivars and deciding which cultivars are appropriate for specific environments. If it is necessary to estimate more than 30% of the data, the results should be analyzed cautiously (i.e., the breeder must be aware that the results shown by the biplot can be significantly changed compared with the information that would be obtained using a complete set of data).

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

We thank CNPq and CAPES for financial support and anonymous reviewers for helpful comments. We also thank Technical Editor Prof. Manjit Kang for his helpful comments for improving the manuscript.

References

Arciniegas-Alarcón, S., M. García-Peña, W.J. Krzanowski, and S. Dias. 2014. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction: Some new aspects. Biom. Lett. 51:75–88. doi:10.2478/bile-2014-0006

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., B 39:1–38.

Do, C.B., and S. Batzoglou. 2008. What is the expectation maximization algorithm? Nat. Biotechnol. 26:897–900. doi:10.1038/nbt1406

Forkman, J., and H.-P. Piepho. 2014. Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. Biometrics 70:639–647. doi:10.1111/biom.12162

Gabriel, K.R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58:453–467. doi:10.1093/biomet/58.3.453

Gauch, H.G. 1992. Statistical analysis of regional yield trials: AMMI analysis of factorial designs. Elsevier, New York.

Gauch, H.G., H.-P. Piepho, and P. Annicchiarico. 2008. Statistical analysis of yield trials by AMMI and GGE: Further considerations. Crop Sci. 48:866–889. doi:10.2135/cropsci2007.09.0513

Gauch, H.G., Jr., and R.W. Zobel. 1990. Imputing missing yield trial data. Theor. Appl. Genet. 79:753–761. doi:10.1007/BF00224240

Gupta, S.K., A. Rathore, O.P. Yadav, K.N. Rai, I.S. Khairwal, B.S. Rajpurohit, and R.R. Das. 2013. Identifying mega-environments and essential test locations for pearl millet cultivar selection in India. Crop Sci. 53:2444–2453. doi:10.2135/cropsci2013.01.0053

Luo, J., Y.B. Pan, Y. Que, H. Zhang, M.P. Grisham, and L. Xu. 2015. Biplot evaluation of test environments and identification of mega-environment for sugarcane cultivars in China. Sci. Rep. 5:15505. doi:10.1038/srep15505

Marasinghe, M.G. 1985. Asymptotic tests and Monte-Carlo studies associated with the multiplicative interaction-model. Comm. Stat. Theory Methods 14:2219–2231. doi:10.1080/03610928508829039

Munaro, L.B., G. Benin, V.S. Marchioro, F. de Assis Franco, R.R. Silva, C.L. da Silva, and E. Beche. 2014. Brazilian spring wheat homogeneous adaptation regions can be dissected in major megaenvironments. Crop Sci. 54:1374–1383. doi:10.2135/cropsci2013.06.0365

Paderewski, J. 2013. An R function for imputation of missing cells in two-way data sets by EM-AMMI algorithm. Comm. Biom. Crop Sci. 8:60–69.

Paderewski, J., and P.C. Rodrigues. 2014. The usefulness of EM-AMMI to study the influence of missing data pattern and application to Polish post-registration winter wheat data. Aust. J. Crop Sci. 8:640–645.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Schott, J.R. 1986. A note on the critical-values used in stepwise tests for multiplicative components of interaction. Comm. Stat. Theory Methods 15:1561–1570. doi:10.1080/03610928608829202

Xu, N.Y., M. Fok, G.W. Zhang, L.I. Jian, and Z.G. Zhou. 2014. The application of GGE biplot analysis for evaluating test locations and mega-environment investigation of cotton regional trials. J. Integr. Agric. 13:1921–1933. doi:10.1016/S2095-3119(13)60656-5

Yan, W. 2013. Biplot analysis of incomplete two-way data. Crop Sci. 53:48–57. doi:10.2135/cropsci2012.05.0301

Yan, W. 2014. Crop variety trials: Data management and analysis. John Wiley & Sons, New York.

Yan, W. 2001. GGEbiplot—A Windows application for graphical analysis of multienvironment trial data and other types of two-way data. Agron. J. 93:1111–1118. doi:10.2134/agronj2001.9351111x

Yan, W. 2015. Mega-environment analysis and test location evaluation based on unbalanced multiyear data. Crop Sci. 55:113–122. doi:10.2135/cropsci2014.03.0203

Yan, W. 2002. Singular value partitioning for biplot analysis of multi-environment trial data. Agron. J. 94:990–996. doi:10.2134/agronj2002.9900

Yan, W., and J.B. Holland. 2010. A heritability-adjusted GGE biplot for test environment evaluation. Euphytica 171:355–369. doi:10.1007/s10681-009-0030-5

Yan, W., and M.S. Kang. 2003. GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists. CRC Press, Boca Raton, FL.

Yan, W., M.S. Kang, B. Ma, S. Woods, and P.L. Cornelius. 2007. GGE biplot vs. AMMI analysis of genotype-by-environment data. Crop Sci. 47:643–653. doi:10.2135/cropsci2006.06.0374

Yang, R.C., J. Crossa, P.L. Cornelius, and J. Burgueño. 2009. Biplot analysis of genotype × environment interaction: Proceed with caution. Crop Sci. 49:1564–1576. doi:10.2135/cropsci2008.11.0665